An Efficient Approach to Multimodal Person Identity Verification by Fusing Face and Voice Information

Hsien-Ting Cheng [1,2], Yi-Hsiang Chao [1,3], Shih-Liang Yeh [1], Chu-Song Chen [1], Hsin-Min Wang [1], and Yi-Ping Hung [1,2]

[1] Institute of Information Science, Academia Sinica, Taipei, Taiwan.
[2] Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan.
[3] Department of Computer and Information Science, National Chiao Tung University, Hsinchu, Taiwan.

Abstract

This paper presents an effective method to combine speech recognition, speaker verification and face verification for biometric authentication. Our method provides a light-weight enrollment process and an easy-to-use verification interface. A multi-face/single-sentence strategy is used to combine voice and face verification modules, and support vector machine is employed for information fusion. Experimental results show that our method can achieve high verification accuracies.

1 Introduction

Fusing multimodal information for biometric-based authentication has been shown an effective way to boost the verification performance of using only a single modality [3][2][9][1]. In this paper, we present a framework of combining face and voice data for person identity verification. In our work, state-of-the-art approaches for face authentication and speaker verification are employed to build individual modules. Then, the face and voice modules are combined to achieve a highly effective person identity verifier.

Our method employs the technique of keyword spotting, so that the identity-claim process becomes easy. Significant characteristics of our individual modules include: (1) We integrate both speech recognition and speaker verification techniques to prevent impostors who try to take the prerecorded speech of a valid speaker from entering in our system. (2) We develop an incremental learning method for kernel Fisher's discriminant, so that the insertion/deletion of a new/old training image becomes more efficient. (3) A light-weight enrollment process is proposed, where the users have only to say very limited words for on-line training data collection.

In the fusion module, since we can grab multiple images during the period of user's saying a sentence, we adopt the strategy of multi-face/single-sentence integration for identity verification. Hence, the face verification module will be used multiple times for an image sequence in a verification task. To generate a single score of the face module for further fusion, we have investigated several strategies and made a useful suggestion. Finally, we use

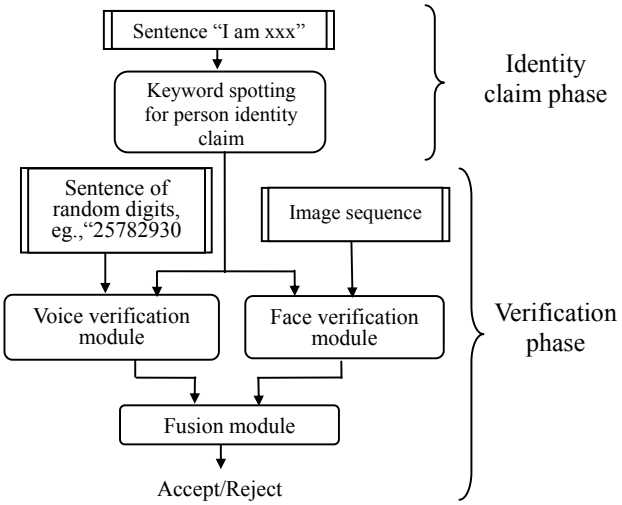

Figure 1. Building blocks of the person identity verification system.

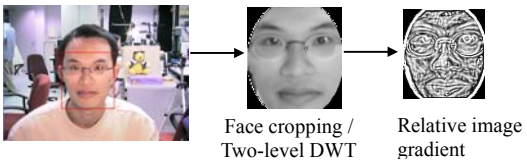

Figure 2. Automatic face cropping and feature extraction.

support vector machine (SVM) as the fusion classifier to make a decision.

2 System Overview

The basic building blocks of our system are introduced in Fig. 1. A person identity verification system shall include both enrollment and verification procedures. In the enrollment phase, a new user has to first type his name and say "I am xxx". If the name the user types is consistent with the name extracted by keyword spotting from what he/she says, the user will be asked to read two digit strings: 0~9 and 9~0. In the meanwhile, an image sequence of the user is captured by the camera. Then, the audio and video features extracted are used to train face and voice classifiers independently. A fusion classifier is trained later for fusing the face and voice classifier outputs. To users, the training process is of light weight since only few sentences have to be read and the required training images

0-7803-9332-5/05/$20.00 ©2005 IEEE

are captured at the same time.

In the on-line verification phase as shown in Fig. 1, a user first claims his/her identity by saying "I am xxx". The system knows the identity of the claimant by keyword spotting, and then asks him/her to read a random 8-digit string, like "25782930." The extracted voice features are then input to the voice classifier. At the same time, the face verification module keeps extracting features until the user finishes reading the digit string, and the face feature vectors are sent to the face classifier. Both voice and face classifiers will output a confidence score ranging from 0 to 1. After collecting both scores, the fusion classifier makes the final decision.

3 Face Verification Module

3.1 Reliable Facial Feature Extraction

There are two main parts in a face verification module – facial feature extraction and face classification. To detect face, we use the cascade-boosting method [15] to find a rough face region at first. However, the detected face may not be aligned well. To achieve a better alignment, we further localize user's eyes by using an eigen-eye technique [14] in corresponding areas of the face region. According to the positions of eyes, a proper geometric normalization (rotation, scaling and cropping) is applied.

After cropping a registered face image, we combine the multilevel two-dimensional discrete wavelet transform and relative image gradient feature [16] to extract lower-dimensional facial features that are insensitive to lighting variations, as shown in Fig. 2.

3.2 Face Verification Framework

In our work, the final face classifier is composed by a series of binary classifiers using the one-against-one scheme. To verify an identity claimed by the user, we first perform 2-class classifications N–1 times, where N is the number of classes. All decision values of the N–1 classifier are normalized to the range [0, 1], and the average decision value is used as the final confidence score.

We use Kernel Fisherface discriminant (KFD) to build face classifiers. KFD is an extension of linear Fisher's discriminant by using kernel trick and has shown its outstanding performances for face recognition [17][10]. Unlike other kernel-based techniques such as SVM, the outputs of KFD lend themselves to probability interpretations [11], and thus the confidence score generated by KFD is suitable for further combination or fusion.

In the enrollment phase, the training images collected for a new user are used to build N binary classifiers for the N users who have registered already. For a user-friendly system, the registered users shall be allowed to add or remove a training image on-line to adjust the classifiers adaptively. However, current KFD frameworks are not suitable for incremental learning when a new (or old) face image of a registered user is added to (or removed from) the training set. Note that incremental learning is useful especially when a verification system has to keep a fixed number of person's recent images as training data. With incremental learning, it can be achieved by adding a new and removing the oldest images of that person. To make such an on-line update feasible, we develop an *incremental KFD* (IKFD) method that can update a trained classifier according to newly input training data.

3.3 Incremental Kernel Fisher Discriminant Learning

For those already existing variations of KFD or kernel linear discriminant analysis, most of them are difficult to have accurate and efficient incremental versions. In our work, the IKFD is constructed by extending Mika's framework [12] since it showed that KFD can be cast into a slightly more general and easily solvable form by transforming its optimization problem into a convex and quadratic one. In this framework, the optimal solution turns out to be the following form:

$$\begin{bmatrix} \boldsymbol{\alpha}^* \\ b^* \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{K}^T\mathbf{K} & \mathbf{K}^T\mathbf{1} \\ \mathbf{1}^T\mathbf{K} & M \end{bmatrix}^{-1}}_{\mathbf{H}} \begin{bmatrix} \mathbf{K}^T\mathbf{y} \\ M_2 - M_1 \end{bmatrix}, \quad (1)$$

where $\mathbf{K}$ is the kernel matrix formed by the training vectors, $\mathbf{y}$ is the label vector, M_1 and M_2 are the number of positive and negative training vectors, respectively, and $M = M_1 + M_2$. The critical step of learning with this method is to compute the inverse of the matrix $\mathbf{H}$.

By extending this framework, we consider that when a training data is added or removed, the matrix $\mathbf{K}^T\mathbf{K}$ is changed and the corresponding $\mathbf{H}^{-1}$ can be represented as according to [11]:

$$\begin{bmatrix} \left(\mathbf{K}^T\mathbf{K}\right)^{-1} + \left(\left(\mathbf{K}^T\mathbf{K}\right)^{-1}\mathbf{K}^T\mathbf{1}\right)\gamma_H\left(\left(\mathbf{K}^T\mathbf{K}\right)^{-1}\mathbf{K}^T\mathbf{1}\right)^T & -\gamma_H\left(\left(\mathbf{K}^T\mathbf{K}\right)^{-1}\mathbf{K}^T\mathbf{1}\right) \\ -\gamma_H\left(\left(\mathbf{K}^T\mathbf{K}\right)^{-1}\mathbf{K}^T\mathbf{1}\right)^T & \gamma_H \end{bmatrix}, \quad (2)$$

where $\gamma_H = \left(M - \mathbf{1}^T\mathbf{K}\left(\mathbf{K}^T\mathbf{K}\right)^{-1}\mathbf{K}^T\mathbf{1}\right)^{-1}$.

$\mathbf{H}^{-1}$ is mostly made up by the matrix $(\mathbf{K}^T\mathbf{K})^{-1}$ that can be updated by the Sherman-Morrison-Woodbury formula [7] when a rank-one change $\Delta\mathbf{k}$ is applied,

$$\left(\mathbf{K}^T\mathbf{K} + \Delta\mathbf{k}^T\Delta\mathbf{k}\right)^{-1} = \left(\mathbf{K}^T\mathbf{K}\right)^{-1} - \frac{1}{1 + \Delta\mathbf{k}\left(\mathbf{K}^T\mathbf{K}\right)^{-1}\Delta\mathbf{k}^T}\left(\mathbf{K}^T\mathbf{K}\right)^{-1}\Delta\mathbf{k}^T\Delta\mathbf{k}\left(\mathbf{K}^T\mathbf{K}\right)^{-1}. \quad (3)$$

By applying Eq. (3), critical computations are updated from previous ones and fast incremental learning is achieved.

4 Voice Verification Module

4.1. Speaker Voice Modeling

By its nature, speech is a temporal signal. In speech recognition or speaker recognition, speech is usually represented as a sequence of feature vectors characterized by some distributions. Techniques based on Gaussian Mixture Models (GMMs) [13] have been successfully applied to speaker recognition in recent years. The most common approach for learning GMMs is the expectation-maximization (EM) algorithm [8]. The number of mixture components is usually decided empirically. In our system, we apply a model-selection-based self-splitting Gaussian mixture learning algorithm, which is able to automatically find an appropriate number of mixture components according to the characteristics of training data [6].

4.2. Verification by Speech

In speaker verification, the identity of the current speaker must be transmitted to the system beforehand and the task is to determine whether the current speaker is the claimed one or not. The log-likelihood ratio (LLR) [13] detector is the most popular method. Given an utterance X and a claimed speaker identity with the corresponding model λ_C, the LLR is calculated as,

$$\Lambda(X) = \log p(X \mid \lambda_C) - \log p(X \mid \lambda_{\overline{C}}) \quad (4)$$

where $\lambda_{\overline{C}}$ is the so-called anti-model. Suppose there are N clients (i.e., speakers) in the system, $\lambda_{\overline{C}}$ is selected as,

$$\lambda_{\overline{C}} = \arg \max_{\lambda_s, 1 \le s \le N, s \ne C} \log p(X \mid \lambda_s). \quad (5)$$

The LLR, $\Lambda(X)$, obtained by Eq. (4) can be further normalized between 0 and 1 via the Sigmoid function.

4.3. Integration of Speech Recognition and Speaker Verification

In general, speaker verification techniques can be classified into *text-dependent* approaches and *text-independent* approaches. The former require the speaker to say certain keywords or sentences with the same textual content for both training and recognition trials, while the latter do not rely on a specific text being spoken. Our voice verification module runs in both modes.

We use the keyword spotting technique [4] for speech recognition. The keyword lexicon contains the names of all clients and 500 random 8-digit strings dynamically generated by the system. In testing, a client first claims himself (or herself) by saying "I am xxx". Then, the system will prompt an 8-digit string and request the client to say it. The client has to obey this strict request and utter the sentence correctly. Otherwise, he or she will be rejected in the speech recognition phase. The advantages of combining speech recognition and speaker verification are twofold: First, the client can claim himself (or herself) by speech conveniently. Second, the impostor taking the recorded speech from the valid speaker will be rejected because the contents of prerecorded speech and the prompted text are mismatched.

5 Fusion Module

The fusion module is the final building block in the system. It has to makes a binary decision of accepting or rejecting the claimed identity via two confident scores output from face and speech voice verification modules. In our work, a multi-face/single-sentence strategy is used, which is defined as combining n face verification outputs with a single speech verification output to make a decision, where n is the number of images taken during the period of saying a specified digit-string, and n is typically 10 in our system. This task can be achieved by training a binary classifier in the (n+1)-dimensional space. However, the classifier might be difficult to train since the face and voice data are unbalanced in such a feature space. In our work, we divide this task into two steps. First, a single score is conducted from the n face verification scores. Then, this score is combined with the voice verification score in the second step to make a decision.

In the first step, we have used some statistical approaches including the selections of the maximum, minimum, median, mean, and trimmed-mean (the mean of the n values except the highest and the lowest ones) to conduct a single score. According to our experience, the trimmed-mean approach achieved the best performance, and is used in our current system. In the second step, making a single decision from the two scores of the face and voice modules is equivalent to doing a two-class classification, and we use SVM to achieve this purpose.

In the past, many works [2][9] also adopted SVM as a fusion classifier. However, why SVM is good for fusion has not been well investigated yet. In [5], we have given a theoretical explanation about why the SVM with radial-basis-function (RBF) kernel is generally better than linear combination for classifier fusion. Based on this study, we choose the RBF kernel for the SVM and perform a complete model selection to conduct the SVM classifier with the minimal cross-validation error for the fusion task.

6 Experimental Results

In our preliminary experiment, both face and voice data of 17 subjects were collected. For each subject, we have collected the training data which are used to train the face and speech classifiers by using our light-weight enrollment process, and evaluation data which are used for training the fusion classifier. The test data are collected in different time periods. The session-1 test data was collected about three hours after the enrollment, and the session-2 test data was collected about one week later. To have some intruders in the test sessions so that the performance can reflect the practical situation, we only used the data of 13 persons for training and evaluation, and the remaining 4 persons who have not been seen in the training and evaluation phases serve as intruders in test sessions 1 and 2. The database is summarized in Table 1.

From Table 2 and the ROC curve in Fig. 3., we can tell that through fusing multimodal information, the performance are indeed with significant improvement compared to using single modality. Also, results show a consistently good performance for both sessions, which reveals a good generalization. From the view of system usability, response time in such a identity verification system is crucial; it's a good thing to see that in our system the response time (calculate from a user finishes his identity claim till the final decision) is just 2.1 seconds.

7. Conclusions

In this paper, we propose an effective framework for combining face and voice data for person identity verification. Without the need to collect too much training data for each person, our method can get high verification accuracies in an acceptable execution time, and efficient incremental learning can be achieved. Our framework inherently rejects the intruders who try to use prerecorded voices to pretend legal users. We have also developed a practically useful strategy to combine multiple face images and a single sentence for conducting a binary decision.

ID	Training	Evaluation	Test session 1	Test session 2
0 ~ 12	30 images and 3 sentences per client	10 images and 10 sentences per client	100 images and 10 sentences per client	100 images and 10 sentences per client
13 ~ 16				

Table 1. Database Configuration

	Face	Voice	Fusion
Session1 EER	9.42%	11.06%	0.46%
Session2 EER	16.05%	10.29%	1.38%

Table 2. Comparison of EER (equal error rate) in face, voice, and fusion approaches

References

[1] S. Bengio, "Multimodal Authentication Using Asynchronous HMMs," *Proc. Intl. Conf. AVBPA, 2003*.

[2] S. Ben-Yacoub, "Multi-modal Data Fusion for Person Authentication using SVM," *Proc. Intl. Conf. AVBPA*, 1999.

[3] R. Brunelli and D. Falavigna, "Person Identity Using Multiple Cues," *IEEE Trans. on PAMI*, 1995.

[4] B. Chen, et al., "A*-admissible key-phrase spotting with sub-syllable level utterance verification," *Proc. ICSLP1998*.

[5] C. S. Chen et al., "Nonlinear Boost," *Technical Report, TR-IIS-04-001*, Inst. Info. Sci., Acad. Sinica, Taiwan, 2004.

[6] S. S. Cheng, H. M. Wang, and H. C. Fu, "A Model-selection-based Self-splitting Gaussian Mixture Learning with Application to Speaker Identification," *EURASIP Journal on Applied Signal Processing*, 17, pp. 1-14, 2004.

[7] G. H. Golub and C. F. van Loan. *Matrix Computations, 3rd edition*, John Hopkins University Press, London, 1996.

[8] X. Huang, A. Acero, H. W. Won, *Spoken Language Processing*, Prentics Hall, New Jersey, 2001.

[9] W. H. Lin, R. Jin, and A. Hauptmann, "Meta-classification for Multimedia Classifiers," *Proc. Intl. Workshop on Knowledge Discov. in Multimedia & Complex Data*, 2002.

[10] Q. Liu, et al., "Face Recognition Using Kernel Based Fisher Discriminant Analysis," *Proc. FG 2002*.

[11] S. Mika, A.J. Smola, and B. Scholkopf, "An improved training algorithm for kernel fisher discriminants," *Proc. AISTATS 2001*.

[12] S. Mika, *Kernel Fisher Discriminants*. PhD thesis, University of Technology, Berlin, October 2002.

[13] D. A. Reynolds, "Speaker Identification and Verification using Gaussian Mixture Speaker Models", *Speech Communication*, 17, pp. 179-192, 1995.

[14] C.-H. Su et al., "A Real-Time Robust Eye Tracking System for Autostereoscopic Displays Using Stereo Cameras," *Proc. IEEE ICRA 2003*.

[15] P. Viola, and M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple." *Proc. IEEE CVPR 2001*.

[16] S.–D. Wei, and S.–H. Lai, "Robust Face Recognition under Lighting Variations," *Proc. IEEE ICPR 2004*.

[17] M. H. Yang, "Face Recognition Using Kernel Methods," *Proc. NIPS 2002*.

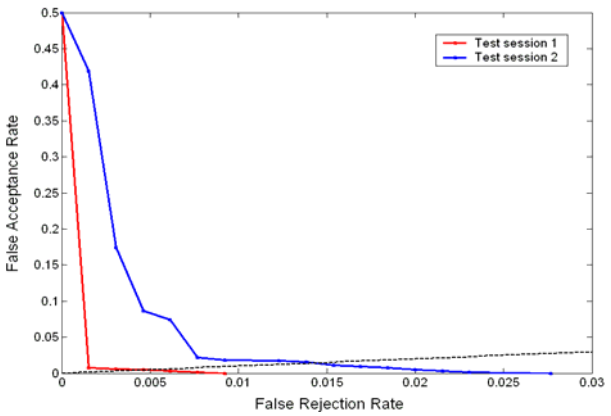

Figure 3. The ROC curves of test sessions 1 (red line) and 2 (blue line). The dashed line denotes the equal error rate.