SALM: Smartphone-based Identity Authentication Using Lip Motion Characteristics

Yaoxuan Yuan*, Jizhong Zhao*, Zhe Zhang*, Wei Xi*, Chen Qian†, Xiaobin Zhang*, Zhi Wang*
* School of Electronic and Information Engineering, Xi'an Jiaotong University
†Department of Computer Engineering, University of California, Santa Cruz

***Abstract*—With rapid development and popularity, smartphones have been of importance in our daily life. Despite of its convenience in communication and computing, smartphones also lead potential security threats to users. Existing methods on smartphones for protecting user's privacy mainly depend on password or fingerprint based authentication. Most smartphone passwords are very simple and easy to guess or crack, and fingerprinting requires extra hardware and hence increases the price of smartphones. In this paper, we present a smartphone-based identity authentication method based on user's lip motion characteristics, called SALM, which can be used as an additional authentication with password. SALM extracts the feature of lip movements as the authentication token, which is unique for each user. We implement SALM using off-the-shelf smartphones and evaluate its performance via extensive experiments. The results show that the overall accuracy of user authentication using SALM (without password) is higher than 96%.**

***Index Terms*—Lip Contour Extraction, Lip Motion Model, User Authentication, Model Fitting.**

I. Introduction

Smartphones are indispensable in many people's daily life. Besides the basic communication functions, smartphones also play important roles in social networking, on-line bank transactions, digital photo album, and other personal applications [15] [16]. Therefore securing smartphone access is crucial for protecting user privacy [8].

There are two popular user authentication methods for smartphones: password based and biological information based solutions. Each of them has disadvantages. Most smartphone passwords are very simple and easy to guess or crack. In addition passwords are vulnerable to overhearing. Biological information, such as the fingerprint, iris, and voice, is unique, measurable and unchangeable for each user. It is much more difficult to be cloned, providing stronger enhancement to the security of smartphones. Since it does not require the user to carry on any extra device, biological information based authentication is especially suitable for building up the secure authentication schemes for smartphones.

Existing biological information based approaches, however, are with shortcomings. Fingerprinting requires extra hardware and hence increases the price of smartphones. The face recognition based authentication may be deceived by using a photo. The iris based authentication requires high resolution cameras, which are also costly or even do not exist on most smartphones. Voice based authentication also suffers from duplicated voice records. In addition, environmental noise would decrease its accuracy.

In this paper, we leverage the lip movement as a user's unique feature and authentication token. The lip movement, similar to other biological information, is uniquely distinguishable among individual users [2]. We design and development a lip movement based authentication scheme, namely SALM. SALM works as follows. We first capture sufficient lip dynamic images from users via the front cameras on their smartphones. The captured images are pre-processed. Two necessary processes are also performed: the lip contour extraction and lip feature extraction. SALM then establishes a lip contour dynamic model. Using this model, SALM can authenticate legitimate users by matching their lip dynamic movement with the data stored in backend systems.

The main contributions of this work are summarized as follows:

- We design a new biological information based authentication approach, namely SALM. SALM is based on user's unique lip movement. SALM is speech-free, *i.e.*, independent to user's voice and speaking content.
- SALM is fully compatible with existing smartphones and does not require any extra hardware. From the user's perspective, the lip movement is dynamic and hard to recorded and replicated. Thus, the security of smartphone and user's private information is highly enhanced.
- SALM consists of several accurate processes, *e.g.*, the ASM based lip contour extraction, which ultimately contribute high authentication accuracy and robustness. We implement SALM on commodity off-the-shelf smartphone platforms. Extensive experiments show that the proposed method achieves high accuracy.

The rest of this paper is organized as follows: Section II briefly reviews the related works. We present our observations on people's talking and their lip movements in Section III. We elaborate SALM design in section IV. In section V, we evaluate the performance of SALM. We conclude our paper in section VI.

II. Related Work

Recently, user authentication for smartphones has attracted much attention. There are some lip-based Authentications studies related to our, The solutions can be classified into two categories: Static lip characteristics-based authentication and dynamic lip characteristics-based authentication.

978-1-5090-6517-2/17/$31.00 ©2017 IEEE

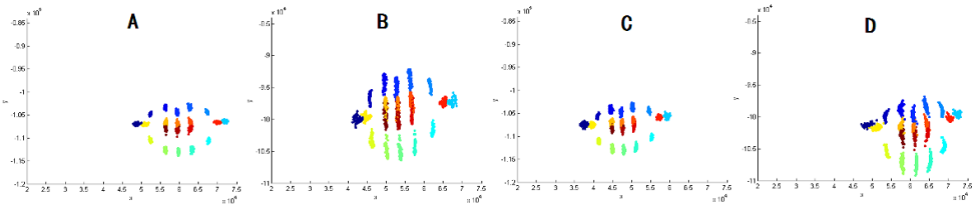

Fig. 1. Different users have different lip movements when they speak a same sentence. Here A, B, C, and D represent four different users, respectively

A. Static lip characteristics-based authentications

Tsuchihasi discovers that lip print characteristics can be used as a new biometric fingerprint [2] [1]. Through the image preprocessing and lip contour extraction technology, using the contour feature of the lip, it has achieved a good success rate [3]. Petajan *et al.* [4] use ASM model to further improve the biometric authentication rate of static lips. The success rate of the lip print certification and recognition is satisfactory, but it requires accurate extraction of lip texture features.

B. Dynamic lip characteristics-based authentication

The literature of dynamic lip characteristics-based authentication is divided into two categories: lipreading recognition and lip moving certification based on dynamic features of lip.

Recognition based on lipreading: The Lip Reading technology [17] was developed by Sumby and Pollack firstly [19]. It uses of visual information technology for language perception of the speaker. Petajan established the first automatic lip reading system [4]. Silsbee et al.use the lipreading technology in speech recognition field to improve the accuracy of the speech recognition system [18].

Authentication based on lip motion: Recently, there is little research using the dynamic features of the lip movement for authentication. The work of Faraj and Bigun [5] combined lip movement and speech features in order to strengthen biometric authentication. However, voice certification needs an environment with small noise, and lip motion features highly dependent on the specific content of the language. This makes its scope having big limitations. A recent work proposes a method depending on lip motion features [6], and uses the B spline model to track the shape of the speaker's lips. Then the DTW method is used to match the same content of the lip before authentication. This method has the following shortcomings: (1) The speaker needs to say a specific content. (2) The accuracy is low because the information contained in the extracted feature is limited.

III. Observations and Verification

We first observe the lip movement in reality and analyze the lip movement during speech. When people are speaking, the brain controls the facial muscles, which changes the outlook of lip.

There are four distinct features in lip movements: 1) when a user is speaking, her lip movement is unique compared to other people. 2) Usually, each user has his/her own speech speed. 3) For each user, the transforming procedure of lip shapes is unique. 4) Each user has his/her unique lip contour due to the diversity of genes.

With above features, it is possible to find differences of lip movement patterns among different users, which acts as the biological feature for authentication.

To verify our claim of the feasibility of using lip movement as an authentication token, we conduct a set of experiments.

1) Four volunteers speak a same sentence for 5 seconds. We record their speeches by a video camera and show the distribution of their lip contour movements in Figure 1. We set several positions in the lip contour. We cluster those position points observed in 5 seconds and plot those for a same position in a same color in Figure 1. Actually, the motion of each position in the lip and as well as the locations outlines the lip movement. It shows that the shape of those clusters and distance among points differ significantly among variant users, even if they speak the same sentence.

2) Each user speaks two different sentences chosen by themselves. Figure 2 shows the distribution of the lip contour points. We discovery two observations. First, the lip contour points distribution is *speech-free*, that the distributions are very similar for the same person despite speaking different sentences. Second, the lip contour points distribution is *user-sensitive*, that the distributions are quite different across different users even though they speak the same sentences.

Results from the above two sets of experiments verify that people have their own unique speech lip movements, which validates that it is feasible to authenticate a user by identifying the characteristics of the user's lip movement.

IV. SALM Design

In this section, we present the overview of SALM. Then we detail its design in four modules.

A. Overall System Design

Our proposed lip movement based authentication system is implemented on commercial off-the-shelf smartphones. A user only needs to speak a few words determined by herself in front of the camera, and no restriction is applied to the speech content. lip-movement biometric based authentication system is secure and easy to use.

Figure 3 shows the architecture of our system. The SALM system works in two phases: 1) Training phase: it includes face region detection, lip contour extraction, lip feature extraction, and lip movement model extraction. 2) Authenticating phase: it includes lip contour dynamic model and gaussian mixture model (GMM) extraction, and final user authentication step.

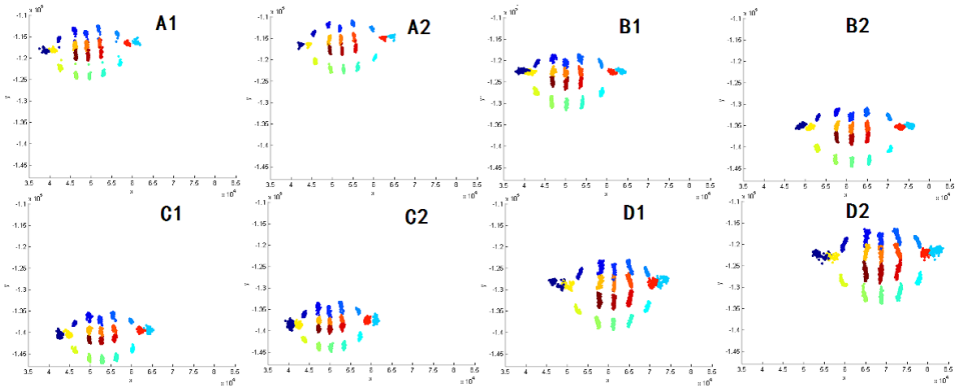

Fig. 2. Different people speaking the same content; A, B, C, D, represent four different people respectively; 1, 2 represent two different sentences respectively

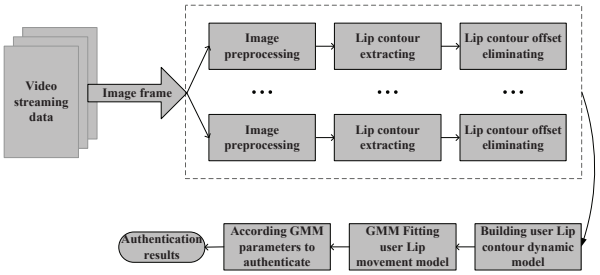

Fig. 3. The work flow of the authentication system

B. Precise Lip Contour Extraction using Enhanced ASM Model

After image preprocessing, we locate the human face region using Viola-Jones detector [27], [31], and extract the lip contour from the face region using our enhanced Active Shape Model (ASM) algorithm.

The mainstream contour extraction approaches can be divided into two categories: parameter optimization based [22], [24] and iterative regression based [23]. The former approach main relies on the optimization of the error function to obtain the optimal shape estimation, while the latter uses a regression function to map the contours directly with the target using machine learning techniques.

We propose an idea of contour extraction: we extract the lip contour based on the parameter optimized ASM model method. However, due the inaccuracy of the ASM model when the shape is reconstructed, we use the method based on iterative regression to calculate the exact lip contour by training the global character of the shape model to minimize the regression factor.

1) Training Patch Model: The patch model represents the user's sub template in the image, which is mainly used to detect the location of each user's feature points [9]. For example, the work in [6] employs a complex model to improve the generality of the detector. However, the calculation of characteristic points becomes very complex. Since the characteristic points of different users are very similar in deformation, we take the advantage of 2D correlation-based feature detectors instead of the one dimensional gradient patch model.

We propose to use sufficient training involving more points to improve the tracking accuracy and robustness [11]. We must carefully select two important parameters: the size of the 2D patch model block L_p and the number of facial feature points N_{ff}. We use 500 images and the corresponding feature points from the Helen face database [30] to train our patch model.

2) Shape Model Training: We use the trained patch model to detect the relative positions among feature points in the face. However, these relative positions must be constrained by the shape model [29]. With it, the shape that composes of the feature points is more consistent with the change of the trained shape.

Suppose each user's profile shape is S. We have $S = \{(x_1, y_1) \ldots, (x_n, y_n)\}$. Consider that (x_i, y_i), $i \in [1, \ldots, n]$ is the coordinates of the i-th feature point. S is a vector of $2n$ dimension. The problem can be solved by decomposing the shape into two parts in the ASM shape model: the variable part and the non-variable part [21].

3) Precise Lip Contour Extraction: We can quickly find a preliminary outline of the lip in the face through the trained shape model. Since we have obtained the initial lip contour by the improved ASM model, we use this lip contour as the initial shape to apply the regression analysis to obtain the accurate contour.

The regression function iteration method is to train a number of weak regression functions $(R^1, \ldots, R^t, \ldots, R^T)$. These functions represent the shape variables' values δS that need to be changed in each iteration. We choose 300 primary weak classifier Fern [10] as the basic element of the weak classifier R. Fern is composed of F features and their corresponding thresholds. It can be divided into 2^F feature space bins. According to [11] [12], the deviation of the current distortion δS_b can be calculated by each bin b.

C. Lip Contour Offset Cancellation

We need to eliminate the error caused by the jitter. These errors are divided into two types. One is the contour angle deviation due to the decline of the head. Another is the head

contour size deviation due to the distance change between the head and the camera.

Here, we choose the center of two eyes as a benchmark. We find the angle between the eye center line and the horizontal line in each frame image $\theta_i(i = 1, 2, \dots N)$, which can be used to solve the first error. About the second error, we take the average value of the two eyes center pixel distance as the user's eyes center distance $\bar{d}$. Next, the eyes center distance of each frame image $d_i(i = 1, 2, \dots N)$ is calculated. Thus, the formula for calculating the size deviation of each frame profile can be expressed as: $S_i = d_i/d(i = 1, 2, \dots N)$. After obtaining the offset angle θ_i of the head and the size deviation S_i of the lip contour, the lip contour offset is eliminated by the rotation and scaling, and the final lip contour data is obtained.

D. Building The Dynamic Model of The Lip Contour

After eliminating the contour deviation, we establish a dynamic model of the lip contour. This process includes:

1) Lip feature extraction. As we know, the lip contour [32] and movement features are not easy to change, so we extract the lip contour feature, and make it the basis for establishing the dynamic model of lip contour. We divide the features of lips into two kinds: geometric feature and shape feature. (1) Geometric feature, the maximum heights of the inner and the outer lips in the vertical direction are: H_{in} and H_{out}. The maximum widths of the inner and outer lips in the horizontal direction are: W_{in} and W_{out}. We extract the average thickness of the upper and lower lips as a feature: T_{up} and T_{dn}. The geometric features mentioned above can fully represent the geometric variation of the lip movement. (2) Shape feature, we take the shape of the basic profile: $Shape = \{(x_i, y_i)\}_{i=1}^{N}$.

2) Establish lip movement model. We build a dynamic model by extracting a video stream of a trainer. Assuming that the video stream have N frame images, based on the above analysis, here we take $N = 300$, for the $i\text{-}th$ frame, we can calculate the feature of his lip as:

$$f_{lip} = \{H_{in}^{(i)}, H_{out}^{(i)}, W_{in}^{(i)}, W_{out}^{(i)}, T_{up}^{(i)}, T_{dn}^{(i)}, Shape^{(i)}, \theta^{(i)}, s^{(i)}\} \tag{1}$$

In order to show the motion feature of the lips, we combine the features of each frame in the video to get a matrix M, M representing the dynamic change of the lip in each frame, we refer to M as the dynamic change matrix of the lips, which is specifically expressed as follows:

$$M = \{(f_{lip}^{(1)})^T, \dots, (f_{lip}^{(i)})^T, \dots, (f_{lip}^{(N)})^T\} \tag{2}$$

In fact, M may include redundant information of the feature and increase the computation cost for the remaining process. Hence we need to reduce the dimension of the matrix M to extract the relevant high dimensional features as the basis for the lip movement model.

After obtaining the dynamic matrix P by feature dimensionality reduction, the feature difference of N frame images is projected onto the feature dimension matrix space. For N frame images, they have N projections in feature dimensionality reduction space. We set these projections to form the user's lip movement model. When the model matching is performed, we need to transform the model coordinate system of the tester. This should make the model coordinate system of the original user consistent with the tester's.

E. User Authentication Based on Lip Dynamic Model

The lip motion model is expressed as a cloud of points in a high dimensional space. Based on our observation, the distribution of this points cloud is similar to the spatial shape of the Gaussian distribution. According to this fact, we use the Gaussian model to fit the distribution of the points cloud. For the diversity of samples, we can also solve it by Gaussian mixture model.

We can get their corresponding Gaussian mixture model through the distribution of points cloud fitting, calculate the Gaussian mixture model's parameters, determine the corresponding parameter interval, and then finish the authentication through the parameters.

Choosing the appropriate K value should be considered together with the authentication accuracy and system efficiency. Here we take $K = 2$ based on empirical resutls. The reason will be introduced in the experiment section.

We define $\overrightarrow{x_1}, \overrightarrow{x_2}, \dots, \overrightarrow{x_N}$ as the points cloud of the lip moving model we have obtained. We assume that these points cloud can be mixed by the K Gaussian models with $K = 2$. Based on the GMM related theories, our Gaussian mixture model is presented as follows:

$$P(\vec{x}|\theta) = \sum_{i=1}^{K} a_k \phi(\vec{x}|\theta_k) \tag{3}$$

1) For any obtained points cloud data $\overrightarrow{x_i}, i = 1, 2, \dots, N$, it is produced in our Gaussian model as the follows. First, according to the probability a_k, we select the $k\text{-}th$ Gaussian model $\phi(\vec{x}|\theta)$. Then the observation data $\overrightarrow{x_i}$ is generated by the probability distribution $\phi(\vec{x}|\theta)$ of the $k\text{-}th$ model. We believe that the observed data $\overrightarrow{x_i}, i = 1, 2, \dots, N$ is known, and the data that reflects the observed data $\overrightarrow{x_i}$ from the $k\text{-}th$ model is unknown, $k = 1, 2, \dots, K$, Here we use the variable γ_{ik} denoted, and it is a random variable of 0 to 1.

So the log likelihood function for the complete points cloud data is:

$$logP(\vec{x}, \gamma|\theta) = \sum_{k=1}^{K} n_k log a_k + \sum_{i=1}^{N} \gamma_{ik} [log(\frac{1}{\sqrt{2\pi}}) - log\sigma_k - \frac{1}{2\sigma_k^2}(\overrightarrow{x_i} - \mu_k)^2] \tag{4}$$

With formula (5), we use the EM algorithm[28] to estimate the parameters.

2) EM algorithm is used to estimate the parameters of Gaussian mixture model.

(1) Select the initial points cloud data, and start iteration.

(2) E step: Based on the current model parameters, calculate the response degree of the sub model k to the point cloud data $\overrightarrow{x_i}$

$$\hat{\gamma_{ik}} = \frac{a_k \phi(\overrightarrow{x_i}|\theta_k)}{\sum_{k=1}^{K} a_k \phi(\overrightarrow{x_i}|\theta_k)} \quad (i = 1, 2, \ldots, N; \quad k = 1, 2, \ldots, K) \tag{5}$$

(3) M step: find the maximum value of the likelihood function of the points cloud data by iterating.

$$\theta^{(i+1)} = \underset{\theta}{\operatorname{argmax}}\, logP(\vec{x}, \gamma|\theta^{(i)}) \tag{6}$$

Here we use $\hat{\mu_k}$, $\hat{\sigma_k}^2$, $\hat{a_k}$, $k = 1, 2, \ldots, K$ to represent the various parameters of $\theta^{(i+1)}$, then we can obtain their results.

(4) Repeat (2) and (3) steps until convergence.

When we get the training data of m users, we take $m = 8$ at this time, Then our user Gaussian mixture model interval can be defined as: $\mu_{kmin} = min\{\mu_{ki}\}_{i=1}^{m}, \sigma_{kmin}^2 = min\{\sigma_{ki}^2\}_{i=1}^{m}, a_{kmin} = min\{a_{ki}\}_{i=1}^{m}, \mu_{kmax} = max\{\mu_{ki}\}_{i=1}^{m}, \sigma_{kmax}^2 = max\{\sigma_{ki}^2\}_{i=1}^{m}, a_{kmax} = max\{a_{ki}\}_{i=1}^{m}$.

After obtaining the lip movement points cloud model of the tested users, the following rules are used for user authentication. If:

$$\begin{cases} \hat{\mu_k} \in [\mu_{kmin}, \mu_{kmax}] \\ \hat{\sigma_k^2} \in [\sigma_{kmin}^2, \sigma_{kmax}^2] \\ \hat{\alpha_k} \in [\alpha_{kmin}, \alpha_{kmax}] \end{cases} \tag{7}$$

the authentication is successful. Otherwise the authentication fails.

V. Experiment And Result Analysis

In this section, we first present the system implementation, the experiment methods, and the methodology for performance evaluation. Then we show the results of real-world experiments to evaluate the performance of the proposed authentication system SALM.

A. Experiment Setup

The hardware equipment used in the experiment was Nexus Google 4. We implement the Android and OpenCV [20] program to retrieve the usage data and process the image data [26].

B. Experimental Data Collection

Since the proposed lip movement authentication system is independent on the content of the speech. We take into account the following fact. Each person has unique lip movement when they speak the same words at each time, and obviously her lip movement is different when she speaks different words. To determine whether the lip movement model is gender-dependent, and whether the model is speech length dependent, etc., we designed the following experimental data collection scheme. We selected 20 different texts as the experimental contents and 20 people including 14 men and 6 women participating in this experiment. The lip movement information of the subjects were captured by the front camera of a smartphone, and each person reads all these 20 paragraphs of text. Each text is repeated by a person for 10 times.

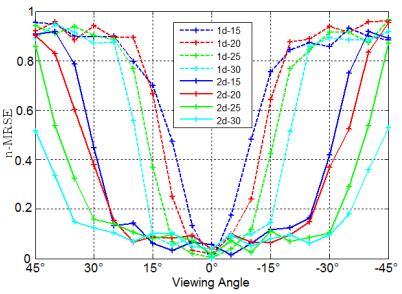

Fig. 4. The template size and n-RMSE in different angles

Through the above experimental scheme, we have a total of 4000 groups of data, including 2800 groups of male speech data and 1200 groups of female speech data.

C. Experimental Results And Analysis

We summarize the results from our real-world experiments.

1) Lip contour extraction experiment: The 2D correlation feature patch model in ASM proposed by us aims to find the location of the Salient points [25]. The standard root mean square error ($RMSE$) of the artificial marked points and the patch model normalized -$RMSE$ were used to evaluate the accuracy of the Salient feature points.

Here we allow 10 participants whose viewing angle with the direction of the phone camera is in the range of $\pm 45°$. A photo is taken each $5°$, resulting in 19 photos for a group. Each person repeats for 5 times, hence a total of 950 photos are collected. We first calculate n-$RMSE$ of the 950 photos by changing the parameter values of the template block size L_p. In order to make a better comparison, we implemented the 1D patch model of [7]. Figure 4 shows the experiment results of n-$RMSE$ by changing the template parameter values L_p from 15 pixels to 30 pixels.

Considering n-$RMSE$, by which the more accurate feature points have been found, we can clearly see the accuracy of the 2D template is higher than the 1D template in Figure 4. For the 2D patch template, we can also see that the accuracy of the larger template block size L_p is always higher than that of the smaller template block. For the ordinary mobile phone users, the angle between the face and the direction of the phone front camera is generally between $\pm 30°$. In the $\pm 30°$ interval, we found that the maximum value of n-$RMSE$ is no more than 0.2 when the size of the model plate is 25px and 30px. We can also see that the error of $L_p = 25$px is smaller than that of $L_p = 30$px in the range of $\pm 15°$. In the range of $[-30°, -15°]$ and $[15°, 30°]$, the error of the $L_p = 25$px and $L_p = 30$px is not very different, hence we set the size of the 2 dimensional patch template value as 25px.

We collected 50 images for each of 20 people and total 1000 pictures from the video data as our experimental data. In order to quantify the accuracy of the lip contour extraction, we

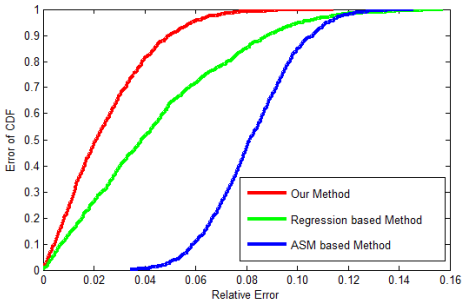

Fig. 5. CDF of lip contour extraction relative error

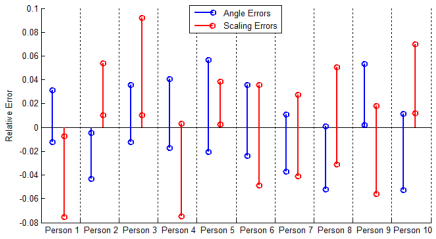

Fig. 6. The relative error range of contour offset elimination

manually marked the contours of these photos as the ground truth. At the same time, we use the traditional ASM model [13] and the regression analysis [14] to extract the lip contour of the 1000 pictures and make a comparison between them. Experimental results are shown in Figure 5.

Figure 5 shows the cumulative distribution of the relative error of the three methods. We can see that the relative error of the proposed lip contour extraction algorithm is less than 0.06 for 95% of cases, and the overall accuracy is significantly higher than the other two algorithms. For the regression analysis, the convergence of the algorithm is not consistent due to the initial location of the lip contour. For the improved ASM model. It improves the patch model of ASM. In the early stage of the lip contour extraction, the location is better than the regression analysis. In the later contour iteration, because it doesn't rely on the limit of the ASM shape model, so it can be more flexible to obtain the precise lip contour.

2) Lip contour offset elimination experiment: The main function of the lip contour offset elimination is to eliminate the jitter and the deviation. In our method, we get the two center points of the two eyes by the shape model of ASM. Then we calculated the angle $\hat{\theta}$ between the connection line of the two center points and the horizontal line. The ratio $\hat{s} = \hat{d}/\bar{d}$ is set as the scaling. $\hat{d}$ is the distance of the two points and $\bar{d}$ is the average distance. The lip contour offset is eliminated by rotation and scaling. Here, we calculate the rectification angle α and the correction ratio γ, and use them to evaluate the contour offset elimination, where, $\alpha = \frac{\theta - \hat{\theta}}{\theta}, \gamma = \frac{s - \hat{s}}{s}$.

In this experiment, we used the data of 10 people, and for each person we collected 100 frames of the lip moving video streaming. The experimental results are shown in Figure 6. We can see that the relative error of the deflection angle is in the range of $\pm 6\%$, the error of the contour scaling is controlled in the range of $[-8\%, +10\%]$, the overall relative error is within $\pm 10\%$, the overall relative error control is better.

3) User authentication test: In this part, we first need to estimate the number of points required to establish the stable points cloud distribution of the lip movement model. Secondly, the model number K of Gaussian mixture model is estimated by the model fitting. Since these two parameters are dependent on each other, we need to estimate the two parameters at the same time. Here, we set the model number K as 1, 2 and 3 respectively. Each user has 100, 200, 300, 400 and 500 points, then we estimate the points cloud distribution based on them. The variance of the parameters $\vec{\mu}, \vec{\Sigma}$ of the Gaussian mixture model is calculated, in order to measure the relationship between the number of the points and the stability of the mixed model parameters.

(1) The lip movement model and parameters estimation. In this part, we collect data from 20 people. Each of them include 50 groups. That is total of 1000 groups data as our experiment data. The experimental results are shown in Table I.

In Table I, $D_1(\vec{\mu})$ and $D_1(\vec{\Sigma})$ are the variance of the parameters when the current points cloud is fitted by the $i\text{-}th$ Gaussian model. Correspondingly, the smaller value of the points cloud, the more stable point cloud distribution. NC denotes that the iteration does not converge. Taking into account the data collection speed and data processing efficiency, we take 300 points as a stable points cloud distribution parameters. Considering the user authentication time constraints, we do not use the model number 3. Comparing model number 1 and 2, we found that the model stability and subsequent authentication accuracy of 2 parameters is higher than that of 1, so we selected 2 as the number of models in the Gaussian mixture model.

(2) Training sample quantity estimation. The relative change rate k of the interval length of the user training data is used as our evaluation parameter. The initial model parameter interval can be formed when $m = 2$. Meanwhile, we set the relative change rate $k = 1$. When $m \geq 3, k = \delta L / L$, where δL is the change value of new training samples to the interval length, L is the length of the current interval.

In this section, we need to identify the training samples size required to determine the stability range. We selected the data from 20 individuals. Each individual has 11 groups, so a total of 220 groups of data were used to carry out the experiment. The relationship between the rate of the relative interval length change and the number of training samples is shown in Figure 7.

When the training samples size changes from 2 to 8, the change rate of the interval length decreases fast. When the number of samples is more than 8, the change rate of the relative range length is stable. Taking into account the stability of the parameters and the easy use of the system when the user is trained, the proposed authentication system takes 8 users data as the number of training samples.

TABLE I
GAUSSIAN MIXED MODEL NUMBER AND THE NUMBER OF POINTS CLOUD ESTIMATION.

Parameter statistics	The number of points cloud				
	100	200	300	400	500
$D_1(\vec{\mu})$	17.5601	10.4064	3.6884	3.0872	2.8931
$D_1(\vec{\Sigma})$	34.8963	25.2083	9.4749	8.1282	7.5214
$D_2(\vec{\mu})$	14.3661	4.8373	1.6116	1.4872	1.1824
$D_2(\vec{\Sigma})$	26.4374	18.3447	5.8235	4.9531	4.3145
$D_3(\vec{\mu})$	NC	13.2349	7.3605	4.1413	1.3270
$D_3(\vec{\Sigma})$	NC	31.0399	23.8834	14.4183	4.3059

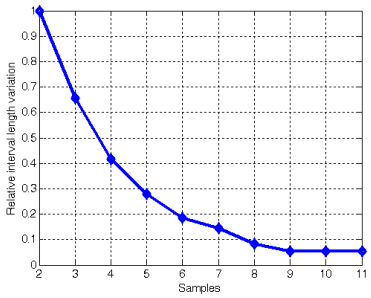

Fig. 7. Sample number and interval length change rate

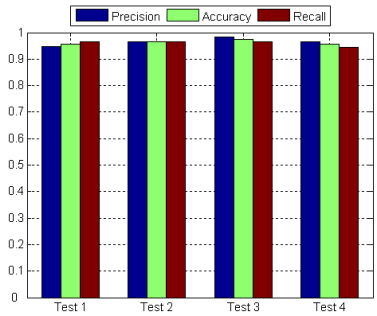

Fig. 8. Authentication accuracy and content dependency

(3) Authentication accuracy test. We used the Precision, Accuracy and Recall to assess the results of the authentication. Precision $P = TP/(TP + FP)$, Accuracy $A = (TP + TN)/(TP+FP+TN+FN)$, Recall $R = TP/(TP+FN)$.

In this part, we first design experiments to verify that the proposed user authentication method is independent of the user's speech content. We designed 4 groups of experiments, 1, for each participant, SALM uses 8 groups having the same text to train the samples, and the authentication is complete through the same speech content; 2, for each participant, SALM uses 8 groups having the same content to train the samples, and the authentication is complete through the different speech content; 3, for each participant, SALM uses 8 groups of data having different content to train the samples, and the authentication is complete through the speech content within the 8 groups of data; 4, for each participant, SALM uses 8 groups of data having different content to train the samples, and the authentication is complete through the speech content from the 8 groups of data;

For the above four experiments, we tested the samples from 20 participants and took 8 groups for training. During the authentication phase of each experiment, we took 57 positive samples, with 3 samples from the remaining 19 individuals, and a total of 57 negative samples. The experiment results are shown in the Figure 8:

From Figure 8, we can see that the overall precision of the system is always maintained at higher than 94.7%, and the accuracy is always maintained at higher than 95.6% up to 97.3%. The user authentication results show that the user training content and the speech content is irrelevant in our proposed system. For all cases, the authentication result has high accuracy, precision and recall.

In order to further evaluate the overall performance of the user authentication system, and test the accuracy and robustness of the system, we designed 3 groups of tests from the following three aspects: Self certification, it uses part of their own data for training, and the data of others for certification; Exclusive certification, it uses their own data for training, and the data of others to authenticate; The actual certification, the positive and negative samples are mixed with a certain proportion, to carry out certification. Here, we use the full data from 20 participants. 8 groups of data for each person are used to train their respective models, and then each person's other 100 groups data are used to complete self-certification with their own training model; Then the data from the other 19 individuals with 20 groups of data are used for the exclusive certification; Finally 100 groups of data from their own, and data from other 19 participants with each of user having 10 groups of data, that is, the proportion of mixed positive and negative samples is about 1:2 for the actual certification. The results are shown in Figure 9. Here, SC, EC and AC denotes Self certification, Exclusive certification and Actual certification respectively.

For self-certification and exclusive certification, because of the particularity of the authentication samples: either all positive or all negative samples, we mainly test the accuracy of the system. It is the judgment ability of the authentication system, that is, it is able to determine the positive samples to be positive, and negative to be negative. We can see from Figure 9, the accuracy for two cases is above 96%, so our system has a high level of discrimination among samples.

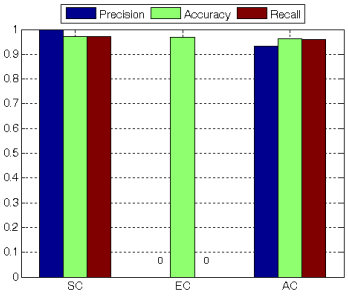

Fig. 9. Authentication results for the three cases

For actual certification, our samples with the proportion of positive and negative about 1:2 are certified. Since there are two kinds of samples, we need to make a comprehensive survey of the certification from precision, accuracy and recall. From Figure 9 we can clearly see that the actual authentication precision is 93.2%, which reflects the high proportion of the true positive samples in our certification results. It means that the system has high positive sample reliability. The accuracy is 96.2%, which shows that the system has a better ability to determine the samples in the actual certification with the positive and negative mixed samples. The Recall is 96%, which reflects the proportion of true positive in positive samples is very high. It means it has a high ability to check the positive sample in the whole sample.

VI. Conclusion

To protect user privacy on the smartphones, a user authentication system is proposed based on the physiological characteristics of lip contour movement. This system relies only on the characteristics of peoples' lip movement during speech. It recognizes the user's physiological feature, which is hard to copy and make a counterfeit. In our approach, we design a lip contour extraction method based on the improved ASM method. After obtaining the precise lip contour extraction, the lip movement feature matrix can be built. The dimensionality of the matrix is reduced to obtain the lip feature space and its projection, and the dynamic lip contour movement model is built upon the distribution of the points cloud. A Gaussian mixture model is used to fit the lip movement model, and then the stable parameters are estimated so that a user is authenticated. Finally, the real-world testing experiments with the real hardware results show high accuracy of the proposed system.

References

[1] J. Kasprzak, B. Leczynska, Cheiloscopy: Human Identification on the Basis of LipPrints (in Polish). CLK KGP Press, Warsaw, (2001) pp. 838-843.

[2] Y. Tsuchihashi, Studies on personal identification by means of lip prints. Forensic Science (1974) 3(3), pp. 233-48.

[3] E. Gomez, C. M. Travieso, J. C. Briceno, et al. Biometric identification system by lip shape. In Proceedings of the IEEE International Carnahan Conference on Security Technology, (2002) pp. 39-42.

[4] E. d. Petajan, B. J. Bischoff, D. A. Bodoff, et al. Proved Automatic Lipread System To Enhance Speech Recognition. Bell Labs Tech. Report TM11251-871012-11, (1987).

[5] Faraj. Maycel-Isaac , Josef Bigun, Audio-visual person authentication using lip-motionfrom orientation maps. Pattern Recognit. Lett (2007) em, pp. 1368-1382.

[6] X. Wang, T. Han, X and S. Yan, An hog-lbp human detectorwith partial occlusion handling. In proceedings of the International Conference on Computer Vision(ICCV), (2009) pp. 32-39.

[7] S. Milborrow, F. Nicolls, Locating Facial Features with an Extended Active Shape Model. In proceedings of the European Conference on Computer Vision(ECCV). Springer-Verlag, (2008) pp. 504-513.

[8] W. Xi, C. Qian, J. Han, K. Zhao, S. Zhong, X. Li, and J. Zhao, Instant and Robust Authentication and Key Agreement among Mobile Devices, in Proceedings of ACM Conference on Computer and Communications Security (CCS), 2016.

[9] Z. Jiang, J. Han, C. Qian, W. Xi, K. Zhao, S. Tang, J. Zhao, and P. Yang, VADS: Visual Attention Detection with a Smartphone, in Proceedings of IEEE Conference on Computer Communications (INFOCOM), 2016.

[10] P. Dollar, P. Welinder, P. Perona, Cascaded pose regression. IEEE (2010), 238(6), pp. 1078-1085.

[11] J. H. Friedman, Greedy function approximation: A gradientboosting machine. The Annals of Statistics. (2001), 29(5), pp. 1189-1232.

[12] M. Ozuysal, M. Calonder, V. Lepetit, et al. Fast keypoint recognition using random ferns. IEEE Trans. Pattern Anal. Mach. Intell. (2010), 32(3), pp. 448-461.

[13] T. F. Cootes, C. J. Taylor, Active shape models. In Proceedings of the British Machine Vision Conference(BMVC), (1995) pp. 266-275.

[14] P. Sauer, T. Cootes, C, Taylor, et al. Accurate Regression Procedures for Active Appearance Models. In Proceedings of the British Machine Vision Conference (BMVC), (2011) pp. 681-685.

[15] C. Qiu, M. Mutka, Cooperation among Smartphones to Improve Indoor Position Information, in Proceedings of IEEE Symposium on a World of Wireless Mobile, and Multimedia Networks (WOWMOM), 2015.

[16] C. Qiu, M. Mutka, iFrame:Dynamic Indoor Map Construction through Automatic Mobile Sensing, in Proceedings of IEEE Conference on Pervasive Computing and Communications (PerCom), 2016.

[17] A. B. Hassanat, Visual Passwords Using Automatic Lip Reading. Int. J. Sci. Baisc. Appl. Res. (2014), 13(1), pp. 218-231.

[18] P. L. Silsbee and A. C. Bovik, Computer Lipreading for Improved Accuracy in Automatic Speech Recognition, in IEEE Transaction on Speech and Audio Processing, 1996.

[19] W. H. Sumby, and I. Pollack, "Visual contribution to speech intelligibility in noise," J. Acoust. Soc. Am. 1954

[20] OpenCV Library [OL] http://www.opencv.org

[21] G. W. Stewart, J. G. Sun, Matrix Perturbation Theory. Applied Mathematical Sciences. (1990) pp. 165-217

[22] K. Seshadri, M. Savvides, Robust modified Active Shape Model for automatic facial landmark annotation of frontal faces. In Proceedings of the IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS), (2009) pp. 1-8.

[23] D. Lee, H. Park, C. D. Yoo, Face alignment using cascade Gaussian process regression trees. In Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR), (2015) PP. 4204-4212

[24] I. Matthews, S. Baker, Active Appearance Models Revisited. Int. J. Comput. Vis. (2004), 60(2), pp. 135-164.

[25] Z. Kalal, K. Mikolajczyk, J. Matas, Tracking-Learning-Detection. IEEE Transactions on Pattern Analysis & Machine Intelligence. (2011), 34(7), pp. 1409-22.

[26] G. R. Bradski, A. Kaehler, Learning OpenCV: Computer Vision with the OpenCV Library, O'REILLY: Sebastopol, CA, USA, (2015).

[27] R. Lienhart, J. Maydt, An extended set of Haar-like features for rapid object detection. In Proceedings of the IEEE International Conference on Image Processing(ICIP), (2002) pp. 900-903.

[28] Hang. Li, Statistical Learning Method; TsingHua University Press: Beijing, China, (2012) pp. 155-170.

[29] D. Cristinacce, T. F. Cootes, Feature Detection and Tracking with Constrained Local Models. In Proceedings of the British Machine Vision Conference(BMVC), (2006) pp. 929-938.

[30] Helen dataset [OL] http://www.ifp.illinois.edu/ vuongle2/helen/

[31] P. Viola, M. J. Jones, Robust Real-Time Face Detection. International Journal of Computer Vision. (2004), 57(2), pp. 137-154.

[32] N. Eveno, A. Caplier, P. Y. Coulon, Accurate and quasi-automatic lip tracking. IEEE Trans. Circuits Syst. video Technol. (2004), 14(5), 706-715.