

Vision Detection System Based on Mixed Recognition of Gesture and Speech

Jun Guo *, Jingjing Hu**, Yuhua Lai**, Kaikai An**, Zhengxiao Zhang**

*Software College, Northeastern University, Shenyang Liaoning Province China

**School of Computer Science and Engineering, Northeastern University, Shenyang Liaoning Province China

guojun@mail.neu.edu.cn, xian236024@163.com, 1101293857@qq.com, 1355113754@qq.com, 2413174539@qq.com

Abstract—In view of the shortage of traditional visual acuity detection, the shortage of visual acuity detection instruments, the time consuming and the lack of integration, we designed and realized the intelligent visual acuity detection system. The system realizes the detection of human vision by gesture and speech mixed recognition technology. The system includes edge acquisition equipment, gesture and speech recognition, cloud-side cooperative data processing and voice broadcast. Through Cloud platform to develop algorithm skills, using edge equipment for model deployment and data acquisition. According to the visual icon on the display screen E the subjects make corresponding gestures or voice answers. The edge device converts the collected data information to the cloud, and the display displays and broadcasts the visual information. The system saves the manual detection of routine medical treatment, is convenient, accurate, time-saving and labor-saving, carries on the archiving and the big data analysis to the visual acuity information, realizes the visual acuity detection integration, the intelligence, simultaneously carries on the dynamic gesture speech interaction, increases the interest, is easier to popularize.

Keywords—Big Data Analysis, Cloud-end Collaboration, Hybrid Recognition, Vision Detection

I. INTRODUCTION

In the traditional vision testing process, schools need to organize students, contact the physical examination center, and then organize the physical examination center. The doctor tested the students' eyesight and recorded the test results in the table. After the physical examination, it was entered into the computer to carry on the statistical analysis, the visual acuity data input, the analysis, the establishment file did not realize the integration[1]. All of these make large-scale vision testing impossible to complete in a short time. To sum up, the traditional visual acuity detection process has the problems of long time period, more human resources, low digital efficiency and so on, which can not meet the social needs.

Therefore, we propose a vision detection method based on gesture and speech hybrid recognition. The intelligent vision detection equipment developed by this project not only has the necessity, but also has the urgency and the market prospect and value potential behind it.

II. RESEARCH BACKGROUND

At present, the high incidence of myopia among children and adolescents, I China's youth vision situation is particularly

severe, juvenile myopia rate has ranked first in the world and is still on the rise[2]. In order to deal with the problem of the younger age of the myopia group, the Ministry of Education issued a circular on the implementation plan of comprehensive prevention and control of children and adolescents to strengthen myopia prevention and control, while the guide on appropriate techniques for the prevention and control of myopia among children and adolescents further pointed out the need for "regular visual monitoring", ensuring one person, one file and real-time transfer with the change of student status", and stipulated that "strict implementation of the system of physical examination of students and the system of visual monitoring twice a semester"[3]. It is very important and urgent to establish a long-term mechanism of comprehensive prevention and control of myopia.

In the market of intelligent vision detection, vision measurement is mainly used in the form of Mini Programs and APP software, and the method of measuring vision is basically the same, that is, finger sliding measurement. Mini Programs includes "check vision", "EYE vision treasure", "vision E finger check", "physical examination treasure", "super vision meter", "vision stick" and so on. There are also a few electronic vision detectors, which have many external equipment, complicated wiring, large volume, and are not easy to carry, which are not conducive to the conduct of vision testing[4]. At present, many companies in China have also announced a large number of autonomous vision testing instruments, but most of them are not put into the market and are widely used. At present, the Internet market has not opened up the intelligent market of school vision detection, and the direction of intelligent vision detection, standard and credible third party evaluation is still a vacancy[5].

Therefore, the system is mainly designed for primary and secondary school visual acuity detection system, dedicated to solve the visual acuity problems of young people due to the uneven equipment, time consumption, low digital efficiency and so on. This design through the construction of intelligent vision detection system, and the establishment of personalized vision files, visual value change trend of big data analysis, in response to national policies, improve the coverage of visual examination services, personalized push visual data to parents, improve parents' understanding of children's vision, improve children's awareness of eye care, and provide protection for adolescent vision health.

III. SYSTEM COMPOSITION

A. Composition of the system

This project will combine visual analysis with machine learning. By changing the size and square of the letter displayed by the touch screen. After the interactive elements of gesture and speech recognition record are sent to the cloud, the meaning of the action is explained and analyzed by Huawei Cloud machine learning algorithm, and the task of visual acuity detection is completed.

The system includes edge acquisition equipment, gesture and speech recognition, cloud-edge collaborative data processing, vision icon conversion and voice broadcast modules. The system composition of the system is shown in figure 1.

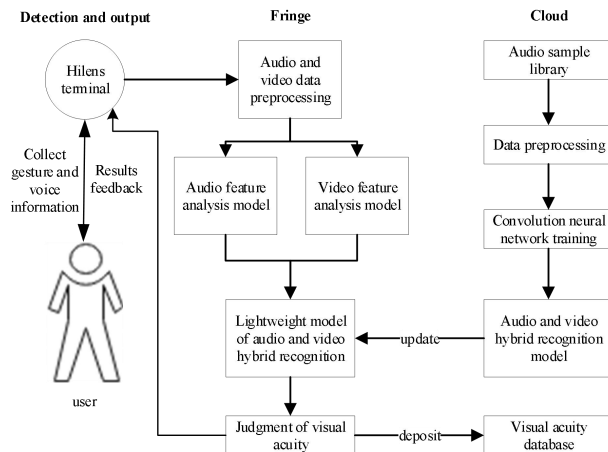


Figure 1. Composition of the System

The display screen is connected HiLens the corresponding interface to display and convert the vision icon "E", as well as the visual information display. The subjects make corresponding gestures or voice answers according to the icon they see, and the display displays and broadcasts the visual information.

HiLens terminal detection equipment is placed in the designated position of the distance display screen, which is directly aligned with the hand of the subject to identify, collect gesture information, conduct gesture analysis, and compare with the uploaded learning sample data set of gesture recognition. Determine whether the standard gesture is met to feedback visual information.

At the same time, the sound recognition hardware is connected to the Hilens to collect the voice information of the detected person, to carry out speech analysis, to recognize the meaning expressed by the detected person, and to compare with the uploaded learning sample data set of speech recognition. Get effective information. According to the judgment result, compare the human visual acuity value corresponding to the medical test standard, so as to obtain the visual acuity information of the tested person and display the final result[6,7].

B. Terminal information acquisition equipment

HiLens hardware preset development framework, encapsulate multimedia library, operator library, model management library, provide audio and video management, external interface management, data automatic acquisition and other capabilities; pre-adaptive ROS system, more convenient development of automatic control class applications; compatible with TensorFlow、Caffe and other mainstream deep learning framework; support Python and C++ mainstream languages.



Figure 2. Huawei HiLens Kit Multi-modal AI development suite

The following table shows the specific parameters of the equipment.

TABLE 1. HiLENS KIT MAIN PARAMETERS

Specifications	Description
NPU	scend 310 processor ,16 ToPS CINT8),8GB memory
CPU	Hayes Hi3559A processor ,4 GB memory
Storage	Support 1 Micro SD card; support SD3.0, maximum support rate SDR50, maximum capacity 2 TB
Camera	200W Pixel 1080P
Operating system	EulerOs
Wireless network	WiFi support
Network interface	GE RJ45*1
Display interface	1 HDMI interface
Audio output interface	Typical power consumption 20 W, maximum power consumption 36 W
Power consumption	Typical power consumption 20 W, maximum power consumption 36 W
Size	55.5*155*110 mm
Weight	400 g

C. Data acquisition and processing

Data processing includes image information processing and audio information processing collected by the terminal, as well as data conversion, transmission and cloud storage, in addition to big data analysis based on visual information database.

The training model is obtained and deployed to the terminal by convolution neural network for gesture recognition. The processing of audio and video data collected by the terminal

adopts cloud-end cooperative processing method. Data preprocessing and feature analysis are carried out at the edge, and the model is lightweight.

On the basis of collecting a large amount of data information, the capacity of the database is continuously expanded. Based on the big data and Huawei Cloud platform, the detected visual value information is archived under the premise of reaching agreement with the corresponding departments and individuals to ensure the security of the data, and the individual data tracking, collective data statistics and the analysis of the visual situation of teenagers in a certain area are carried out.

D. Visual acuity detection based on cloud-end collaboration

In order to reduce the network delay and bandwidth of video data in the transmission process, improve the speed of vision detection, enhance the real-time of vision detection, and improve the accuracy of vision detection, this project intends to study the framework of gesture recognition vision detection under cloud cooperative computing system.

1) Edge end:

Gesture recognition of the image data collected by the Hilens platform, further action recognition, and establishment of scoring mechanism combined with time and other factors, real-time evaluation of the detected results, and send the results to the cloud.

2) Cloud:

Store a large number of detected data and fuzzy processing data, and carry out big data analysis of fuzzy processing data; at the same time, establish a rich vision detection model and send it to the edge.

Using cloud to do fuzzy data analysis, the load pressure at the edge end is alleviated. The edge end evaluates the visual acuity of the detected person in real time, realizes the independent decision of the edge end, and ensures the real-time data processing. The timeliness and reliability of the system are improved by cloud edge collaborative calculation.

IV. HAND AND SPEECH HYBRID RECOGNITION

A. Composition of the system

Because the single mode RGB data is vulnerable to illumination, angle of view and other factors, the recognition rate is low. The video of gesture action is collected, the model training is carried out by deep learning, the convolution neural network realizes the sharing of convolution kernel, eliminates the requirement of artificial feature extraction, and then judges the visual value instead of the traditional remote control or key. To achieve a more simple and intelligent man-machine communication.

Because the recognized sound is only a few isolated words, and there is often noise interference in the test, we collect a certain number of subjects' sounds, and use the hmm-gmm to analyze the isolated words to get the model. The direction of the sound can be obtained by comparison, and the visual value can be judged instead of the traditional remote control or key.

We mainly use the open source code provided by the CMU in gesture recognition openpose, and transform it on its basis. Because of the lack of data set, we can only use the model provided by it. Because of the use of its model, it is difficult to modify its model, its environment is difficult to deal with when using openpose, and its processing speed is somewhat deficient. Therefore, after light weighting the model, The original processing video efficiency is changed from 1~3 fps to 5~9. At the same time, we also change the original video stream processing to image stream processing to increase its efficiency and accuracy, but it is still different in speed than expected. Therefore, we should further lightweight the model to reduce its processing time[8].

For speech recognition, because isolated words are recognized and there are fewer types of recognition, the HMM-GMM matching method is adopted. This method has high accuracy in identifying isolated words, and the model is easy to construct and process quickly. Therefore, we use iFLYTEK technology to build the model and increase the noise processing effect.

B. Terminal information acquisition equipment

1) Training data::

The training data set is *coco*, the data set contains the coordinates and type information of human *keypoint*. To generate *part* and *paf* training samples based on *keypoint* information[9].

Part training sample: let $S^*_{j,k}$ be the confidence map of j part of person k . $x_{j,k}$ is the ground truth coordinate information of the j th part of the k th person. The value of $S^*_{j,k}$ position P is defined as follows:

$$S^*_{j,k}(P) = \exp\left(-\frac{\|P - X_{j,k}\|_2^2}{\sigma^2}\right), \quad (1)$$

σ is the parameter used to control the distribution, and the *groundtruth confidence map* used to train *part* is to maximize the *confidence map* of each person.

$$S^*_j(P) = \max_k S^*_{j,k}(P). \quad (2)$$

In this way, if two different people have the same joint close to each other, their maximum value will not be affected (because the non maximum inhibition (NMS) algorithm is used to obtain the peak value as the joint position). In this way, we finally get the *groundtruth confidence map* $S^* \in R_w * h * 19$ as the training sample of the part. It contains 18 *part* and 1 background.

paf training sample: a total of 19 connections are defined. For each connection, the information of x direction and y direction is included. Let $L^*_{c,k} \in R_w * h * 2$ be the ground truth PAF of the climb of person K . The limb corresponds to part j_1 and part j_2 . For position P on $L^*_{c,k}$, if P is on limb C , it is defined as follows:

$$L^*_{c,k}[x_p, y_p, 2 * c] = \frac{(x_2 - x_1)}{[(x_2 - x_1)^2 + (y_2 - y_1)^2]^{\frac{1}{2}}} \quad (3)$$

$$L^*c, k[x_p, y_p, 2^*c+1] = \frac{(y_2 - y_1)}{\sqrt{[(x_2 - x_1)^2 + (y_2 - y_1)^2]^{\frac{1}{2}}}} \quad (4)$$

When p are not on the *limb* c :

$$L^*c, k[x_p, y_p, 2^*c] = 0, \quad (5)$$

$$L^*c, k[x_p, y_p, 2^*c+1] = 0 \quad (6)$$

The *ground truth paf* eventually used to train *paf* is to average the *paf* of each person.

$$L_c^*(P) = \frac{1}{n_c(P)} \sum_k L_{c,k}^*(P), \quad (7)$$

This way, if there are two different people with the same *limb* overlap, in the overlap part of the average, can take into account the occlusion *limb* and the information of the occluded *limb*. the final obtained groundtruth *paf* were used $L^*c, k \in R_w * h * 38$ as *paf* training samples.

Using the above training samples to train the neural network, the final output dimension of the network is $w * h * 75$, including *part heatmap* and *paf*.

2) Algorithm:

The algorithm consists of the following four steps:

(1) Neural Network Prediction

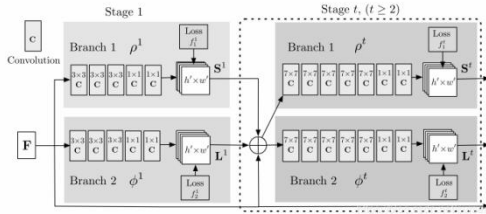


Figure 3. Neural network structure

The network structure is shown in the figure. Firstly, the feature is extracted by *vgg-19*, and then the multi-stage CNN with two branches is used. The first branch is used to predict the *confidence map st*, and the second branch is used to predict *PAFS Lt*. After each stage, the prediction and image features from the two branches are connected for the next stage [10].

In order to increase the receptive field while the local can not be blurred, the original image zooms into the neural network with four fixed *scale*, and the four predicted values are averaged.

In this way, we input $image \in R_w * h * 3$, output $heatmap_avg \in R_w * h * 19$ and $paf_avg \in R_w * h * 38$.

(2) found the joint

By Through the first step, we obtain the *heatmap* of the joint. In this step, the specific position of the joint is extracted from the *Heatmap* [11]. The NMS algorithm is used to obtain the peak value in the *heatmap*, and the confidence value is used as the *score* of the *part*. The output is:

$$all_peaks = [((h_0, w_0, s_0, 0), (h_1, w_1, s_1, 1), \dots)] \quad (8)$$

All values of the first part

$$[((h_i, w_i, s_i, i), (h_{i+1}, w_{i+1}, s_{i+1}, i+1), \dots)] \quad (9)$$

All values of the second part

(3) find a limb connection

In the second step, we obtain the joint information (position, score), in this step, we use the joint information and *paf* to obtain the limb connection. The model consists of 19 limbs, and the two *parts* and *paf* corresponding to each limb are determined. The *paf* information between the two *part* is integrated to obtain the confidence degree of the limb [12].

For example, to find all the left upper arm, there is already *part* information for the left shoulder and elbow, and *paf* information for the left upper arm. All the *part* in the left shoulder are matched with all the parts of the left elbow, and the integral of *paf* information between the two parts is calculated as the *score* of the connected *part*. Finally, all scores are sorted to determine whether the *part* are connected [13].

So we get all the connection information, and each connection can be seen as a limb.

(4) assembling the body

Having acquired all the limbs will have the same joint the body is regarded as the body of the same person.

3) Gesture recognition:

The gesture recognition method based on convolution neural network can accurately locate the selected joint point. The joint point coordinates output by the gesture recognition algorithm are converted, and the angle of the specific joint point is selected to judge the direction of the current gesture. Match the result to the "E" displayed on the current display to determine the visual value[14].

By Huawei Cloud the built-in machine learning algorithm to complete the collection of information processing. Combined with the experience accumulated in practical application, the project team combines the written gesture recognition algorithm with the Huawei Cloud open source gesture recognition and speech recognition algorithm to replace the variables and add interfaces to adjust the machine learning algorithm to improve the accuracy of gesture recognition and speech recognition.

V. SYSTEM TESTING

A. Experimental environment

All the experimental data in this paper are from Lenovo y700, to test the accuracy of gesture and speech recognition for different scenes. In mixed recognition, only when speech and gesture recognition are the same, otherwise the test fails to carry out the next test, and the exact number represents the same number as the test in the passed example.

B. Experimental results

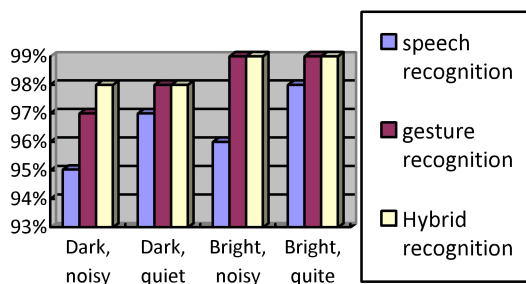


Figure 4. Experimental Results

C. Experimental Conclusions

It can be seen from the above table that the accuracy of gesture recognition or speech recognition will be greatly reduced when the environment changes slightly, but after the combination of these two recognition methods, the recognition accuracy will be greatly improved. A hybrid recognition method based on convolution neural network and hmm-gmm proposed in this paper can greatly improve the accuracy of recognition in actual situations, even in the case of poor environment. In this paper, when recognizing gestures, we do not directly use pictures to recognize gestures, but recognize human joints in pictures to calculate the direction of gestures, and only recognize isolated words in their speech recognition. It is difficult to recognize many times and can be trained by neural network to obtain more accurate model.

VI. CONCLUDING REMARKS

The intelligent vision detection system based on gesture and speech mixed recognition developed in this study is a highly intelligent, convenient and integrated visual acuity detection equipment with high performance and price ratio suitable for the public. The design is unique and the basic functions of the design are expected to be realized. The equipment can be used for visual examination in primary and secondary schools, and can also be used in various medical institutions. The equipment has good completeness, high efficiency and safety. If the equipment can be put into use, students can always grasp their visual health, parents will also pay attention to, must be able to deal with early prevention and treatment to avoid myopia or hyperopia and other eye diseases. This will also effectively alleviate the problem of vision decline of young people in our country, and on this basis, the vision can be further improved.

REFERENCES

- [1] Autonomous Visual Inspection Instrument System [D]; and Round. Shandong Construction University, 2012
- [2] Jianke network community. World Health Organization reports Chinese myopia 47% [EB/OL]. of Population <https://www.jianke.com/ykpd/2339986.html>
- [3] Circular of the General Office of the State Health and Health Commission on the Pilot Work on Appropriate Technologies for Prevention and Control of Myopia among Children and Adolescents [R]; and Beijing: National Health Commission of the People's Republic of China, 2019
- [4] Zhang Qin. The Design and FPGA Implementation of Autonomous Vision Detector Based on Gesture Recognition 2015.
- [5] Real-time Vision Monitoring and Health Technology and its Application Based on Intelligent J. Equipment Zhou Zhilong. China Hi-tech. 2019(03)
- [6] Zhou Zhilong. Health-care technology for real-time vision monitoring based on intelligent devices and its application [J]; and China Hi-tech. 2019(03):80-84.
- [7] GB 11533-2011, Standard Logistic Watch [S]. Table Beijing: China National Standardization Administration Committee, 2011.
- [8] Detection of Substation Hazard Behavior Based on OpenPose Human Attitude Recognition Zhu Jianbao, Xu Zhilong, Sun Yuwei, Ma Qingshan. Automation and instrumentation. 2020(02)
- [9] Research on Classification and Recognition of γ Photon Images Based on Deep Convolutional Neural Network [J]. 10 Wu Rong, Zhao Min, Sun Tong, Xu Jun, Yao Min. Mechanical Manufacturing and Automation. 2020(05)
- [10] Ryuji Tanaka, Chika Oshima, Koichi Nakayama. Intention inference from 2D poses of preliminary action using OpenPose [P]. Saga University, Saga, Japan; Saga University, Saga, Japan; Saga University, Saga, Japan, 2019.
- [11] ULDuan Junchen, Liang Meixiang, Wang Rui. Human pose recognition based on human skeletal point detection and multilayer perceptron [J]. Electronic Measurement Technology, 2020, 43(12): 168-172.
- [12] Ding Peifu, Zhan Lingchao, Hu Tianmin, Peng Haocheng, Zheng Yuchao. Study on Pedestrian Abnormal Posture Based on OpenPose [J]; and Digital Technology and Applications, 2019, 37(02): 107+109.
- [13] Zhu Hongkun, Yin Jiawei, Feng Wenyu, Hua Liang, Fei Keen, Zhang Kun. A Study and Application of Lightweight Real-time Human Posture Detection Model [J]. for Human Body Journal of system Simulation, 1-14.
- [14] Zhang Qin. Design [J]. of Vision Detection System Based on Gesture Recognition Computers and Telecommunications, 2019(03): 5-8+12.



Jun Guo, born in 1974, Ph.D., is an associate professor at Northeastern University. His research interests include edge computing, service computing and cloud computing.



Jingjing Hu, study as an undergraduate student in major of Computer Science and Engineering from Northeastern University, Shenyang, China, from 2018 to 2022. Her research interests include edge computing, Computer Graphics and so on.



Yuhua Lai, study as an undergraduate student in major of Computer Science and Engineering from Northeastern University, Shenyang, China, from 2018 to 2022. His research interests include Network Communications, Computer Graphics and so on.



Kaikai An, study as an undergraduate student in Software Engineering from Northeastern University, Shenyang, China, from 2018 to 2022. His research interests include Network Communications, Deep learning technology in image recognition and so on.



Zhengxiao Zhang, study as an undergraduate student majoring in the Internet of things engineering at Northeastern University, China. He entered Northeastern University in August 2018. His research interests include The hardware and software design and realization of wireless sensor network, application of wireless sensor network and so on