IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 5, 2020, accepted May 27, 2020, date of publication June 1, 2020, date of current version June 11, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2999115

An Efficient Android-Based Multimodal Biometric Authentication System With Face and Voice

XINMAN ZHANG[1], DONGXU CHENG[1], PUKUN JIA[2], YIXUAN DAI[1], AND XUEBIN XU[3]

[1]MOE Key Laboratory for Intelligent Networks and Network Security, School of Automation Science and Engineering, Xi'an Jiaotong University, Xi'an 710049, China
[2]Graduate School of Information, Production and Systems, Waseda University, Fukuoka 808-0135, Japan
[3]Guangdong Xi'an Jiaotong University Academy, Foshan 528000, China

Corresponding author: Xinman Zhang (zhangxinman@mail.xjtu.edu.cn)

This work was supported in part by the National Natural Science Foundation under Grant 61673316, and in part by the Major Science and Technology Project of Guangdong Province under Grant 2015B010104002.

ABSTRACT Multimodal biometric authentication method can conquer the defects of the unimodal biometric authentication technology. In this paper, we design and develop an efficient Android-based multimodal biometric authentication system with face and voice. Considering the hardware performance restriction of the smart terminal, including the random access memory (RAM), central processing unit (CPU) and graphics processor unit (GPU), etc., which cannot efficiently accomplish the tasks of storing and quickly processing the large amount of data, a face detection method is introduced to efficiently discard the redundant background of the image and reduce the unnecessary information. Furthermore, an improved local binary pattern (LBP) coding method is presented to improve the robustness of the extracted face feature. We also improve the conventional endpoint detection technology, i.e. the voice activity detection (VAD) method, which can efficiently increase the detection accuracy of the voice mute and transition information and boost the voice matching effectiveness. To boost the authentication accuracy and effectiveness, we present an adaptive fusion strategy which organically integrates the merits of the face and voice biometrics simultaneously. The cross-validation experiments with public databases demonstrate encouraging authentication performances compared with some state-of-the-art methods. Extensive testing experiments on Android-based smart terminal show that the developed multimodal biometric authentication system achieves perfect authentication effect and can efficiently content the practical requirements.

INDEX TERMS Multimodal biometric authentication, Android-based smart terminal, improved LBP, improved VAD, adaptive fusion. strategy.

I. INTRODUCTION

With the rapid development of science and technology, mobile terminal-based payments, for example Alipay and WeChat pay, etc., are becoming more and more popular. Correspondingly, the conventional authentication mode is no longer fit for the increasing security requirements on the smart terminal. These problems have attracted more and more researchers to study the access control and authentication management for the smart terminal. The conventional authentication systems usually adopt the password or unimodal biometrics to realize the authentication task. However, they are limited to be applied in many practical fields, because they are easy to be stolen, simulated or even forgot by users themselves.

Recently, the multimodal biometric authentication methods [1]–[4] have been proposed as they were expected to be more reliable and convenient than conventional unimodal biometrics [5]–[7], such as the electrocardiogram (ECG), fingerprint, voice, iris and face. The multimodal biometric authentication system may operate in one of three different modes, i.e., the serial mode, parallel mode and hierarchical mode. How to select a suitable fusion strategy is becoming a critical role for the development of an effective multimodal biometric authentication system. The commonly employed fusion strategy can be categorized into the sensor level, feature level, match score level, rank level, and decision level fusion. In the past decade, lots of research results have been

The associate editor coordinating the review of this manuscript and approving it for publication was Andrea F. Abate.

achieved in this area. Fierrez-Aguilar *et al.* [8] proposed a multi-biometric authentication system by integrating the biometric information of face, fingerprint and signature based on the score level fusion. Kumar *et al.* [9] proposed a multimodal algorithm by combining the palm print and hand geometry biometrics where the concatenation operation and the max rule are adopted to implement the biometric fusion at the feature level. Toh *et al.* [10] proposed a multimodal method by integrating the biometrics of hand geometry, fingerprint and voice at the matching score level. Recently, the sparse representation-based multimodal recognition methods [11], [12], which utilized the sparse linear combination of the training samples to represent the testing sample, have attracted substantial attention and achieved some encouraging performances.

The acquisitions of the face image and voice signal are non-mandatory and easy to be implemented. In recent years, many researchers have begun to pay attention to these fields. Zhang *et al.* [13] presented a multimodal biometric authentication system for smartphone based on face and voice. Since the LBP [14]–[16] coding-based feature has the merits of simple to be implemented, better representation of the image texture features, good insensitivity to the variation of the illumination condition, and gray intensity. As a consequence, it has attracted many researchers to conduct a thorough study and present lots of improved versions. Duan *et al.* [14] proposed a context-aware local binary feature learning (CA-LBFL) method for face recognition which exploited the context-aware information of adjacent bits by constraining the number of shifts from different binary bits. Yang *et al.* [15] introduced an adaptive local ternary pattern (ALTP) feature descriptor based on the automatic strategy to select the threshold and proposed a center-symmetric ALTP (CS-ALTP) feature description method to implement face recognition. Wei *et al.* [16] proposed a new robust feature, named block Gabor directed derivative layer local radii-changed binary pattern, to implement face recognition. On the other hand, the Mel-frequency cepstrum coefficient (MFCC) [17]–[19] is a well-known and commonly used feature descriptor and it has been widely utilized to implement the feature extraction task for voice biometrics. For the voice authentication task, GMM [20]–[23] is an applicable method and it has been widely studied by many researchers. For example, Nakagawa *et al.* [17] proposed a phase information extraction method which normalized the variation in the phase according to the frame position of the inputted voice and combined the phase information with MFCCs in the text-independent speaker identification methods. You *et al.* [21] introduced a Bhattacharyya-based GMM distance measurement and exploited the universal background model (UBM) with the maximum a posteriori (MAP) criterion to implement voice authentication.

Since the multimodal biometric authentication approach employs richer information, we believe that it will provide better performance compared with the conventional unimodal biometric authentication method. Inspired by this, an efficient multimodal biometric authentication system with face and voice is developed to realize the improvement of the authentication efficiency. The main novelties and contributions are summarized as follows:

TABLE 1. Notations used in this paper.

Notation	Description
(x_c, y_c)	coordinate of the neighborhood center
P	pixel number of the neighborhood
i_p	gray value of the pixel p
i_c	gray value of the neighborhood center
N	number of the sample points for each frame
f	actual frequency of the voice signal
f_{Mel}	Mel frequency of the voice signal
K	number of Gaussian functions

(1) Taking into full consideration the performance of the smart terminal, we introduce an efficient face detection algorithm, named as Haar cascades and AdaBoost algorithm, to improve the effectiveness and accuracy of face detection.

(2) An improved LBP coding method is presented to boost the robustness of the face feature, which can effectively enhance the face matching efficiency of the developed authentication system.

(3) We improve the conventional VAD method to enhance the detection accuracy of the mute and transition information in the voice stream and boost the voice matching efficiency.

(4) By integrating the merits of face and voice biometrics, an adaptive fusion strategy is presented to realize the user identity authentication.

(5) Extensive experiments show that the developed multimodal biometric authentication system is compatible and can be easily installed on the Android-based smart terminal. It can conquer the defects of the unimodal biometric authentication system and meet the practical application requirements.

The rest of this paper is organized as follows. The basic principle of our developed multimodal biometric authentication system is introduced in section 2. Section 3 describes the algorithm framework of the system in detail. Simulation experiments and testing experiments for the developed Android-based multimodal biometric authentication system are demonstrated in section 4. At last, this paper is concluded by section 5.

For simplicity, we summarize some of the frequently used notations in Table 1.

II. PRINCIPLE OF THE ANDROID-BASED MULTIMODAL BIOMETRIC AUTHENTICATION SYETEM

A. BASIC PRINCIPLE OF THE MULTIMODAL BIOMETRIC AUTHENTICATION SYSTEM

Fig. 1 demonstrates the basic principle of the developed Android-based multimodal biometric authentication system. It is mainly consisted of the data collection module and the information processing module. For the data collection module, it collects the user's face image and voice signal by

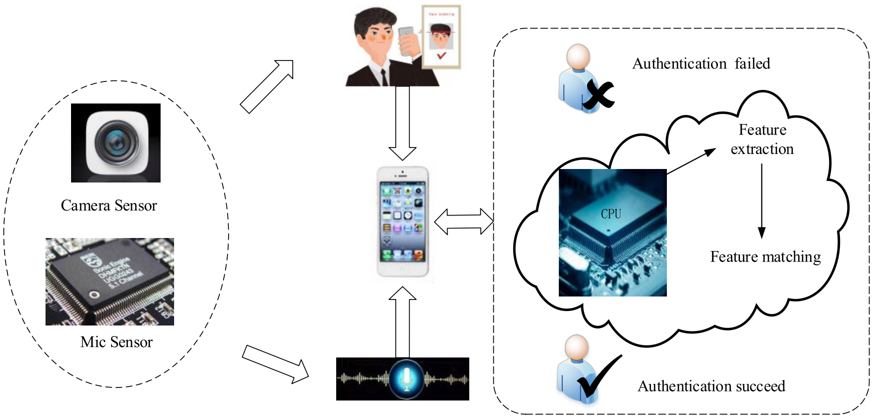


FIGURE 1. Principle of the multimodal biometric authentication system for the smart terminal.

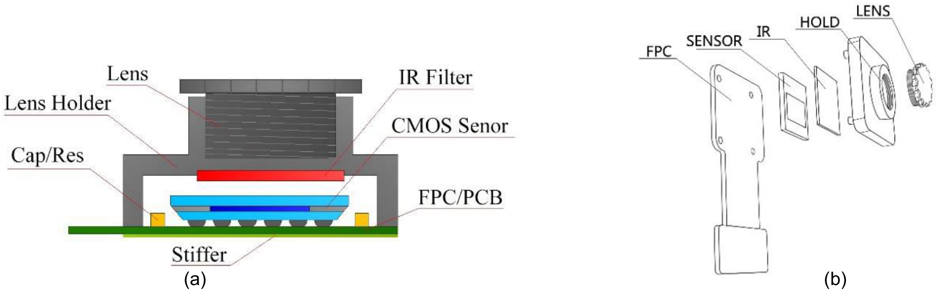


FIGURE 2. Structure diagram of the smart terminal camera module. (a) Overall schematic diagram. (b) Decomposition diagram.

calling the camera and microphone devices of the smart terminal. After the data collection, data processing will be implemented by employing the data interaction and processing devices (such as CPU and GPU) of the smart terminal. In this paper, we use the Android-based smartphone to demonstrate and test our developed multimodal biometric authentication system.

B. HARDWARE OF THE SYSTEM

The hardware of the multimodal biometric authentication system mainly involves the data acquisition devices (such as camera and microphone) and processer devices (such as CPU and GPU).

1) DATA ACQUISITION DEVICES

Fig. 2 illustrates the structure of the smart terminal camera module. The camera module of the smart terminal is mainly composed of FPC/PCB, CMOS sensor, IR filter, lens holder, and lens. CMOS sensor is also called the image sensor. It is mainly used as a stack CMOS chip, which integrates the image signal amplifier, signal readout circuit, A/D conversion circuit, image signal processor, and controller into one chip. So a single CMOS chip can realize all of the camera's basic functions. IR filter is also known as an infrared filter. The process of image capturing with a camera of the smart terminal is described as follows. Firstly, the optical image of the object generated by the lens is projected onto the surface of the image sensor. Then it is converted into electrical signals. After the A/D conversion, the digital image signal is sent to the processing chip. At last, the final digital image is obtained.

The voice device of the smart terminal is essentially a microphone chip, which collects the voice signal through the sensor, and then gets the voice signal through a series of filtering processing and real-time signal processing.

2) CONTROLLERS

The face image captured by the camera and the voice signal obtained by the microphone need to be transmitted to the CPU and GPU of the smart terminal for the further processing. CPU is a large scale integrated circuit, which is the core and control unit of the smart terminal. CPU mainly includes the arithmetic logic unit (ALU), the high-speed buffer memory (Cache), and the data bus. The main function of CPU is to interpret the instructions of the smart terminal and handle the data in the terminal APP. GPU is also known as a graphics processor unit and the core cell of smart terminal display. GPU integrates the display chip, a kind of image processing operation microprocessor of the smart terminal, and helps the CPU to complete the image processing work. Fig. 3 describes the working principle of CPU and GPU.

III. ALGORITHM FRAMEWORK

To deal with the problem of instability, increase the authentication accuracy and system reliability, we design

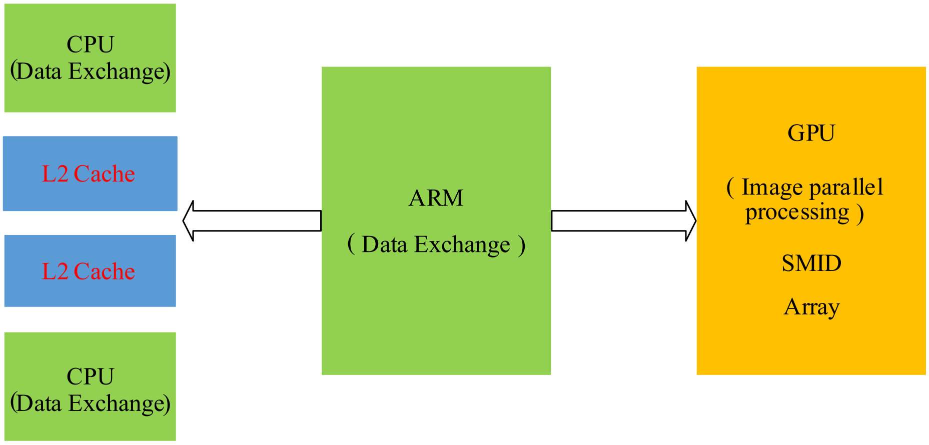


FIGURE 3. Working principle of CPU and GPU.

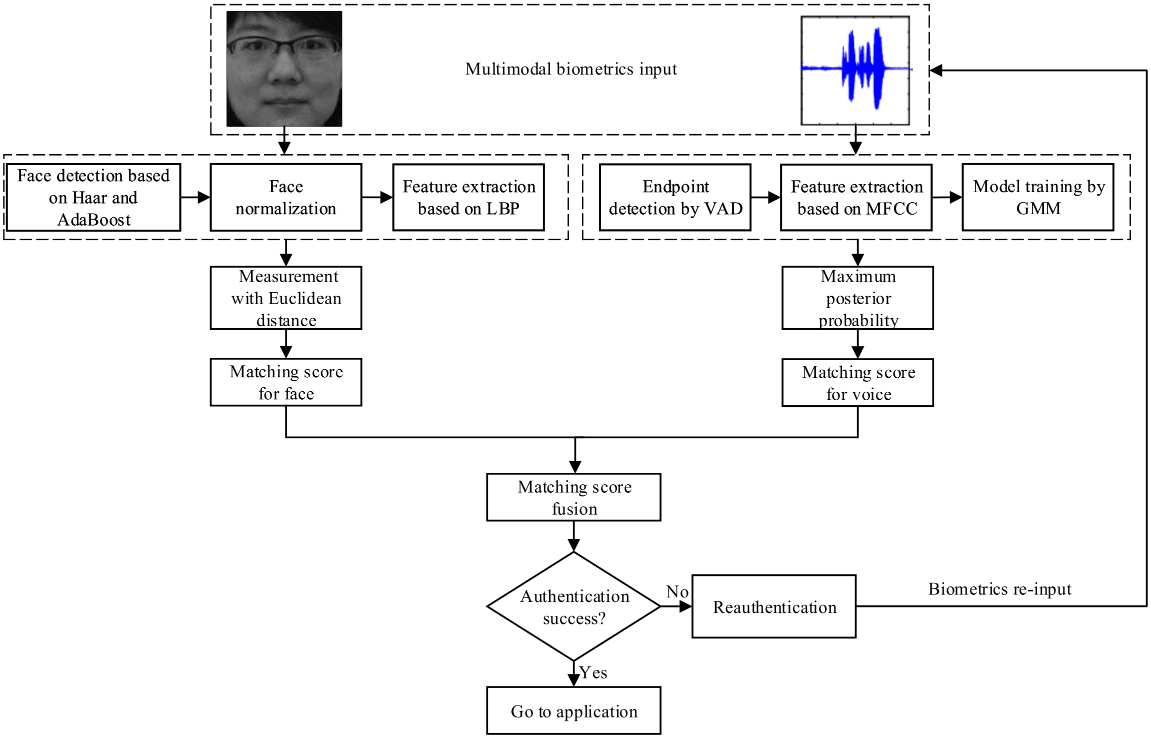

Multimodal biometrics input
Face detection based on Haar and AdaBoost
Face normalization
Feature extraction based on LBP
Endpoint detection by VAD
Feature extraction based on MFCC
Model training by GMM
Measurement with Euclidean distance
Maximum posterior probability
Matching score for face
Matching score for voice
Matching score fusion
Authentication success?
No
Reauthentication
Biometrics re-input
Yes
Go to application

FIGURE 4. Flow chart of the multimodal biometric authentication system.

and develop a novel efficient Android-based multimodal biometric authentication system by integrating the face and voice biometric features. Fig. 4 illustrates the flow chart of the developed system. Obviously, it can be considered as consisted of the following processes, i.e., face matching, voice matching, and fusion authentication. Next, each process will be separately discussed in detail.

A. FACE MATCHING PROCESS

The face matching process includes the pre-model training stage and post-testing matching stage. Fig. 5 illustrates the flow chart of this process. It is easy to see that the face matching process can be separated into the stages of image preprocessing, face detection, image normalization, facial feature extraction, model establishment, and feature matching.

For the face training stage, the image collection is easy to be impacted by different environmental factors (such as uneven illumination and ambient noise.) or deficiency of hardware equipments. In consequence, the collected images often suffer different interference problems, such as the noise contamination and low contrast. To address these problems, we employ the preprocessing procedure, including histogram equalization, image gray normalizing, and image filtering, to enhance the face matching efficiency.

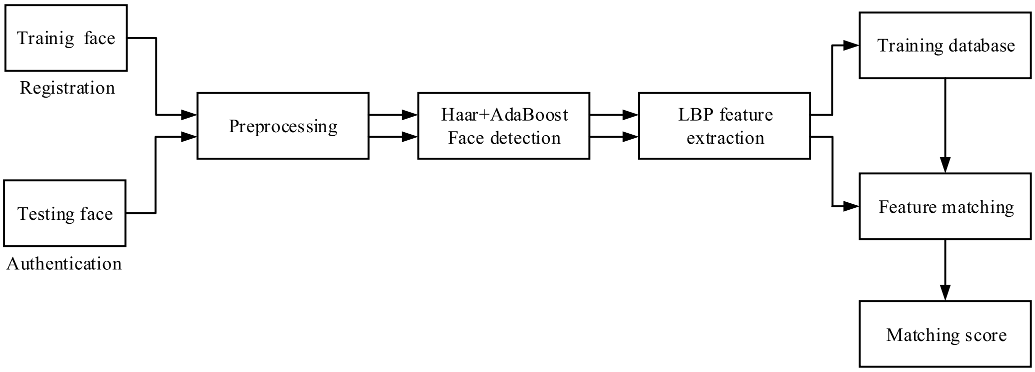


FIGURE 5. Flow chart of the face matching process.

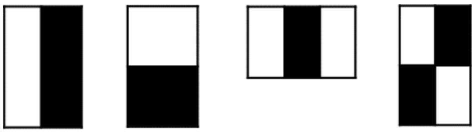


FIGURE 6. Some representative Haar feature templates.

1) FACE DETECTION BASED ON HAAR CASCADES AND ADABOOST ALGORITHM

With the image preprocessing procedure accomplished, we carry out the face detection process to determine whether the image includes the face. If there is a face, output its position and size, and then cut out the background part to obtain the face region. Since the distance between the face and the camera device is variable, the position and proportion of the detected face are unfixed. In order to ensure the consistency of the detected face image size, we employ the geometric normalization to resize it with a uniform size.

Feature-based face detection methods [24], [25] possess good detection efficiecy and accuracy and have been widely used in many practical applications. In consequence, we employ Haar cascades and AdaBoost algorithm to realize the face detection task. Haar features are generated in accordance with the difference of image pixel gray values, which can reflect the gray variance of the face image well. Fig. 6 shows four representative classes of the basic Haar feature template.

Haar cascades and AdaBoost algorithm is a precise classification method based on the cascade of multiple weaker classifiers. The basic principle is to concatenate several classifiers with weaker classification performance one by one to form a classifier with excellent performance. This algorithm can effectively increase the discrimination accuracy by using the feedback mechanism among different classifiers. For each weaker classifier, if the detected target is diagnosed as the negative sample, it is abandoned and no longer inputted into the next level classifier. Otherwise, it is inputted into the next level classifier for further discrimination. This not only greatly reduces the time consumption, but also lowers the false positive probability of the final output positive sample. For a face image, it initially detects and discards a large number of negative samples, which can greatly reduce a mass of data and speed up the efficiency of the algorithm. Fig. 7 shows the basic principle of the Haar cascades and AdaBoost algorithm.

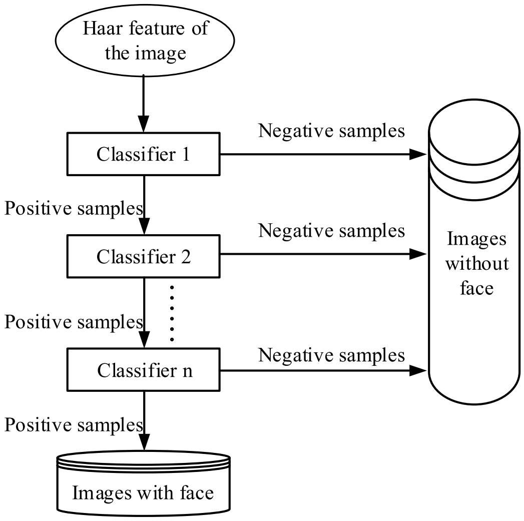


FIGURE 7. Basic principle of Haar cascades and AdaBoost algorithm.

TABLE 2. Testing results of the face detection methods.

Evaluation metrics	Hog+SVM	Haar cascades and Adaboost
Accurate rate	92.65%	96.60%
Misdetection rate	6.76%	2.51%
Repeated detection rate	3.24%	0.92%

To evaluate the effectiveness of the adopted face detection method, we implement experiments with the face images of XJTU multimodal database, and the results are demonstrated in Table 2. It is easy to see that the employed face detection method achieves a higher detection accuracy (i.e. 96.60%) compared with the method of histograms of oriented gradients (HOG) plus support vector machine (SVM) and it can effectively meet the practical application requirements.

2) FACIAL FEATURE EXTRACTION BASED ON IMPROVED LBP ALGORITHM

Since the original data information contained in the detected face image is large and redundant, a feature extraction process is employed to reduce the information redundancy and

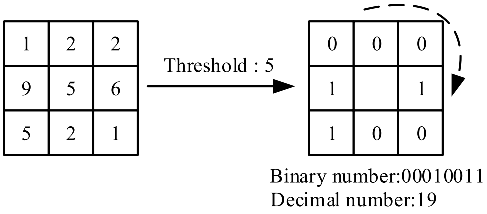


FIGURE 8. Diagram of LBP coding.

improve the feature discrimination in the low-dimensional space. LBP method has the merits of simplicity for extracting, perfect representation for image texture features, and insensitivity to illumination and gray variations. In this paper, an improved LBP method is presented to implement the feature extraction task.

By comparing the gray value of each pixel with the pixels in its neighborhood, we can obtain the LBP code based on the comparison results (see Fig. 8).

As a consequence, the LBP coding principle can be formulated as follows

$$\mathrm{LBP}(x_c, y_c) = \sum_{p=0}^{P-1} 2^p s(i_p - i_c) \tag{1}$$

where, $s(x)$ is the sign function, i.e., if $x \geq 0, s(x) = 1$; otherwise, $s(x) = 0$.

Since the binary encoding of the conventional LBP algorithm relies on the direction and start point, when there is a rotation, the value of LBP code will change accordingly. In consequence, the conventional LBP code does not possess the property of rotation invariance. To address this issue, researchers extended the square neighborhood to a circular region and proposed the cyclic shift coding method to improve the conventional LBP algorithm. The specific description of the novel rotation-invariant LBP [26] algorithm is summarized as follows.

First of all, the neighborhood of the conventional LBP operator is extended to a circular region and the radius is set as R. Then, P sampling points on the circumference are sampled to construct a circular neighborhood. The coordinates of the sampled points can be calculated as follows

$$x_p = x_c + R\cos\left(\frac{2\pi p}{P}\right), \quad y_p = y_c - R\sin\left(\frac{2\pi p}{P}\right). \tag{2}$$

Since the obtained coordinate values of the sampling points may be non-integers, we calculate the corresponding gray values of the sampled points by utilizing the bilinear interpolation method.

For the conventional LBP code of each pixel, when the jumping time, i.e. it jumps from 1 to 0 or from 0 to 1, is not more than two (such as 00000000 (zero jump), 00000111 (one jump) and 10001111 (two jumps)), it is categorized in the first class, called the equivalent mode (see Fig. 9). When the jumping time is greater than 2 (such as 10010111, which jumps 4 times), it is categorized in the other class, called the mixed mode. The category number of the extracted LBP feature by using the above improved method can be effectively reduced. When the number of sampling points is P, the category number of LBP code reduces from 2^P to $p \times (p-1) + 2$.

To enable the obtained LBP code to have the property of rotation invariance, we introduce the cyclic shift theory into the conventional LBP coding operator. The minimal code value generated by this method is set as the final LBP code. Fig. 10 illustrates an example of the rotation invariant LBP operator.

By using the above improved LBP operator, the proposed feature extraction algorithm can be summarized as follows:

(1) Employ the improved LBP operator, which possesses the equivalent mode and rotation invariance, to calculate the LBP code and obtain the feature image.

(2) Divide the LBP feature image into blocks with a uniform size.

(3) Calculate the LBP code histogram of each feature image block.

(4) Concatenate the histogram features of each block according to the spatial order to compose the LBP feature vector.

3) FACE MATCHING METRIC RULE

After the feature of the face image extracted, Euclidean distance is employed to measure the difference between the authenticating image and the registered images.

Assuming that the feature vectors are denoted as $X = (x_1, x_2, \cdots, x_n)^T$ and $Y = (y_1, y_2, \cdots, y_n)^T$, respectively. Calculate the distance by

$$dist(X, Y) = \left(\sum_{i=1}^{n} (x_i - y_i)^2\right)^{1/2}. \tag{3}$$

B. VOICE MATCHING PROCESS

The voice matching process consists of the registration stage and matching stage. In order to improve the reliability, voice content recognition is pre-completed before the voice matching process. If the voice content is consistent with the preset content, subsequent steps, including voice preprocessing, feature extraction, model training, and voice matching, will be carried out. Otherwise, the user needs to re-collect the voice signal. Fig. 11 illustrates the flow chart of the voice matching process.

1) VOICE CONTENT RECOGNITION AND VOICE PROCESSING

Since our authentication system is developed for Android-based smart terminal which contains a built-in voice recognition class, namely RecognizerIntent, we can use this Intent mechanism to start the voice activity and realize the voice content recognition.

To improve the efficiency of voice matching, we employ the discrete wavelet transform (DWT) to implement the voice denoising process. Based on this method, we can decompose

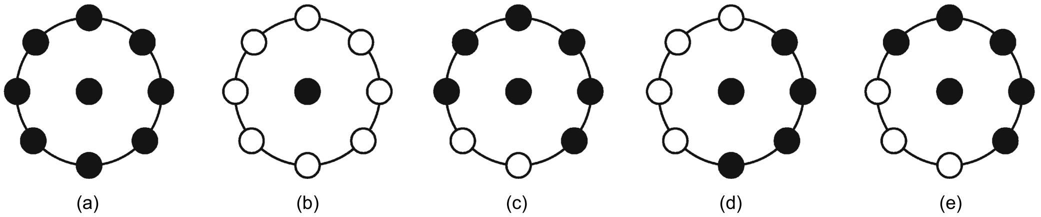


FIGURE 9. The equivalent mode of LBP.

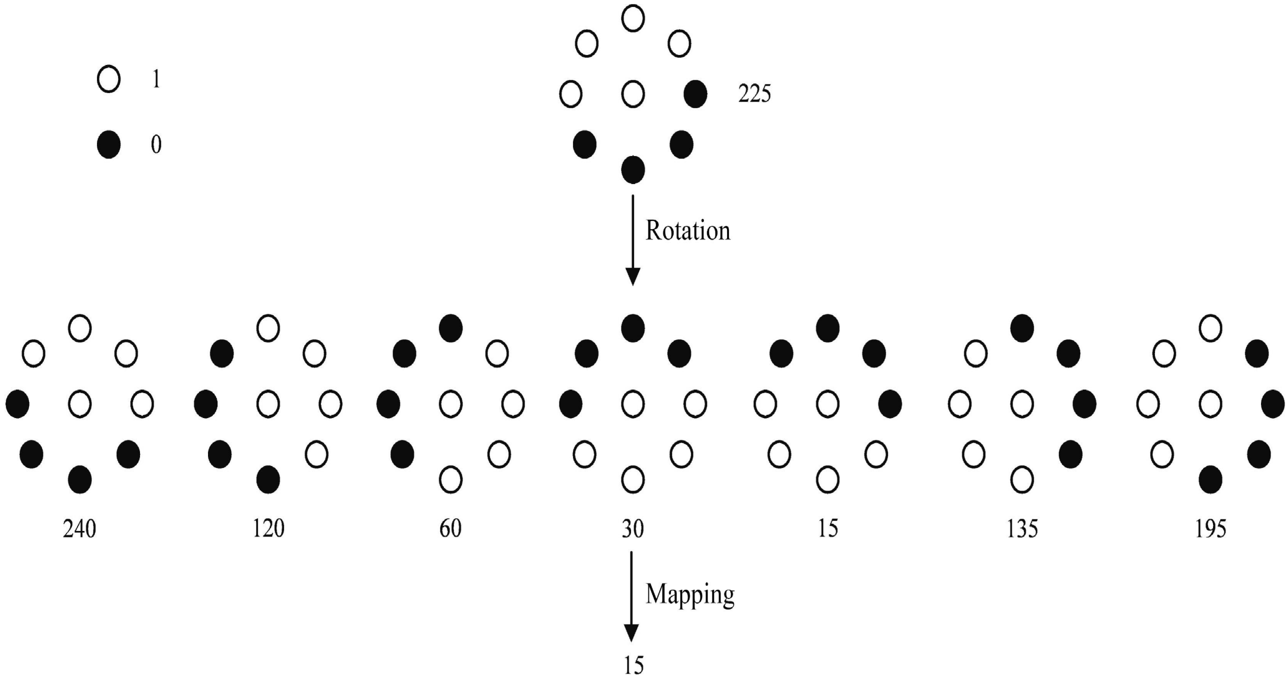


FIGURE 10. Rotation invariant LBP operator.

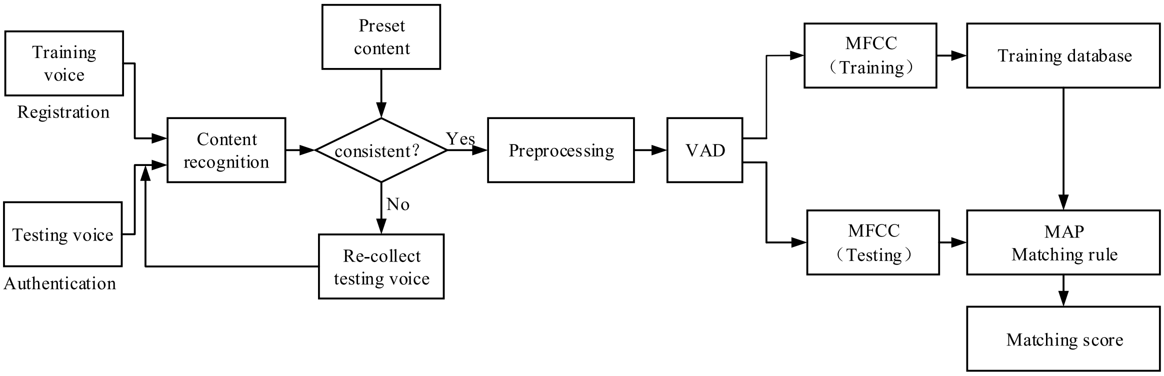


FIGURE 11. Flowchart of the voice matching process.

the original voice signal into different frequencies. Since the intensity distribution of the wavelet decomposition coefficients has significant characteristics and the difference between the useful signal and noise is significant, we can efficiently discard the wavelet coefficients according to the noise contamination. Finally, we utilize the remained wavelet coefficients to reconstruct the noise-free voice signal approximately.

Table 3 illustrates the voice denoising results by DWT method. In the experiment, we add different intensity noises into the voice signal of the XJTU multimodal database. From Table 3, it is clear that the denoising process can efficiently improve the signal to noise ratio (SNR) of the voice signal. This means that the denoising processing can alleviate the noise interference and boost the voice matching efficiency.

TABLE 3. **SNR comparison for voice denoising by DWT method.**

SNR(db) of the noised voice	SNR (db) of the denoised voice
1	8.49
3	10.14
5	11.07
7	12.03
9	13.64
11	15.75
13	17.13
15	18.65

2) ENDPOINT DETECTION BASED ON THE IMPROVED VAD METHOD

VAD is a classical and practical voice processing technology which can detect and intercept the signal segment from the original continuous voice stream. VAD method can achieve the following functions, such as interrupt voice segment automatically, remove the silent segment, retain the effective voice data, and so on. The conventional VAD method divides the original voice signal into the silent segment, transition segment, and effective signal segment. Based on the characteristic that the amplitude or energy of voice signal changes slowly with time, we can divide it into some equal frames to implement the following processing. For each frame traversing of the original voice signal, we can calculate the corresponding parameters of the short-time energy and the short-time average zero-crossing rate. Based on this, we can estimate the corresponding status of each frame and determine the starting and ending points of the useful voice segments in the original voice stream. The concepts of the short-time energy and short-time average zero-crossing rate are described as follows.

- Short-time energy: For a given signal $x(n)$, there are N sampling points for each frame. In order to avoid the large difference between two adjacent frames, an overlapping region containing M sampling points is set. In general, M is set as 1/2 or 1/3 of N. Then the short-time energy at the n-th sampling point can be calculated as follows

$$E_n = \sum_{m=n-(N-1)}^{N-1} [x\,(m)\,\omega\,(n-m)]^2, \tag{4}$$

where $\omega(n-m)$ is the window function. According to the needs of practical application and analysis, we select Hamming window function to deal with the voice signal within a frame, which can ensure the continuity of the left and the right sides of the endpoint. The Hamming window function is defined as follows

$$\omega(n) = \begin{cases} 0.54 - 0.46\cos[2\pi n/(N-1)], & 0 \le n \le N-1 \\ 0, & \text{otherwise.} \end{cases}$$

- Short-time average zero-crossing rate: For a sampled discrete voice signal, the time that the waveform passes through the horizontal coordinate axis is defined as the short-time zero-crossing rate. It can be calculated by the changing number of the sign between the adjacent sampling points. For a given signal $x(n)$, the short-time average zero-crossing rate can be calculated as follows

$$\begin{aligned} Z_n &= \sum_{m=-\infty}^{+\infty} |\mathrm{sgn}[x(m)] - \mathrm{sgn}[x(m-1)]|\,\omega\,(n-m) \\ &= |\mathrm{sgn}[x(n)] - \mathrm{sgn}[x(n-1)]| * \omega(n), \end{aligned} \tag{5}$$

where m represents the sampling point.

$\mathrm{sgn}[x(n)] = \begin{cases} 1, x(n) \ge 0 \\ -1, x(n) < 0 \end{cases}$ denotes the sign function.

The conventional VAD method has serious misjudgment and leakage issues for the voice signal when its peak signal-to-noise ratio (PSNR) is low. Since the smart terminal devices are easily influenced by the transmission channel and environment factors, the conventional VAD method is no longer fit for practical applications.

To enhance efficiency and universality, we improve the conventional VAD algorithm as follows:

(1) Set a single threshold for the zero-crossing rate. When PSNR of the voice segment is low, the conventional dual-threshold VAD method is no longer applicable. Since it has no positive effect on the determination of the effective segment and silent segment, instead, it slows the calculation speed. To address this issue, in our method, we employ only one threshold for the zero-crossing rate.

(2) Introduce a median filter into the conventional VAD method. With the influence of noise, the short-time zero-crossing rate and short-time energy curves fluctuate greatly. To address this issue, a median filter is added into the conventional VAD to make these curves relatively smooth.

(3) Set a minimal length to determine the end of voice. In order to reduce the misjudgment ratio for the silent voice region, a minimal voice length, denoted by the symbol of ‘‘Stime’’, is introduced to judge the endpoint of the voice signal. To judge whether the voice ends or not, the length of the effective voice segment should be greater than the minimal length Stime. By using this parameter, the ending length of the user's voice segment can be effectively extended and the misjudgment ratio of the silent region can be dramatically reduced.

Through the above analysis, the improved VAD algorithm can be summarized as follows. (1) Select a high short-term energy threshold T1 and a low short-term energy threshold T2 in accordance with the average energy of the voice signal and the background noise, respectively. (2) Employ the selected T1 and T2 to determine the preliminary starting and ending points approximately. (3) On the basis of the average zero-crossing rate of the background noise, we estimate the low zero-crossing threshold T3 and detect the more precise voice endpoints.

For example, given a voice segment, calculate the average short-time energy, the average short-time zero-crossing rate, and the threshold value T3, firstly. Then, traverse the voice

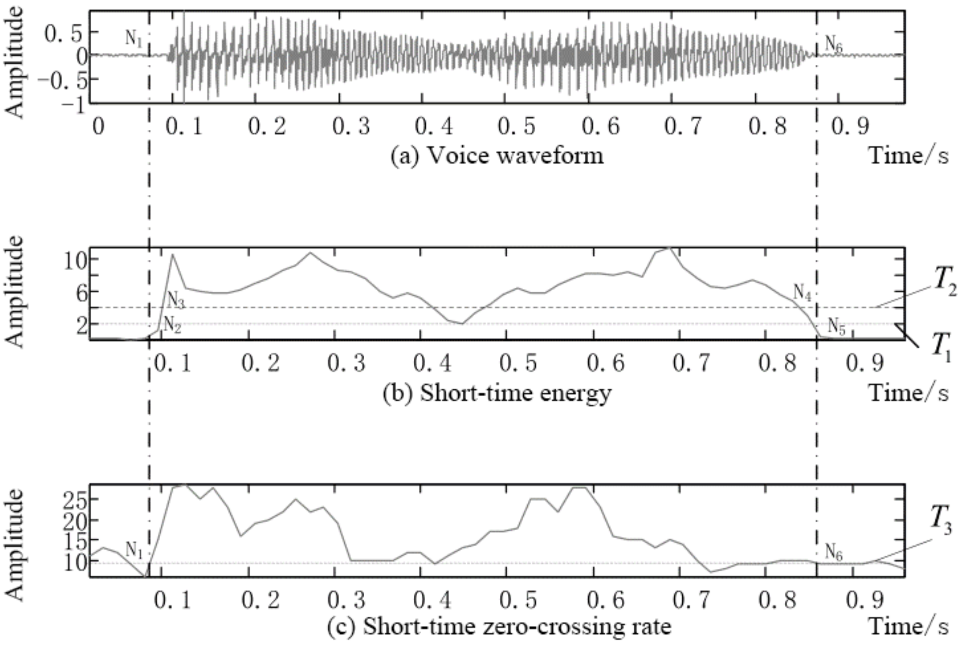


FIGURE 12. Principle of the improved VAD method.

stream and seek out the first frame whose energy value is greater than T1, and set it as the initial starting point N1. Successively, compare and judge whether the energy amplitude of the subsequent frames is greater than T2 and whether the zero-crossing rate is less than T3, respectively. If not, regard this frame as the endpoint of the voice segment and denote the location as N2. Calculate the frame length between the starting point and the ending point and compare it with the minimal voice segment length Stime. If the difference is not less than zero, update the Stime as the next frame after N2, then repeat the above process and judge the effective voice segment from the current position. If the difference is less than zero, discard the afore-setting value N1 and set the subsequent frame whose energy amplitude is greater than T1 as N1. Repeat this process and terminate until the frame whose energy amplitude is greater than T2. Finally, output the obtained starting and ending points, respectively. Fig. 12 illustrates the principle of the improved VAD method.

3) FEATURE EXTRACTION BASED ON MFCC

In practice, the collected voice information is often redundant. We implement the feature extraction processing to reduce the data redundancy and enhance the authentication robustness and accuracy. MFCC is a widely used voice feature, which considers the human auditory characteristic and simulates the human ear perception, simultaneously. It can well describe the energy distribution of the voice signal in the Mel frequency domain. The Mel frequency transformation is formulated as follows

$$f_{\mathrm{Mel}} = 2595 \log(1 + f/700). \tag{6}$$

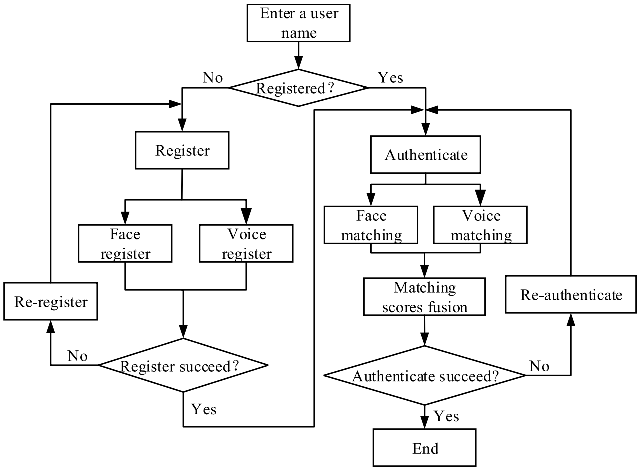


FIGURE 13. Flow chart of the multimodal biometric authentication system.

4) VOICE MATCHING MODEL BASED ON GMM

GMM employs the probability function to calculate the relationship measurement of the samples, which has the merits of high accuracy and excellent efficiency. Many researchers have made in-depth studies and applied it in practical applications [21]–[23]. Generally, the GMM model consists of several Gaussian functions, each of which has the parameters $\{c, \mu, \sigma\}$. For any given random variable x, GMM is formulated as follows

$$p(x) = \sum_{k=1}^{K} c_k \varphi(x|\theta_k), \tag{7}$$

where $\theta_k = (\mu_k, \sigma_k^2)$. c_k represents the weight value associated with the k-th Gaussian function and satisfies

$\sum_{k=1}^{K} c_k = 1$. $\varphi(x\,|\theta_k\,)$ represents the k-th Gaussian probability density function, its expectation and standard deviation are μ_k and σ_k, respectively. $\varphi(x\,|\theta_k\,)$ is formulated as follows

$$\varphi(x\,|\theta_k\,) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(x-\mu_k)^2}{2\sigma_k^2}\right). \tag{8}$$

The goal of GMM probability density estimation is to find the parameter values of each Gaussian function.

For the voice matching of the multimodal authentication system, we employ MAP method to estimate the voice matching score. MAP model is defined as follows

$$P(\theta\,|x_0\,) = \frac{P(x_0\,|\theta\,)P\,(\theta)}{P\,(x_0)}. \tag{9}$$

Since the sample x_0 is given, probability $P(x_0)$ is known. Then we can obtain the estimated value θ by maximizing the likelihood function value of $P(x_0|\theta)$.

C. MULTIMODAL BIOMETRIC FUSION METHOD FOR FACE AND VOICE

The commonly used fusion method of multimodal biometric authentication system can be categorized into the feature level, matching level, and decision level fusion. Due to the great difference between the biometric features of the human face and voice, it is difficult to fuse these feature vectors. Therefore, the fusion method based on the feature level is not suitable for our system. On the other hand, the fusion method based on the decision level is regarded as the highest grade and the implementation difficulty is the smallest. However, this method only refers to the result of the unimodal biometric authentication, which utilizes less information about the original biometric characteristics, and its reliability is lower than the fusion method based on matching level and feature level fusion strategies. The matching level based fusion strategy is a kind of way between the feature and decision level fusion. Its implementation difficulty is inferior to the feature level, and it considers the information with different biometrics more sufficiently than the decision level. As a consequence, it could effectively enhance the reliability of the authentication system. Based on this, we employ the fusion strategy based on the matching scores to accommodate the development requirements of the authentication system for Android-based smart terminal.

In consideration of the differences between the biometric features of face image and voice signal, we employ two completely different matching rules to implement each matching sub-module respectively, i.e., the distance measurement for face feature and MAP method for voice feature. Obviously, there are considerable differences between these matching scores. To deal with this problem, the min-max method is adopted to normalize the matching scores. We use f_score and v_score to denote the normalized scores of the face and voice biometrics, respectively.

In practice, the weighted sum rule and product rule are usually adopted to implement the fusion task, i.e.

$$t_1 = a \cdot v\_score + (1-a) \cdot f\_score, \tag{10}$$

$$t_2 = (a \cdot v\_score) \cdot ((1-a) \cdot f\_score). \tag{11}$$

where a is the weighted value.

Since the voice signal is easy to be influenced by time, environment, and equipment device, etc. We proposed an adaptive method to assign the weighted value a. The specific weighted value is calculated based on SNR of the voice signal, and a piecewise processing method is adopted to carry out the assignment. Particularly, when the SNR of the voice is greater than 45 dB, the voice is considered as noise-free and the weighted values of the two biometrics are both assigned as 0.5. With the decrease of SNR, the weighted factor a decreases gradually. The specific assignment of the weighted value is illustrated in Table 4.

With the weighted value assigned, we can employ equations (10) and (11) to calculate the fusion scores in accordance with the sum rule and product rule, respectively. In practical application, if one of the unimodal biometric matching score and the corresponding weighted value is both bigger, the other match score and weighted value are smaller simultaneously, the conventional fusion method probably generates a successful authentication result. This is contrary to reality, i.e., the authentication result is incorrect and it will lower the reliability of the multimodal authentication system. To address this issue, an adaptive weighted fusion strategy is presented as follows

$$f_{decision} = f\,(t_1) \cdot f\,(t_2). \tag{12}$$

where $f\,(t_1)$ and $f\,(t_2)$ represent the sum and product decisions, respectively. The specific expressions are formulated as follows

$$f\,(t_1) = \begin{cases} 1, & t_1 \geq t_{sum} \\ 0, & t_1 < t_{sum}, \end{cases} \quad f\,(t_2) = \begin{cases} 1, & t_2 \geq t_{pro} \\ 0, & t_2 < t_{pro}. \end{cases}$$

where t_{sum} and t_{pro} represent the threshold values for the sum and product fusion scores, respectively. By utilizing equations (10) and (11), we can employ equation (12) to make the final authentication decision.

By taking advantage of the proposed adaptive fusion method, the matching measurements of two biometric features can be effectively preserved and the environment influence is fully considered, which significantly improves the authentication accuracy.

IV. SIMULATION EXPERIMENTS AND THE ANDROID-BASED MULTIMODAL AUTHENTICATION SYSTEM

Since the time consumption of validating a large amount of cross-testing experiment is large, in this section, we firstly implement the cross validation scheme on PC, and then introduce our developed multimodal authentication system for Android-based smart terminal in detail. To validate the efficiency of the developed system, we implement extensive testing experiments on Android-based smart terminal and illustrate the result analysis.

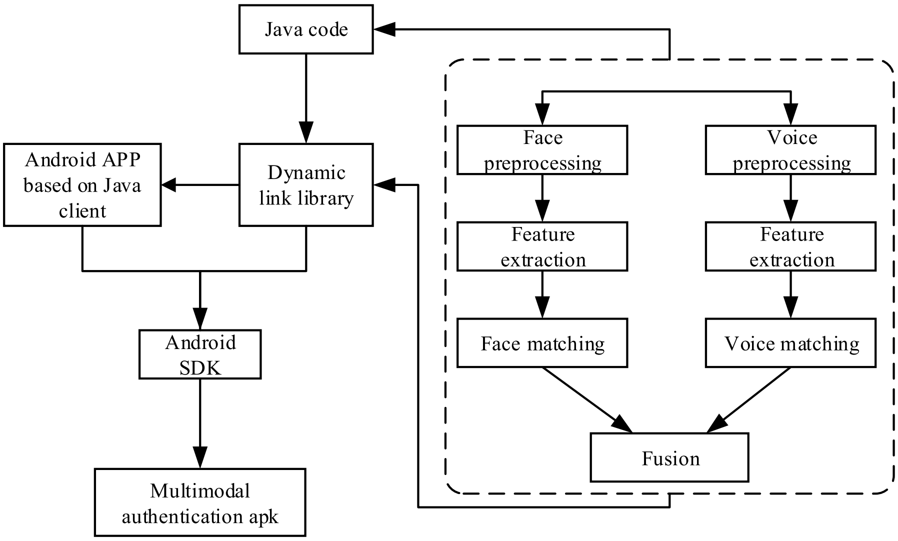


FIGURE 14. Flow chart of the development process.

TABLE 4. Specific assignment of the weighted value *a*.

SNR(db) of the voice signal	(45,+∞]	(40,45]	(35,40]	(30,35]	(25,30]	(20,25]	(15,20]	(10,15]	(0,10]
Weighted value assignment	0.50	0.45	0.40	0.35	0.30	0.25	0.20	0.15	0.10

TABLE 5. Cross-validation results compared with the presented method, face authentication and voice authentication methods.

Authentication methods	TAR(%)	FRR(%)	FAR(%)	MT(s)
Face authentication	98.78	1.22	0.98	0.302
Voice authentication	89.02	10.98	1.09	0.241
Presented method	100.00	0.00	0.00	0.341

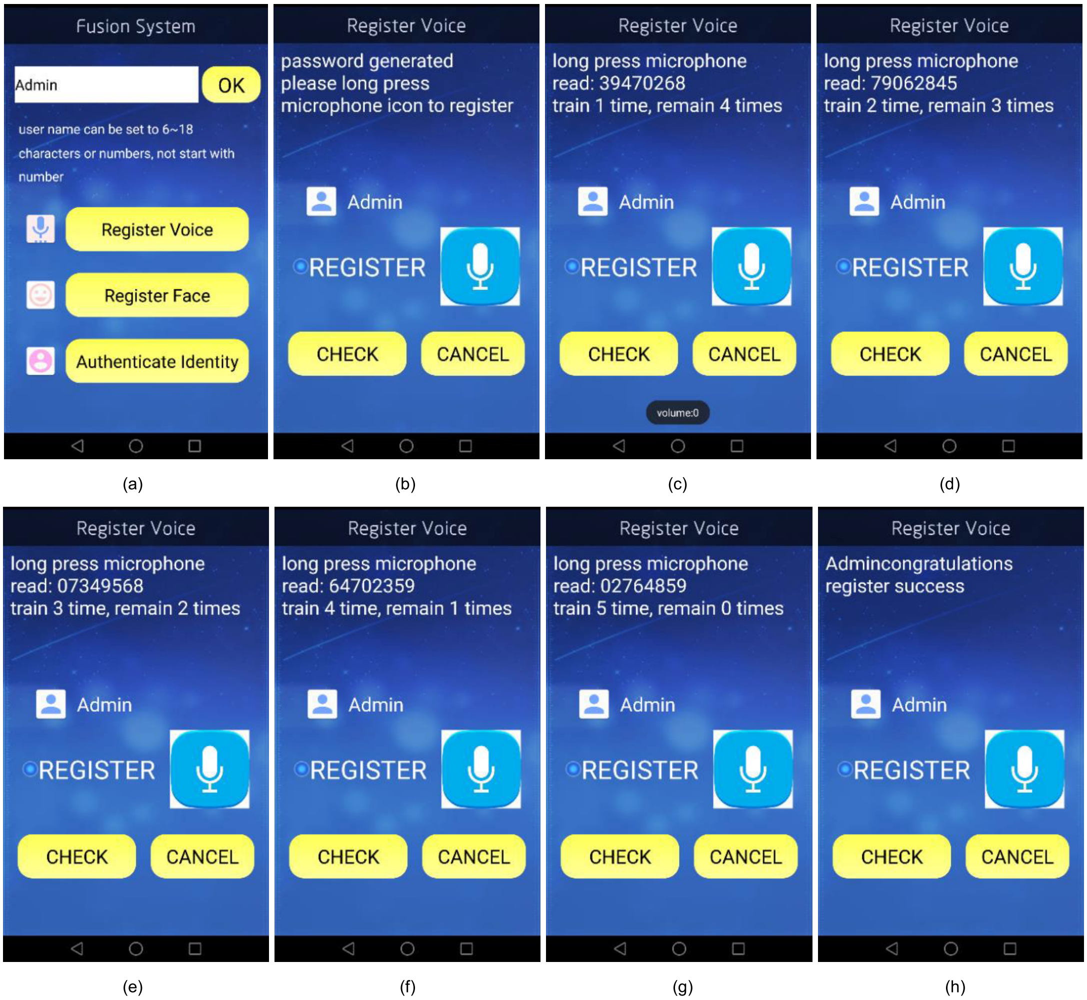
A. SIMULATION EXPERIMENTS

To validate the effectiveness of the presented multimodal authentication method, we firstly implement the cross-validation experiments with the XJTU multimodal database on PC. The XJTU multimodal database contains the face images and voice streams collected from 102 volunteers. For each volunteer, we select five face images and voice streams to form the training set, and the remainder samples constitute the testing set. We employ the evaluation metrics of the true accept rate (TAR), false reject rate (FRR), and false accept rate (FAR) to evaluate the effectiveness of the presented method. We regard the testing samples of the next individual as the negative testing samples of the current individual. We simulate the presented multimodal authentication method and compare it with the face authentication method by utilizing LBP feature and the voice authentication method by utilizing the VAD and GMM methods, respectively. The detailed experiment results are demonstrated in Table 5. In order to evaluate the time consumption of the proposed multimodal authentication methods, the mean time (MT) of each testing sample is calculated and the corresponding result is also listed in Table 5.

It is easy to find that the presented multimodal authentication method achieves excellent accuracy (i.e., TAR is 100%, FRR is 0%, and FAR is 0%). This means that the presented multimodal authentication method is efficient for identity authentication. Compared with the face and voice unimodal authentication results, the FRR of the presented authentication method increases about 1.22% and 10.98%, respectively. Although the time consumption of the presented method is greater than the unimodal authentication cases, the difference is very small (i.e., it is 0.039s compared with the face authentication case), and it is acceptable.

To further validate the effectiveness of the proposed authentication method, we demonstrate the cross-validation experiments with a public face and voice database (i.e., GT_DB and TIMIT databases, provided by Georgia Institute of Technology, Texas Instruments, Massachusetts Institute of Technology and Stanford Research Institute), and compare it with some state-of-the-art multimodal authentication methods. Table 6 illustrates the specific experimental results. It is easy to find that the presented method achieves a considerable authentication accuracy, which is 1.14% and 0.85% higher compared with the method in [3] for TAR and FAR respectively. Although the authentication performance of our method is lower than [12], the gap is trivial, and the authentication speed of our method is much faster than that of [12]. This implies that our method greatly improves the efficiency on the premise of ensuring authentication effectiveness.

Through the above experiment result analysis, the presented multimodal biometric authentication method has excellent comprehensive performance and it can efficiently satisfy the practical application requirements.



Fusion System

Admin

user name can be set to 6~18 characters or numbers, not start with number

Register Voice

password generated please long press microphone icon to register

long press microphone read: 39470268 train 1 time, remain 4 times

long press microphone read: 79062845 train 2 time, remain 3 times

long press microphone read: 07349568 train 3 time, remain 2 times

long press microphone read: 64702359 train 4 time, remain 1 times

long press microphone read: 02764859 train 5 time, remain 0 times

Admincongratulations register success

volume:0

(a) (b) (c) (d) (e) (f) (g) (h)

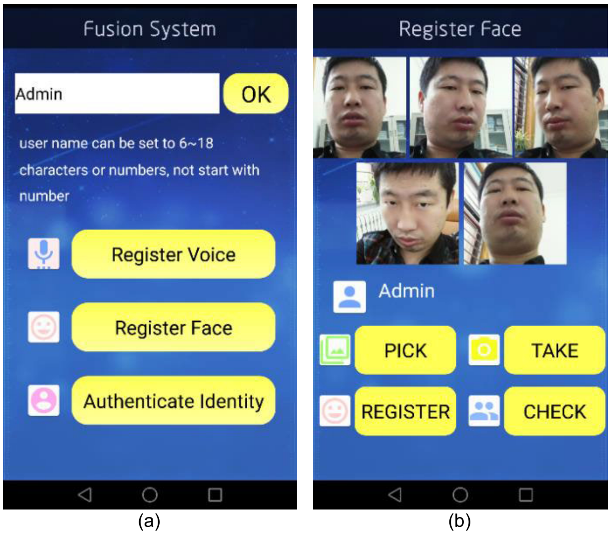
FIGURE 15. Voice registration process. (a) Main interface. (b)-(h) Voice registering process.

TABLE 6. Cross-validation results for public databases in comparison with some state-of-the-art multimodal authentication methods.

Database	Authentication methods	TAR(%)	FRR(%)	FAR(%)	MT(s)
XJTU database	Ref [3] Ref [12] Our method	99.61 100.00 100.00	0.39 0.00 0.00	0.20 0.00 0.00	0.354 0.852 0.341
GT_DB and TIMIT databases	Ref [3] Ref [12] Our method	89.14 90.43 90.28	10.86 9.57 9.72	10.14 9.14 9.29	0.349 0.827 0.339

B. FRAMEWORK FOR SYSTEM DESIGNING AND DEVELOPMENT

The developed system consists of two main modules, i.e., the registration module and the authentication module. For any user, firstly, input the user's name to determine whether it has been registered or not. If not, collect its face and voice to implement the registration procedure. Otherwise, the user should execute the authentication step. Then we can input the face and voice biometric information to implement the matching fusion, and determine whether the authentication is passed or not based on the developed authentication system. The overall architecture flow chart of the developed



Fusion System

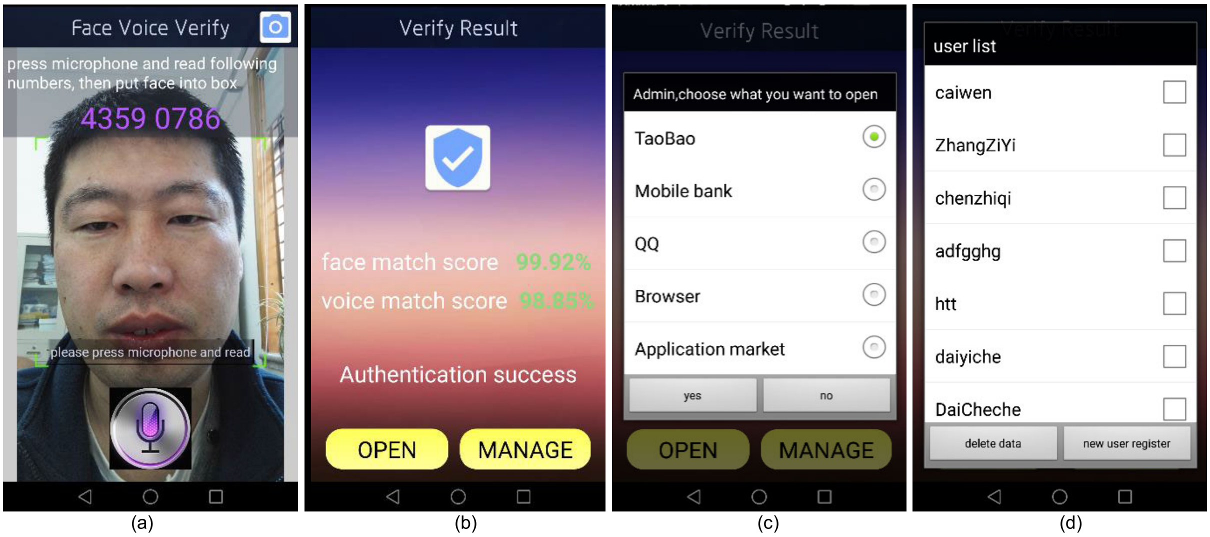
Admin

user name can be set to 6~18 characters or numbers, not start with number

Register Face

(a) (b)

FIGURE 16. Face registration process. (a) Main interface. (b) Face register interface.



(a)

Face Voice Verify

press microphone and read following numbers, then put face into box

4359 0786

please press microphone and read

(b)

Verify Result

face match score 99.92%

voice match score 98.85%

Authentication success

(c)

Admin,choose what you want to open

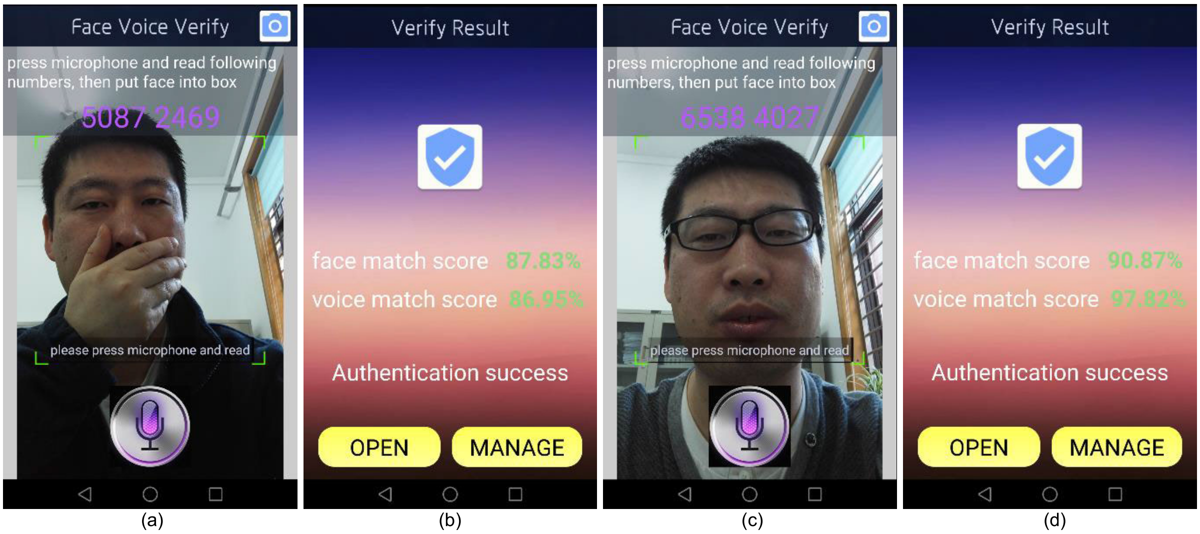
- ● TaoBao
- ○ Mobile bank
- ○ QQ
- ○ Browser
- ○ Application market

(d)

user list

- □ caiwen
- □ ZhangZiYi
- □ chenzhiqi
- □ adfgghg
- □ htt
- □ daiyiche
- □ DaiCheche

FIGURE 17. Authentication process. (a) Feature collection. (b) Authentication result. (c) Application list. (d) User database.



(a)

Face Voice Verify

press microphone and read following numbers, then put face into box

5087 2469

please press microphone and read

(b)

Verify Result

face match score 87.83%

voice match score 86.95%

Authentication success

(c)

Face Voice Verify

press microphone and read following numbers, then put face into box

6538 4027

please press microphone and read

(d)

Verify Result

face match score 90.87%

voice match score 97.82%

Authentication success

FIGURE 18. Authentication instance with disguise. (a) Covered disguise. (b) Authentication result. (c) Glass disguise. (d) Authentication result.

multimodal biometric authentication system is illustrated in Fig. 13.

Based on the authentication system framework illustrated in Fig.13, we adopt the Java platform to accomplish system development.

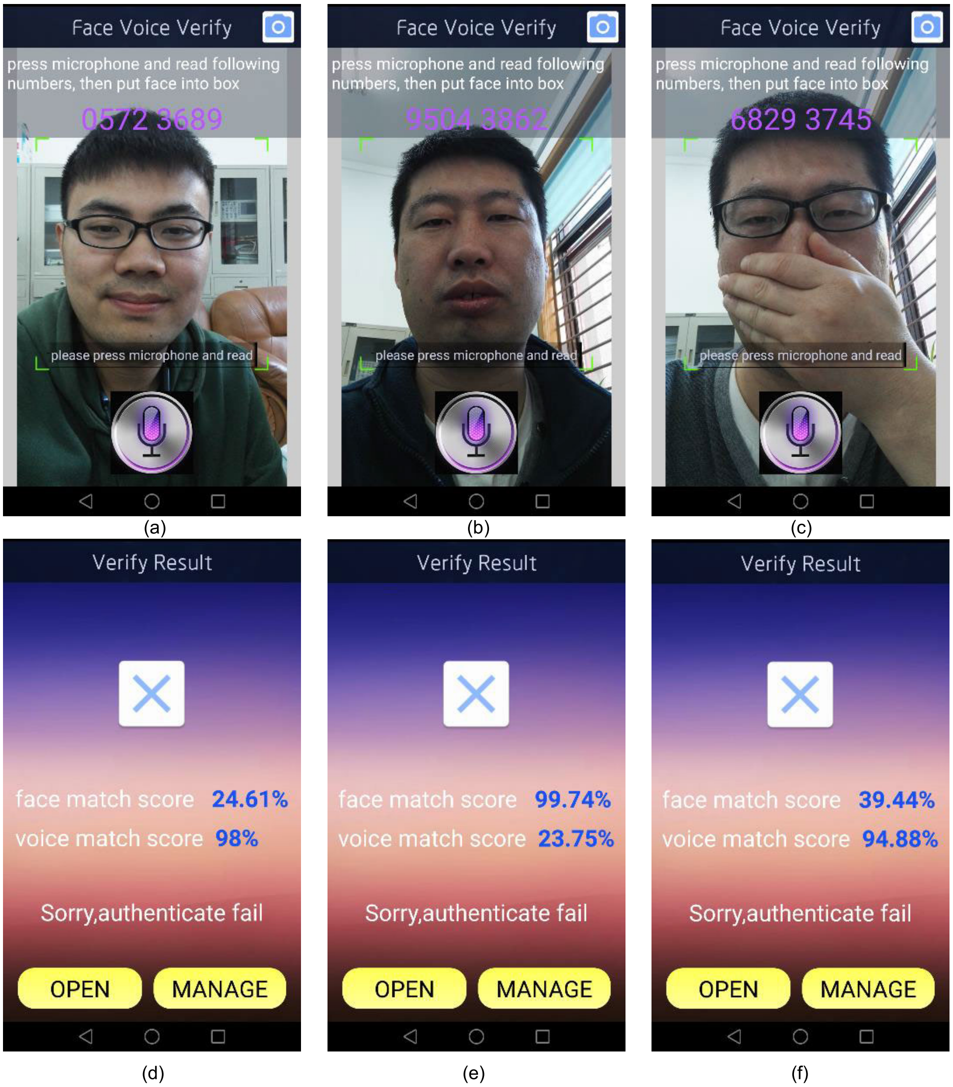
Firstly, the processes of biometric information collection (i.e. face and voice biometrics), feature extraction, feature matching, and fusion authentication are programmed. Then, the code is debugged on the Android-based studio platform, and the SDK tool is used to compile and generate the dynamic link library. Finally, the Java client is used to generate the Android-based SDK, and Android-based application for the multimodal authentication system in apk format is encapsulated and generated. The flow chart of the specific process is illustrated in Fig. 14.

C. INTRODUCTION AND TESTING EXPERIMENTS FOR THE MULTIMODAL AUTHENTICATION SYSTEM

1) INTRODUCTION OF THE MULTIMODAL AUTHENTICATION SYSTEM

In this subsection, we introduce our developed Android-based multimodal authentication system. Fig. 15 (a) illustrates the main interface of the system. Since it is composed of the registration module and the identity authentication module, we will introduce them separately.

(1) Registration. As illustrated in Fig. 15(a), input a new user's name and then click the "OK" button to add it. Then users can click the "Register Voice" button to initiate the voice registration procedure and enter into the voice registration interface (as shown in Fig. 15(b)). Click the "REGISTER" button in Fig. 15(b), the system will generate



Face Voice Verify
press microphone and read following numbers, then put face into box
0572 3689
please press microphone and read

(a)

Face Voice Verify
press microphone and read following numbers, then put face into box
9504 3862
please press microphone and read

(b)

Face Voice Verify
press microphone and read following numbers, then put face into box
6829 3745
please press microphone and read

(c)

Verify Result
face match score 24.61%
voice match score 98%
Sorry,authenticate fail

(d)

Verify Result
face match score 99.74%
voice match score 23.75%
Sorry,authenticate fail

(e)

Verify Result
face match score 39.44%
voice match score 94.88%
Sorry,authenticate fail

(f)

FIGURE 19. **Failure instances of fusion authentication system. (a)-(c) Cases of imitated face, imitated voice and severe occlusion. (d)-(f) Results of cases (a), (b) and (c), respectively.**

five random number registration code and a successive messages with content "password generated, please long press microphone icon to register" (Fig. 15 (b)). According to this prompt, the user can complete the input of 5 training voice (see Fig. 15 (c)-(g)). If there is an error or omission in reading, the system will prompt the error message at the top of the screen, and the user needs to repeat it according to the prompt. With the five voice streams recorded accurately, the prompt message will display "XXX, congratulations, register success" at the top of the screen (see Fig. 15(h)), which indicates that the voice registration procedure has been completed. Where, "XXX" represents the name of the current registering user. After the voice registration, the authentication system will return to the main interface. Then users can click the "Register Face" (Fig. 16(a)) button to turn into the face registration interface. The face registration images can be captured by the system camera or selected from the system album. Click the "PICK" or "TAKE" button to generate five training samples (Fig. 16(b)), and click the "REGISTER" button to complete the face registration.

(2) Authentication. With the success of face and voice biometric registration, the user's face and voice biometric information will be enrolled into the system database and the authentication process can be executed.

For the authentication procedure, the developed system integrates the face and voice collection in the same step. After entering the authentication interface, the system will randomly generate 8-bit number as the voice authentication code (see Fig. 17(a)). Then press the microphone button and read the authentication code according to the prompt information in the interface. At the same time, put the face into the picture collection box and release the microphone button to trigger the camera shooting operation. After taking the photo, the system skips into the authentication result interface (see Fig. 17(b)) and the authentication results are displayed on the screen.

If the matching degrees of face and voice both achieve their thresholds respectively, the authentication succeeds. Then the fusion authentication system gives users the permission for further operation. For example, the user can click on the "OPEN" button to jump into the higher security required applications (such as Taobao, Mobile bank, QQ, and Application market, etc. See Fig. 17(c)). Users can also click on the button of "MANAGE" to enter the user's information management interface (Fig. 17(d)).

2) TESTING EXPERIMENTS FOR THE MULTIMODAL AUTHENTICATION SYSTEM

Fig. 17 gives an authentication example in the case of without occlusion or disguise, where the face matching score is 99.92%, the voice matching score is 98.85%, and the authentication succeeds. In this case, our system achieves an excellent authentication effect.

To verify the effectiveness of the developed multimodal authentication system, we test some other cases, such as face disguise and voice imitation. Fig. 18(a) and (c) show the authentication instances when the user is disguised by cover and glasses. Fig. 18(b) and (d) illustrate the corresponding authentication results, respectively. Fig. 18(b) shows that, in the case of cover disguise, the face matching score of our system is 87.83%, the voice matching score is 86.95%, and the fusion authentication succeeds. In this case, cover disguise generates some influences on the user's voice, therefore the matching score of the voice is lower than the normal case, but an ideal fusion authentication result can still be obtained. Similarly, the glasses also have an impact on the face matching score, i.e., it decreases about 9%, but the system authenticates exactly. The experimental results show that our authentication system is robust to face disguise and slight changes of voice.

Fig. 19 demonstrates some failure instances of the authentication test. Fig. 19(a) shows the authentication with the user's voice and imitated face, Fig. 19(b) shows the case of imitated voice and the user's face, and Fig. 19(c) shows the case of severe occlusion. Meanwhile, Fig. 19(d), (e) and (f) list the authentication results corresponding to the above three cases, respectively.

As demonstrated in Fig. 19, for the case of face imitating (see Fig. 19(a) and (d)), although the voice matching score is still very high, which is up to 98%, the face matching score is very low (only 24.61%), our authentication system can still give the result of authentication failure. Similar results can be obtained for the case of voice imitating (see Fig. 19(b) and (e)). In addition, when the face disguise is too severe (see Fig. 19(c) and (f)), our authentication system also gives the result of no passing. This is because excessive disguise significantly reduces the score of face matching, although the voice matching score is high, the system still gives the wrong authentication result. It can be found from the above tests that, in the cases of high single feature matching scores, the developed multimodal authentication system still gives the failure authentication results.

When the matching scores of the two biometric features are both higher than the thresholds, the authentication succeeds and users can enter some applications which equipped with higher security levels. This characteristic perfectly meets the high security requirements of identity authentication in mobile payment and financial services.

V. CONCLUSION

In this paper, an efficient Android-based multimodal biometric authentication system with face and voice biometrics is developed. In this system, an improved LBP coding-based feature extraction method is introduced to decrease the time and space complexity. We also present an improved VAD method to lower the misjudgment ratio for the voice endpoint, discard the invalid voice segment, and boost the algorithm effectiveness in the low SNR case. Considering the hardware performance of the Android-based smart terminal, we present an adaptive fusion strategy to implement the multimodal biometric fusion authentication, which overcomes the shortcomings of the unimodal biometric authentication and effectively improves the authentication performance. Experimental results show that the developed authentication system can well implement the identity authentication under various scenarios, and realize the management operation with high-security applications.

Although the developed multimodal authentication system has many merits, there still exist some disadvantages. For example, only the face and voice biometrics are considered to realize identity authentication. Many other biometrics are not studied. The registering process needs too much training data, which will lower the user's experience. For the matching processes, although some improvements have been studied, there are still many improvements need to be studied. These are also the focuses and directions of our work. In future works, we will research some other biometrics, such as the ECG, fingerprint, and so on, to implement multimodal authentication, study the deep learning framework based on the mobile terminal and try our best to boost the authentication accuracy while reducing the training data.

ACKNOWLEDGMENT

This article was presented at the IEEE 4th International Conference on Computer and Communications (ICCC 2018).

REFERENCES

[1] S. Soviany and M. Jurian, "Multimodal biometric securing methods for informatic systems," in *Proc. 34th Int. Spring Seminar Electron. Technol. (ISSE)*, May 2011, pp. 447–450.

[2] M. Hammad, Y. Liu, and K. Wang, "Multimodal biometric authentication systems using convolution neural network based on different level fusion of ECG and fingerprint," *IEEE Access*, vol. 7, pp. 26527–26542, 2019.

[3] H. Aronowitz, M. Li, O. Toledo-Ronen, S. Harary, A. Geva, S. Ben-David, A. Rendel, R. Hoory, N. Ratha, S. Pankanti, and D. Nahamoo, "Multimodal biometrics for mobile authentication," in *Proc. IEEE Int. Joint Conf. Biometrics*, Sep. 2014, pp. 1–8.

[4] M. Hammad and K. Wang, "Parallel score fusion of ECG and fingerprint for human authentication based on convolution neural network," *Comput. Secur.*, vol. 81, pp. 107–122, Mar. 2019.

[5] D. Valdes-Ramirez, M. A. Medina-Perez, R. Monroy, O. Loyola-Gonzalez, J. Rodriguez, A. Morales, and F. Herrera, "A review of fingerprint feature representations and their applications for latent fingerprint identification: Trends and evaluation," *IEEE Access*, vol. 7, pp. 48484–48499, 2019.

[6] D.-J. Kim, K.-W. Chung, and K.-S. Hong, "Person authentication using face, teeth and voice modalities for mobile device security," *IEEE Trans. Consum. Electron.*, vol. 56, no. 4, pp. 2678–2685, Nov. 2010.

[7] S. Thavalengal, P. Bigioi, and P. Corcoran, "Iris authentication in handheld devices–considerations for constraint-free acquisition," *IEEE Trans. Consum. Electron.*, vol. 61, no. 2, pp. 245–253, May 2015.

[8] J. Fierrez-Aguilar, J. Ortega-Garcia, and J. Gonzalez-Rodriguez, "Fusion strategies in multimodal biometric verification," in *Proc. Int. Conf. Multimedia Expo. ICME*, Jul. 2003, pp. 5–8.

[9] A. Kumar, D. C. M. Wong, H. C. Shen, and A. K. Jain, "Personal verification using palmprint and hand geometry biometric," in *Proc. Int. Conf. Audio-Video-Based Biometric Person Authentication (AVBPA)*, in Lecture Notes in Computer Science, vol. 2688. Berlin, Germany: Springer, Jun. 2003, pp. 668–678.

[10] K.-A. Toh, X. Jiang, and W.-Y. Yau, "Exploiting global and local decisions for multimodal biometrics verification," *IEEE Trans. Signal Process.*, vol. 52, no. 10, pp. 3059–3072, Oct. 2004.

[11] S. Shekhar, V. M. Patel, N. M. Nasrabadi, and R. Chellappa, "Joint sparse representation for robust multimodal biometrics recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 113–126, Jan. 2014.

[12] H. Zhang, V. M. Patel, and R. Chellappa, "Low-rank and joint sparse representations for multi-modal recognition," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4741–4752, Oct. 2017.

[13] X. Zhang, D. Cheng, Y. Dai, and X. Xu, "Multimodal biometric authentication system for smartphone based on face and voice using matching level fusion," in *Proc. IEEE 4th Int. Conf. Comput. Commun. (ICCC)*, Dec. 2018, pp. 1468–1472.

[14] Y. Duan, J. Lu, J. Feng, and J. Zhou, "Context-aware local binary feature learning for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1139–1153, May 2018.

[15] W. Yang, Z. Wang, and B. Zhang, "Face recognition using adaptive local ternary patterns method," *Neurocomputing*, vol. 213, pp. 183–190, Nov. 2016.

[16] W. Xia, S. Yin, and P. Ouyang, "A high precision feature based on LBP and Gabor theory for face recognition," *Sensors*, vol. 13, no. 4, pp. 4499–4513, Apr. 2013.

[17] S. Nakagawa, L. Wang, and S. Ohtsuka, "Speaker identification and verification by combining MFCC and phase information," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 4, pp. 1085–1095, May 2012.

[18] T. Kinnunen, R. Saeidi, F. Sedlak, K. A. Lee, J. Sandberg, M. Hansson-Sandsten, and H. Li, "Low-variance multitaper MFCC features: A case study in robust speaker verification," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 7, pp. 1990–2001, Sep. 2012.

[19] J. Jo, H. Yoo, and I.-C. Park, "Energy-efficient floating-point MFCC extraction architecture for speech recognition systems," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 24, no. 2, pp. 754–758, Feb. 2016.

[20] Y. H. Chao, W. H. Tsai, and H. M. Wang, "Improving GMM-UBM speaker verification using discriminative feedback adaptation," *Comput. Speech Lang.*, vol. 23, no. 3, pp. 376–388, 2009.

[21] C. H. You, A. L. Kong, and H. Li, "GMM-SVM kernel with a Bhattacharyya-based distance for speaker recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 6, pp. 1300–1312, Aug. 2010.

[22] P. K. Nayana, D. Mathew, and A. Thomas, "Comparison of text independent speaker identification systems using GMM and i-vector methods," *Procedia Comput. Sci.*, vol. 115, pp. 47–54, Jan. 2017.

[23] Q. Wang, W. L. Woo, and S. S. Dlay, "Informed single-channel speech separation using HMM–GMM user-generated exemplar source," *IEEE Trans. Audio, Speech, Language Process.*, vol. 22, no. 12, pp. 2087–2100, Dec. 2014.

[24] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 886–893.

[25] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. CVPR*, Dec. 2001, pp. 511–518.

[26] C. Zhu and R. Wang, "Local multiple patterns based multiresolution gray-scale and rotation invariant texture classification," *Inf. Sci.*, vol. 187, pp. 93–108, Mar. 2012.

XINMAN ZHANG received the Ph.D. degree in control science and engineering from Xi'an Jiaotong University, Xi'an, Shaanxi, China, in 2004.

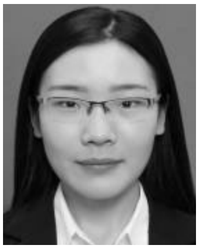
From 2004 to 2020, he was a Teacher with the MOE Key Laboratory for Intelligent Networks and Network Security, School of Automation Science and Engineering, Xi'an Jiaotong University. His research interests include biometric recognition, machine vision, and video analysis.

DONGXU CHENG received the B.S. degree in information and computing science and the M.S. degree in applied mathematics from the Xi'an University of Technology, Xi'an, Shaanxi, China. He is currently pursuing the Ph.D. degree with the School of Automation Science and Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi.

His research interests include biometric recognition, machine vision, and video analysis.

PUKUN JIA received the B.S. degree from the School of Automation, Chongqing University of Posts and Telecommunications, Chongqing, China.

His research interests include biometric recognition, machine vision, and video analysis.



YIXUAN DAI received the B.S. and M.S. degrees from the School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi, China.

Her research interests include biometric recognition, machine vision, and video analysis.

XUEBIN XU received the Ph.D. degree from the School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi, China.

His research interests include 5G/6G and the Internet-of-things technology, deep learning/artificial intelligence, and brain functional science/bioinformatics.