

ON CONSISTENT FUSION OF MULTIMODAL BIOMETRICS

S. Y. Kung

Dept. of Electrical Engineering
Princeton University, USA

Man-Wai Mak

Dept. of Electronic and Information Engineering
The Hong Kong Polytechnic University, Hong Kong SAR

ABSTRACT

Audio-visual (AV) biometrics offer complementary information sources, and the use of both voice and facial images for biometric authentication has recently become economically feasible. Therefore, multi-modality adaptive fusion, combining audio and visual information, offers an efficient tool for substantially improving the classification performance. In terms of implementation, we propose to integrate an audio classifier (based on Gaussian mixture models) and a visual classifier (based on FaceIT, a commercially available software) into a well-established mixture-of-expert fusion architecture. In addition, a consistent fusion strategy is introduced as a baseline fusion scheme, which establishes the lower bound of the “consistent region” in the FAR-FRR ROC. Our simulation results indicate that the prediction performance of the proposed adaptive fusion schemes fall in the consistent region. More importantly, the notion of consistent fusion can also facilitate the selection of the best modalities to fuse.

1. INTRODUCTION

Audio-visual (AV) biometrics has long been an active area of research, primarily because of the promise it can bring to practical applications. These two independent and complementary information sources are ideal candidates for enhancing biometric system reliability. With the recent introduction of third-generation mobile services, the use of both voice and facial images for biometric authentication has become practical and economically viable.

Multi-modality adaptive fusion offers an efficient tool for improving the classification performance of AV biometric systems. For example, voice biometrics can suffer severe performance degradation under a noisy acoustic environment, but facial images are unaffected. Conversely, facial image quality can be severely affected in poor lighting conditions, but lighting has no effect on voice quality. This paper explains how consistent fusion can benefit audio-visual biometric authentication. Accordingly, a Mixture-of-Expert (MOE) type of fusion network as shown in Figure 1 is proposed. Several variants of fusion networks are studied and their performance compared.

2. AUDIO AND VISUAL MODALITIES

The XM2VTSDB corpus [1] was used to demonstrate the benefit of consistent fusion. The corpus consists of the audio and video recordings of 295 subjects taken over a period of four months. Each subject participated in four recording sessions, each with two utterances and two video shots.

This work was in part supported by The Research Grant Council of the Hong Kong SAR (Project Nos. UPolyU 5230/05E and A-PF54).

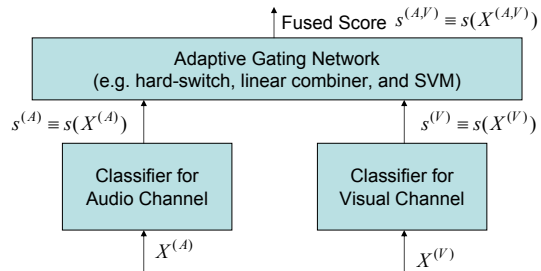


Fig. 1. Fusion network in an MOE (mixture-of-expert) architecture. Each vector sequence is compressed into a local score. The local scores are then fused by a gating network.

Configuration II of the corpus was adopted 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 the 70 impostors using the audio and video data of the four sessions.

Audio Modality. Because the original audio files were captured in a quiet, controlled environment using a high-quality microphone, the equal error rate using the audio data alone is very low (about 0.7%); as a result, performing audio-visual fusion was unnecessary. To simulate a more realistic acoustic environment, GSM codec distortion and factory noise [2] at an SNR of 4dB were introduced to the sound files (see [3] for details). Twelve MFCCs and their time derivative were extracted from the noisy, transcoded files using a 28ms Hamming window at a rate of 71Hz. Cepstral mean subtraction was performed on all MFCCs to remove linear channel effects.

The training sessions of 200 client speakers in the speaker set were used to create a 128-center GMM background model. The background model was then adapted to speaker models using MAP adaptation [4]. As defined in Configuration II of XM2VTSDB, two sessions (i.e., four utterances) per speaker were used for model training.

Visual Modality. Similar to audio files, the quality of video files in the corpus is also very good, making AV fusion unnecessary (as face verification on the original video data already approaches 0% EER). Distortion was introduced to the images of the video sequences using PhotoShop Version 7.0 (see [3] for details). The noise-added image sequences were input to Identix’s Face Verification SDK [5] 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.

Normalization of Scores. Because there were only 200 client subjects with two verification utterances per subject, a client-independent decision threshold was used to increase the resolution of the error rates. Specifically, the 400 client scores (200 clients \times 2 utterances per client) were lumped together and the scores were compared against the 120,000 impostor scores (200 clients \times 75 impostors per client \times 8 utterances per impostor) to obtain a client-independent EER and a DET plot.

Systems that use client-independent thresholds must ensure that the single threshold falls into the right range of all client and impostor score distributions. This can be achieved by using Z-norm [6]: $s_{norm}^{(m)} = (s^{(m)} - \mu_b^{(m)})/\sigma_b^{(m)}$, $m \in \{A, V\}$, where $s^{(m)}$ is the mean of claimant’s audio scores (when $m = A$) or visual scores (when $m = V$), and $\mu_b^{(m)}$ and $\sigma_b^{(m)}$ are the mean and standard derivation of the client-dependent impostor scores, respectively. These impostor scores can be obtained during training by testing a client model against pseudo-impostor attempts. In this work, the impostor observations were obtained from the 25 pseudo-impostors defined in Configuration II of the XM2VTSDB corpus.

3. FUSION OF AUDIO AND VISUAL SCORES

One utterance and one video shot from the claimant were obtained in a verification session. Then, the utterance and the video shot were divided into two equal-length subutterances and two equal-length subvideo shots. Feeding these subutterances and subvideo shots to the speaker verification system and the face verification system gives two streams of audio scores and two streams of visual scores. Multisample fusion [3, 7] was applied to the two audio score streams and also to the two visual score streams independently to obtain the mean of the fused audio scores $s^{(A)}$ and the mean of the fused visual scores $s^{(V)}$.

3.1. DET Performance Based on Single Modality

Let the distribution of the client scores and impostor scores from the audio (or visual) channel be $p(s(X)|\Lambda_c)$ and $p(s(X)|\Lambda_i)$, respectively, where X represents a sequence of feature vectors derived from an utterance or a video shot.¹ A test sequence X from a claimant is classified as coming from the true client if

$$\log p(s(X)|\Lambda_c) > \log p(s(X)|\Lambda_i) + \eta, \quad (1)$$

otherwise it will be classified as coming from an impostor. By counting the number of misclassified test sequences, we can compute the FAR and FRR (or precision, specificity, and sensitivity in other applications) corresponding to a single point on the ROC curve or DET curve [8]. To produce the entire spectrum of FAR and FRR, we can gradually adjust the running variable η to change from small to large values. For example, we set $\eta > 0$ (resp. $\eta < 0$) when a lower FAR (resp. FRR) is desired. Note that because $s(X)$ ’s are scalars, the DET or ROC can also be obtained by sweeping a decision threshold ζ from the minimum to the maximum value of the test scores $s(X)$ in the following decision rule:

$$\text{If } s(X) \begin{cases} > \zeta & X \text{ is from a client} \\ \leq \zeta & X \text{ is from an impostor.} \end{cases} \quad (2)$$

¹For clarity, the superscript in $X^{(A)}$ and $X^{(V)}$ were omitted.

Figure 2(a) shows the FRR against FAR of face-only, voice-only, and face plus voice systems. Evidently, the crossing point of the two DET curves allows us to choose the best system in different applications. For example, the visual features have lower FRRs in the low FAR region, whereas the audio features have lower FRRs in the high FAR region. This provides very crucial information for the fusion strategy proposed in the next section.

3.2. DET Performance Based on Fusion of Multi-modalities

The DET (or ROC) corresponding to the fusion of multi-modalities can be obtained by extending the aforementioned idea to multi-modality cases. More specifically, $s(X)$ ’s in Eq. 1 become two-dimensional vectors $s(X^{(A)}, X^{(V)}) = [s(X^{(A)}) \ s(X^{(V)})]^T$ comprising the scores derived from two modalities, and Λ_c and Λ_i become 2-D Gaussian mixture models representing the score distributions of the client and impostor classes, respectively. By counting the number of test scores $s(X^{(A)}, X^{(V)})$ falling on the wrong side of the decision boundary, we can compute the FAR and FRR (or precision, specificity, and sensitivity) corresponding to a single point on the DET curve. The entire spectrum of FAR and FRR and their corresponding decision boundaries (see Figure 4) can then be obtained by adjusting the value of η in Eq. 1.

3.3. Consistent Fusion

A minimum objective of fusion is naturally to deliver a *consistent fusion* [9], which by definition has an equal or better performance than any individual modalities in the entire FRR/FAR region.

As to which modalities to fuse, a natural selection criteria is to adopt the modalities that offer most complementary information. Again, an important clue can be obtained by examining the DET curves, as exemplified by Figure 2(a). Specifically, the audio modality shown in Figure 2(a) has a relatively lower FRR in the high-FAR region but a relatively higher FRR in the low-FAR region. In contrast, the visual modality has just the opposite performance. In this case, the two DET curves have a crossover point. Therefore, these two modalities are truly complementary to each other and can serve as ideal fusion candidates.

4. FUSION NETWORK IN MIXTURE-OF-EXPERT (MOE) ARCHITECTURE

Once we know which modalities to fuse, the next question to address is the fusion strategy. One possibility is to consider a direct fusion scheme, where the original feature vectors are concatenated to form an expanded vector to be collectively processed in the fusion layer. However, our prior experiences suggest that the direct feature fusion has consistently inferior performance compared with other fusion approaches. This may be attributed to the exceedingly large vector dimension after feature concatenation. In particular, when there exists a limited amount of training data, it is usually more difficult to model the distribution of high-dimensional vectors and consequently the performance deteriorates severely.

Based on this experience, we propose an indirect fusion scheme illustrated in Figure 1, where each feature vector is processed by a local expert and result in a local score. From the architectural design perspective, we adopt a well established Mixture-of-Expert (MOE) architecture comprising two layers: (a) the lower layer contains several local experts, each of which produces a local score

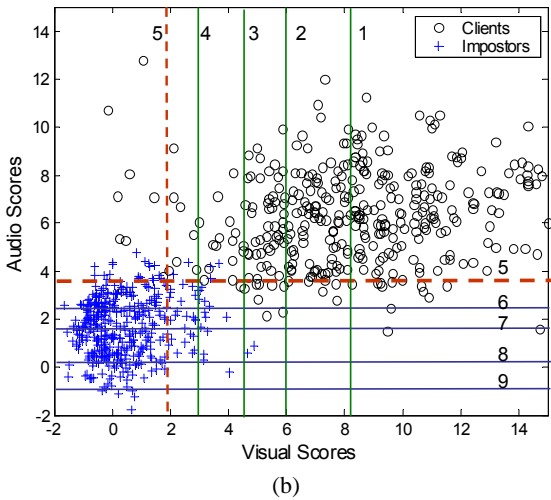
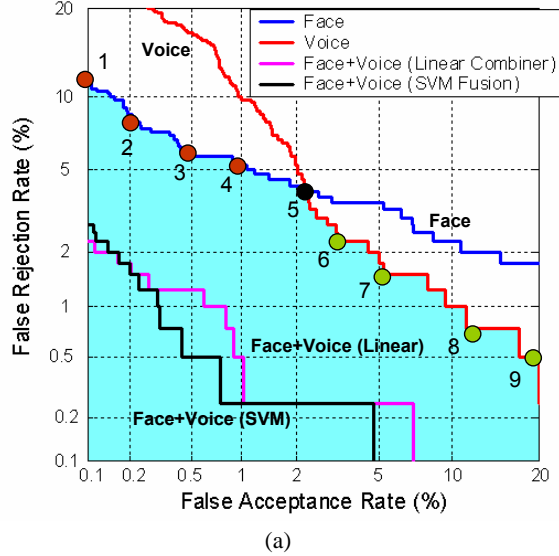


Fig. 2. Diagram illustrating the concept of consistent fusion. (a) Region of consistent fusion (the light blue area). (b) The adaptive hard switching fusion scheme guarantees a consistent fusion result. Note that at a specific FAR, the vertical decision boundaries based on the visual modality – boundaries #1, #2, #3, and #4 – have an FRR lower than that of the audio modality; whereas at a specific FRR the horizontal decision boundaries based on the audio modality – boundaries #6, #7, #8, and #9 – have an FAR lower than that of the visual modality. Therefore, the boundaries are switched from vertical to horizontal ones around the crossover point. At the crossover point, the boundary (#5) can be either horizontal or vertical as they deliver exactly the same FRR/FAR performance.

based on a single modality; and (b) the upper layer contains a fusion (or gating) network whose function will be elaborated below.

There are many ways to combine the audio and visual scores. Typical examples include (1) hard switching fusion network, (2) sum rule and product rule in rule-based fusion, and (3) support vector machines, multilayer perceptrons, and binary decision trees in learning-based fusion. Research has shown that the *sum rule* and *support vector machines* are generally superior [10, 11]. They

will be addressed in the subsequent discussions.

4.1. Adaptive Hard-Switching Networks

A hard-switching network can be implemented by the following scheme.

1. Determine the crossover point of DET A and DET B. Denote the FRR and FAR at the crossover point as FRR_{cr} and FAR_{cr} , respectively. For example, in Figure 2(a), the crossover of the audio modality and the visual modality is at the point $FRR_{cr} = 4.0\%$ and $FAR_{cr} = 2.5\%$.
2. If we want to guarantee an FAR to remain lower than or equal to $FAR_{cr} = 2.5\%$ while keeping the FRR to a minimum, then the decision boundary pertaining to Model B should be adopted, i.e., the visual modality in Figure 1. On the other hand, if we want to be sure of an FRR no higher than $FRR_{cr} = 4.0\%$ while keeping the FAR to a minimum, then we should adopt Model A, i.e., the audio modality.

The hard-switching scheme will yield basically a lower bound performance of any consistent fusion, as shown in the light-blue region in Figure 2).

Note that this scheme requires an additional data set (called held-out set) for determining the thresholds corresponding to the crossover point, because the true labels of the test data are supposed to be unknown. Once the crossover point is found, the DET corresponding to the test set can be obtained by applying hard-switching at the crossover point. Under such experimental procedure, a consistent performance can continue to hold up only under the additional assumption that the held-out data set shares the same statistics as the testing set.

4.2. Adaptive Weighted Combination Networks

Mathematically, denote the fusion score as

$$s^{(A,V)} \equiv s(X^{(A,V)}) = \alpha s(X^{(A)}) + \beta s(X^{(V)}). \quad (3)$$

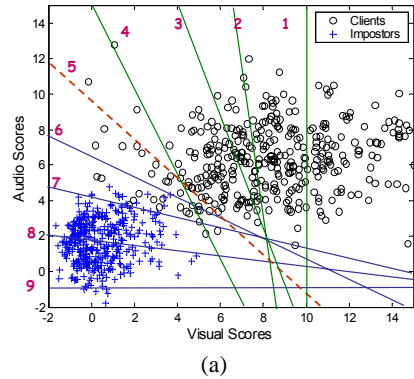
In the hard-switching scheme, we have either $\alpha = 1, \beta = 0$ or $\alpha = 0, \beta = 1$. In contrast, one may adopt a linear soft fusion scheme such that $0 \leq \alpha, \beta \leq 1$ and $\alpha + \beta = 1$. In many cases, such a soft fusion scheme can lead to better-than-lower-bound performance, as explained in Figure 3. The optimal values of α and β can better be derived via prominent machine learning techniques, such as Fisher classifiers and support vector machines (SVMs) with a linear kernel [12]. Unfortunately, it is known that linear classifiers often have limited discriminating power.

4.3. Adaptive Nonlinear Fusion Networks

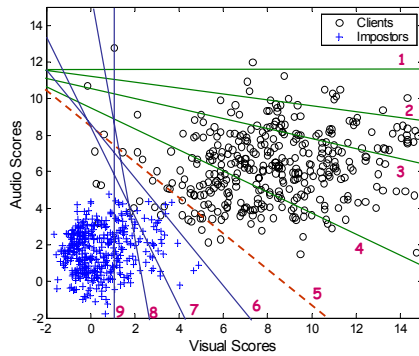
The hard-switching and linear combination schemes described earlier can only produce linear decision boundaries to separate the positive (client) and negative (impostor) classes in the score space. To allow more flexible decision boundaries, we can use nonlinear classifiers (e.g., SVMs [12] or decision-based neural networks [3]) to combine the local scores. Specifically, a 2-input SVM can be trained to compute the fused score

$$s^{(A,V)} = \sum_{j \in \mathcal{S}} \alpha_j y_j K(\mathbf{s}, \mathbf{s}_j) + b, \quad (4)$$

given the input $\mathbf{s} = [s(X^{(A)}) \ s(X^{(V)})]^T$. In Eq. 4, $\alpha_j, j \in \mathcal{S}$, are the Lagrange multipliers, \mathcal{S} contains the indexes to the support



(a)



(b)

Fig. 3. (a) On one hand, with carefully designed gating network, a proper weighting could lead to performance better than that of consistent fusion. (b) On the other hand, an improper weighting could lead to inconsistent fusion results.

vectors $s_j, y_j \in \{-1, +1\}$, b is a bias term, and $K(s, s_j)$ is a kernel function. The most common kernels are polynomial kernels and radial basis function kernels. Ben-Yacoub et al. [11] obtained the best results using the polynomial kernel.

Figure 4 illustrates the audio and visual scores and the decision boundaries created by a polynomial SVM and a weighted linear combiner ($\alpha = \beta = 0.5$ in Eq. 3) for the verification of claimants in XM2VTSDB. Evidently, the SVM is more capable in separating the client scores from the impostor scores.

Let us now take a closer look at the cross-validation accuracies in terms of FAR and FRR. Figure 2(a) shows the DET performance based on hard switching, linear fusion, and adaptive nonlinear fusion (SVM). Evidently, SVM fusion attains the best performance.

5. CONCLUSIONS

In this paper, we have proposed the notion of consistent fusion and demonstrated its applicability via audio-visual biometric authentication experiments. The consistent fusion framework leads nicely to several adaptive fusion schemes, namely hard-switching, linear combination, and adaptive nonlinear fusion using SVMs. Results have further justified that consistent fusion can benefit audio-visual biometric authentication. Moreover, we advocate the conjecture that it is not uncommon to have different performance requirement in different applications, i.e., some applications may opt for low FAR and some may opt for low FRR. Our results suggest that

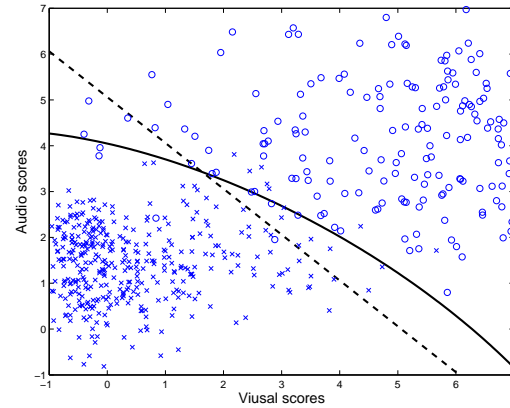


Fig. 4. Diagram illustrating the concept of Adaptive Nonlinear Fusion. Usually, it can lead to results better than the consistent fusion performance.

the notion of consistent fusion provides a valuable framework for choosing and fusing different modalities in multimodal biometric authentication. More importantly, in case there are more than two modalities, the same framework can also facilitate the selection of the best modalities to fuse (see [13]).

6. REFERENCES

- [1] K. Messer, J. Matas, J. Kittler, J. Luetttin, and G. Maitre, "XM2VTSDB: The extended M2VTS database," in *Proc. AVBPA'99*, Washington DC, 1999.
- [2] http://spib.rice.edu/spib/select_noise.html.
- [3] S. Y. Kung, M. W. Mak, and S. H. Lin, *Biometric Authentication: A Machine Learning Approach*, Prentice Hall, Upper Saddle River, New Jersey, 2005.
- [4] 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.
- [5] <http://www.identix.com>.
- [6] 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.
- [7] 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. ICASSP*, 2003, pp. 745–748.
- [8] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. Eurospeech'97*, 1997, pp. 1895–1898.
- [9] S. Y. Kung and M. W. Mak, "A machine learning approach to dna microarray biclustering analysis," in *2005 IEEE International Workshop on Machine Learning for Signal Processing*, Mystic, Connecticut, Sept. 2005.
- [10] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, 1998.
- [11] 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.
- [12] V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.
- [13] S. Y. Kung and M. W. Mak, "Machine Learning for Multi-Modality Genomic Signal Processing," *IEEE Signal Processing Magazine*, May 2006.