Multimodal Person Authentication System Using Features of Utterance

Qian Shi*, Takeshi Nishino and Yoshinobu Kajikawa†
E-mail: *k045127@kansai-u.ac.jp, †kaji@kansai-u.ac.jp
*†Faculty of Engineering Sceince, Kansai University, 3–3–35, Yamate-cho, Suita-si, Osaka 564–8680, Japan

***Abstract*—In this paper, we propose a biometrics authentication method using multimodal features in utterance. The multimodal features in utterance consists of lip shape (physical trait), lip motion pattern and voice pattern(behavioral trait). Therefore, the proposed method can be constructed with only a camera extracting lip area and voice without special equipment like other personal authentication methods. Moreover, the utterance phrase itself has a role of a key function by setting up an utterance phrase arbitrarily, and then the robustness of the authentication increases according to the phrase recognition which can reject an imposter with the feature similar to a registrant. In the proposed method, lip shape and voice features are extracted as edge or texture in the lip image and pitch or spectrum envelope in the voice signal. Experimental results demonstrate that the proposed method can improve the authentication accuracy compared with other methods based on the single modal.**

I. Introduction

Recently, the increase in the amount of information leakage and the number of incidents of theft by electronic trespassing have led to the need for enhanced security and improved convenience [1]–[3]. With this background, biometrics authentication technologies that use a physiological and/or behavioral trait have been studied by many researchers. A physiological trait is related to innate characteristics of humans, such as the fingerprints, iris, palm form, ear form, and face. On the other hand, a behavioral trait is related to human activities, such as voiceprint, signature, and keystroke. These traits have both advantages and disadvantages.

In this paper, we focus on a multimodal biometrics authentication method. The multimodal features consists of lip shape (a physiological trait) and variations of lip shape and voice during an utterance (a behavioral characteristic). Moreover, the proposed method can be realized with only a camera and microphone to extract the lip area and voice without the special equipment used in other personal authentication methods.

We have already proposed a biometrics authentication method using images [4]. In this method, the lip shape and its changes were used as features for person authentications. This method can achieve high authentication accuracy with an authentication rate of about 99.5 %. However, this method does not have robustness to variations of image acquisition environments.

Therefore, we have also proposed an authentication method using features related to both the image and voice [5]. In this method, the authentication accuracy is maintained even if the image acquisition environment varies because both image and voice features are used for the authentication, that is, the weights of the image and voice classifiers are adjusted according to the image acquisition environment.

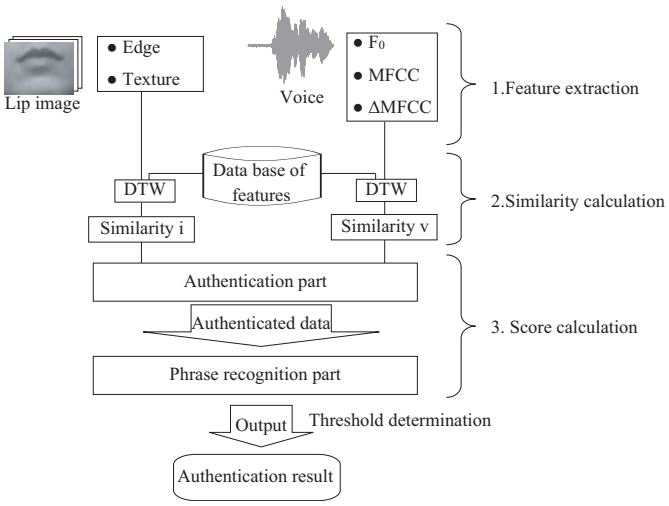


Fig. 1. Overview of the proposed authentication system.

Another multimodal biometrics authentication method using both the image and voice was proposed in [3]. In this literature, both voice and dental geometric information associated with the vowel /i/ was used as features for authentication, and the authentication accuracy was very high with an authentication rate of 98.36 %. However, the utterance phrase was limited to the vowel /i/.

On the other hand, the proposed authentication method can register an arbitrary utterance phrase and consequently provide a key function to the registered phrase. In this paper, we examine the effectiveness of the proposed method through some experimental results in comparison with other methods and clarify some problems. Moreover, we propose a novel classifier arrangement to solve the problems.

II. Proposed Authentication Method

Figure 1 shows an overview of the proposed authentication method. As shown in Fig. 1, the proposed authentication method consists of voice- and image-processing parts. Moreover, each processing part consists of three steps: the extraction of features, the computation of similarity, and the computation of decision scores from classifiers. Furthermore, the classifiers are independently constructed for person and

phrase authentications, that is, their outputs correspond to the decision scores for the person and phrase authentications.

The decision score used for the person authentication is an index for distinguishing whether a speaker is a registrant in the system, and the decision score for phrase authentication is an index for distinguishing whether an utterance phrase is a registered one. Moreover, the decision scores obtained from the image- and voice-processing parts are integrated according to the data acquisition environment.

The phrase authentication is carried out only using data accepted by the person authentication process. Hence, the phrase authentication decides whether the correct (registered) phrase was uttered regardless of the registrant, that is, if a registrant utters a different (unregistered) phrase, the proposed method rejects the request. This function is effective for safeguarding a system against imposters with similar features in both image and voice signals.

As stated above, the proposed authentication method is robust against variations in the data acquisition environments and imposters. We consider that the proposed authentication method will be suitable for the entrance of private houses and small companies.

III. Features of Image

A. Lip detection

In the proposed authentication method, lip detection follows face detection, pupil detection, the normalization of a face image, and lip region detection. The lip detection part consists of the following steps.

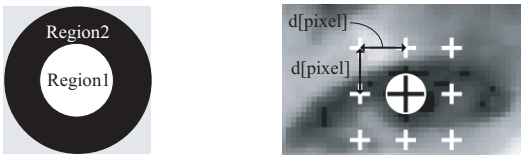
(1) The face region is detected from each frame of a moving image using the face detector proposed by Viola and Jones [6].

(2) The eye region is detected from the upper half of the face region using the method in step 1.

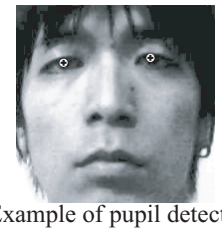
(3) The pupil positions are determined in the eye region as follows. First, the lengths and circumference ratios of all arc shapes in the eye region are obtained by the Hough transform, and some likely positions are selected from all candidate pupil positions on the basis of a threshold length to circumference ratio of the arcs. Next, the separability of the selected positions is calculated using the circular separability filter [7] shown in Fig. 2(a). Finally, the pupil positions are determined from the ratio between the average luminance values of a candidate pupil position selected on the basis of the separability and the neighbor pixels d pixels away from the candidate pupil position, as shown in Fig. 2(b). The ratio is calculated as

$$\beta = \frac{L_p}{L_n}, \tag{1}$$

where L_p is the luminance value of the candidate pupil position and L_n is the average luminance value of the neighbor pixels. Moreover, d is also the diameter of the circle obtained using the circular separability filter. Figure 2(c) shows an example of pupil detection.

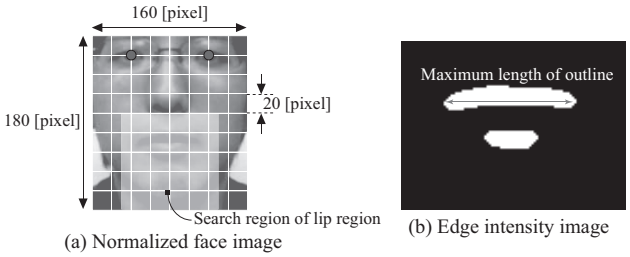


(a) Circular separability filter (b) Comparison of luminance values

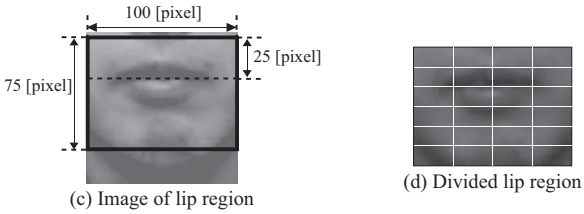


(c) Example of pupil detection

Fig. 2. Principle of pupil detection.



(a) Normalized face image (b) Edge intensity image



(c) Image of lip region (d) Divided lip region

Fig. 3. Principle of lip detection.

(4) The rotation correction and normalization of the face image are executed on the basis of the pupil positions. In the normalization of the face image, the length of one side of the lattice is set to one quarter of the distance between the pupils. Figure 3(a) shows a normalized face image.

(5) The search region of the lip region shown in Fig. 3(a) is determined on the basis of the horizontal edge obtained using the Gabor filter. It is assumed that the maximum length of the outline is along the center line between the lips, as shown in Figs. 3(b) and (c). Finally, the lip region is divided into cells. Preliminary experimental results demonstrate that twenty-four cells (Fig. 3(d)) yield the highest authentication rate. Hence, the lip region is divided into twenty-four (24) cells in this paper.

B. Edge-based feature

The edge-based feature is obtained using the following procedure. First, the horizontal and vertical direction edge intensities, $G_x(x, y)$ and $G_y(x, y)$, are calculated using the Sobel filter, and the edge intensity $G(x, y)$ and edge direction

$\theta(x, y)$ of each pixel are calculated as follows (Fig. 4(a)):

$$G(x,y) = \sqrt{G_x(x,y)^2 + G_y(x,y)^2}. \quad (2)$$

$$\theta(x,y) = \arctan\left(\frac{G_x(x,y)}{G_y(x,y)}\right). \quad (3)$$

Next, the edge direction $\theta(x, y)$ is quantized into K steps (e.g, bin_k indicates the kth step) and the edge intensity of direction k is determined as follows (Fig. 4(b)):

$$\psi_k(x,y) = \begin{cases} G(x,y), & \text{if} \quad \theta(x,y) \in bin_k \\ 0, & \text{otherwise} \end{cases}. \quad (4)$$

The edge direction histogram of the ith cell and fth frame is calculated as

$$E^f(i) = [E_1^f(i), E_2^f(i), \cdots, E_k^f(i), \cdots, E_K^f(i)], \quad (5)$$

where $E_k^f(i)$ denotes the cumulative edge intensity of direction k and is given as

$$E_k^f(i) = \sum_{(x,y)\in i} \psi_k(x,y). \quad (6)$$

Finally, the edge-based feature vector is given by $E^f(i)$ over all frames and cells.

$$E = [E(1), E(2), \cdots, E(f), \cdots, E(F)], \quad (7)$$

where F is the frame number, and the number of quantizations K is set to 8 in this paper [8].

C. Texture-based feature

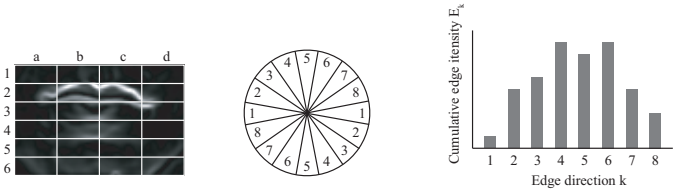
The local binary pattern (LBP) [9] is used as the texture-based feature, which expresses the texture pattern of each cell. The LBP can express $2^8 = 256$ texture patterns using binary images (3×3) and is obtained by the following procedure. First, eight pixels around the center pixel in a 3×3 window are binarized according to whether or not the surrounding pixels have a larger value than the center pixel, as shown in Figs. 5(a) and (b). Next, the product of the binary image shown in Fig. 5(b) and the weight coefficient shown in Fig. 5(c) for every pixel yields the result shown in Fig. 5(d). Finally, the value of the LBP is given by the sum of the products. In the proposed method, the value of the LBP is limited to eight groups including the rotations of $\pm\pi/2$ and π[rad], as shown in Fig. 6. The LBP histogram of the fth frame is constructed by counting the number of LBPs belonging to each group as

$$L^f(i) = [L_1^f(i), L_2^f(i), \cdots, L_g^f(i), \cdots, L_G^f(i)], \quad (8)$$

where $L_g^f(i)$ is the number of LBPs belonging to group g, as shown in Fig. 6, and is defined as

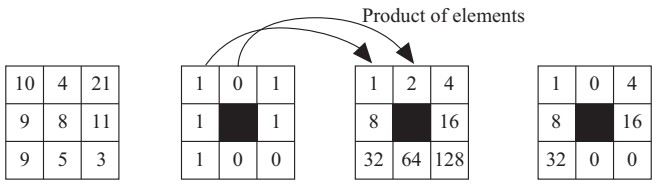
$$L_g^f(i) = \sum_{(x,y)\in i} \zeta_g(x,y), \quad (9)$$

where ζ_g is the binary image of group g. Finally, the texture-based feature vector is given by $L^f(i)$ over all frames and cells.



(a) Edge intensities (b) Quantization of edge direction (c) Edge direction histogram

Fig. 4. Edge feature.



(a) Original image (b) Binary image (c) Weight coefficient (d) Output

Fig. 5. Principle of local binary pattern (LBP).

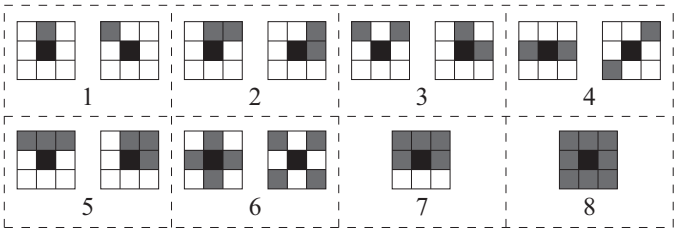


Fig. 6. LBP classification.

IV. Features of Voice

In the section, we describe the features of the voice. The proposed method uses three features of the voice: fundamental frequency (F_0), 12 dimensions of the MFCC (Mel–Frequency Cepstrum Coefficient), and 12 dimensions of the ΔMFCC.

The parameters of the MFCC are standardized by the ETSI (European Telecommunications Standards Institute) [10]. In this paper, the parameters of the MFCC standardized by the ETSI are used as one of the voice features. The MFCC can specify vowels accurately, but cannot specify consonants accurately. Therefore, in the proposed method, ΔMFCC, which is the time variation of the MFCC, is used as one of the voice features to handle the phrase authentication appropriately.

V. Similarity and Score

A. Calculation of similarities for each feature

As the features of the image and voice are extracted from every frame, the features are defined as time series vectors. The vector length of each feature depends on the utterance time of the speaker, that is, the vector length varies with the individual and from trial to trial. The similarity between two vectors of each feature is accordingly calculated by DTW (Dynamic Time Warping) [11]. DTW is a technique developed in the field of speech recognition and can obtain the distance between data with different vector lengths using dynamic programming

with comparatively low computational complexity. Moreover, the similarities between image features are calculated for each cell of the divided lip region.

B. Integration of final decision scores in image and voice authentication parts

In the proposed method, two features for the image and three features for the voice are extracted. Because all features are independent, the similarities are calculated for each feature. The authentication classifiers are generated by AdaBoost [11], and a label (+1 or -1) is assigned to a correct or incorrect pair. In the proposed method, weak classifiers are generated for each feature, and the similarity for each feature is input into the corresponding weak classifier. Hence, more useful features more strongly affect the authentication performance.

A strong classifier is generated for each registrant; therefore, a classifier specific to each registrant can be generated. The numbers of weak classifiers for the image- and voice- processing parts are defined as m_i and m_v, respectively. In other words, the total number of weak classifiers for image and voice processing parts are $(m_i \times 2)$ and $(m_v \times 3)$, respectively. Moreover, m_i and m_v for the person and phrase authentications are denoted by m_{i1} and m_{v1}, and m_{i2} and m_{v2}, respectively. Strong classifiers are independently generated for image- and voice-processing parts. In the image-processing part, strong classifiers are generated for each cell, because the lip region is divided into cells. Therefore, the final decision score for the image-processing part is obtained by averaging all decision scores of the strong classifiers for each cell.

Furthermore, the final decision scores for the image and voice parts are integrated using the formula

$$S_m = \alpha S_i + (1 - \alpha) S_v, \tag{10}$$

where S_i and S_v are the final decision scores for the image and voice parts, respectively, and α is the weight used to determine the relative importance of the scores. If α is close to 1, then the decision score of the image part is much more highly valued and vice versa. The values of S_m and α for the person and phrase authentications are denoted as S_{m1} and α_1, and S_{m2} and α_2, respectively. In this paper, α_1 and α_2 are set to 0.5.

VI. Experimental Results

In this section, we discuss the effectiveness of the proposed method through the results of an authentication experiment.

A. Experimental conditions

Table I shows the experimental conditions. Other conditions and information on the experimental environment are as follows:

- All subjects are in their twenties
- The distance between the subject and the camera is 40 cm
- Subjects close their mouths when not making an utterance
- The duration of filming is three months
- Voice activity is manually detected

TABLE I
EXPERIMENTAL CONDITIONS.

Authentication phrase		"ohayougozaimasu"
Subjects	Registrants	5 persons
	Non-registrants	6 persons
	Total subjects	11 persons
Movie	Frame rate	30 fps
	Image size	640 × 480 pixels
Voice	Sampling frequency	8 kHz
	Points of FFT	256
	Frame length	30 ms
	Frame interval	10 ms
Equipment	Web camera	UCAM-DLC300T

TABLE II
SETUP OF EACH PARAMETER.

Parameter	Authentication	Phrase
Number of cells	4	36
m_i	3	14
m_v	13	1

The authentication accuracy is evaluated in terms of the false acceptance rate (FAR) and false rejection rate (FRR). FAR is the probability that another person is falsely recognized as a registrant and FRR is the probability that a registrant is falsely rejected. The equal error rate (EER), the value for which FAR and FRR are equal (trade-off point), is utilized to determine the threshold in the authentication. In this paper, we denote FAR and FRR at the EER as $\mathrm{FAR_E}$ and $\mathrm{FRR_E}$, respectively. Table II shows the parameter values in each part of the authentication system. These parameters were empirically determined through some preliminary experiments. Table III gives details of the training and evaluation data sets.

B. Results of experimental in person authentication

Table IV shows the person authentication accuracies in the cases where only the image part, only the voice part, and both parts are applied. It can be seen from Table IV that the authentication error of the multimodel-processing part is lowest. Therefore, the multimodal-processing is valid for the personal authentication system.

On the other hand, for the registrants A and B, the authentication error of the voice-processing part is 13.9 % and 22.6 % , which are much higher than those of other registrants. One of the reasons is considered that the registered sample of the registrants A and B had different intonation, and thus the distance between the registered and evaluated samples was large.

C. Effect of intonation

In this section, we examine the effect of the intonation, which is one of the reasons that yields high authentication error in the voice processing part. From the reliability value of voice processing part in Fig 7, we can found that the relability value of MFCC is highest. Thus, we examine the effect of the intonation through the MFCC values in this section. Moreover,

TABLE III
DATA SETS USED IN EXPERIMENTS.

		True	False	Total
Data sets A @	Training sample	45 per person	500 per person	545 per person
	Evaluation sample	50 per person	500 per person	550 per person
Data sets B @	Training sample	45 per person	100 per person	145 per person
	Experimental sample	50 per person	100 per person	150 per person

TABLE IV
PERSON AUTHENTICATION ACCURACIES.

Registrant	Image	Voice	Multimodal
A	2.1	13.9	1.4
B	3.7	26.6	0.0
C	7.2	6.7	1.1
D	5.4	6.8	1.6
E	2.6	1.2	0.0
Average	4.2	11.0	0.8

we examine the similarities obtained from DTW, which is the input of weak classifiers.

Figures 8–12 show the changes in the similarities of each registrant. In these figures, the combination numbers 1 to 45 indicate the cases of Identity-Identity and the combination number 46 to 445 indicate the cases of Identity-Others, respectively. From Fig. 12, we can find clear difference between both cases in the registrant E. Hence, the authentication error of the registrant E is low as shown in Talbe IV because it is easy to determine an appropriate threshold. On the other hand, we can see from Figs. 8–10 that the similarities for the registrants A and B vary largely. Hence, it is difficult to determine an appropriate threshold for these registrants because the differences are very small for both cases. As a result, the authentication errors of the registrants A and B are high as shown in Table IV. From the foregoing examinations, it is considered that the intonation of voice caused the degradation of authentication accuracy. However, we see a possibility that a particular dimension of MFCC has the adverse impact to the authentication accuracy.

D. An improvement of authentication accuracy

In this section, we try to improve the authentication accuracy due to the intonation of voice. Figure 13 shows the overview of an improved system where the similarities between samples for each element are calculated and conveyed to the corresponding classifiers. It is considered that the authentication accuracy can be improved by reducing some particular elements which has adverse impact to the authentication accuracy.

Table V shows the experimental results of the improved system on the authentication. From Table V, we can see the authentication errors of every registrant are 0%. It is therefore believed that the proposed system can improve the authentication accuracy because it is able to realize a more detailed identification for each registrant by using each element of MFCC as a feature.

E. Results of experimental in phrase authentication

Table VI shows the phrase authentication accuracies in the cases where only the image part, only the voice part, and both parts are applied, respectively. It can be seen from Table VI that the authentication error of the multimodel-processing part is lowest. Therefore, the multimodal-processing is valid for the personal authentication system.

On the other hand, the authentication error of the image-processing part is higher than the voice-processing part for every registrants. One of the reasons is considered that the quaility of image samples is poor.

Figure 14 shows an example of the failure in the lip detection which we used for the phrase authentication. It can be seen from Fig. 14 that the upper part of the lip is not included in the image. It is consequently considered that the inadequate lip images caused the degradation of authentication accuracy, because the block matching can not be properly executed by using them. From the foregoing examinations, it is therefore believed that the authentication accuracy of the proposed system can be improved by improving the detection accuracy of the lip region.

VII. CONCLUSION

In this paper, we examined the effectiveness of the proposed multimodal authentication method and clarified the problems. Moreover, we propose an improvement method. In the experimental results of person authentication, the multimodal authentication error was 0.0 %; therefore, the proposed method can realize high authentication accuracy. In the experimental results of the phrase authentication, we found that the inadequate lip images caused the degradation of authentication accuracy.

However, the number of samples used in the experiments was small. Hence, we will increase the number of samples to demonstrate the effectiveness of the proposed authentication method in the future.

TABLE V
PERSON AUTHENTICATION ACCURACIES.(IMPROVED)

Registrant	Before			After		
	FAR_E	FRR_E	HTER	FAR_E	FRR_E	HTER
A	13.8%	14.0 %	13.9 %	0.0 %	0.0 %	0.0 %
B	27.1%	26.0 %	26.6 %	0.0 %	0.0 %	0.0 %
C	5.3 %	8.0 %	6.7 %	0.0 %	0.0 %	0.0 %
D	5.6 %	8.0 %	6.8 %	0.0 %	0.0 %	0.0 %
E	0.4 %	1.2 %	1.2 %	0.0 %	0.0 %	0.0 %
Average	4.2 %	11.6 %	11.0 %	0.0 %	0.0 %	0.0 %

TABLE VI
PHRASE AUTHENTICATION ACCURACIES.

Registrant	Image	Voice	Multimodal
A	14.5	10.0	7.0
B	10.0	8.0	3.0
C	10.0	0.0	0.0
D	5.0	0.0	0.0
E	10.0	3.2	2.0
Average	10.0	3.2	2.0

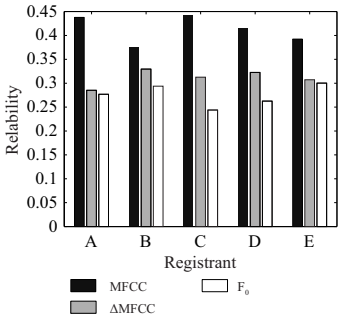


Fig. 7. Relability of voice features (Authentication)

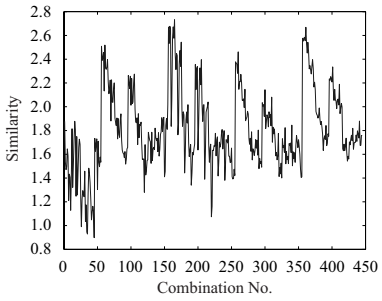


Fig. 8. Variations of similarities(Registrant A)

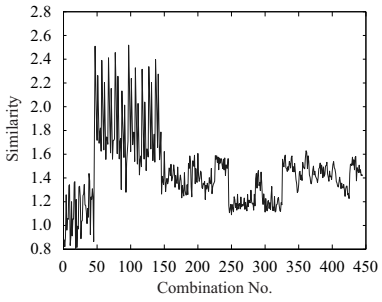


Fig. 9. Variations of similarities(Registrant B)

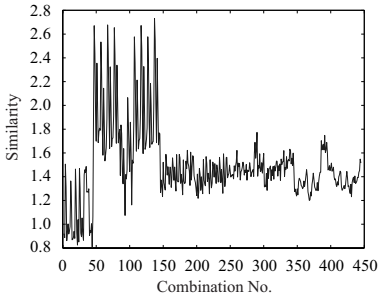


Fig. 10. Variations of similarities(Registrant C)

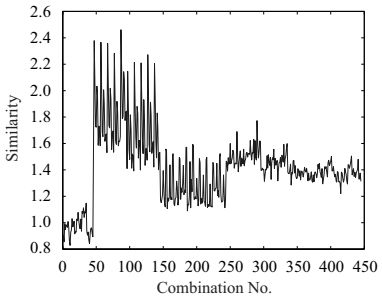


Fig. 11. Variations of similarities(Registrant D)

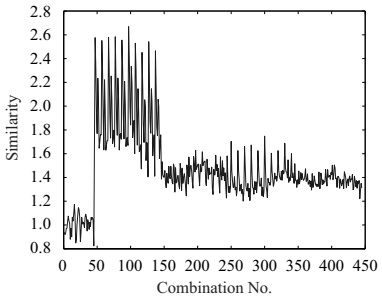


Fig. 12. Variations of similarities(Registrant E)

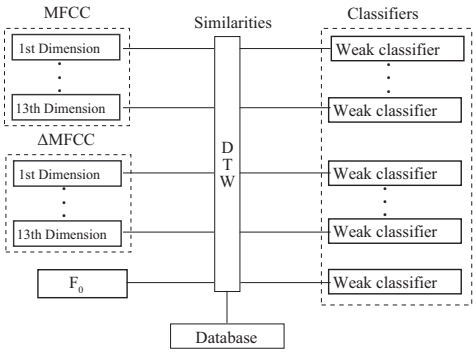


Fig. 13. Overview of an improved system for the authentication accuracy

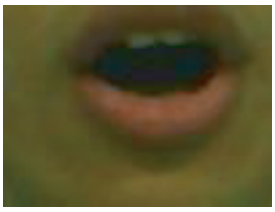


Fig. 14. An example of the failure in lip detection

REFERENCES

[1] M. Faundez–Zanuy, "Biometric security technology," IEEE A&E Syst. Mag, vol. 21, no. 6, pp. 15–26, Jun. 2006.

[2] A. K. Jain, J. Feng, K. Nandakumar, "Fingerprint matching," IEEE Computer, vol. 43, no. 2, pp. 36–44, Feb. 2010.

[3] D. J. Kim, K. W. Chung, K. S. Hong, "Person authentication using face, teeth and voice modalities for mobile device security," IEEE Transactions on Consumer Electronics, vol. 56, no. 4, pp. 2678–2685, Nov. 2010.

[4] A. Sayo, Y. Kajikawa, M. Muneyasu, "Biometrics authentication method using lip motion in utterance," Proc. of 8th International Conference on Information, Communications and Signal Processing, Singapore, Dec. 2011.

[5] T. Nishino, Y. Kajikawa, M. Muneyasu, "Multimodal Person Authentication System Using Features of Utterance," 2012 IEEE International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS 2012), pp. 43-47, Sep. 2012.

[6] P. Viola, M. Jones, "Robust real-time face detection," International Journal of Computer Vision, vol. 57, pp. 137–154, May 2004.

[7] K. Fukui, O. Yamaguchi, "Facial feature point extraction method based on combination of shape extraction and pattern matching," Systems and Computers in Japan, vol. 29, no. 6, pp. 2170–2177, 1998.

[8] K. Levi, Y. Weiss, "Learning object detection from a small number of examples: the importance of good features," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. II53–II60, Jun. 2004.

[9] S. Zhiping, Y. Fei, "Symmetrical invariant LBP texture descriptor and application for image retrieval," Proc. of the IEEE International Congress on Image and Signal Processing, vol. 2, pp. 825–829, May 2008.

[10] "Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms," ETSI ES 202 050, vol. 1.1.5, Jun. 2007.

[11] D. J. Berndt, J. Clifford, "Finding patterns in time series: A dynamic programming approach," Advances in Knowledge Discovery and Data Mining, pp. 229–248, AAAI, 1996.

[12] R. E. Schapire, Y. Singer, "Improved boosting algorithms using confidence-rated predictions," Machine Learning, vol. 37, no. 3, pp. 297–336, 1999.