Audio-Visual Person Identification on the XM2VTS Database

Roland Hu and R. I. Damper

School of Electronics and Computer Science
University of Southampton
Southampton SO17 1BJ, UK
{hh03r|rid}@ecs.soton.ac.uk

Abstract

This paper presents a multimodal person identification system based on combination of audio and visual classifiers. The audio classifier was built by using mel-frequency cepstrum coefficient features and Gaussian mixture models. The visual classifier was implemented by Haar-like features and AdaBoost algorithm for face detection, and principal component analysis for identification. A new method is proposed to estimate the optimal weighting parameter based on probability density function estimation under Gaussian assumptions. Simulations indicate that the proposed method obtains slightly better results than the frequently-used empirical method of optimising on held-out training data.

Index Terms: Face recognition, speaker recognition, person identification, weighted sum rule

1. Introduction

Several theoretical and practical studies have confirmed the potential of multiple biometrics for person identification [1, 2, 3]. A particular instance is audio-visual person identification. Face and voice are very easy to obtain, and combination of these two sources of information makes the recognition system more robust to noise and complex environments.

This paper presents our research work on this promising area. First, we have built individual speaker identification and face identification systems. The text-independent speaker identification system is based on mel-frequency cepstral coefficients [4] and Gaussian mixture models [5]. The face identification system consists of modules both for face detection and identification. The face detection module is implemented using Haar-like features and the AdaBoost algorithm [6]; the face identification module is based on principal component analysis [7].

After obtaining identification results for the separate speaker and face identification systems, a late fusion algorithm (proposed in our previous work [8, 9]) was used to combine the scores of both classifiers to give a final identification result. Simulations indicate that a high identification rate of 97.95% is achieved by using the above algorithms.

2. XM2VTS Database

In this paper, we use the XM2VTS database [10] to test algorithms on audio-visual person identification. This database, intended for research into multimodal person recognition, is obtainable from the Centre for Vision, Speech and Signal Processing at the University of Surrey, UK. It contains high-quality colour images, 32 kHz 16-bit sound and video files for 295 subjects. These files are recorded in 4 sessions, over a period of 4 months. Each session consists of 6 sentences uttered by each subject. The third sentence ("Joe took father's green shoe bench out") in each session was used for this research work. Thus, we obtain 4 video files for each subject, which were recorded in 4 sessions. However, several subjects have only 3 video files because of errors in the recording procedure or unavailability of these subjects.

3. Speaker Identification

After silence is removed from the speech signal (see [8] for details), we use mel-frequency cepstral coefficients [4] as features. The magnitude spectrum from a 20 ms short-time segment of speech is pre-emphasised and processed by a mel-scale filter bank, then the log-energy filter outputs are cosine transformed to produce the cepstral coefficients. We use the first 20 coefficients excluding the zeroth coefficient, plus the first 20 delta coefficients as the feature set. This process occurs every 10 ms, producing 100 features per second.

Speaker modelling is based on the Gaussian mixture model (GMM) [5]. A Gaussian mixture density is a weighted sum of M component densities:

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \tag{1}$$

where $\vec{x}$ is a D-dimensional random vector, $b_i(\vec{x})$ is the component density of the ith mixture and p_i is the weight of the ith mixture. Each component density is a D-variate Gaussian function of the form:

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} e^{\left(-\frac{1}{2}(\vec{x}-\vec{\mu_i})^T \Sigma_i^{-1}(\vec{x}-\vec{\mu_i})\right)}$$

with mean vector $\vec{\mu_i}$ and covariance matrix Σ_i. The mixture weights satisfy the constraint that $\sum_{i=1}^{M} p_i = 1$. The complete Gaussian mixture density is parameterised by the mean vectors, covariance matrices and mixture weights from all component densities. These parameters are collectively represented as the 3-tuple:

$$\lambda = \{p_i, \vec{\mu_i}, \Sigma_i\} \quad i = 1, 2, \ldots, M$$

Gaussian mixtures consisting of 64 component densities are used in the simulations ($M = 64$). The GMM model for each speaker is trained (i.e., the parameters of λ_s are estimated) using the EM (expectation-maximisation) algorithm [11].

Suppose there are K speakers to be identified. Then λ_k, $k = 1, 2, \ldots, K$ is the model corresponding to the kth enrolled speaker. The goal of speaker identification is to find the one

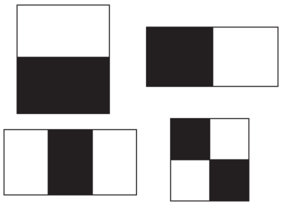


Figure 1: Four types of rectangular Haar wavelet-like features. A feature is a scalar calculated by summing up the pixels in the white region and subtracting those in the dark region.

among these K models that best matches the test data represented by a sequence of F frames, $X = \{\vec{x_1}, \vec{x_2}, \ldots, \vec{x_F}\}$. In making the decision, we use the following frame-base weighted likelihood distance measure, d_k, which refers to the distance from the test data to the kth speaker model:

$$d_k = \frac{1}{F} \sum_{f=1}^{F} \log p(\vec{x_f}|\lambda_k)$$

in which $p(\vec{x_t}|\lambda_k)$ is given in (1). The normalisation by F is necessary as each token will, in general, have a different length and, therefore, a different number of frames.

Suppose the task of a classifier is to assign an input sequence X to one of K classes $\omega_1, \omega_2, \ldots, \omega_K$. We can then identify speaker s according to the rule:

$$\text{decide } X \in \omega_s \text{ if } s = \arg \max_{i=1..K} d_i$$

4. Face Recognition

The recognition process can be divided into two steps: face *detection* localises the face in the image, and is followed by *identification* of the detected face. Benchmark detection and identification algorithms are implemented in this paper.

4.1. Face Detection

Recently, Haar-like features and the AdaBoost method for face detection have been received much attention, because so far they are the most successful ones in terms of accuracy and speed. The features used in this method are Haar basis functions as used by [12]. The method uses three kinds of features. The value of a *two-rectangle feature* is the difference between the sum of the pixels within two rectangular regions. The regions have the same size and shape and are horizontally or vertically adjacent. A *three-rectangle feature* computes the sum within two outside rectangles subtracted from the sum in a centre rectangle. Finally a *four-rectangle feature* computes the difference between diagonal pairs of rectangles (as shown in Figure 1). These features are located in a subregion of a subwindow and vary in location and size inside the subwindow.

Viola [6] has proposed the AdaBoost algorithm to select from these features which are suitable for face detection. A strong classifier is constructed based on the selected features. In this paper, the AdaBoost face detection algorithm is taken from the Intel Open Source Computer Vision Library (OpenCV), which was developed by the Intel Corporation [13]. The classifier consists of 20 stages, and for each stage, 10–15 features are selected. Figure 2 shows the first five Haar-like features in the first stage, which can be regarded as the most powerful features chosen by the algorithm to discriminate face and non-face images. It can be seen that these five features have very direct meanings for face detection.

(a) Face Image

(b) The First Five Haar-like Features

Figure 2: One face image and the first five Haar-like features overlapped on it. We can see that these five features have very direct meanings for face detection. The first feature measures the difference in intensity between the region of the eyes and a region across the upper cheeks. This feature capitalises on the observation that the eye region is often darker than the cheeks. The second feature measures the difference in intensity between the region of the right pupil and the region of the right cheek. Similar analysis could be applied on the third, fourth and fifth features.

Simulations indicate that the AdaBoost face detection algorithm can achieve around 98% detection rate on images of the XM2VTS database. However, because video files consist of many frames, it can not be guaranteed that the detection system will succeed for every frame. A face tracking algorithm based on dynamic programming (DP) is implemented to solve the problem of detecting faces in video files. The face tracking method is very similar to [14]. First, an energy function is defined for face image candidates in each frame. Then DP is used to obtain the optimal path which minimises the energy function. This face tracking algorithm is applied to all the video files and 100% detection rate is achieved based on the criteria that (1) the detected face contains the elbows, eyes, and mouth, and (2) no false detection occurs.

4.2. Face Identification

The face identification system is constructed by using the principal component analysis (PCA) method. PCA [15] is a dimensionality-reduction technique based on extracting the desired number of principal components of high-dimensional data. Implementing PCA for face analysis and representation starts from the ground-breaking work of [16]. Their paper was followed by [7], the first application of PCA to face recognition.

Suppose a face image $I(x, y)$ is a two-dimensional $N \times N$ array of intensity values, or a vector of dimension N^2. For example, a 128×128 image describes a vector of dimension 16384. Let the training set of face images be $\Gamma_1, \Gamma_2, \ldots, \Gamma_M$. The average face of the set is defined by the size-N^2 vector:

$$\Psi = \frac{1}{M} \sum_{i=1}^{M} \Gamma_i$$

so that each face differs from the average by the vector $\Phi_i = \Gamma_i - \Psi$. The eigenvectors are obtained by solving the eigenvalue problem:

$$\Lambda = U^T C U$$

where Λ is a diagonal matrix, with eigenvalues on the main diagonal; U is an orthogonal matrix, which means that $U^{-1} = U^T$, or $U^T U = I$; and C is the covariance matrix of the data:

$$C = \sum_{i=1}^{M} \Phi_i \Phi_i^T = AA^T$$

where the matrix $A = [\Phi_1\ \Phi_2\ \dots\ \Phi_M]$, is a N^2 by M matrix. The matrix C, however, is N^2 by N^2, and determining the N^2 eigenvectors and eigenvalues is an intractable task for typical image sizes. Fortunately, it is possible to solve this problem by first solving a much smaller M by M matrix problem, and taking linear combinations of the resulting vectors ([17, Section 6.7]). Because the obtained eigenvectors have the same dimension as the face image, they are also called 'eigenfaces'.

Once the eigenfaces are created, identification becomes a pattern recognition task. The M' significant eigenfaces of the matrix AA^T are chosen as those with the largest associated eigenvalues. The eigenfaces span an M'-dimensional subspace of the original N^2 image space. A test face image Γ is transformed into its eigenface components by a simple operation, $p_k = u_k^T(\Gamma - \Psi)$, for $k = 1, 2, \dots, M'$ (u_k is the kth significant eigenvectors of AA^T). This describes a set of point-by-point image multiplications and summations. These weights form a vector $P^T = [p_1, p_2, \dots, p_{M'}]$ that describes the contribution of each eigenface in representing the input face image.

Suppose $P_1, P_2, \dots, P_M$ are weight vectors which are generated by the M training face images $\Gamma_1, \Gamma_2, \dots, \Gamma_M$. For a testing video file which contains N face images, we assume these face images are $\Lambda_1, \Lambda_2, \dots, \Lambda_N$, and their corresponding weighting vectors are $Q_1, Q_2, \dots, Q_N$. To identify a face, we find that face image in the training set that is most similar to the test image according to:

$$\text{find } u, v = \arg \max_{\substack{i=1..M \\ j=1..N}} \left(\frac{|P_i \cdot Q_j|}{\|P_i\| \, \|Q_j\|} \right)$$

Then we classify the test video file as belonging to the class which the training image Γ_u belongs to.

5. Combining Audio and Visual Classifiers

After obtaining identification scores of the audio and visual classifiers, the next step is to combine these with a view to obtaining better identification results. Some well-known simple fixed rules for combining the set of base classifiers, such as product rule, sum rule, maximum rule, minimum rule and median rule, are described in [2, 18]. However, fixed rules are sub-optimal [18] and there exist rules which need a training set to adjust some parameters so as to obtain better identification. One well-known example is the weighted sum rule, which we use here. Note that the method is previously published in [8, 9].

Suppose each of the audio and video classifiers consists of K discriminant functions, $f^1(X)$, $f^2(X)$,...,$f^K(X)$. The decision rule in terms of discriminant functions is:

$$\text{decide } X \in \omega_s \text{ if } s = \arg \max_{i=1..K} f^i(X) \tag{2}$$

Here, we denote by $f_1^1(X), f_1^2(X), \dots, f_1^K(X)$ the scores obtained from the video classifier (face identification), and by $f_2^1(X), f_2^2(X), \dots, f_2^K(X)$ the scores obtained from the audio classifier (speaker identification). The weighted-sum rule is described as:

$$f_{\text{comb}}^k(X, \alpha) = \alpha f_1^k(X) + (1-\alpha) f_2^k(X),\ k = 1, 2, \dots, K \tag{3}$$

The parameter α should be selected according to the relative reliability of the M classifiers. One way to do this is to optimise α so as to maximise the identification rate on some training data [19, 18], but this carries the danger of over-fitting, so reducing the ability to generalise to unseen data. In this section, we propose a new method for accurately choosing the weighting parameter that directly minimises the estimated correct identification rate. We simplify the notation for discriminant functions by dropping arguments X and α, *except* when it is necessary to distinguish among different values of these arguments.

The first step of our method is to normalise the scores of the training data. We use so-called z-score normalisation, which is calculated using the arithmetic mean and standard deviation of the given data. Refer to [20] for an overview of score normalisation techniques in multimodal biometric systems.

The normalisation process can be divided into two steps. In the first step, all scores of both audio and video classifiers have their mean subtracted and the result is then divided by their variance:

$$\begin{aligned} \overline{f_m^k(X_i)} &= \frac{f_m^k(X_i) - \mu_m}{\sigma_m} \\ \text{where} \quad \mu_m &= \frac{\sum_{i=1}^{I} \sum_{k=1}^{K} f_m^k(X_i)}{I \times K} \\ \text{and} \quad \sigma_m &= \frac{\sum_{i=1}^{I} \sum_{k=1}^{K} \left(f_m^k(X_i) - \mu_m \right)^2}{I \times K} \end{aligned} \tag{4}$$

Here, I is the number of training data points, K is the number of classes, and $m \in \{1, 2\}$.

The second step of normalisation is to make the correct score (i.e., that for the correct person) zero. This gives us a known reference point from which to assess scores, and simplifies the derivation of an appropriate mathematical model under Gaussian assumptions—see below. If we set the weighting parameter α to a constant value, we can obtain the combined scores $\overline{f_{\text{comb}}^1}, \overline{f_{\text{comb}}^2}, \dots, \overline{f_{\text{comb}}^K}$ by equations (3) and (4). The second step of the normalisation process is:

$$\text{if } X \in w_i \text{ then } F_{\text{comb}}^k = \overline{f_{\text{comb}}^k} - \overline{f_{\text{comb}}^i},\ k = 1, 2, \dots, K \tag{5}$$

Equation (5) is used to make the correct score zero. We can see from the decision rule, equation (2), that these two steps of normalisation do not change the identification result because the new scores in (5) are obtained only by subtracting and dividing the same number from the original scores, which does not influence the rank of the scores.

After normalisation, the next step is to estimate the probability distribution of the scores. We assume that the values of the score functions are independent. That is:

$$P\left(F^1_{\text{comb}}, \ldots, F^{i-1}_{\text{comb}}, F^{i+1}_{\text{comb}}, \ldots, F^K_{\text{comb}} | X \in w_i\right)$$
$$= \prod_{k=1, k\neq i}^{K} P\left(F^k_{\text{comb}} | X \in w_i\right)$$

The reason why $k \neq i$ is that, after the normalisation, F^i_{comb} always equals zero if $X \in w_i$. We denote the correct identification rate (the probability of correct identification) when $X \in w_i$ as $C_i(\alpha)$. Since $F^i_{\text{comb}} \equiv 0$ when $X \in w_i$ after the normalisation process, we can calculate $C_i(\alpha)$ on the basis of equation (2) as:

$$C_i(\alpha) = \prod_{k=1, k\neq i}^{K} P\left(F^k_{\text{comb}} < 0 | X \in w_i\right) \tag{6}$$

5.1. Probability Density Estimation

To calculate the probability $P\left(F^k_{\text{comb}} < 0 | X \in w_i\right)$ for each $k = 1, 2, \ldots, K$, we first have to estimate the probability distribution $P\left(F^k_{\text{comb}} | X \in w_i\right)$ from the training data in the form of a Gaussian mixture model. But a problem of sparse data arises when we try to model the distribution this way. In essence, it is hard to estimate the density of a multimodal data distribution reliably.

Our approach to this problem is to break the available training data up into 'sections', and to treat each of these as a unimodal Gaussian, and then to combine them. Suppose there are M training data available for deciding the weighting parameter α. Among these M files, there are M_1 files belonging to class ω_1, M_2 files belonging to class ω_2, $\ldots$, and finally M_K files belonging to class ω_K ($M_1 + M_2 + \ldots + M_K = M$).

We denote the M_i training data belonging to class ω_i as $X_1, X_2, \ldots, X_{M_i}$. The Gaussian mixture is then:

$$P\left(F^k_{\text{comb}} | X \in w_i\right)$$
$$= \frac{1}{M_i} \sum_{j=1}^{M_i} \frac{1}{\sqrt{2\pi} A} e^{\left(-\frac{(F^k_{\text{comb}} - \mu_{kj})^2}{2A^2}\right)} \tag{7}$$

where A is a parameter controlling the variance(s).

The component means μ_{kj} are obtained as $\mu_{kj} = F^k_{\text{comb}}(X_j, \alpha)$, $j = 1, 2, \ldots, M_i$. From this, we see that the means of the mixture components are the scores of the training data. When A is large, the variance of each mixture component is large; when it is small, the variance is small. In the extreme case when A becomes zero, the probability density shrinks to a series of impulse functions.

5.2. Estimating Correct Identification Rate

Using the estimated probability density function (7), we now calculate the probability that $F^k_{\text{comb}}(X)$ is less than zero as:

$$P\left(F^k_{\text{comb}} < 0 \mid X \in w_i\right)$$
$$= \frac{1}{M_i} \sum_{j=1}^{M_i} \frac{1}{\sqrt{2\pi} A} \int_{-\infty}^{0} e^{\left(-\frac{\left(F^k_{\text{comb}} - \mu_{kj}\right)^2}{2A^2}\right)} \mathrm{d}\left(F^k_{\text{comb}}\right)$$

Training Session	Test Session	Identification Rate (%)
1,2,3	4	94.92
1,2,4	3	97.25
1,3,4	2	96.26
2,3,4	1	89.83

Table 1: Simulation results for speaker identification in video files.

Training Session	Test Session	Identification Rate (%)
1,2,3	4	62.71
1,2,4	3	63.23
1,3,4	2	67.35
2,3,4	1	71.19

Table 2: Simulation results for face identification in video files.

$$= \frac{1}{M_i} \sum_{j=1}^{M_i} \Phi\left(-\frac{\mu_{kj}}{A}\right)$$

where $\Phi(x)$ is the integral of the Gaussian distribution:

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{\left(-\frac{x^2}{2}\right)} \mathrm{d}x$$

From equation (6), we can finally obtain $C_i(\alpha)$, which is the correct identification rate for a specified α when $X \in w_i$, as:

$$C_i(\alpha) = \frac{1}{M_i^{K-1}} \prod_{k=1, k\neq i}^{K} \left(\sum_{j=1}^{M_i} \Phi\left(-\frac{\mu_{kj}}{A}\right)\right)$$

The overall correct identification rate, denoted $C_{\text{prop}}(\alpha)$, is given as:

$$C_{\text{prop}}(\alpha) = \sum_{i=1}^{K} C_i(\alpha) P(X \in w_i)$$

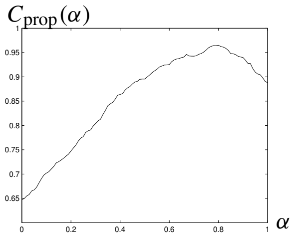
where $P(X \in w_i)$ can be estimated as $\frac{M_i}{M}$ with M_i equal to the number of training data points that belong to class ω_i, and M equal to the total number of training data points. Thus, we have transformed the problem of choosing weighting parameter α for combining two classifiers to a problem of maximising the correct identification rate $C_{\text{prop}}(\alpha)$:

$$\text{decide } \alpha = \alpha_{\text{opt}} \text{ if } \alpha = \arg\max_{\alpha} C_{\text{prop}}(\alpha)$$

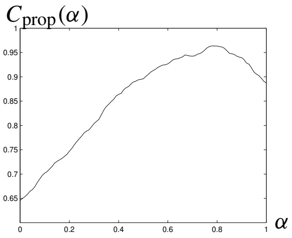
6. Simulation Results

First, we divide the video files into training and testing files. A cross validation test is used on these video files. The audio and visual classifiers are trained by video files from three sessions, and then tested on data from the remaining session. The process iterates for these four sessions. The audio and visual identification results are summarised in Tables 1 and 2, respectively.

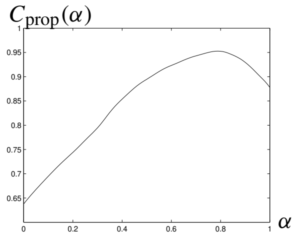
We use the combination method proposed in Section 5 to combine the scores of audio and visual classifiers. To obtain the optimal weighting parameter α, we use the scores



(a) $A = 0.001$



(b) $A = 0.01$



(c) $A = 0.1$

Figure 3: Estimated correct identification rate $C_{\text{prop}}(\alpha)$ using the proposed method as α varies from 0 to 1 using data from sessions 1 and 2: (a) $A = 0.001$; (b) $A = 0.01$: (c) $A = 0.1$.

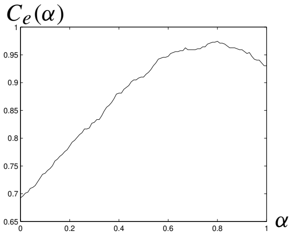


Figure 4: Empirical correct identification rate using the training data in sessions 1 and 2, with α varying from 0 to 1.

in sessions 1 and 2 for training, and scores in sessions 3 and 4 for testing. We can obtain the estimated identification curves by using the method discussed in Section 5. Figure 3 shows the estimated identification rate when A equals 0.001, 0.01 and 0.1. The corresponding optimal α values are 0.80, 0.78 and 0.79, respectively.

We can, of course, also plot the *empirical identification rate* curve $C_e(\alpha)$ as α increases from 0 to 1 by first determining the individual scores of the audio and visual classifiers, then calculating the combination scores using equation (3), and finally using these for identification. This is shown in Figure 4. In the empirical identification curve, the identification rate takes its maximum (97.45%) when α equals 0.80.

We obtain three candidates for α_{opt}, which are 0.78, 0.79 and 0.80. We need to see which candidate performs best on the testing data. Figure 5 shows the curve of empirical identification rate of sessions 3 and 4. Table 3 indicates the different identification rates of sessions 3 and 4 when α equals 0.78, 0.79 and 0.80.

These three candidates for α achieve similar identification rate, with $\alpha = 0.78$ slightly better than the other two. Considering this candidate was generated by the proposed method of estimating correct identification rate when $A = 0.01$, we can say that this method may be slightly better than the empirical

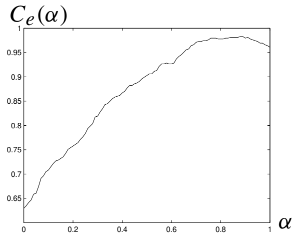


Figure 5: Empirical correct identification rate using the test data from sessions 3 and 4, with α varying from 0 to 1.

α	Identification Rate (%)
0.78	97.95
0.79	97.78
0.80	97.78

Table 3: Identification rates of sessions 3 and 4 when α equals 0.78, 0.79 and 0.80.

method. However, the optimal α for sessions 3 and 4 is 0.88, where the identification rate achieves its maximum of 98.29%. Both the proposed method and the empirical method fail to find the optimal α for the testing files. We have discussed elsewhere [9] that the proposed method is better than the empirical method in average performance, but not in all cases.

7. Comparison with Related Publications

Another paper by Fox et al. [21], published in 2003, also discusses the audio-visual person identification problem using the XM2VTS database. They built a text-*dependent* speaker identification classifier based on hidden Markov models. For face identification, they used a commercial software package FaceIt. Both classifiers are tested on 291 subjects of the XM2VTS database, contrasting to 295 subjects in our simulations. They use sessions 1, 2 and 3 for training, and session 4 for testing. Their speaker identification system achieves 98.01% identification rate, which is some 3 percentage points higher than our 94.92% (first line of Table 1) but higher rates are only to be expected for the easier text-dependent task. Because they used commercial face recognition software contrasting to our preliminary algorithms, their face identification rate is 93.23%, which is much higher than our 62.71% result. By combination of these two classifiers, they achieved 100% identification rate.

Fox et al. built another classifier for lip motion, and combined it with the audio and visual classifiers. They also devised a fusion method to combine scores of different classifiers when the audio signal was contaminated with noise. They tested the three-classifier system in different signal-to-noise ratio (SNR), and obtained good results even when immense noise is added to the audio signal. For example, by using the three classifiers, they can achieve 96.81% identification rate when SNR equals 0.

8. Conclusions

In this paper, we have built an audio-visual person identification system, and tested it on the whole XM2VTS database. A high identification rate of almost 98% is achieved on 295 subjects, which indicates the potentiality of combining biometrics for person identification.

A new method is proposed to estimate the optimal weighting parameter for combining scores of the audio and visual classifiers. This method is tested on the database and obtained slightly better results than the empirical method.

Our future work lies in two parts. First, more research needs to be carried out to improve the performance of the face identification system. Second, visual features, especially the lip movement information, need to be combined with the audio and visual classifiers.

9. References

[1] L. Xu, A. Krzyzak, and C. Y. Suen, "Methods of combining multiple classifiers and their applications to handwriting recognition," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 22, no. 3, pp. 418–435, 1992.

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

[3] K. A. Toh and W. Y. Yau, "Combination of hyperbolic functions for multimodal biometrics data fusion," *IEEE Transactions on System, Man, and Cybernetics – Part B: Cybernetics*, vol. 34, no. 2, pp. 1196–1209, 2004.

[4] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[5] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[6] P. Viola, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[7] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *Proceedings of the International Conference on Pattern Recognition, ICPR'91*, Atlantic City, NJ, 1991, pp. 586–591.

[8] R. Hu and R. I. Damper, "Fusion of two classifiers for speaker identification: Removing and not removing silence," in *Proceedings of Eighth International Conference on Information Fusion*, Philadelphia, PA, 2005, pp. 1250–1253.

[9] ——, "Optimal weighting of biomodal biometric information with specific application to audio-visual person identification," submitted to *Information Fusion*.

[10] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The extended M2VTS database," in *Proceedings of 2nd International Conference on Audio and Video-based Biometric Person Authentication, AVBPA'99*, Washington, DC, 1999, pp. 72–77.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[12] C. Papageorgiou, M. Oren, and T. Poggio, "A general framework for object detection," in *International Conference on Computer Vision, ICCV98*, Bombay, India, 1998, pp. 555–562.

[13] G. R. Bradski and V. Pisarevsky, "Intel's computer vision library: Applications in calibration, stereo, segmentation, tracking, gesture, face and object recognition," in *IEEE International Conference on Computer Vision and Pattern Recognition, CVPR'00*, Hilton Head Island, SC, 2000, pp. 796–797.

[14] P. Lappas, J. N. Carter, and R. I. Damper, "Robust evidence-based object tracking," *Pattern Recognition Letters*, vol. 23, no. 1–2, pp. 253–260, 2002.

[15] I. T. Jolliffe, *Principal Component Analysis*. New York, NY: Springer-Verlag, 1986.

[16] M. Kirby and L. Sirovich, "Application of the Karhunen-Loève procedure for the characterization of human faces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 1, pp. 103–108, 1990.

[17] G. Strang, *Introduction to Linear Algebra*. Wellesley-Cambridge Press, 1998.

[18] R. P. W. Duin, "The combining classifier: To train or not to train?" in *Proceedings of the International Conference on Pattern Recognition, ICPR'02*, vol. 2, Quebec, Canada, 2002, pp. 765–770.

[19] B. Maison, C. Neti, and A. Senior, "Audio-visual speaker recognition for video broadcast news: Some fusion techniques," in *Proceedings of the IEEE Conference on Multimedia Signal Processing*, Copenhagen, Denmark, 1999, pp. 161–167.

[20] A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern Recognition*, vol. 38, no. 12, pp. 2270–2285, 2005.

[21] N. A. Fox, R. Gross, P. de Chazal, J. F. Cohn, and R. B. Reilly, "Person identification using automatic integration of speech, lip, and face experts," in *Proceedings of the ACM SIGMM Multimedia Biometrics Methods and Applications Workshop*, Berkeley, CA, 2003, pp. 25–32.