A TWO-LEVEL FUSION APPROACH TO MULTIMODAL BIOMETRIC VERIFICATION

Ming-Cheung Cheung, Man-Wai Mak

Center for Multimedia Signal Processing
Dept. of Electronic and Information Engineering
The Hong Kong Polytechnic University, China

Sun-Yuan Kung

Dept. of Electrical Engineering
Princeton University
USA

ABSTRACT

This paper proposes a two-level fusion strategy for audio-visual biometric authentication. Specifically, fusion is performed at two levels: intramodal and intermodal. In intramodal fusion, the scores of multiple samples (e.g. utterances or video shots) obtained from the same modality are linearly combined, where the combination weights depend on the difference between the score values and a client-dependent reference score obtained during enrollment. This is followed by intermodal fusion in which the means of intramodal fused scores obtained from different modalities are either linearly combined or fused by a support vector machine (SVM). Experimental results based on the XM2VTSDB corpus show that intramodal and intermodal fusion are complementary to each other and that SVM-based intermodal fusion is superior to linear combination.

1. INTRODUCTION

Various researches have suggested that no single modality can provide an adequate solution for high-security applications. These studies agree that it is vital to use multiple modalities such as visual, infrared, acoustic, chemical sensors, and so on.

To cope with the limitations of individual biometrics, researchers have proposed using multiple biometric traits concurrently for verification. Such systems are commonly known as multimodal verification systems [1]. Multicue biometrics helps improve system reliability. For instance, while background noise has a detrimental effect on the performance of voice biometrics, it does not have any influence on face biometrics. Conversely, although the performance of face recognition systems depends on lighting conditions, lighting does not have any effect on voice quality. As a result, audio and visual (AV) biometrics has attracted a great deal of attention in recent years.

Generally, AV fusion can be treated as either a classifier combination problem or a pattern classification problem. Kittler et. al. [1] proposed a set of fusion rules to combine classifiers. For example, for those systems that can only provide decisions, a majority voting method can be used. If the outputs of classifiers are compatible (e.g., in the form of posterior probabilities), they can be linearly combined (sum rule) or multiplied together (product rule). Besides these combination methods, researches have also suggested to consider the outputs of individual classifiers as feature vectors and use a classifier such as support vector machines, binary decision trees, and radial basis function networks to classify the vectors [2, 3].

This work was supported by the Research Grant Council of Hong Kong SAR (Project Nos. PolyU 5131/02E and GT860).

This paper extends our recently proposed multi-sample fusion technique [4] to audio-visual biometric authentication systems. The technique is different from the conventional ones in that it divides the fusion process into two stages. In the first stage, the method assigns a larger weight to the more reliable scores in a frame-by-frame basis. This weight assignment process is performed on the audio and visual modalities independently. In the second stage, the weighted sum of the frame-based scores from the audio and visual modalities are further fused by either the sum rule or a support vector machine.

The remainder of this paper is organized as follows. Section 2 details the approach to computing the optimal weights for individual scores, based on the score distribution of independent samples and the prior knowledge of the score statistics. Section 3 discusses two types of intermodal fusion: sum rule and support vector machines. Evaluations of the proposed multi-sample fusion technique on speaker verification, face verification, and audio-visual biometric authentication are presented in Sections 4 and 5. Concluding remarks are provided in the Section 6.

2. INTRAMODAL MULTI-SAMPLE DECISION FUSION

We assume that in each verification session, $T^{(m)}$ normalized scores [5] can be obtained from modality m, that is

$$\mathcal{S}^{(m)} = \{s_t^{(m)} \in \Re; t = 1, \ldots, T^{(m)}\}, \tag{1}$$

where t is the frame index. In the *equal-weight* fusion approach [6], the mean score

$$\bar{s}^{(m)} = \frac{1}{T^{(m)}} \sum_{t=1}^{T^{(m)}} s_t^{(m)} \tag{2}$$

is used for decision making.

Instead of assigning an equal weight to all scores, Mak et al. [4] proposed a *zero-sum* intramodal fusion approach in which different weights are assigned to different scores. The approach splits a score sequence into K subsequences:

$$\mathcal{S}^{(m,k)} = \{s_t^{(m,k)} \in \Re; t = 1, \ldots, T^{(m)}/K\} \quad k = 1, \ldots, K. \tag{3}$$

The frame-level fused scores are then computed as

$$\hat{s}_t^{(m)} = \sum\nolimits_{k=1}^{K} \alpha_t^{(m,k)} s_t^{(m,k)}, \tag{4}$$

where $t = 1, \ldots, T^{(m)}/K$, and $\alpha_t^{(m,k)} \in [0, 1]$ represents the confidence (reliability) of the score $s_t^{(m,k)}$. The fusion weights

$\alpha_t^{(m,k)}$ are made dependent on both the training data (prior information) and recognition data (scores):

$$\alpha_t^{(m,k)} = \frac{\exp\{(s_t^{(m,k)} - \tilde{\mu}_p^{(m)})^2/2(\tilde{\sigma}_p^{(m)})^2\}}{\sum_{l=1}^{K}\exp\{(s_t^{(m,l)} - \tilde{\mu}_p^{(m)})^2/2(\tilde{\sigma}_p^{(m)})^2\}}, \quad (5)$$

where $t = 1, \ldots, T^{(m)}/K$ and $k = 1, \ldots, K$. By using enrollment data, the user-dependent prior score $\tilde{\mu}_p^{(m)}$ and prior variance $(\tilde{\sigma}_p^{(m)})^2$ are computed as follows:

$$\tilde{\mu}_p^{(m)} = \frac{K_c\tilde{\mu}_c^{(m)} + K_b\tilde{\mu}_b^{(m)}}{K_c + K_b} \quad (6)$$

and

$$(\tilde{\sigma}_p^{(m)})^2 = \frac{1}{K_c + K_b}\sum_{k=1}^{K_c+K_b}\left[\bar{s}^{(m,k)} - \tilde{\mu}_p^{(m)}\right]^2, \quad (7)$$

where K_c and K_b are respectively the numbers of client's enrollment utterances and pseudo-impostors' utterances, $\tilde{\mu}_c^{(m)}$ and $\tilde{\mu}_b^{(m)}$ are respectively the score means of client's and pseudo-impostors' utterances, and $\bar{s}^{(m,k)}$ denotes the mean score of the k-th enrollment utterance. Finally, the mean fused score

$$\hat{s}^{(m)} = \frac{K}{T^{(m)}}\sum\nolimits_{t=1}^{T^{(m)}/K}\hat{s}_t^{(m)} \quad (8)$$

is used for decision making.

A system that uses a single client-independent decision threshold must ensure that all client and impostor scores have values comparable to the threshold. This requirement can be fulfilled by normalizing the scores so that they fall into a predefined range. One possible approach (called Z-norm [7]) shifts and scales the impostor scores so that their mean and variance become zero and unity, respectively. More specifically, the claimant's score $\bar{s}^{(m)}$ in Eq. 2 or $\hat{s}^{(m)}$ in Eq. 8 is normalized:

$$s_{norm}^{(m)} = \frac{s^{(m)} - \mu_b^{(m)}}{\sigma_b^{(m)}} \qquad m \in \{A, V\} \text{ and } s \in \{\bar{s}, \hat{s}\}, \quad (9)$$

where $\mu_b^{(m)}$ and $\sigma_b^{(m)}$ are respectively the mean and standard deviation of client-dependent impostor scores. These impostor scores can be obtained during training by testing a client model against nontarget observations.

3. INTERMODAL DECISION FUSION

There are many ways to combine the scores of multiple modalities. Typical examples include (1) sum rule and product rule in rule-based fusion; and (2) support vector machines, multilayer perceptrons, binary decision trees in learning-based fusion. Research has shown that that the *sum rule* and *support vector machines* are generally superior [2, 3, 8, 9].

3.1. Sum Rule

Given the audio score $s^{(A)}$ and visual score $s^{(V)}$, the audio-visual score s is obtained by linearly combining the two scores:

$$s = \beta s^{(A)} + (1 - \beta)s^{(V)}, \quad (10)$$

where β is a combination weight that can be computed using training data or made dependent on the quality of audio or visual data [10–12]. The audio and visual scores must have the same range for the fusion to be meaningful. This can be achieved by normalizing the scores, as in Eq. 9.

3.2. Support Vector Machines

A support vector machine (SVM) [13] is a binary classifier that maps input patterns $\mathbf{x}_i \in \Re^d$ to output labels $y_i \in \{-1, 1\}$, where $i = 1, \ldots, l$, and l is the number of patterns. Generally, an SVM has the form

$$f(\mathbf{x}) = \sum_{j\in\Omega}\alpha_j y_j K(\mathbf{x}, \mathbf{x}_j) + b, \quad (11)$$

where α_j are the Lagrange multipliers, Ω contains the indexes to the support vectors for which $\alpha_j \neq 0$, b is a bias term, $\mathbf{x}$ is an input vector to be classified, and $K(\mathbf{x}, \mathbf{x}_j)$ is a kernel function. The most common kernels are:

- Polynomial:

$$K(\mathbf{x}, \mathbf{x}_j) = (\mathbf{x} \cdot \mathbf{x}_j + 1)^p, \qquad p > 0; \quad (12)$$

- Radial Basis Function:

$$K(\mathbf{x}, \mathbf{x}_j) = \exp(-\|\mathbf{x} - \mathbf{x}_j\|^2/2\sigma^2). \quad (13)$$

For audio-visual biometrics, $\mathbf{x}$ is typically composed of audio and visual scores (i.e. $\mathbf{x} = [s_{norm}^{(A)} \quad s_{norm}^{(V)}]^T$) and decisions are based on whether the value $f(\mathbf{x})$ is above or below a threshold. Research has found that the polynomial kernel is superior to the RBF kernel for audio-visual fusion [2].

4. EXPERIMENTS

4.1. Audio-Visual Data Sets

The XM2VTSDB corpus [14] was used in the evaluations. XM2VTSDB is an audio-visual corpus designed for biometric research. The corpus consists of the audio and video recordings of 295 subjects taken over a period of four months. We adopted Configuration II as specified in [14] in the evaluation. More precisely, the database was divided into 200 clients, 70 impostors (part of the 95 impostors in DVD003b) for testing, and 25 pseudo-impostors (the remaining impostors in DVD003b) for finding decision thresholds or other system parameters. For each client, the first two sessions were used for training, and the last session was used for testing. Each client was impersonated by 70 impostors using the audio and video data of the four sessions.

4.2. Preprocessing of Audio Files

Because the original audio files were captured in a quiet, controlled environment using a high-quality microphone, the equal error rate (EER) using the audio data alone is very low (about 0.7%); as a result, performing audio-visual fusion was unnecessary. Therefore, coder distortion and factory noise were introduced to the sound files in an attempt to simulate a more realistic acoustic environment.

The audio files in the corpus were down-sampled from 32kHz to 8kHz. The down-sampled PCM files were transcoded by a GSM

codec. Factory noise ("factory1.wav" of the NOISE92 database) was added to the test files at a signal-to-noise ratio of 1dB. Note that the addition of noise was applied to the test files only; the training files were only down-sampled and GSM-transcoded. This introduces acoustic mismatch between the training and testing files. Nineteen MFCCs and their time derivative (delta MFCCs) were extracted from the files using a 28ms Hamming window at a rate of 71Hz.

We used the training sessions of 200 client speakers in the speaker set to create a 128-center background model. The background model was then adapted to speaker models using MAP adaptation [5]. As defined in Configuration II of XM2VTSDB, two sessions (i.e., four utterances) per speaker were used for model training. Cepstral mean subtraction (CMS) was performed on all MFCCs before they were used for training, testing, and evaluation.

4.3. Preprocessing of Video Files

Similar to audio files, the quality of video files in the corpus was also very good, making audio-visual fusion unnecessary (as face verification on the original video data already approaches 0% EER). As a result, distortion was introduced to the video sequences using PhotoShop Version 7.0 as follows. First, each of the AVI files in the corpus was converted into a sequence of high-quality JPEG files with 720 $\times$ 576 pixels. Second, the frame rate was reduced to one frame per second; for each frame, the JPEG images were down-sampled to 176 $\times$ 144 pixels. Third, the images was blurred by setting the "Gaussian Blur" of PhotoShop to 1.0. Finally, Gaussian noise was added to the image by setting the "Gaussian Noise" of PhotoShop to 3.0.

The noise-added image sequences were input to an Identix's Face Verification SDK to locate the head and compute the scores, which have a range of 0 to 10. The higher the score, the more likely the claimant is genuine.

4.4. Audio-Visual Multi-Sample Fusion

We assumed that one utterance and one video shot can be obtained from the claimant in a verification session. The utterance and the video shot were divided into two equal-length subutterances and two equal-length subvideo shots, i.e., $K = 2$ and $m \in \{A, V\}$ in Eq. 3, where A represents the audio channel and V the video channel. Feeding these subutterances and subvideo shots to the speaker verification system and the face verification system (FaceIT) gives two streams of audio scores and two streams of visual scores. We applied intramodal fusion to the two audio score streams and also to the two visual score streams independently to obtain the mean of the fused audio scores, $\hat{s}^{(A)}$, and the mean of the fused visual scores, $\hat{s}^{(V)}$. These scores were further normalized according to Eq. 9 to ensure that they have the same range. The client-dependent fusion parameters, including the prior scores and prior variances ($\tilde{\mu}_p^{(m)}$,$(\tilde{\sigma}_p^{(m)})^2$; $m \in \{A, V\}$), were obtained by feeding the utterances and video shots of 25 pseudo-impostors to the client and background models.

The normalized scores of every clients and of the 25 pseudo-impostors were used for training a second-degree polynomial SVM, i.e. $p = 2$ in Eq. 12. We used the training data of all clients and all pseudo-impostors to obtain 400 clients scores (2×200) and $40,000$ pseudo-impostors scores $(8 \times 25 \times 200)$ and used these scores to train a client-independent SVM. Because a client-independent SVM was used, Z-norm was applied to all of the audio and visual scores to ensure that the decision threshold is appropriate for all clients.

During verification, a total of 400 client trials (200 clients $\times$ 2 utterances per client) and 120,000 impostor attempts (200 clients $\times$ 75 impostors per client $\times$ 8 utterances per impostor) were used to test the system.

5. RESULTS AND DISCUSSIONS

Table 1 shows the EERs of speaker verification and face verification using different types of intramodal multi-sample fusion techniques described in Section 3. The results show that (1) zero-sum fusion generally performs better than equal-weight fusion and (2) Z-norm helps lower the EER of both type of fusion.

Table 2 summarizes the error rates obtained by applying audio-visual multi-sample fusion with β in Eq. 10 set to 0.5, and Fig. 1 plots the corresponding DET curves. The results show that zero-sum fusion always performs better than equal-weight fusion. The results also show that applying intermodal fusion (either linear combination or SVMs) on intramodal fused scores can further reduce the EERs. Figure 2 plots the decision boundary created by the sum rule and the second-degree polynomial SVM. It shows that the SVM can create a nonlinear boundary to separate the client scores from the impostor scores, which results in lower error rates.

It is of interest to compare the proposed two-level fusion approach with the multi-frame–multi-expert system proposed in Czyz et al. [15]. In their system, two different face verification algorithms were applied to the same facial sequence to create two set of scores. The frame-based scores were then fused using the minimum rule to obtain two minimum scores, which were subsequently fused using an SVM. The multi-frame part of [15] can be considered as a special case of our multisample fusion because when $\tilde{\mu}_p^{(m)}$ in Eq. 5 is large, $\alpha_t^{(m,k)}$ in Eq. 4 will be close to 1 for small scores and close to 0 for large scores, which has the effect similar to the minimum rule.

6. CONCLUDING REMARKS

This paper has presented an audio-visual biometric authentication system. A novel two-level fusion technique that fuses the scores obtained from speaker and face models was detailed. The proposed technique is general and is applicable to multimodal biometric systems. This is evident by the encouraging experimental results based on the XM2VTSDB database. It was found that an error rate reduction of up to 89% can be achieved when the proposed fusion technique is applied to fuse the scores derived from speaker models and face models.

7. REFERENCES

[1] J. Kittler, G. Matas, K. Jonsson, and M. Sánchez, "Combining evidence in personal identity verification systems," *Pattern Recognition Letters*, vol. 18, no. 9, pp. 845–852, Sept. 1997.

[2] S. Ben-Yacoub, Y. Abdeljaoued, and E. Mayoraz, "Fusion of face and speech data for person identity verification," *IEEE Trans. on Neural Networks*, vol. 10, no. 5, pp. 1065–1074, 1999.

[3] V. Chatzis, A.G. Bors, and I. Pitas, "Multimodal decision-level fusion for person authentication," *IEEE Trans. on Systems, Man and Cybernetics-Part A: Systems and Humans*, vol. 29, no. 6, pp. 674–680, 1999.

Fusion Method	Voice	Rel. Red. (w.r.t. equal-weight fusion)	Face	Rel. Red. (w.r.t. equal-weight fusion)
Equal-weight	14.52%	N.A.	9.64%	N.A.
Equal-weight+Znorm	10.29%	29.13%	5.23%	45.75%
Zero-sum	9.73%	32.99%	8.51%	11.72%
Zero-sum+Znorm	8.50%	41.80%	4.77%	50.52%

Table 1. Equal Error rates (EERs) and relative reduction (Rel. Red.) with respective to equal-weight fusion achieved by the speaker and face verification systems using intramodal multi-sample fusion. Note that fusion takes place only within the audio and visual scores, not between them. *Equal-weight+Znorm* (*Zero-sum+Znorm*) means that Z-norm was performed on the mean fused scores.

Intramodal Fusion Type	Intermodal Fusion Type	EER	Rel. Red. of EER (w.r.t. voice only)
EW+Znorm	Sum Rule	2.56%	82.37%
ZS+Znorm	Sum Rule	1.77%	87.81%
EW+Znorm	SVM	1.93%	86.71%
ZS+Znorm	SVM	1.51%	89.60%

Table 2. Error rates and relative error reduction with respect to the EER of speaker verification (Table 1) obtained by linearly combining the means of intramodal fused scores and by polynomial SVMs. The combination weight β in Eq. 10 was set to 0.5. *EW* and *ZS* stand for equal-weight and zero-sum, respectively.

[4] M. W. Mak, M. C. Cheung, and S. Y. Kung, "Robust speaker verification from GSM-transcoded speech based on decision fusion and feature transformation," in *Proc. IEEE ICASSP'03*, 2003, pp. II745–II748.

[5] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[6] N. Poh, S. Bengio, and J. Korczak, "A multi-sample multi-source model for biometric authentication," in *Proc. IEEE 12th Workshop on Neural Networks for Signal Processing*, 2002, pp. 375–384.

[7] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, pp. 42–54, 2000.

[8] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. on Pattern Anal. Machine Intell.*, vol. 20, no. 3, pp. 226–239, 1998.

[9] J. A. Fierrez, J. G. Ortega, D. R. Garcia, and J. R. Gonzalez, "A comparative evaluation of fusion strategies for multimodal biometric verification," in *AVBPA 2003*, 2003, pp. 830–836.

[10] U. Meier, W. Hurst, and P. Duchnowski, "Adaptive bimodal sensor fusion for automatic speech reading," in *Proc. ICASSP'96*, 1996, pp. 833–836.

[11] C. Neti et al., "Audio-visual speech recognition," in *Final Workshop 2000 Report*, Center for Language and Speech Processing, The Johns Hopkins University, Baltimore, 2000.

[12] C. Sanderson and K. K. Paliwal, "Noise compensation in a person verification system using face and multiple speech features," *Pattern Recognition*, vol. 36, pp. 293–302, 2003.

[13] T. Joachims, "Making large-scale SVM learning practical," in *Advances in Kernel Methods - Support Vector Learning*. B. Scholkopf, C. Burges and A. Smola (ed.) MIT-Press, 1999.

[14] J. Luettin and G. Maitre, "Evaluation protocol for the extended M2VTS database," Tech. Rep., IDIAP, Martigny, Valais, Switzerland, Oct. 1998.

[15] J. Czyz, M. Sadeghi, J. Kittler, and L. Vandendorpe, "Decision fusion for face authentication," in *ICBA'04*, 2004, pp. 686–693.

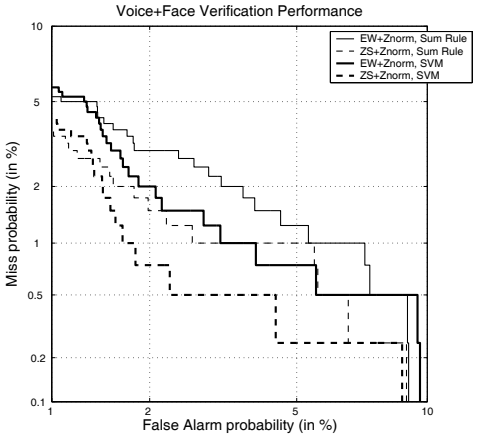

Fig. 1. DET plots showing the performance of the sum rule and SVMs in fusing the audio and visual scores. *EW* stands for equal-weight fusion and *ZS* stands for zero-sum fusion. *EW+Znorm* means that Z-norm was performed on the fused scores using equal weight fusion. A similar definition applies to *ZS+Znorm*.

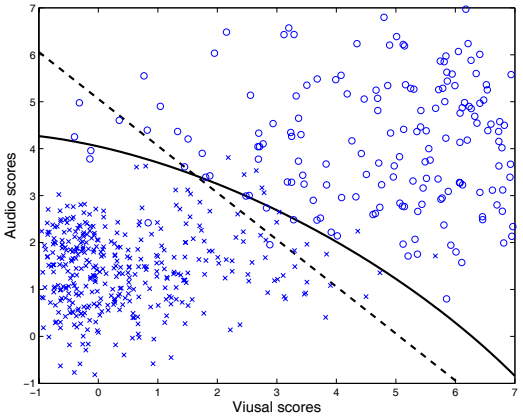

Fig. 2. Decision boundary created by the sum rule (dotted) and a second-degree polynomial SVM (solid). Circles (○) and Crosses (×) represent clients' and impostors' attempts, respectively.