Intelligent Epidemic Prevention System Based on Voice and Gesture

Haiguang Chen, Mingxing Liu, Susheng He
Shanghai Normal University

Abstract---In the end of 2019, the new crown epidemic has outbreak, and medical personnel have become the most scarce resources in all countries. Protecting their lives is an urgent task. This paper takes the principle of contactlessness as the core, the speech recognition module based on the deep learning framework GRU+CTC and the gesture recognition module based on KNN and SVM as the core algorithm, and develops an intelligent epidemic prevention system based on voice and gestures. Experimental results show that whether it is a voice module or a gesture module, its accuracy and efficiency are relatively high.

Keywords: CNN+GRU+CTC, KNN, SVM, Voice, Gesture Recognition

1. INTRODUCTION

Since December last year, the novel coronavirus disease(COVID-19 for short) has spread on a large scale around the world, and the productivity of the "contactless economy" represented by artificial intelligence and machine learning has been indispensable in this war, which is widely used in community material transportation, automatic temperature measurement in public places, disinfection and antivirus, and hotel intelligent services. The principle of "no contact" is separated by a thick layer of glass inside and outside the epidemic isolation ward. Medical staff wear heavy protective clothing, masks and goggles. It is very inconvenient to communicate inside and outside. The use of walkie-talkies at night may also affect the isolation ward. The patient's rest inside.

This intelligent epidemic prevention system is based on the principle of "no contact" combined with artificial intelligence to design a gesture voice intelligent epidemic prevention system. Some routine tasks(such as injections, medicine feeding, etc.) are designed into simple gestures, and medical staff in the isolation ward make gestures to pass The system recognizes and converts into voice and sends it to the medical staff in the duty room in time, and transmits work information while keeping the isolation ward quiet. At the same time, the system also has a voice dialogue function, so that patients can communicate with the system. In addition, the algorithm framework of this paper is shown in Figure 1. The main contributions of this paper are:

(1) The intelligent epidemic prevention system developed by us is highly usable in the current epidemic environment;

(2) We combine the gesture module and the voice module to study, combining the advantages of the two, and also have the function of improving efficiency and saving time under normal circumstances.

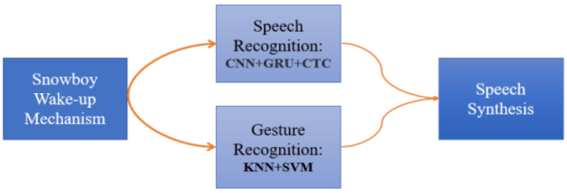

Figure 1 The basic flow of the algorithm

2. RELATED WORK

With the development of deep learning in recent years, speech recognition technology has also made considerable progress. DNN, CNN, and LSTM are some of the more mainstream directions. CNN[1] is introduced into speech recognition in 2012 by Ossama Abdel-Hamid. Initially, the convolutional layer and the pooling layer appear alternately, and the scale of the convolution kernel is relatively large, and the number of CNN layers is also Not much, mainly used to process and process features, so that they can be better used for DNN classification, as CNN shines in the image field, the application of VGGNet[2], GoogleNet[3] and ResNet[4,5], for CNN in speech Recognition provides more ideas, such as multi-layer convolution followed by pooling layer, reducing the size of the convolution kernel can enable us to train deeper and better CNN models. But CNN has no memory ability and can only handle a specific visual task, and cannot handle new tasks based on previous memories.

Recurrent Neural Network (RNN) is based on the idea of memory model. It is expected that the network can remember the features that appeared before, and infer the subsequent results based on the features, and the overall network structure is continuously circulating, hence the name Recurrent Neural Network The internet. Recurrent neural networks currently use the two most variants: LSTM[6] and GRU[7]. LSTM is a long short-term memory network. It is known from the literal meaning that it still solves the problem of short-term memory, but this short-term memory is relatively long and can solve the problem of long-term dependence to a certain extent. GRU will forget the door It synthesizes an "update gate" with the input gate. At the same time, the network no longer gives an additional memory state, but transmits the output result as a memory state continuously backwards. The input and output of the network become particularly simple.

Connectionist Temporal Classification(CTC)[8] is a temporal classification algorithm proposed by Graves et al. in 2006. Unlike the Cross Entropy Loss method commonly used by some traditional models, CTC allows the model to learn alignment operations by itself, thereby saving time and improving efficiency. Therefore, this paper is based on a speech recognition method based on a hybrid model of CNN+GRU+CTC.

At the same time, another major module of the system-gesture recognition module. The term gesture recognition refers to the entire process of tracking human gestures, recognizing their representations, and converting them into semantically meaningful commands[9]. Generally speaking, whether the way of collecting information from gesture interaction is contact or non-contact, the gesture interaction system can be divided into two types: contact-based sensors and non-contact-based sensors.

Gesture recognition based on touch sensors is usually based on technologies such as data gloves, accelerometers, and multi-touch screens that use multiple sensors. In 2004, Kevin[10] and others designed a wireless instrument glove "CyberGlove II" for gesture recognition; in 2007, Bourke[11] and others proposed a recognition system with an accelerometer to detect normal gestures in our daily activities. Gesture recognition based on non-contact sensors is usually based on the use of optical sensing, radar detection and other technologies. In 2002, Bretzner[12] and others proposed gesture recognition using a camera to collect multi-scale color features.

This article adopts non-contact gesture recognition, uses the camera to collect 10 kinds of gestures, and then processes the gesture pictures based on KNN and SVM algorithms to give the gestures different meanings, and speak the meanings by voice.

In today's global outbreak of the new crown pneumonia epidemic, the design of this intelligent epidemic prevention system can create a data set of some common questions and input it into the framework of a deep neural network for training, which is used to answer patient questions and reduce the pressure on medical staff; It is easy to contact and needs to be kept quiet to realize the function of transmitting information with gestures; of course, it can also realize the function of ordinary chat, which can be widely used in special epidemic environments or ordinary occasions.

3. ALGORITHM

This intelligent epidemic prevention system combines a voice module and a gesture module. The voice module first needs a wake-up module, which is implemented by the Snowboy wake-up module. Then the core algorithm of the voice module is a speech recognition algorithm based on the CNN+GRU+CTC framework. The gesture module is mainly implemented based on KNN and SVM. The main modules are described in detail below.

3.1 Snowboy Wake-up Mechanism

Snowboy is a highly customizable wake word detection engine that can freely create and train your own wake words; it can be used in real-time embedded systems, and it can always monitor even when offline, with high accuracy and low latency. Protect privacy at all times. Snowboy supports all Raspberry Pi (equipped with Debian Jessie 8.0), 64-bit MAC OS X, 64-bit Ubuntu (12.04 and 14.04), iOS, Android (ARMv7 CPU), etc. It can run a complete automatic speech recognition (ASR, Automatic Speech Recognition) to perform hot word detection, and can perform simple commands and control actions when triggered.

The intelligent anti-epidemic assistant designed in this project is developed using the Raspberry Pi 4B board and is equipped with Snowboy to configure voice wake-up. Since the native audio device of the Raspberry Pi cannot record and does not support voice input, choose to connect a driver-free USB microphone and a speaker to amplify the response. After preparing the voice wake-up model, we can install sox and port-audio's python binding to use the microphone, download snowboy and its python package, and successfully implement the voice wake-up function. Modify the callback function in the decoder main program to achieve the corresponding effect.

3.2 CNN+GRU+CTC Framework

Based on the original memory function of LSTM, GRU turns the input gate, forget gate, and output gate into two gates: update gate and reset gate. The function of the update gate is to control the state information of the previous processing moment and the current processing state. The degree of relevance, the larger the value, the greater the degree of relevance between the state information at the previous processing time and the current; the role of the reset gate is to control the degree of forgetting the state information at the previous processing time, and the smaller the value, the more the state information is forgotten. The structure of a GRU unit is shown in Figure 2

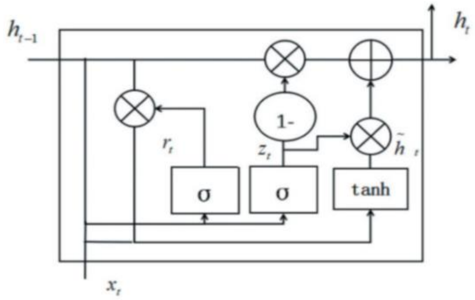

Figure 2 Unit structure of GRU

The state information of the update gate and the reset gate are obtained through the state h_{t-1} transmitted at time t-1 and the input x_t of the current unit at time t. The calculation process of obtaining the state information of the two doors is as follows, where z_t is the update gate, r_t is the update gate, σ is the Sigmoid function, and W is the weight:

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t])$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t])$$

After obtaining the state information of the two gates, first use the reset gate rt to get the reset $h_{t-1} * r_t$, then connect with the input x_t, get h_t through the tanh activation function, and then selectively add it to the current In the state, this process is the process of memory and forgetting. The value range of the update gate z_t is 0~1. The closer the value is to 1, the more information is memorized. The closer the value is to 0, the more information is forgotten. , And finally get the status information h of the current unit:

$$\widetilde{h_t} = \tanh\,(W \cdot [r_t * h_{t-1}, x_t]$$
$$h_t = (1 - z_t) * h_{t-1} + z_t * \widetilde{h_t}$$

CTC can be trained based on an input sequence and output sequence, and directly The most prominent point of the probability of outputting the predicted sequence is the introduction of the blank node, which is mainly used to model the silent, pause, and other parts without effective information to indicate the output state of the network when predicting uncertain information. There are also CTCs. An F transform. If an output sequence of the network can be mapped to the correct label sequence through F transform, then the output sequence is a CTC path. The process of F transform is: first remove the repeated labels between adjacent blank nodes in the sequence , And then remove the blank node. For example, the following transformation methods:

$$F(\otimes XY \otimes\otimes YY \otimes ZZ) = XYYZ(\text{where } \otimes \text{ represents blank node})$$

CNNs[13] are exceptionally good at capturing high level features in spatial domain and have demonstrated unparalleled success in computer vision related tasks. One natural advantage of using CNN is that it's invariant against translations of the variations in frequencies, which are common observed across speaker with different pitch due to their age or gender.

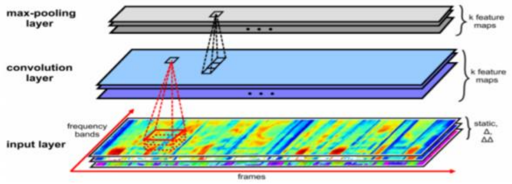

Figure 3: Convolutional network applied upon input features

3.3 Gesture Recognition

The gesture recognition module is the highlight module of this article. The system acquires gesture graphics through the camera, and preprocesses the image after denoising, skin color detection, binarization, morphological processing, and contour extraction operations. The gestures are collected first for data enhancement, and the image is pre-processed. The contours extracted in the processing are used to establish a gesture feature library; then as a training set for KNN and SVM model training; finally, the test set is used to verify the recognition effect and calculate the accuracy rate. The gesture recognition process can extract specific gestures, and then assign specific meanings to the gestures, and then return to the voice module after matching, and then output through speech synthesis

The image preprocessing part is mainly for perfect skin color detection, which requires complete and accurate recognition of the hand position in the image in a complex background. Skin color YCbCr color space is a commonly used color model for skin color detection, where Y represents brightness, Cr represents the red component in the light source, and Cb represents the blue component in the light source. The difference in appearance of human skin color is caused by chroma, and the skin color of different people is concentrated in a small area. Convert the image to the YCbCr space and project it on the CbCr plane, then the skin color sample points can be collected. Testing the recognition results, it is found that the skin color in the YCrCb color space will not be changed by the brightness of the light, and the skin color area can be identified more clearly and accurately. After the skin color is extracted, further morphological processing is performed on the image, isolated small dots and burrs are removed by opening operation, and small holes and small gaps are filled in by closing operation.

The next step is mainly to extract the features of the image data after preprocessing, mainly using Fourier descriptors, which can describe the contour features well, and only a small number of descriptors (that is, the number in the vector does not need too many) Can roughly represent the entire outline. Secondly, after a simple normalization operation on the Fourier description character, the descriptor can be translated, rotated, and scale-invariant, that is, it is not affected by the position, angle, and scaling of the contour in the image. Great image features.

The main task of the final model training and testing is to use the existing sample library to train the algorithm model and save it. The system mainly uses two algorithms: KNN and SVM. The KNN algorithm classifies by measuring the distance between different feature values. First, a training sample set is obtained, and the correspondence between each data in the sample set and its label is obtained. After entering the new data without labels, compare each feature of the new data with the corresponding features of the data in the sample set, select the first k most similar data in the sample data set, and count the category with the most occurrences among the k data as the new Classification of data. The SVM algorithm finds the closest points to the separating hyperplane to ensure that they are as far away as possible from the separating surface. The support vector is the points closest to the separating hyperplane. By maximizing the distance between the support vector and the separating plane, the optimization operation of the algorithm is realized.

4. EXPERIMENTS

4.1 Introduction to Dataset

A total of two data sets are required in this system, one is the data set of the speech recognition module; the other one is the data set of the gesture recognition module.

The voice data set is a knowledge system for the current novel coronavirus pneumonia, and the data comes from the data updated in real time on the current network.

The gesture module recognizes a total of 10 gestures, takes 10 gesture pictures from 1 to 10, and then uses data augmentation to expand the data set. The main method is: rotating each gesture picture at different angles, and then flipping the picture. Expanding each gesture picture (that is, each gesture) to 500 pictures, then there are 10 gestures in the data set, a total of 5000 pictures can be used for training or testing.

4.2 Analysis of experimental results

The experimental results show that the system can achieve the expected functions, designing injections, medications, temperature measurement, rescue, bottle changing, blood drawing and other tasks into simple gestures. The medical staff in the isolation ward make gestures and convert them into voice through robot recognition. It is sent to the medical staff in the duty room in time to deliver work information and improve work efficiency while keeping the ward quiet. At the same time, it has the function of voice question and answer, and simple chats can be carried out. In addition, the following is a demonstration of the algorithm training of the gesture recognition part.

In order to enhance the accuracy of the data, the algorithms of KNN and SVM are used to train the original gesture outline and the gesture outline processed by the elliptic fourier descriptor respectively, and the four results of the current gesture are obtained by using the four trained models. Then, we set the way to vote on the four judgment results obtained, and the larger number is the final output result. The accuracy of gesture recognition is detailed after using the four algorithms in Table 1.

Table 1: Comparison of Experimental Results(Gesture Recognition)

Methods	Model Accuracy
KNN	79%
SVM	83%
Fourier Descriptor + KNN	94%
Fourier Descriptor + SVM	99%

5. CONCLUSIONS

This paper implements an intelligent epidemic prevention system based on voice and gestures. The gesture module uses the Fourier descriptor + svm algorithm to achieve an accuracy rate of 99%. The voice module is based on the GRU+CTC deep learning framework, and the accuracy and efficiency of speech recognition Have improved a lot.

However, there is room for improvement in the overall model. Since the data set is self-made, it currently only supports Chinese voice chat. If it is to be widely used, the model still has room for improvement.

6. REFERENCE

[1] O. Abdel-Hamid, A. Mohamed, H. Jiang and G. Penn, "Applying Convolutional Neural Networks concepts to hybrid NN-HMM model for speech recognition," *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, 2012, pp. 4277-4280, doi: 10.1109/ICASSP.2012.6288864.

[2] Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR, abs/1409.1556*.

[3] C. Szegedy et al., "Going deeper with convolutions," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 2015, pp. 1-9, doi: 10.1109/CVPR.2015.7298594.

[4] He K., Zhang X., Ren S., Sun J. (2016) Identity Mappings in Deep Residual Networks. In: Leibe B., Matas J., Sebe N., Welling M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol 9908. Springer, Cham.

[5] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[6] Sepp Hochreiter, Jürgen Schmidhuber. Long Short-Term Memory. Neural Comput. 9(8): 1735-1780 (1997).

[7] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. CoRR abs/1406.1078 (2014)

[8] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd international conference on Machine learning(ICML '06). Association for Computing Machinery, New York, NY, USA, 369–376. DOI:https://doi.org/10.1145/1143844.1143891.

[9] Rautaray, Siddharth S. Vision based hand gesture recognition for human computer interaction: a survey [J]. ARTIFICIAL INTELLIGENCE REVIEW, 2015, 43(1):1-54.

[10] Kevin NYY, Ranganath S, Ghosh D. Trajectory modeling in gesture recognition using cybergloves and magnetic trackers [J]. IEEE TENCON, 2004, 10:571–574.

[11] Bourke A, O'Brien J, Lyons G. Evaluation of a threshold-based tri-axial accelerometer fall detection algorithm [J]. Gait & Posture, 2007, 26(2):194–199.

[12] Bretzner L, Laptev I, Lindeberg T. Hand gesture recognition using multi-scale colour features, hierarchical models and particle filtering [J]. Fifth IEEE international conference on automatic face and gesture recognition, 2002:405–410.

[13] Zhang, Y., Pezeshki, M., Brakel, P., Zhang, S., Laurent, C., Bengio, Y., Courville, A. (2016) Towards End-to-End Speech Recognition with Deep Convolutional Neural Networks. Proc. Interspeech 2016, 410-414.

Authors

Haiguang Chen

is Ph.D , a professor of ShangHai Normal University. During 2006-2007, he was a visiting scholar in Dept. of IST at Weber State University,
UT, USA His research interests include Wireless Sensor Networks, Mesh networks
and the security of networks.