Research on real-time distance measurement of mobile eye tracking system based on neural network

Ling Hu[1,2] ,Jiarui Gao[1,*]

1. Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China;

2. University of Chinese Academy of Sciences, Beijing, China

13718287002@163.com, gaojiarui@opt.ac.cn

***Abstract*—With the development and application of eye-tracking technology, mobile eye-tracking systems have become more widely used due to their safety and portability. We combine eye-tracking systems with real-time object detection using machine learning. We propose a method of wearing an eye tracker in daily life to obtain the distance between the eye tracking system and the gaze target in real time. During the visual interaction of the eye tracking system, in order to obtain the distance from the eyeball fixation target to the eyeball in real time, the world camera of the mobile eye tracking system pupil labs first collects the position and scale information of the detected target image in real time, and uses camera calibration principle, pinhole camera model and camera distortion model to establish a ranging equation, and then the feasibility of the real-time ranging equation is verified through a specified distance experiment. The total average relative error after de-distortion at the position of 50cm-75cm is reduced to 1.25%, and the highest accuracy-0.9182cm distance measurement can be achieved within the effective distance.**

***Keywords*—mobile eye-tracking system; neural network; pinhole camera; distance measurement**

I. INTRODUCTION

The eye tracking system is mainly used to study human behavior and understand the cognitive process[1]. It is mainly divided into two types: desktop fixed and mobile eye tracking system. The advent and development of wearable mobile eye-tracking systems has opened up new directions for tracking and analyzing human visual behavior during everyday activities.

With the rapid development of human-computer visual interaction needs, the current mainstream commercial mobile eye tracking systems mainly include tobii glasses II developed by the Swedish Tobii company[2], the smi mobile eye tracker designed and manufactured by the German SMI company[3], and Pupil labs developed by the Eye tracking laboratory in Berlin, Germany[4], these eye tracking system can obtain the angle information of the line of sight in the three-dimensional space through different line-of-sight estimation algorithms and the accuracy of the line-of-sight tracking can reach 0.5^{o}. Eye distance has become another requirement for the development of eye tracking systems.

We use the eye tracker developed by the German Eye Tracking Laboratory-Pupil labs. The camera parameters are shown in Table I. This eye-tracking system is based on binocular recognition. The main configuration includes two 850nm band near-infrared cameras with eye recognition and a world camera that captures the wearer's field of view, of which the near-infrared camera has a frame rate of 120 frames per second and an image resolution of 320 × 240 pixels; the world camera corresponds to a variety of sampling frequencies, 30hz @ 1080p, 60hz @ 720p, 120hz @ 320 × 240 pixels, etc. Here, in order to get a wider field of view of the wearer, we choose the highest image resolution in the world camera.

TABLE I. PUPIL LABS PARAMETERS

World camera	Eye camera	Eyeball model	Gaze precision	Gaze accuracy
$30hz@1080p$	$120hz@320\times240$ *pixels*	*3d model ignoring kappa corner*	0.08°	$1.5\sim2.5^{\circ}$

Real-time object detection is based on the object recognition application programming library API (application programming interface) in the artificial intelligence framework tensorflow released by Google[5]. This library includes at least 25 model network structures trained on the coco training set, such as SSD + MobileNet, SSD + Inception, R-FCN + ResNet101, Faster RCNN + ResNet101, Faster RCNN + Inception_ResNet, etc. In order to meet the processing speed requirements of the eye tracking system and the lightweight machine learning network model, we use ssd_mobilnet_v1_coco training model.

The Coco training set contains 80 categories in daily life, such as people, flower pots, books, etc. We use setting target categories and increasing target matching thresholds to set the detection of achieving a specific goal in daily life scenarios. The threshold is set here to 0.8.

II. METHOD

We introduce the basic principles of eye tracking in the

978-1-7281-4323-1/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: SUNY AT STONY BROOK. Downloaded on August 11,2020 at 16:28:02 UTC from IEEE Xplore. Restrictions apply.

3D mode of the eye tracking system and how to combine target object detection to achieve the real-time estimation of the distance between the person wearing the eye tracking system and the viewpoint.

A. Pupil labs eye tracking system

Pupil labs eye tracking system is a 3D eyeball estimation based on ignoring the kappa angle between the visual axis and the optical axis[6]. The basic principle of eye tracking eye estimation is shown in Figure 1, where A is the position where the target object takes the image through the world camera, B is the position of the pupil on the eye recognition camera imaging surface. By establishing a mapping model of A to B, the position of the viewpoint in the world image when the gaze target pupil position changes can be obtained Variety.

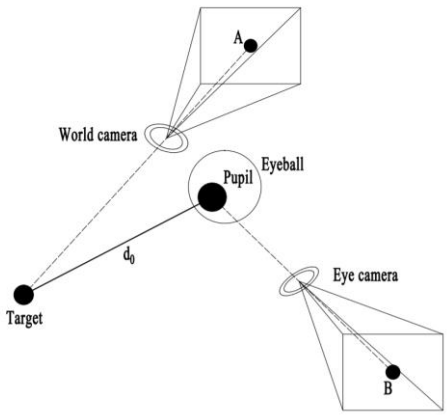

Fig. 1. principles of eye tracking eye tracking

Through equation (1), A and B can be transformed into the same world coordinate system, where (x_A, y_A, z_A) and (x_B, y_B, z_B) are the A、B coordinates in the world coordinate system, respectively, and the mode $|d_0|$ is the distance from the eyeball to the target object.

$$d_0 = (x_A - x_B, y_A - y_B, z_A - z_B) \tag{1}$$

The position of the world camera to the eyeball of the eye tracking system is a fixed value Δd, so the distance $|d_0|$ from the eyeball to the target object can be converted into the distance d from the world camera to the target object, including:

$$d = d_0 - \Delta d \tag{2}$$

B. Real-time target detection

With the rapid development of deep learning technologies, deep learning-based object detection frameworks have gradually become mainstream[7]. Among them, classic deep learning frameworks include R-CNN , YOLO, SSD, etc.

The SSD model is a method proposed by Wei Liu and Dragomir Anguelov[8] et al. Using a single deep neural network to detect targets in an image. The structure of the model is shown in Figure 2.The first five layers are the convolutional layers of the neural network. The sixth and seventh fully connected layers are converted into two convolutional layers. Finally, three more convolutional layers and an average pool are added. The layer model obtains the position and scale information of the target object through regression method, and uses the characteristics of multiple layers to obtain the actual area of the best target object. Therefore, the final mark detection frame can include all the information of the target object.

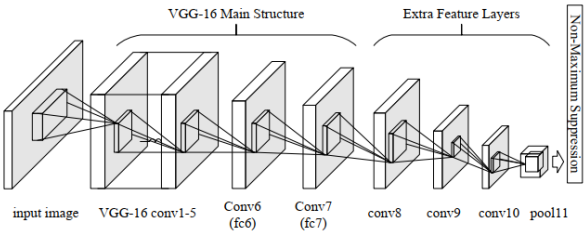

Fig. 2. Structure of the SSD model

The model SSD_Mobilenet is to replace the network with the network as the basic network structure. The network is a lightweight, low-latency and efficient embedded vision model proposed by Andrew G.Howard [9]et al., Which can meet the real-time requirements of target object detection. The target detection plug-in tensorflow api is built into the eye tracking system. At the same time, the eye tracking system pupil labs is worn in daily life scenarios. The world camera captures the specific gaze target image and obtains the recorded gaze target detection accuracy ρ and the target in the image. Scale information w.

C. Image distance modeling

The process of the world camera mapping the coordinate points in the three-dimensional world to the two-dimensional image plane can be described by a pinhole camera model, and its frame is shown in Figure 3.

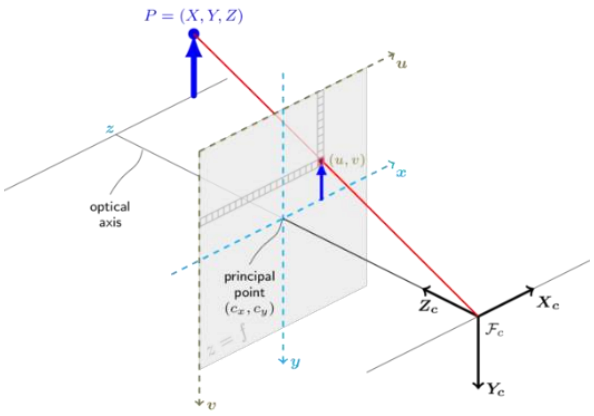

Fig. 3. Pinhole camera model

Equations (3) and (4) describe the relationship from the

Authorized licensed use limited to: SUNY AT STONY BROOK. Downloaded on August 11,2020 at 16:28:02 UTC from IEEE Xplore. Restrictions apply.

three-dimensional point $M=[X,Y,Z]^T$ of the world coordinate system to the two-dimensional point $m=[u,v]^T$ of the pixel coordinate system.

$$s\tilde{m}=A[R\ \ t]\tilde{M} \tag{3}$$

or

$$s\begin{bmatrix}u\\v\\1\end{bmatrix}=\begin{bmatrix}f_x & 0 & c_x\\0 & f_y & c_y\\0 & 0 & 1\end{bmatrix}\begin{bmatrix}r_{11} & r_{12} & r_{13} & t_1\\r_{21} & r_{22} & r_{23} & t_2\\r_{31} & r_{32} & r_{33} & t_3\end{bmatrix}\begin{bmatrix}X\\Y\\Z\\1\end{bmatrix} \tag{4}$$

Where s is the zoom factor, A is the camera's internal parameter matrix, $[R|t]$ is the camera's external parameter matrix, $\tilde{m}$ and $\tilde{M}$ are the homogeneous coordinates corresponding to m and M respectively, (c_x,c_y) are the main points of the image, (f_x,f_y) are the focal lengths x and y in the camera and directions (unit: pixel).

Equation (5) shows the joint rotation $[R|t]$ transformation matrix, which can convert the points (X,Y,Z) in the world coordinate system to the camera coordinate system

$$\begin{bmatrix}x\\y\\z\end{bmatrix}=R\begin{bmatrix}X\\Y\\Z\end{bmatrix}+t\longrightarrow\begin{cases}u=f_x\bullet\dfrac{x}{z}+c_x\\v=f_y\bullet\dfrac{y}{z}+c_x\end{cases} \tag{5}$$

Lenses generally produce distortions[10]. Lens distortions are mainly divided into radial distortion and tangential distortion, as shown in equation (6).

$$\begin{cases}x'=x(1+k_1r^2+k_2r^4+k_3r^6)+2p_1xy+p_2(r^2+2x^2)\\y'=y(1+k_1r^2+k_2r^4+k_3r^6)+2p_2xy+p_1(r^2+2y^2)\\r^2=x^2+y^2\end{cases} \tag{6}$$

Among them (k_1,k_2,k_3) and (p_1,p_2) are the radial and tangential distortion parameters of the lens distortion, respectively. According to the pinhole camera model and the triangle similarity equation, the relationship between w and d in inverse proportion can be obtained under ideal conditions

$$\frac{d}{f_x}=\frac{w'}{w} \tag{7}$$

Among them w' is the scale information of the target object in the three-dimensional world.

III. EXPERIMENT

We measured on a laptop with a CPU of i7-3720QM, a main frequency of 2.60GZHz, and a 8GB memory, and a computer screen plus a shell width of 37.2cm. The mobile eye tracking system has a world camera resolution of 1080p and a maximum viewing angle of 100 °. The experimental process is based on the recommended distance of the eye and the computer screen in [11]. Six groups of experiments are planned at a distance of 5 cm between 50-75 cm. Open the built-in real-time target detection API in the pupil labs interface. In order to reduce the experimental error, record the computer screen detection accuracy ρ (unit %) and scale information w (unit cm) in 100 groups of world camera images at a specific distance per group. Figure 4 is the experimental scenario.

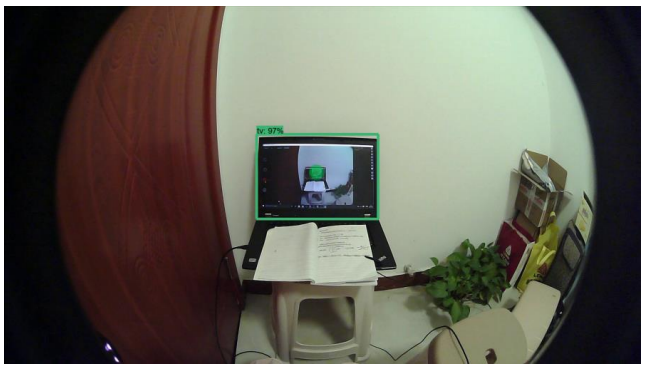

Fig. 4. Real-time experimental scenarios

A. Image distance equation

Due to barrel distortion of the world camera of pupil labs, we need to perform distortion processing on the computer screen scale information. The camera_calibration module in the matlab software platform is used to calibrate the world camera[12], and the internal parameters of the camera are obtained at an average reprojection error of 0.16 pixel. The matrix and distortion parameters are as follows:

$$\begin{bmatrix}f_x & 0 & c_x\\0 & f_y & c_y\\0 & 0 & 1\end{bmatrix}=\begin{bmatrix}804.99 & 0 & 957.21\\0 & 804.31 & 557.91\\0 & 0 & 1\end{bmatrix} \tag{8}$$

$$\begin{bmatrix}k_1\\k_2\\p_1\\p_2\\k_3\end{bmatrix}=\begin{bmatrix}-0.4587\\0.1934\\-0.0014\\0.0014\\0\end{bmatrix} \tag{9}$$

The detection frame of the target object recognition will contain all the information of the computer screen after distortion. Therefore, the maximum value of the computer screen is the width of the target detection frame. The coordinates of the image captured by the world camera can be obtained $r^2<1$, $k_1<1$, and $|k_1|>|k_2|$, through normalization processing. The maximum value w obtains at $y=0$, according to the screen width measurement obtained in the figure, the distortion is reduced to the corresponding specified distance, and the non-linear image distance equation is

Authorized licensed use limited to: SUNY AT STONY BROOK. Downloaded on August 11,2020 at 16:28:02 UTC from IEEE Xplore. Restrictions apply.

$$d = \frac{24056.55}{w} \tag{10}$$

The image distance equation without distortion processing is:

$$d' = \frac{29945.63}{w} \tag{11}$$

B. Analysis of experimental results

Using the built-in object detection api framework in pupil labs, the computer screen scale measurements obtained are shown in Figure 5:

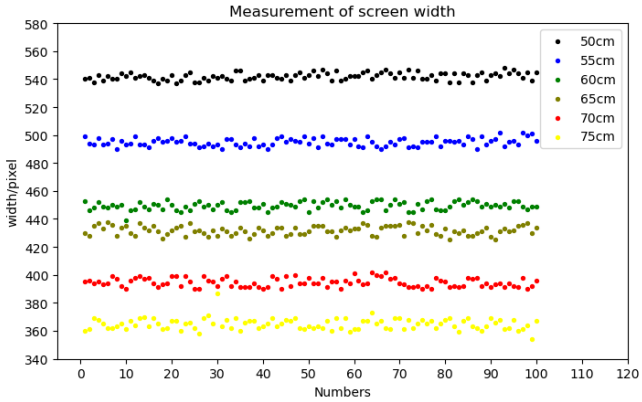

Fig. 5. Measurement of screen width

We use the average absolute error and relative error to analyze the accuracy of the image distance equation, where $N = 100$, Table II lists the experimental data and the error comparison before and after the distortion correction.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|e_i| \tag{12}$$

TABLE II. EXPERIMENTAL DATA AND ERROR ANALYSIS

d_0	50	55	60	65	70	75
$\bar{w}$	542.02	494.97	449.63	431.82	394.54	364.83
$\bar{\rho}$	91.7	93.6	89.8	90.3	91.4	94.3
MAE	0.5379	0.3917	0.9236	1.5632	0.6263	0.6581
MAE'	5.2495	5.5019	6.6034	4.3513	5.9051	7.0913
MRE	1.08	0.71	1.54	2.40	0.89	0.87
MRE'	10.50	10.00	11.01	6.69	8.44	9.46

Where $\bar{w}$ is the average screen width (unit cm), $\tilde{\rho}$ (unit %) is the average accuracy of target detection, MAE' and MRE' is the average absolute error and average relative error (unit %) without distortion processing.

According to Table II, after the distortion correction, the total average relative error of the data decreased from 9.35% before correction to 1.25% after correction.

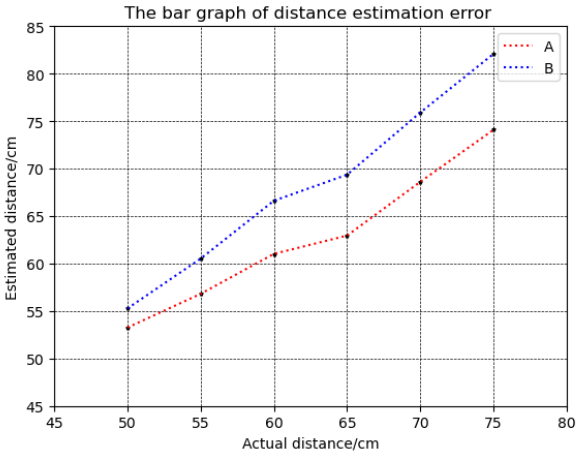

Fig. 6. The bar graph of distance estimation error

Figure 6 is an error bar graph of the estimated distance, where A and B is the estimated distance obtained from the image distance modeling equations (10) and (11), and it can be seen from the figure 6 that the estimated distance accuracy obtained by the image distance equation after distortion correction is high , The ranging accuracy reached ± 3.2367 / 2.0668cm, and the highest accuracy reached -0.9182cm.

IV. CONCLUSIONS

Based on the research of head-mounted eye tracking system, we propose a measurement method that uses the combination of tensorflow api framework target detection and eye tracking system to obtain the distance between the eye tracking point and the wearer in real time. Auxiliary equipment is required, and distance measurement has better accuracy. The system is first assembled with a tensorflow-based object detection plug-in, followed by camera calibration, the distortion model and image distance equation are established, and finally the distance measurement is completed through six sets of distance experiments. The experimental results show that when the target detection framework is incorporated into the pupil labs, the total average accuracy of the target detection is 91.85%, the total average relative error after de-distortion is reduced to 1.25%, and the highest accuracy of the ranging is -0.9182 cm.

ACKNOWLEDGMENT

This work was supported by the Youth Innovation Promotion Foundation of Chinese Academy of Sciences.

REFERENCES

[1] Wade, N., and Tatler, B. W. The moving tablet of the eye: The origins of modern eye movement research. Oxford University Press, 2005.
[2] Tobii Technology, Danderyd, Sweden. Eye Track- ing Products. http://www.tobii.com
[3] SensoMotoric Instruments, Teltow, Germany. SMI Gaze & EYE Tracking Systems. http://www.smivision.com
[4] Pupillabs, https://pupil-labs.com
[5] ABADI M, AGARWAL A, BARHAM P, et al. Tensorflow: Large-scale machine learning onheterogeneous distributed systems [J]. Computer Vision, arXiv: 1603. 04467, 2016.

Authorized licensed use limited to: SUNY AT STONY BROOK. Downloaded on August 11,2020 at 16:28:02 UTC from IEEE Xplore. Restrictions apply.

[6] M. Kassner, W. Patera, and A. Bulling. 2014. Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication (UbiComp '14 Adjunct). ACM, New York,NY, USA, pp. 1151-1160.

[7] Kou Dalei, Quan Jichuan, Zhang Zhongwei. Research on progress of target detection framework based on deep learning [J]. Computer Engineering and Applications, 2019, 55 (11): 25-34.

[8] Liu W, Anguelov D, Erhan D, *et al*. SSD: Single shot multibox detector. arXiv preprint arXiv:1512.02325, 2015

[9] Howard AG, Zhu ML, Chen B, *et al*. MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861, 2017.

[10] Shu Na. Research on camera calibration method [D]. Nanjing University of Science and Technology, 2014.

[11] R EMPEL D， WILLMS K， ANSHEL J， et al. The effects of visual display distance on eye accommodation， head posture， and vision and neck symptoms [J] . Human Factors， 2007， 49(5) : 830-838.

[12] Bouguet J Y. Camera calibration toolbox for Matlab [EB/OL]. (2004-10-13)[2007-06-26]. http://www.vision.caltech.edu/bouguetj/calib_doc/index.html.

Authorized licensed use limited to: SUNY AT STONY BROOK. Downloaded on August 11,2020 at 16:28:02 UTC from IEEE Xplore. Restrictions apply.