Combining Voiceprint and Face Biometrics for Speaker Identification Using SDWS

Dongdong Li, Yingchun Yang and Zhaohui Wu

Department of Computer Science and Technology
Zhejiang University, Hangzhou, P.R.China, 310027
{lidd,yyc,wzh}@zju.edu.cn

Abstract

The biometric system that uses multiple biometric traits promises higher identification accuracy than identification in either individual domain. To reach this goal, special attention should be paid to the strategies for combining voiceprint and face experts. We propose an improved weighted sum rule based on the scores difference (SDWS) between the genuine speaker class and the mistaken speaker class labeled by each classifier, and demonstrate that the performance of multi-biometric system can be further improved by SDWS. The tests were conducted on a multi-modal database with 54 users We compare our approach with other existing methods and show that SDWS improved performance by about 7.8-13.3%, much better than the others.

1. Introduction

The authentication of a person is a complex task with high performance and robustness. Biometric verification has attracted attention recently because it is more secure than knowledge- or token-based verification techniques. Voiceprint and face recognition systems are among the top choices: they are non-invasive and friendly and therefore more popular with the users. However, when used separately, each modality reaches some limitations or shows a lack of robustness.

The system that combines different authentication modules is motivated by the fact that fusing several experts can increase identification rates over mono-modal biometrics. Also it has been suggested to fuse voiceprint and face these two easily accepted biometric traits could achieve an acceptable level of distinctiveness and user friendliness at the same time.

Several decision-level fusion rules have been developed over the last ten years [1-3]. According to the fusing rules, there are two different strategies [4]. One strategy is fixed fusion methods, such as Min, Max, Average, Majority Vote, and Sum. The second strategy is trained fusion methods, such as Dempster-Shafer[5], Behavior-knowledge space [6], and Naive Bayes. The bad quality and/ or the limited size of training sets quickly cancel the theoretical advantages of optimal trained rules [7]. Otherwise, the significant different pair-wise of each individual classifier affects the performance of the fixed rule [4,8].

The weighted sum rule is a trained fusion method and training data of small size can qualify the task. It is one of the most widely used fusion techniques in multi-modal system, for its simplification, convenience and excellent result. We further improve system performance by learning the weights for individual biometric expert according to the scores difference between genuine class and mistaken class assigned by the expert.

This paper is organized as follows: we give a brief introduction to the architecture of the fusion system in Section 2. In Section 3, we propose details of Scores Difference-Based Weighted Sum Rule (SDWS). The data set and the individual expert are presented in Section 4 with the experimental comparison and discussion between SDWS and other classical methods such as Majority Vote, DS, BKS. Finally, we give a conclusion in Section 5.

2. Fusion architecture

2.1 Identification architecture

With the identification problem, we assume that only enrolled persons would access the system. Therefore, identification is concerned with determining that person from a closed-set, whose features best match the features of the claimer to identify.

The Voiceprint-face fusion classifier incorporates two domain-specific experts, one for the acoustic speaker, and the other for the visual speaker domain. The final decision for closed-set person identification is implemented as high support decision procedure applied to the fusion module outputs, see Fig. 1.

2.2 Classifier fusion

The identification problem using a combination of classifiers can be formulated as follows:

Supposed $x \in \Re^n$ is an input feature vector and $DP(x)$ is a decision profile that consists of the outputs of the voiceprint and face experts [5].

$$DP(x) = \begin{bmatrix} d_{1,1}(x), d_{1,2}(x), ..., d_{1,c}(x) \\ d_{2,1}(x), d_{2,2}(x), ..., d_{2,c}(x) \end{bmatrix}$$

$d_{i,j}(x)$ is the degree of support (i.e. scores) delivered by classifier D_i to the hypothesis that x comes from class ω_j (most often an estimate of the posterior probability $P(\omega_j \mid x)$). $\{\omega_1, \omega_2, ..., \omega_c\}$ is the label set of c classes.

The outputs of these experts are taken as the input to a second-level classifier in some intermediate feature space. We design a new classifier for the second (combination) level. The fusion result is denoted by

$$\mu_D(x) = \Im(DP(x)) = [\mu_D^1(x), ..., \mu_D^C(x)]^T \quad (1)$$

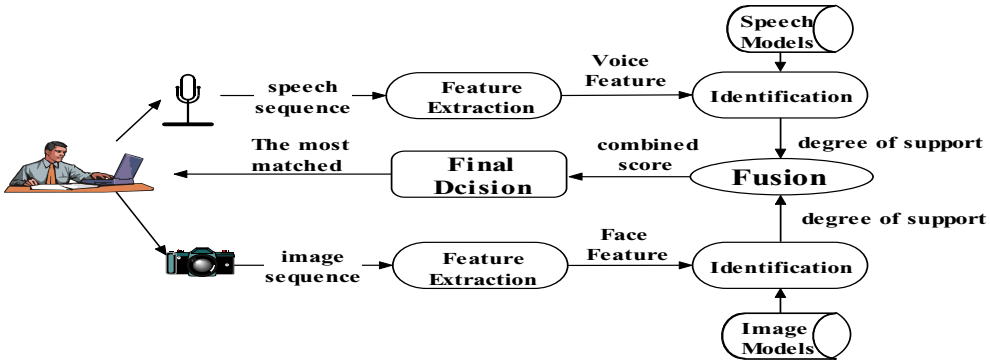

Figure 1: *The Voiceprint-face identification architecture.*

$\mu_D^i(x) \in [0,1]$. $\mu_D^i(x) = P(i \mid x)$ is the posterior probabilities for the class i, given x.

3. Scores Difference-Based Weighed Sum

3.1 Sum rule

This scheme [9,10] is one of the classic and convenient methods in all the fusion techniques. It operates as simply as the Min, Max rule, but always results in better performance. The sum rule can be expressed as:

$$\mu_D^j(x) = \sum_{i=1}^{2} d_{i,j}(x) \tag{2}$$

where i is the individual expert and j is the class label.

3.2 Weighted sum rule

The Weighted Sum Rule [11] has been developed from the sum rule for the reason that the performances of different classifiers are different. For this purpose, we give each classifier a weight to exhibit its importance or confidence. The scores received from the individual expert are combined by a weighted sum. The best-scoring candidate is chosen then. The combined score is defined as:

$$\mu_j(x) = \sum_{i=1}^{2} W_i d_{i,j}(x) \tag{3}$$

In the traditional weighted sum rule, the weights are calculated usually as follows:

$$W_i = \frac{1-2E_i}{2-2\sum_{i=1}^{2} E_i}, \text{ i=1,2, j=1,2 , i} \neq \text{j} \tag{4}$$

where E_i is the ER(Error Rate) expert i. Obviously, $\sum_{i=1}^{2} W_i = 1$

3.3 Scores Difference-based Weighted Sum Rule

There are two classifier behaviors. One kind is the classifiers which delivers similar degree of support to each class, namely classifiers with little scores difference, such as voiceprint expert; the other kind is the classifier with significant scores difference which gives a quite high degree of support to the class which it assigns the input feature vector to, such as face expert.

In fact, the traditional weighted sum rule has little effect if some classifier labels the wrong class with the degree of support much higher than the genuine one for a given input.

Learning the scores difference between the wrong class and the genuine class helps reduce the error rates, thereby improves the performance of the system.

We proposed the Scores Difference-Based Weighted Sum Rule (SDWS) to eliminate the effect of the scores difference that the expert gives to the mistaken class label between the genuine one.

Definition1. Let $D = \{D_1, D_2, ..., D_L\}$ be a set of classifiers and $\Omega = \{\omega_1, ..., \omega_c\}$ be a set of class labels. Each classifier gets as its input a feature vector $x \in \Re^n$ and assigns it to a class label from, i.e., $D_i : \Re^n \to \Omega$, or equivalently, $D_i(x) \in \Omega, i = 1,..., L$. Then, the output for the training data set with N elements $X = \{x_1, x_2, ..., x_N\}$ is demoted as :

$$S(X) = \begin{bmatrix} s_{1,1}(X),...,s_{1,L}(X) \\ \vdots \\ s_{j,1}(X),...,s_{j,L}(X) \\ \vdots \\ s_{N,1}(X),...,s_{N,L}(X) \end{bmatrix} \tag{5}$$

where $s_{j,i}$ is the class labels assigned to x_j by expert D_i . This is typically done by the maximum membership rule:

$$\begin{aligned} s_{j,i} &= D_i(x_j) \\ &= s \Leftrightarrow d_{i,s}(x_j) = \max_{o=1,2,...,c} \{d_{i,o}(x_j)\} \end{aligned} \tag{6}$$

Here, j=1,…,N is the number of the training data, and i=1,…L is the number of the experts. c is the number of classes, and represents the number of people in closed system.

The genuine class labels of X are $L(X) = [k_{1,...,} k_N]^T$ i.e., $L : \Re^n \to \Omega$.

Definition2. The scores difference $SD_i(X)$ of expert i can be calculated as follows:

$$SD_i(X) = \sum_{j=1}^{N} SD_i^j(x_j) = \sum_{j=1}^{N} \sum_{s_{j,i} \neq k_j} |d_{i,k_j}(x_j) - d_{i,s_{j,i}}(x_j)| \quad (7)$$

$SD_i(X)$ is the scores difference between the wrong class label and the genuine class label when $s_{j,i} \neq k_j$. $d_{i,j}(x)$ is the element in $DP(x)$.

Now, we can utilize the difference between the genuine speaker and the mistaken speaker of each expert as weights.

$$W_i = \frac{SD_i(X)^{-1}}{\sum_{i=1}^{L} SD_i(X)^{-1}} \quad (8)$$

The following example helps to clarify why SDWS has better performance. Let c=3, L=2, and N=5, and let the DP_j ,j=1,…5, obtained from the classifiers be:

$$DP_1 = \begin{bmatrix} 0.15 & 0.60 & 0.25 \\ 0.25 & 0.48 & 0.27 \end{bmatrix} \quad DP_2 = \begin{bmatrix} 0.73 & 0.14 & 0.13 \\ 0.38 & 0.17 & 0.45 \end{bmatrix}$$

$$DP_3 = \begin{bmatrix} 0.11 & 0.60 & 0.29 \\ 0.46 & 0.42 & 0.12 \end{bmatrix} \quad DP_4 = \begin{bmatrix} 0.55 & 0.19 & 0.26 \\ 0.40 & 0.35 & 0.25 \end{bmatrix}$$

$$DP_5 = \begin{bmatrix} 0.58 & 0.22 & 0.20 \\ 0.33 & 0.31 & 0.36 \end{bmatrix}$$

The genuine class label of the five training data is [2,3,2,1,1], so the error rate of expert 1 is 20% and the expert 2 is 40%.

Assumed that for an input x, the following decision profile has been obtained

$$DP(x) = \begin{bmatrix} 0.14 & 0.61 & 0.25 \\ 0.43 & 0.26 & 0.31 \end{bmatrix}$$

The genuine class label of x is 1, which means that x belongs to class 1.

Applying each of the operators columnwise, we obtain as the final soft class label $\mu_D(x)$

Sum Rule = $(0.57,0.87,0.56)^T$;

Weighted Sum Rule = $(0.21,0.52,0.27)^T$;

SDWS Rule = $(0.41,0.28,0.31)^T$;

Here, SDWS draws the correct conclusion by,

1. calculating the scores difference between the genuine class and the mistaken class.

$$SD_1 = 0.73 - 0.13 = 0.60$$

$$SD_2 = (0.46 - 0.42) + (0.36 - 0.33) = 0.04$$

2. getting the SDSW weight:

$$W_1 = \frac{SD_1(X)^{-1}}{\sum_{i=1}^{2} SD_i(X)^{-1}} = 0.06 \quad W_2 = \frac{SD_2(X)^{-1}}{\sum_{i=1}^{2} SD_i(X)^{-1}} = 0.94$$

SDWS is proposed based on the difference between the degrees of support which here we refer to as scores that delivered by classifiers to class sets.

In many cases, the individual expert would assign a score to the class label which it classifies the input to, distinctly higher than scores it delivered to other classes, even when the expert makes wrong decisions. SDWS reports transcendent performance in such fusion system with classifiers mentioned above.

4. Experiment results

4.1 Database

The data we used for the experiments were extracted from a multi-modal corpus [12]. The database consists in 54 different persons (17 females and 37males) ands contains asynchronous images and speech data.

The Utterance set is divided into seven sessions, which are personal information, mandarin digits, dialect digits, English digits, province phrase, paragraph and free talking. The personal information and paragraph sessions which are used for training include each visitor's registered data and a subsection from a famous Chinese essay. In the other five sessions, 10 prompts are asked to read per session, which are used for testing.

Each person is present in 4 different shots, 2 frontal ones, and 2 side profiles. We use one frontal image for training and the other for testing.

The recording environment is an office with low level of acoustic noise and sufficient lighting. In this way, we obtain a corpus of 54 subjects, with 4 face images and 54 voiceprint sentences per subject.

4.2 The individual biometric expert

The voiceprint identification expert is based on Dynamic Bayesian Network [13]. DBN is a new statistical approach from the perspective of Bayesian networks proposed for time series data modeling. It has become a promising way to modelize the speaker variability, for more details see [14].

The face identification algorithm used to compute the image scores is based on the Eigenface method. Our face identification system is a standard Principal Component Analysis classifier (PCA).

4.3 Results and discussion

In the voiceprint feature extraction, the hamming window is 32 mms and the frame shift is 16mms. The silence and unvoiced segments are discarded based on an energy threshold. The feature vectors are composed by 16 MFCC and their delta coefficients.

In order to investigate whether the method is robust under different speech content of different speech type, we execute experiments on some subsets of our multi-modal data corpus: Mandarin, Dialect, English, Phrase, and Free talk. We also compare the SDWS method with other classical fusion methods. The results are listed in table 1.

Table 1: *This is an example of a table Experimental results under different speech type and content of test sets. IR is the abbreviation for identification rate; Man for Mandarin type; Eng for English type; Dia for dialect type; Phrfor phrase type and FTlk for free talk.*

Fusion Method	IR for each speech content and type (%)					
	Man	Dia	Eng	Phr	FTlk	Ave rage
Voice Only	84.6	85.6	91.1	87.8	87.8	87.4
Face Only	85.18					
Sum	85.4	85.2	86.1	85.2	85	85.4
Weighted Sum	85.4	85.2	86.7	85.2	85	85.5
SDWS	98.0	98.0	98.9	99.3	98.3	98.5
Majority Vote	85.2	85.2	85.2	85.2	85.2	85.2
BKS	89.2	89.7	92.3	90.2	88.1	89.9
DS	95.5	96.0	98.4	96.3	96.8	96.6

The results show clearly that the SDWS outperforms the other approaches and that it leads to an identification error rate 9 times smaller than the traditional weighted sum rule in average.

We notice that the sum rule and the weighted sum rule make little effect. These two rules can't remedy the wrong classification given by the face expert with significant scores difference, while the voiceprint expert nearly gives every class the same scores, only a little higher for the genuine one.

The Majority Vote rule does the worst here, because only two experts fused. This method operates on the crisp decision profile. Once the two expert give the different answer, the rule choose one of them randomly.

The BKS, DS rule indeed enhance the performance of the system, while the SDWS significantly outperforms any of them. The look-up table of BKS needs large data sets to be properly trained. The calculation that DS involved however, are more complex than any of the other schemes.

The considerable performance achieved in the test shows that it is a promising way of using SDWS in multi-modal fusion problems.

5. Conclusions

This paper presents an approach of fusing the voiceprint and face experts in speaker identification. We discuss how to fuse two classifiers for speaker identification, and propose a new technique Score Different-Based Sum rule (SDWS) to calculate the weight for the sum rule. SDWS is based on the score difference between the assigned class delivered expert and the genuine class the input **x** essentially belongs to. Encouraging results of experiments on multi-modal corpus compared with 5similar measures demonstrate that SDWS is a promising way for multi-classifiers fusion techniques.

Although the presented work combines only two modalities, it can be extended to any number of modalities. Further experiments on a larger database (295 people) will focus on robustness with respect to missing data.

6. Acknowledgements

This work is supported by National Natural Science Foundation of P.R.China (No.60273059), National High Technology Research & Development Programme (863) of P.R.China (No.2001AA4180), Zhejiang Provincial Natural Science Foundation for Young Scientist of P.R.China (No.RC01058), Zhejiang Provincial Education Office Foundation (20020721), and Zhejiang Provincial Doctoral Subject Foundation (20020335025).

7. References

[1] Benoit Duc.el., "Fusion of audio and video information for multimodal person authentication", Pattern Recogniton Letters 18 (1997) 835-843.

[2] Verlinde P. and Chollet. G., "Comparing decision fusion paradigms using k-NN based classifiers, decision trees and logistic regression in a multi-modal identity verification application", In: Proc. 2nd Int.l Conf. on Audio- and Video-Based Biometric Person Authentication. Washingtion D.C. (1999) 188-193.

[3] Ben-Yacoub. S., Abdeljaoued. Y. and Mayoraz. E., "Fusion of face and speech data for person identity verification", IEEE Transactions on Neural Networks. (1999) 1065-1074.

[4] Roli. F., Kittler. J., Fumera. G. and Muntoni. D., "An Experimental Comparison of Classifier Fusion Rules for Multimodal Personal Identity Verification Systems", Multiple Classifier Systems (2002) 325-336

[5] Kuncheva L.I., J.C. Bezdek and R.P.W. Duin., "Decision templates for multiple classifier fusion: An Experimental Comparison. Pattern Recognition", 34 (2). (2001) 299-314.

[6] Huang. Y.S. and Suen. C.Y., "A method of combining multiple experts for the recognition of unconstrained handwritten numerals", Pattern Analysis and Machine Intelligence. IEEE Transactions on. Volume: 17. Issue: 1. (1995) 90 – 94.

[7] Roli. F. and Fumera. G., "Analysis of linear and order statistics combiners for fusion of imbalanced classifiers", 3rd Int. Workshop on Multiple Classifier Systems. Cagliari. Italy. (2002)

[8] Roli. F., Raudys. S. and Marcialis. G.L., "An experimental comparison of fixed and trained fusion rules for crisp classifier outputs", 3rd Int. Workshop on Multiple Classifier Systems (MCS 2002). Cagliari. Italy. (2002)

[9] Kittler. J., Hatef. M., Duin. R.P.W. and Matas. J., "On combining classifiers", Pattern Analysis and Machine Intelligence. IEEE Transactions on. Volume: 20. Issue: 3. March (1998) 226-239

[10] Ross. A. Jain. A. K. and Qian. Jian Zhong, "Information Fusion in Biometrics", Proc. 3rd International Conference on Audio- and Video-Based Person Authentication (AVBPA). Sweden. (2001) 354-359

[11] Jain A. K. and Ross. A., "Learning User-specific Parameters in a Multibiometric System", Proc. International Conference on Image Processing. Rochester. New York. (2002) 57-60

[12] Dongdong Li, LiFeng Sang, Yingchun Yang and Zhaohui Wu.: Bimodal Speaker Identification Using Dynamic Bayesian Network. To be appear in Lecture Notes in Computer Science, Vol. 3338. Springer-Verlag, Berlin Heidelberg New York (2004)

[13] Murphy. K., "Dynamic Bayesian Networks: Representation. Inference and Learning", Ph.D. thesis. U.C. Berkeley. (2002)

[14] Lifeng Sang. Zhaohui Wu. Yingchun Yang. Wanfeng Zhang.: Automatic Speaker Recognition Using Dynamic Bayesian Network. IEEE ICASSP 2003. Vol.1. (2003) 188-191.