An Identity Authentication Method Based on Multi-modal Feature Fusion

Jiabo Chen¹, Linxi Cai¹, Yuyang Tu¹, Rui Dong¹, Diankun An¹, Baili Zhang*¹

¹College of Software Engineering, Southeast University, Nanjing, Jiangsu, China
*Corresponding author’s e-mail: 101010783@seu.edu.cn

Abstract. At present, many judicial organs have adopted the daily attendance system based on face recognition to strengthen the supervision of community correction personnel. In order to prevent a few personnel from using pre-prepared photos and videos to deceive the face recognition system, this paper proposes an identity verification scheme with liveness detection based on dynamic combination of multimodal features. The main idea is as follows. Firstly, during face verification, the user is required to read out random numbers on the screen. Secondly, generating dynamic combination of speech, voiceprint, lip-reading and other verification methods according to the user's risk personas, so as to achieve a balance between convenience and security. In addition, in view of the low accuracy of lip-reading recognition in practice, this paper changes the traditional lip-reading recognition based on morpheme to lip-reading recognition based on classifier. By optimizing the interactive content, the distinction of pronunciation between different Chinese characters is increased, and the accuracy of lip-reading recognition is significantly improved.

1. Introduction

At present, Chinese judicial organs manage nearly seven hundred thousand community correction personnel nationwide. Some of these community correction personnel need to report their thoughts and whereabouts on a regular basis after authenticating in the "unmanned correction pavilion" as required. When authenticating, if only the traditional face recognition is used, users will deceive the system by using photos, recorded videos and other methods. In order to solve this problem as well as consider the convenience of users and the limited construction funds, we need to build a convenient and reliable authentication system at a low cost.

The basic idea of biometric authentication is to collect the biological characteristics of the users. If the users’ biological characteristics match the recorded biological characteristics, the authentication will be passed. Traditional biological characteristics include face and fingerprint, but these two characteristics are easy to be forged by fingerprint stickers, photos, models and other items, which can be used to deceive the recognition system. Therefore, this paper selects face, voiceprint and lip-reading as the biological characteristics for authentication to prevent these common deception methods. If only the face is checked, the system may be attacked by a photo, a video or a face model. If only the voiceprint is checked, the system may be attacked by a voice changer. If the authentication is not dynamic, the system may be attacked by a pre-recorded video.

Therefore, this paper proposes a multimodal feature fusion authentication method by comprehensively using face recognition, lip-reading recognition, speech recognition and voiceprint recognition. This paper dynamically combines face recognition, lip-reading recognition, speech recognition and voiceprint recognition. Through this method, the system can effectively resist common
attacks such as photo attacks, video attacks and 3D model attacks, which improves the security of authentication with a lower cost and a more convenient way.

2. Identity Verification Scheme
This paper randomly combines the methods of face recognition, voiceprint recognition, lip-reading recognition and speech recognition to solve the above problems. In order to simplify the complexity of using the system under the premise of ensuring security, this paper introduces the persona technology. For users with good performance in the past, the system only carries out simple authentication (choose one or two methods for authentication). For users with poor performance in the past, the system will carry out complex and rigorous authentication, so as to keep balance between efficiency and security.

This method includes four recognition modules, namely, lip-reading recognition, speech recognition, face recognition and voiceprint recognition. The whole process of authentication includes logging in, reading persona, randomly combining authentication methods, authenticating, and recording the result. The recognition process of combining all authentication methods is shown in figure 1.

![Figure 1. Specific process of the scheme](image)

The specific process is as follows.

1) The user is required to authenticate the face;
2) The user needs to look at the camera and read out a series of random numbers/Chinese characters given on the system screen;
3) According to the user’s persona, the system dynamically combines the following recognition methods. (a) Voiceprint recognition: verifying whether the speaker’s voiceprint matches the registered voiceprint. (b) Speech recognition: verifying whether the number read is the random number on the screen. (c) Lip-reading recognition: judging whether the user’s lip actions match the correct lip actions when reading the number.

The numbers that appear on the screen are randomly generated and the number of probable combinations is huge. Due to the time limit for authentication, attackers cannot record cheating videos in advance and cannot quickly make pictures or videos for deception in a low-cost way during authentication. Therefore, it can effectively resist various common attacks.

3. Key Modules

3.1 Lip-reading recognition module

3.1.1 Algorithm
Due to the low accuracy of traditional lip-reading recognition algorithm which is 40-50% in specific applications, this paper proposes a new lip-reading recognition algorithm based on classifier. The main idea is as follows. Firstly, the video is divided into segments which each segment contains only one tone
by the corresponding audio. Because each Chinese character has only tone, these segments contain pronunciation of only character. Secondly, the segmented videos are preprocessed as the input of the neural network classifier. After the classifier identifies the content, it will check whether the number read by the user is the random number on the screen. The identification process is shown in figure 2.

![Figure 2. Lip-reading recognition flow chart](image)

### 3.1.2 Interactive content

One of the reasons for the difficulty of lip-reading recognition is that there are many words with similar lip shapes. Lip-reading is not like speech. The lip shapes of many sounds are very similar, such as "ling", "yi", "si" and "qi" (Chinese pronunciation of the numbers "zero", "one", "four", and "seven"). Therefore, in this paper, we remove three of the four numbers “zero”, “one”, “four” and “seven” with similar lip shapes from numbers zero to nine, and add another three highly distinguishable Chinese characters "bian", "ma" and "da" to the interactive content.

### 3.1.3 Data set

The data set of lip-reading for training is a set of 240 videos which are collected by our team independently and have no obvious relative motion between the speaker and the camera. The speech consists of ten Chinese characters. Each video lasts about 0.5 seconds and the interval between different character pronunciations is more than 0.1 seconds.

### 3.1.4 Preprocessing of lip motion video

The preprocessing of lip-reading recognition data set includes the extraction of key frames in lip motion video, the detection of face region, and the location and extraction of lip region.

Through speech analysis, the start time and end time of each independent pronunciation unit are accurately located on the audio signal. In this paper, a fixed-length sequence containing 10 pictures is sampled from each independent speech video, which is called a key frame.

For the detection of the face region, this paper uses Dlib to detect 68 feature points of the face [1], as shown in Fig. 3. For the location and extraction of the lip region, the 49th-68th feature points are used to locate the lip region. Points 49, 51, 53, 55 and 58 are respectively the left and right corners of the lip, the two highest points of the upper lip and the lowest point of the lower lip. These five feature points can be used to determine the lip boundary in a picture [2].
After the positions of the corresponding lip features in each frame has been located, the system will find the maximum values of their length and width respectively to cut the mentioned key frame sequence. The image sequence will be cut into the shape of $n \times 224 \times 224 \times 1$ as the input [3].

The extracted image sequence should be normalized, and the processing formula is shown in equation (1).

$$X = \frac{X - \hat{X}}{X_{\text{max}} - X_{\text{min}}}$$  (1)

### 3.1.5 Construct training set

After getting the lip features, this paper constructs a self-defined dataset class MyDataSet, uses the RandomResizedCrop function to fix the whole size of the image to $3 \times 224 \times 224$, and imports the corresponding image sequence at one time. Each 10 images constitute a corresponding image sequence, namely $< (10 \times 3 \times 224 \times 224), y_i >$, where $y_i$ represents the corresponding correct answer, $y_i \in \{0, 1, 2...9\}$.

### 3.1.6 Lip-reading recognition training model

The whole process of lip-reading recognition training can be abstracted as a multiple classification problem, namely $y_i = f(X_i)$, where $X_i$ represents the tensor of the corresponding image sequence $10 \times 3 \times 224 \times 224$, $y_i$ represents the corresponding classification.

The VGG_LSTM model is constructed. Some of the parameters of vgg16 network model and LSTM network model are modified: inputsize= $512 \times 7 \times 7$ (represent the dimension of the input data), hiddensize=56 (represent the size of hidden vector), numlayers=1 (represent the layer of RNN). The lip feature data can be obtained from the upper layer of the final classification layer of the loading model.

This paper adopts Adam function as the optimization algorithm. The Adam function replaces the traditional stochastic gradient descent (SGD) process, and it can update the weights of neural network iteratively and automatically based on the training data. This paper adopts ReLU() as the activation function to better mine the relevant features and fit the training data even when the training data is not sufficient. The loss function adopts the cross entropy function nn.CrossEntropyLoss(). The cross entropy function of PyTorch is shown in equation (2).
\[ H(p, q) = -\sum_x (p(x) \log(q(x))) \] (2)

The output layer uses the linear1 function to ensure that the output is a \(1 \times 10\) vector. The maximum value is selected as the final result of prediction.

3.2 Other recognition modules
The modules described in this section use mature technologies. Due to limited space, this paper will briefly introduce.

The face collection module uses the face detector Haar\(^4\) of OpenCV to detect the face in the images and gray images. The face recognition module uses the classic alexnet\(^5\) network model in pytorch, and uses Adam\(^6\) as the optimizer. In recognition, the probability that the face image belongs to each category is obtained by softmax\(^7\).

This paper uses keras to build speech recognition model. The speech recognition model consists of ten 2D convolution layers and two fully connected layers, among which mixed with pooling layers and dropout layers which are used to interrupt some connections to prevent overfitting\(^8\). At the end of the model, by the type of ctc loss\(^9\), CTC algorithm\(^10\) is used to prevent overlapping phonemes. At last, Adam optimization algorithm is used to train the model. The data set used in the speech recognition training model in this paper is THCHS30 Chinese speech data set of Tsinghua University. In recognition, only an audio file in wav format is input, and the program will output the phonetic symbol of the voice contained in the audio by combining with the corresponding phonetic dictionary after obtaining the corresponding index list through data processing and model prediction.

The characteristic parameters extracted by voiceprint recognition module is MFCC feature. This paper uses liborosa to achieve the extraction process, and the input audios are converted to numpy matrix through the librosa.feature.melspectrogram() function. This paper also uses the data augmentation techniques. If the voice length is too long, the program will be cut into smaller pieces. After that, a Resnet50 convolutional neural network classification model is built\(^11\). In recognition, the feature data of speech can be extracted by loading the upper layer of the final classification layer of the model. This paper programs the loading data function and the execution prediction function and inputs two pieces of voice. Then their feature data can be obtained through the prediction function. Their diagonal cosine value can be obtained by calculating their feature data. This value can represent the similarity between these two pieces of voice. If the value is greater than 0.7, the system will judge that the two samples are from the same person, otherwise, they are judged to be from different people.

3.3 Dynamic combination of authentication methods based on persona
In this paper, the persona technology is introduced to simplify the authentication process of legitimate users on the premise of ensuring the security and the accuracy of living detection. Persona is a user model abstracted from users’ behaviors. In this system, if a user's past performance is good, he or she can be simply verified; if a user's past performance is not good, the system will extract a variety of high security authentication methods to verify his identity. The specific algorithm is as follows.

The system will memorize the user's past 20 records. If the pass rate is: a) more than 80%, the system will randomly select an authentication method with low false rejection rate for authentication. b) between 70% and 80%, two authentication methods with low false rejection rate will be randomly selected. c) 60% ~ 70%, three verification methods will be randomly selected. The lip-reading authentication will be Monosyllabic. d) less than 60%, all verification methods need to be passed. For users with less than 20 records, they will be treated as situation b).

In this way, on the premise of ensuring security, this system ensures the simplicity of the authentication process for legal users.
4. System Building and Experiment

4.1 System introduction
In order to verify the effectiveness of the model and algorithm established in Section 2, this section will build and test the authentication system.

People who use the system should register in advance. First, input the ID of the person to be added in the text box, and click add button. Then the system turns on the camera and automatically collects face photos and voiceprint information. After the collection, the information is transmitted to the server for training of face recognition model and voiceprint recognition model.

After registration, users can use this apps. First, input the number of the person to be verified (automatically verify the number, if the number does not exist, an error will be alerted), and then collect a face of the person to be verified. If the face verification score is greater than 0.8, other authentication will be carried out, otherwise, attendance failure will be alerted. Other authentications: the label "Read the number string on the right side" appears below. On the right side of the reminding label is a label containing four random numbers. The user should read out the numbers in three seconds, after which the audio and video will be sent to the server for authentication, and the server will return the attendance results.

4.2 Experiment and result analysis

4.2.1 False rejection rate experiment
False rejection rate (FRR) is one of the important indicators of biometric authentication system, which is defined as the percentage of false rejection in actual authentication.

This paper collected 400 correct samples for the false rejection rate experiment, and the number of samples finally passed is 73. Thus, the false rejection rate of this algorithm is 18.25%.

4.2.2 Error acceptance rate experiment
False acceptance rate (FAR) is also one of the important indicators of biometric authentication system, which is defined as the percentage of false acceptance in actual authentication.

This paper collected 50 lip-reading false samples, 50 voiceprint false samples, 50 face false samples, 30 speech false samples and 400 correct samples for the false acceptance rate experiment to test the system which has been registered by 20 people. Each sample was authenticated for three times. In 1200 experiments, the error sample only passed three times. That is to say, the false acceptance rate of this system is only 0.25%. The error rate mainly occurs in voiceprint error samples. In fact, the probability that the attacker only has voiceprint mismatch is very low, and there will not be so many error samples. In other words, the actual false acceptance rate should be much lower than the experimental result of 0.25%.

| Experiment    | Error times | Error rate(%) |
|---------------|-------------|---------------|
| FRR 400       | 73          | 18.25%        |
| FAR 1200      | 3           | 0.25%         |

5. Conclusions
This paper proposes a high security authentication method based on multimodal feature fusion, establishes the authentication model, studies its key technologies, and gives the authentication process. The conception of persona is applied to the program, which ensures the simplicity of the authentication process on the premise of ensuring the safety and the accuracy of living detection. In order to solve the problem of low lip-reading recognition rate in the environment mentioned in this paper, the traditional lip-reading recognition based on morpheme recognition is changed to lip-reading recognition based on...
classifier. By optimizing the interactive content, the distinction between different pronunciations of interactive content is increased, and the accuracy of lip-reading recognition is significantly improved.

In the experiment, the error acceptance rate is 0.25% and the error rejection rate is 18.25%, which shows that this method has obvious advantages in accuracy and reliability, and has strong competence in avoid cheating.

Acknowledgements

This work was partly supported by Rapid Support Project (61406190120), National Student Research Training Program project (202010286166), the National Key R&D Program of China (2018YFC0830200) and Open Research Fund from Key Laboratory of Computer Network and Information Integration in Southeast University, Ministry of Education, China.

References

[1] Z. Wu, Research on Facial Feature Points Extraction and Relevant Application [D], Shanghai Jiaotong University, 2007.
[2] J. Yu, Lipreading Technology Based On Lip Motion Feature [D], Guangdong University of Technology, 2008.
[3] J. Yan, Research on Lip Reading Method Based on Deep Learning [D], North China University of Technology, 2019.
[4] C.P. Papageorgiou, M. Oren and T. Poggio, A General Framework for Object Detection [C], ICCV.1998.710772.
[5] A. Krizhevsky, I. Sutskever and G.E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, Commun. ACM, 60 (2017) 84-90
[6] D. P. Kingma, and J. Ba, Adam: A method for stochastic optimization, 3rd International Conference for Learning Representations, San Diego, 2015.
[7] W. Liu, Y. Wen and Z. Yu, Large-Margin Softmax Loss for Convolutional Neural Networks [C], Proceedings of the 33rd International Conference on Machine Learning, 2016: 507-516.
[8] M. Gales and S. Young, The Application of Hidden Markov Models in Speech Recognition, Signal Processing, 1 (2007) 195–304.
[9] A. Graves, S. Fernández, F.J. Gomez and J. Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks, ICML, 2006: 369-376.
[10] I. Sutskever, O. Vinyals and Q.V. Le. Sequence to Sequence Learning with Neural Networks. NIPS, 2014.
[11] K. He, X. Zhang, S. Ren and J. Sun. Deep Residual Learning for Image Recognition [C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 770-778.