Volume: 14 Issue: 05 Pages: 5651 - 5657 (2023) ISSN: 0975-0290

Driver's Gaze Estimation Algorithm Based on Deep Learning

College of Computer and Communication, Hunan Institute of Engineering,Xiangtan 411104, China
DENGAo
Email:1548613569@qq.com
YANGMeng-hao
Email:1169352664@qq.com
ZENG Gui
Email:3209959946@qq.com
LUOYin
Email:571816459@qq.com
HUYing
Email: huying1983@hnie.edu.cn

----------**ABSTRACT**----------
In traffic safety accidents, 80% of accidents are caused by driver distraction,there is an important mapping relationship between the direction of human sight and the focus of attention,studying the driver's gaze estimation algorithm can reduce the incidence of vehicle traffic accidents.In this paper, a driver's gaze estimation system based on deep learning technology is designed.For human eye feature localization, a multi-level localization method is adopted,firstly, the detection network is used to locate the face position, and then the key points of 68 faces are located based on the face image.In the aspect of the expression of the human eye's line of sight, a method of the expression of the human eye's line of sight based on the central point of the pupil and the direction vector of the human eye's line of sight is adopted,a convolution neural network is used to return the line of sight direction.On the MPIIGaze public data set and self-made data set, the experimental results show that the method can accurately estimate the human eye line of sight.

Keywords **:Convolutional neural network,CycleGAN, Face detection, Facial landmark detection,Gaze estimation**

Date of Submission: February 7, 2023 Date of Acceptance: March 1,2023

I. INTRODUCTION

During the driving process, the driver's distracted behavior will reduce the driver's vigilance, and thus affect the driver's handling of the driving conditions and emergency situations.The direction of sight has an important mapping relationship with the focus of attention, when the sight frequently leaves the front direction of driving, it is easy to cause dangerous operations such as lane line deviation, rear-end collision and emergency braking.Therefore, for the distracted driving behavior of the driver, an effective method is to capture and analyze the driver's attention through some method, and then judge whether the driver is distracted to make decisions to remind the driver of safe driving.

In this paper, a driver's gaze estimation system based on deep learning technology is designed, when the driver's line of sight leaves the normal driving area for a long time, the system will give the driver warning information, which can effectively reduce the occurrence of distracted driving behavior and improve the driving safety.The contributions are mainly as follows:

(1) Aiming at the problem that human eye view data is difficult to be labeled manually on a large scale, a method of human eye view data generation based on UnityEyes software generation and CycleGAN(Cycle Generation Adversary Networks) model rendering is proposed.

(2)For human eye feature location, a multi-level location method is adopted. Firstly, the face location is located using the detection network, and then the key points of 68 human faces are located based on the face image.

(3)In terms of the representation of the human eye's line of sight, compared with the rough representation of the human eye's line of sight such as the region method, a more accurate representation of the human eye's line of sight based on the pupil center point and the direction vector of the human eye's line of sight is adopted, and the direction of the line of sight is returned through a convolution neural network(CNN).

II. RELATED WORK

The driver's gaze estimation task is of great significance to improve the safety of the driver's driving process. In recent years, the relevant research has gradually increased. The driver's gaze estimation task mainly includes two methods: gaze estimation based on wearable devices and head-eye features.The wearable gaze estimation method mainly infers the eye movement information of the tester by measuring the changes of eye and facial pressure and electrical signals[1][2].The main work of the gaze estimation method of head-eye feature fusion focuses on pupil center detection and head-eye feature fusion.The early pupil feature detection method mainly depends on the bright pupil response of the near-infrared illuminator to determine the exact position of the pupil[3], but this method

Volume: 14 Issue: 05 Pages: 5651 - 5657 (2023) ISSN: 0975-0290

is vulnerable to the influence of natural light in the daytime.Tawari et al. [4]proposed a detection model that can train iris center using HoG work,Vicente et al. [5] detected eye feature points including eye contour points and pupil points through SDN tracker, and then estimated eye features through 3D eye model, but the detection effect of this method decreased significantly when head rotation was large.In order to process eye images with large head rotation, Yafei Wang et al. [6]combined the eye featuresobtained by sparse coding with the driver's head features.

In recent years, there are more and more researches on the fusion of head-eye features with depth learning algorithm to estimate the direction of vision of the tester.Choi et al. [7] completed the task of facial feature detection through the combination of Haar feature facial detector and MOSS tracker, and then input the facial image into the five-layer CNN to classify the driver's nine vision areas.Haopin Deng et al. [8] used two CNN networks to train the head posture and eye movement respectively, and then combined the head posture and eye network through the staring transformation layer to finally obtain the unconstrained line of sight direction.Yihua Cheng et al. [9] evaluated the gaze direction of both eyes respectively through asymmetric regression network, and proposed corresponding adaptive adjustment strategies according to the evaluation effect of both eyes during the optimization process.The method based on facial features uses face detection algorithm and facial key point location algorithm to determine driving status[10][11],The state detection based on facial features has high accuracy, low cost and easy implementation, which is the most concerned method at present.

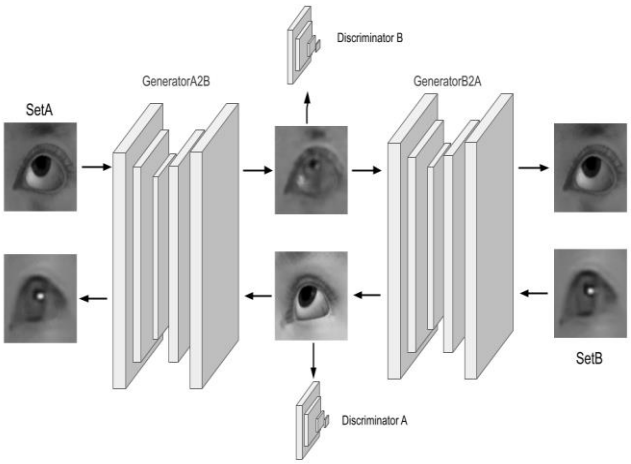
III. PROPOSED WORK

3.1 System architecture

The system only uses the video image collected by a single near-infrared camera of non-intrusive equipment as input data, so that the system can adapt to a variety of scenarios.The data collection method is software data generation plus CycleGAN rendering,and manual annotation of real human eye data.Firstly, we use the CNN to locate and intercept the face in the image, face location is the basis of processing face information.The intercepted face image is sent into the CNN to return the key points of the face, including eyes, nose, mouth, cheeks and other key points,In this way, the position information of the human eye can be extracted from the key point information of the face, and the location of the human eye from the whole image can be realized, this also provides accurate human eye image support for the human eye line of sight estimation model to achieve high-precision line of sight estimation.Finally, theCNN is used to regress the human eye image.

3.2 Data generation and rendering

It is difficult to manually label the direction data of human eyes on a large scale,the method of generating human eye data through UnityEyes software can provide a largenumber of human eye data with accurate human eye direction for the vision estimation model, and solve the problem that the human eye direction data is difficult to label.The style of the human eye data generated by this method is similar to the style of color camera collection, which is very different from the style of near-infrared camera.In order to unify the style of training data and application scenarios, and increase the accuracy of the human eye direction estimation model,aCycleGAN model is designed to realize the conversion from color style to near-infrared style[12],the color style of the generated human eye image is converted to the near-infrared style as the training data of the line-of-sight estimation model.Fig.1



shows the CycleGANnetworkmodel structure.

Figure 1. CycleGANnetwork model structure diagram

The CycleGAN model consists of a pair of GAN models, including two generators and two discriminators. The two generators realize the mutual conversion between the two styles, and the role of the two discriminators is to implement constraints on the two generators.Set A color style image is converted into set B near-infrared style data through GeneratorA2B,The discriminator determines whether the generated set B data belongs to the set A distribution, Then it is required that the generated data of set B can be restored to the data of set A through GeneratorB2A.At the same time, the data of set B should also go through the same steps to train GeneratorA2B and GeneratorB2A to achieve the function of image style conversion between different sets.When the discriminator is unable to determine whether the data generated by the generator is real data, the CycleGAN model reaches nash equilibrium and the training is completed.

3.3 Driver's eye positioning

The driver's eye location module is composed of a face detection model and a face key point model, in order to extract the human eye image from the whole image. Firstly, locate the position information of the face from the whole image, and then locate the key points of the face from the captured face image to obtain the position information of the human eye.

3.3.1 Driver face detection

The face images collected in the real world may present different sizes due to different distances from the camera,in order to improve the detection rate of various sizes of faces as much as possible,the SSD model with multi-scale fusion

Volume: 14 Issue: 05 Pages: 5651 - 5657 (2023) ISSN: 0975-0290

is used as the model for face detection[13].The SSD network model uses VGG10 as the basic model, input resolutionis 256×144.Feature fusion is carried out between three feature maps sampled at 8 times, 16 times, and 32 times, and target detection is carried out on three feature maps fused with multi-resolution location information and multi-layer semantic information, ensuring the detection effect of the model.Convert conv_ 3, conv_ 4, conv_ 5 features are fused with each other, and the fused output features will be used for detection and regression.On the network anchor setting, set anchor of different sizes for each output feature map,the anchor size are[25:28 30:30 35:40], [60:70 80:80 80:90], [110:120 120:120 130:140].Fig.2 shows the face detection network structure.

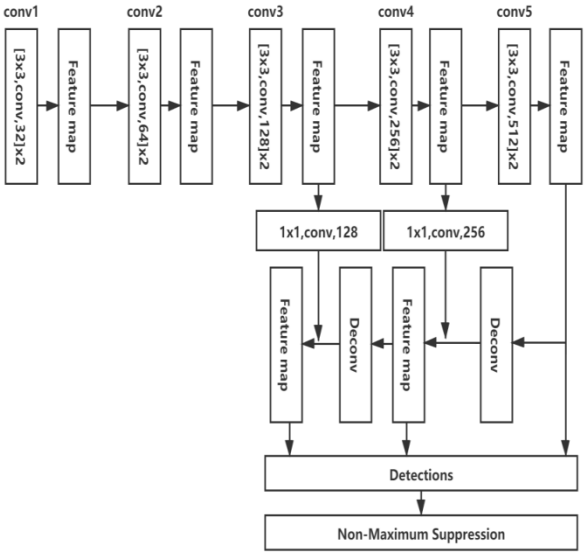


Figure 2.Schematic diagram of face detection network

3.3.2 Driver face key point location

The purpose of the face key point location model is to calculate the position information of each key point in the face from the input face image, which is essentially a regression model.ResNet is used as the model of face key point location of the basic network, ResNet network is connected by residual, which can deepen the network's ability to obtain stronger feature extraction and ensure that the model can learn well[14].The face key location model is composed of multiple bottlenecks,Fig.3 showsResNetnetwork structure.

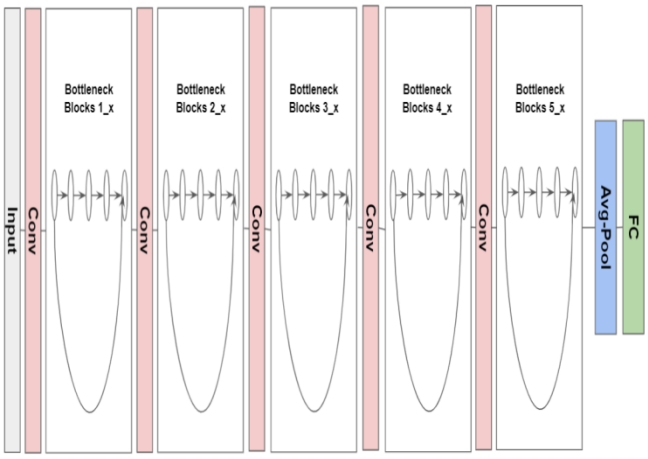


Figure 3.ResNetnetwork modelstructure diagram

In the first volume layer, the network adopts 7 × 7 convolution check input image for feature extraction,while the convolution kernel of the large receptive field is used as much as possible, the input features are down-sampled to reduce the size of the features and minimize the amount of network parameters.Then two bottleneck structures are used to extract features, the bottleneck structure does not down-sample the input features,bottleneck passes the input feature through 1 × 1 convolution reduction,then pass 1 × 1 operation of convolution dimension,make the parameters of the network model composed of multiple bottlenecks stacked in an acceptable range.After every two bottlenecks, the feature map will be down-sampled once to reduce the size of the feature map. At the same time, the number of channels of the feature map will be increased to increase the dimension of the extracted information.After the main part of the network feature extraction is calculated, the output feature map is expanded into a one-dimensional feature vector through a mean pooling layer, and then the position information of 68 key points of the face is regressed through a full connection layer. Parameter configuration of ResNetnetworkare shown in Table 1.

Table 1: Parameter configuration of ResNet network

layer name	output size	model
Conv-1	64×64	7×7, 64, stride 2
Conv-1-x	64×64	$\begin{bmatrix}1\times1,16,\text{stride}1\\3\times3,16,\text{stride}1\\1\times1,64,\text{stride}1\end{bmatrix}\times2$
Conv-2	32×32	3×3, 128, stride 2
Conv-2-x	32×32	$\begin{bmatrix}1\times1,32,\text{stride}1\\3\times3,32,\text{stride}1\\1\times1,128,\text{stride}1\end{bmatrix}\times2$
Conv-3	16×16	3×3, 128, stride 2
Conv-3-x	16×16	$\begin{bmatrix}1\times1,32,\text{stride}1\\3\times3,32,\text{stride}1\\1\times1,128,\text{stride}1\end{bmatrix}\times2$
Conv-3	8×8	3×3, 256, stride 2
Conv-4-x	8×8	$\begin{bmatrix}1\times1,64,\text{stride}1\\3\times3,64,\text{stride}1\\1\times1,256,\text{stride}1\end{bmatrix}\times2$
Conv-5	4×4	3×3, 256, stride 2
Conv-5-x	4×4	$\begin{bmatrix}1\times1,64,\text{stride}1\\3\times3,64,\text{stride}1\\1\times1,256,\text{stride}1\end{bmatrix}\times2$
Ave-Pooling	1×1	4×4, 256, stride 1
FC	136	

Volume: 14 Issue: 05 Pages: 5651 - 5657 (2023) ISSN: 0975-0290

Considering that the beginning of face key point training is that the regression of the model to the key points is relatively rough, the deviation of the key points is relatively large, and the loss of the network is relatively large and changes greatly, but with the further training of the model, the regression of the key points will be more accurate, the loss of the network is relatively small, and the change is also relatively small.In order to avoid loss explosion due to large error at the beginning of training, Wing loss is used as the loss function.Wing loss function uses ln x functionto enhance the influence of small error, and the derivative of function gradually increases as it approaches 0[15].For large error data, Wing loss is similar to L1 loss,Wing loss is defined as:

$$wing(x) = \begin{cases} wln\left(1 + \frac{|x|}{\varepsilon}\right) & if|x| < w \\ |x| - C & otherwise \end{cases} \quad (1)$$

3.4 Gaze estimation

Gaze estimationis to retrieve the position of the central point of the pupil and the line of sight direction vector from the image of the human eye.In this paper, SE-ResNet network is used to realize the regression of the direction of human eye sight,SE module has spatial attention mechanism[16],embedding SE module into ResNet network structure can make the region of interest of the network contain more useful information and make the network training more efficient.Fig.4 showsSE-ResNetnetwork structure.

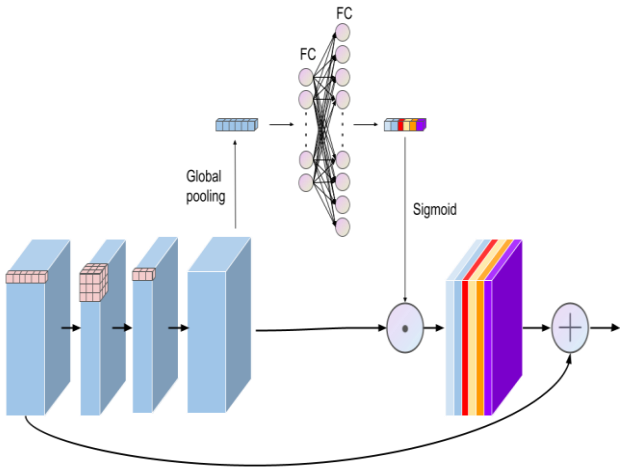


Figure 4.SE-ResNetnetwork modelstructure diagram

Using single eye image as model input, compared with using two eyes as model input, single eye as model input can shield some useless information between two eyes, which increases the efficiency of extracting effective information from the model.Take the64×64 size eye picture as input,the backbone network is composed of multiple SE-ResNet modules, and each SE-ResNet block uses a convolution layer with a step size of 2 to reduce the size of the feature map,then the extracted feature map is pooled to reduce the feature dimension, and finally the line of sight is returned through a full connection layer.Parameter configuration of SE-ResNetnetwork are shown in Table 2.

Table 2: Parameter configuration of SE-ResNet network

layer name	output size	model
Conv-1	32×32	7×7, 64, stride 2
SE-Res1-x	32×32	$\begin{bmatrix} Conv, 1\times1, 16 \\ Conv, 3\times3, 16 \\ Conv, 1\times1, 64 \\ FC, [16, 64] \end{bmatrix} \times 2$
Conv-2	16×16	3×3, 128, stride 2
SE-Res2-x	16×16	$\begin{bmatrix} Conv, 1\times1, 32 \\ Conv, 3\times3, 32 \\ Conv, 1\times1, 128 \\ FC, [32, 128] \end{bmatrix} \times 4$
Conv-3	8×8	3×3, 128, stride 2
SE-Res3-x	8×8	$\begin{bmatrix} Conv, 1\times1, 32 \\ Conv, 3\times3, 32 \\ Conv, 1\times1, 128 \\ FC, [32, 128] \end{bmatrix} \times 6$
Conv-4	4×4	3×3, 256, stride 2
SE-Res4-x	4×4	$\begin{bmatrix} Conv, 1\times1, 64 \\ Conv, 3\times3, 64 \\ Conv, 1\times1, 256 \\ FC, [64, 256] \end{bmatrix} \times 4$
Conv-5	2×2	3×3, 256, stride 2
SE-Res5-x	2×2	$\begin{bmatrix} Conv, 1\times1, 64 \\ Conv, 3\times3, 64 \\ Conv, 1\times1, 256 \\ FC, [64, 256] \end{bmatrix} \times 2$
Max-Pooling	1×1	2×2, 256, stride 1
FC	4	

The direction of the line of sight can be determined by the spatial position of the pupil and eye center in the camera coordinate system.The deviation angle between the predicted gaze direction vector and the real gaze direction vector in 3D space is used as the evaluation index of the gaze estimation model.The SE-ResNet network outputs a four-dimensional vector, including the abscissa of the x pupil center point, the ordinate of the y pupil center point, the pitch angle and the yaw angle, which represent the rotation angle about the x axis and the y axis, respectively.The center point of the pupil is taken as the origin of three-dimensional space coordinates, and the direction vector of the human eye line of sight is an arbitrary unit vector starting from the origin of coordinates.The calculation formula of line of sight direction vector defined as:

Volume: 14 Issue: 05 Pages: 5651 - 5657 (2023) ISSN: 0975-0290

$\vec{d}= (x, y, z) = (pitch \times 1, yaw \times 1, \sqrt{1 - pitch^2 - yaw^2})$(2)

The pitch and yaw are the angles of the line-of-sight direction vector.

IV. EXPERIMENT RESULTS AND DISCUSSION

4.1 Data set

The gaze estimation training data set consists of two parts of data generated by UnityEyes software and manually labeled data.Due to the difficulty and low efficiency of manual annotation, only 4000 pieces manual annotation data were collected.UnityEyes software can efficiently generate multi-angle human eye data. 40000 pieces generated 3D human eye data were used in this paper, and renders it into near-infrared style through the CycleGAN model, which is similar to the human eye data collected by near-infrared cameras in the real world.2000 pieces of data are randomly selected from 4000 pieces of real human eye sight direction data as the test set, and the rest 42000 pieces of data are used as training data.The data of 2000 pieces test sets are divided into three subsets:frontal collection data, side collection data and glasses data.

Figure 5.True line-of-sight direction data

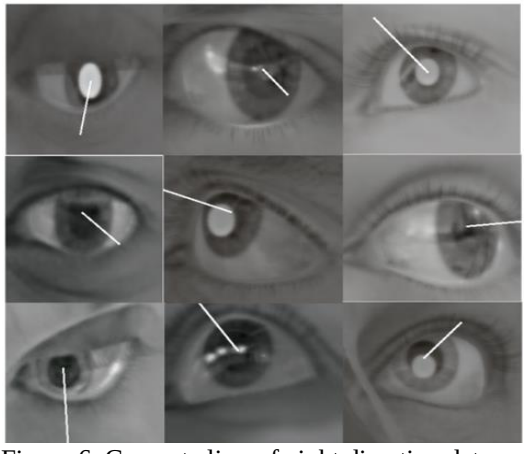


Figure 6. Generate line-of-sight direction data

4.2 Training parameters

To make the model converge better, the batchsize is set to 128, 240000 iterations are trained, and the learning rate attenuationis set at 120000 and 200000 iterations, and the attenuation parameter is 0.1.Configuration parameters for model trainingare shown in Table 3.

Table 3: Configuration parameters for model training

Initialize learning rate	0.05
Optimization function	Adam
Momentum	0.9
Gamma	0.1
Weight decay	0.0005

4.3 Experiment analysis

Figure 7 shows loss value change curve,it can be seen from the loss decline curve that when the first learning rate of 120000 iterations decays, the loss curve has an obvious drop jitter, and the fluctuation range of loss has been very slight at 200000 iterations, indicating that the model has fully converged at this time.

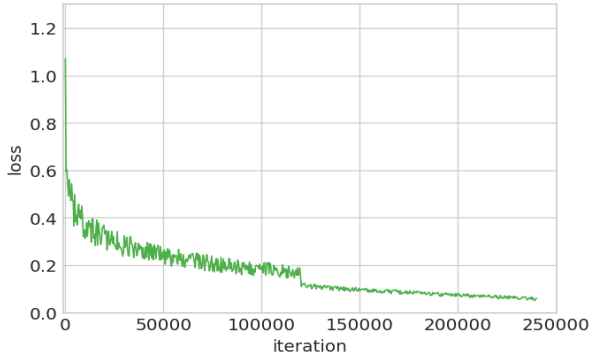


Figure 7.Loss value graph

In this paper, the angle between the real line of sight vector and the predicted line of sight vector in the three-dimensional space is used as the evaluation index.The larger the angle between the two vectors, the more inaccurate the prediction of the line of sight direction. The smaller the angle between the two vectors, the closer the predicted line of sight direction is to the real line of sight direction.The deviation of the line of sight angle for the front data collection, the side data collection and the data collection with glasses are respectively 7.567, 7.584, 10.517.In order to verify the feasibility of the model proposed in this paper, comparative experiments were carried out with EOG[17] and CNN [18] underMPIIGaze public dataset. The experimental results are shown in Table 4, The method proposed in this paper introduces the spatial attention mechanism, which can complete the direction of vision estimation under the interference of different skin color, light, image blur, etc.the experimental results verify the effectiveness of the model proposed in this paper.

Table 4: Comparison of different training algorithms

Method	Deviation angle of sight direction
EOG	10.800
CNN	10.210
Proposed work	9.221

Volume: 14 Issue: 05 Pages: 5651 - 5657 (2023) ISSN: 0975-0290

V. CONCLUSION AND FUTURE WORK

A driver's gaze estimation system based on deep learning is designed in this paper, which can accurately estimate the direction of the driver's line of sight, thus providing technical support for the corresponding warning of dangerous driving actions such as distracted driving.In order to improve the fidelity of the data, the CycleGAN model is used to realize the function of converting high-quality color images to near-infrared images.The multi-scale feature fusion method is used to detect the face of the whole image, and the accurate location of human eye information is realized through the face key point location model.A ResNet network structure with spatial attention mechanism is proposed, which can effectively regress the pupil center point and the line of sight direction vector.The experimental results verify that the method proposed in this paper has a high accuracy rate for gaze estimation, but the method proposed in this article still has some shortcomings, The future work mainly focuses on the following two aspects:(1)Combining the head posture information and the human eye information, the human eye line of sight is estimated. There is a strong correlation between the direction of the human eye's line of sight and the head posture, introduction of the head posture information will enable a more accurate estimation of the human eye's line of sight.(2)Get more large-scale human vision data,exploring a more efficient and accurate method for human eye line of sight direction labeling,generate data closer to the real near-infrared human eye image.

ACKNOWLEDGEMENTS

This study was supported by 2021 National College Student innovation and entrepreneurship training program project"Research and implementation of driver's gazeestimation system based on deep learning", 2022 National College Student innovation and entrepreneurship training program project"Research and implementation of traffic sign recognition in complex urban scenes",2021 Hunan Provincial Science and Technology Innovation Talents Plan Undergraduate Science and Technology Innovation and Entrepreneurship Project "Hunan Institute of Engineering New Engineering Talents Science and Technology Innovation and Entrepreneurship Ability Training Base"(XKJ [2021] No. 40, Project No.: 2021RC1011),2022 Scientific research project of Hunan Provincial Department of Education(22B0735)"Research on Deep Learning Method of Sign Language Recognition Based on Attention Mechanism".

REFERENCES

[1]. EC Lee, RP Kang, CW Min, J Park,Robust Gaze Tracking Method for Stereoscopic Virtual Systems,*HCI Intelligent Multimodal Interaction Environments*, Beijing,China,2007,700–709.

[2]. Zhou X R. *Research on eye tracking algorithm based on headworn eye tracker*,doctoral diss.,Harbin University of Technology, Harbin ,2019.

[3]. Ji Q, Yang X J, Real-Time Eye, Gaze, and Face PoseTracking for Monitoring Driver Vigilance, *Real-Timelmaging*,*8*(5), 2002,357-377.

[4]. Tawari A, Chen K H, Trivedi M M, Where is the driver looking: Analysis of Head,Eye and Iris for Robust Gaze Zone Estimation,*IEEE International Conference on Intelligent Transportation Systems*, Qing Dao,China,2014,988-994.

[5]. Vicente F, Huang Z, Xiong X, etal, Driver Gaze Tracking and Eyes Off the Road Detection System, *IEEE Transactions on Itelligent Transportation Systems, 16(4)*,2015,2014-2027.

[6]. Wang Y, Zhao T, Ding X, et al, Head pose-free eye gaze prediction for driver attention study,*2017 IEEE International Conference on Big Data and Smart Computing*, Jeju,KR,2017,42- 46.

[7]. Choi I H, Hong S K, Kim Y G, Real-time categorization of driver's gaze zone using the deep learning techniques, *International Conference on Big Data and Smart Computing*, Hong Kong,China,2016,143-148.

[8]. Deng H, Zhu W, Monocular Free-Head 3D Gaze Tracking with Deep Learning and Geometry Constraints,*2017 IEEE International Conference on Computer Vision*，Venice, Italy,2017,3162-3171.

[9]. Cheng Y, Lu F, Zhang X, Appearance-based gaze estimation via evaluation-guided asymmetric regression,*Proceedings of the European Conference on Computer Vision*, Munich, Germany, 2018,100-115.

[10]. Chen Z L,*Design and implementation of fatigue driving detection system based on facial features*, master diss.,Xi'an Technological University, Xi'an,2022.

[11]. Ma X T, Fei S M, Research on fatigue driving state detection based on facial features and deep learning, *Electronic test, 1(11)*,2021,33-36.

[12]. ZhuJ Y, Park T, Isola P, etal, Unpaired Image-to-Image Translation using Cycle-ConsistentAdversarialNetworks,*2017 IEEE International Conference on Computer Vision*,Venice, Italy,2017,2223–2232.

[13]. Lin TP,Girshick R,et al, Feature pyramid networks for object detection, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*,Hawaii US, 2017, 2117-2125.

[14]. Zhang H W,HuY,Zou Y J,et al,Fingerspelling Identification for American Sign Language Based on Resnet-18,*Int. J. Advanced Networking and Applications,1(13)*,2021,4816-4820.

[15]. ZhenH F,Josef K,Muhammad A,et a, Wing Loss for Robust Facial Landmark Localisation with Convolutional Neural Networks,*IEEE Conference on Computer Vision and Pattern Recognition*, UT, USA, 2018,2235-2245.

[16]. Hu J, Shen L, Sun G, et al, Squeeze-and-excitation networks, *IEEE Conference on Computer Vision and Pattern Recognition*, UT, USA, 2018,7132-7141.

[17]. Yan B,Yu S C, et al,Gaze Estimation Method Based on EOG Signals,*IEEE2016 Sixth International*

Volume: 14 Issue: 05 Pages: 5651 - 5657 (2023) ISSN: 0975-0290

Conference on Instrumentation & Measurement, Computer, Communication and Control, Harbin, China,2016,443-448.

[18]. Zhuang Y Y, Zhang Y C, et al, Appearance based gaze estimation using separable convolution neural networks, *IEEE 2021 Advanced Information Technology, Electronic and Automation Control Conference*,Chongqing,China, 2021,609-612.

Biographies and Photographs

DENGAo is currently pursuing his Bachelor's degree (Communication Engineering) from Hunan Institute of Engineering, Xiangtan, China. His current research areas are: Computer Vision, pattern recognition, machine learning and Deep learning.