Audio-Visual Person Verification*

S. Ben-Yacoub, J. Lüttin

IDIAP
CP 592, 1920 Martigny
Switzerland
{sby,luettin}@idiap.ch

K. Jonsson, J. Matas and J. Kittler

University of Surrey
Guildfor Surrey GU2 5XH
United Kingdom
{ee1sjk,ee1kj,ees2gm}@ee.surrey.ac.uk

Abstract

In this paper we investigate benefits of classifier combination (fusion) for a multimodal system for personal identity verification. The system uses frontal face images and speech. We show that a sophisticated fusion strategy enables the system to outperform its facial and vocal modules when taken seperately. We show that both trained linear weighted schemes and fusion by Support Vector Machine classifier leads to a significant reduction of total error rates. The complete system is tested on data from a publicly available audio-visual database (XM2VTS, 295 subjects) according to a published protocol.

1 Introduction

Recognition systems based on biometric features (face, voice, iris, etc ...) have received a lot of attention in recent years Most of the proposed approaches focus on mono-modal **identification**. The system uses a single modality to find the closest person to the user in a database. Relatively high recognition rates were obtained for different modalities like face recognition and speaker recognition [21, 8]. Verification of person identity based on biometric informations is important for many security applications. Examples include access control to buildings, surveillance and intrusion detection. In person identity verification, the user claims a certain client identity and the system decides to accept or reject the claim. Only very low error rates can be tolerated in many of the above mentioned applications. It has been shown that combining different modalities leads to more robust systems with better performance [5].

One of the remaining questions is what strategy should be adopted for combining different modalities. In order to assess the performance of a method and compare it to other approaches, a large database and an evaluation protocol are necessary. Most of the work done in multi-modal verification [7, 12, 14, 6] was tested and evaluated on small databases (less than 40 persons) or medium-sized (less than 100 persons in [5]).

We describe and evaluate in this paper a **complete multi-modal user verification system** based on facial and vocal modalities. Each module of the system (face, voice, fusion) is tested and evaluated on a large database (XM2VTS database[1] with 295 people) according to a published protocol[2].

The rest of the paper is organised as follows: face and speech verification modules are described in Section 2 and 3. The multi-modal data fusion issue is presented in Section 4. The XM2VTS database and its evaluation protocol are described in Section 5. The results and different experiments are presented in section 6.

2 Face Verification

The face verification method used is based on robust correlation [11]. Registration is achieved by direct minimization of the robustified correlation score over a multi-dimensional search space. The search space is defined by the set of all valid geometric and photometric transformations. In the current implementation method the geometric transformations are translation, scaling and rotation. Given a weak affine transformation $T_{\vec{a}}$

$$T_{\vec{a}}(x, y) = (a_1 x - a_2 y + a_3, a_2 x + a_1 y + a_4) \quad (1)$$

the error function expressing the intensity difference between a pixel s in the model image I_m and its projection in the probe image I_p is defined as

$$\epsilon(s, \vec{a}) = I_m(s) - I_p(T_{\vec{a}}(s)) \quad (2)$$

*This work was supported by the ACTS-M2VTS project and the Swiss Federal Office for Education and Science.

[1] From ACTS-M2VTS project, available at http://www.ee.surrey.ac.uk/Research/VSSP/xm2vts

[2] Available with the XM2VTS database

0-7695-0149-4/99 $10.00 © 1999 IEEE

The score function used to evaluate a match between the transformed model image and the probe image is

$$S(\mathcal{R}, \vec{a}) = \frac{1}{|\mathcal{R}| \cdot \rho_{\max}} \sum_{s \in \mathcal{R}} \rho(\epsilon(s, \vec{a})) \qquad (3)$$

where ρ denotes a robust kernel. The function is the average percentage of the maximum kernel response taken over some set of pixels R. Possible kernel functions are the Huber Minimax and the Hampel (1,1,2) [9]. Experiments reported in [4] showed that the choice of kernel is not critical.

In Equation (3), parameters of the score function are purely geometrical and intensity values are not transformed. In our previous work [17], we included parameters for affine compensation of global illumination changes (gain, offset) into the search space. For efficiency reasons, we decided to adopt a less sophisticated approach in which we shift (for each point in the search space) the histogram of residual errors using the median error.

To find the global extremum of the score function we employ a stochastic search technique incorporating gradient information. The gradient-based search is implemented using steepest descent on a discrete grid. Resolution of the grid is changed during the optimization (multi-resolution in the parameter domain) following a predefined schedule. The different components of the gradient (the partial derivatives with respect to the affine coefficients) are

$$\begin{aligned}
\frac{\partial S(\mathcal{R},\vec{a})}{\partial a_1} &= -\textstyle\sum_{s\in\mathcal{R}} \Psi(\epsilon) \left(\frac{\partial I_p(s')}{\partial x} x + \frac{\partial I_p(s')}{\partial y} y \right) \\
\frac{\partial S(\mathcal{R},\vec{a})}{\partial a_2} &= -\textstyle\sum_{s\in\mathcal{R}} \Psi(\epsilon) \left(-\frac{\partial I_p(s')}{\partial x} y + \frac{\partial I_p(s')}{\partial y} x \right) \\
\frac{\partial S(\mathcal{R},\vec{a})}{\partial a_3} &= -\textstyle\sum_{s\in\mathcal{R}} \Psi(\epsilon) \left(\frac{\partial I_p(s')}{\partial x} \right) \\
\frac{\partial S(\mathcal{R},\vec{a})}{\partial a_4} &= -\textstyle\sum_{s\in\mathcal{R}} \Psi(\epsilon) \left(\frac{\partial I_p(s')}{\partial y} \right)
\end{aligned}$$

where Ψ denotes the influence function of the robust kernel (obtained by differentiating the kernel) and $s' = T_{\vec{a}}(s)$. To escape from local maxima, stochastic search is performed by adding a random vector drawn from an exponential distribution (this optimization technique is effectively a special case of simulated annealing [13]).

To meet real-time requirements of the verification scenario, we adopt a multi-resolution scheme in the spatial domain. This is achieved by applying the combined gradient-based and stochastic optimization described above to each level of a Gaussian pyramid. The estimate obtained on one level is used to initialize the search at the next level. In addition to the speed-up, the multi-resolution search also has the benefit of removing local optima from the search space and thus effectively improving the convergence characteristics of the method.

In the training phase we employ a feature selection procedure based on minimizing the intra-class variance and at the same time maximizing the inter-class variance. A feature criterion is evaluated for each pixel and the subset of pixels that best discriminates a given client from other clients in the database (effectively modeling the impostor distribution) are selected. This feature subset is then used in verification allowing efficient identification of the probe image.

The presented system runs in real-time on a high-end PC.

3 Speaker Verification

Speaker verification methods can be classified into text-independent and text-dependent methods. The latter usually requires that the utterances used for verification are the same as for training. These methods can exploit text-dependent voice individuality and therefore often outperform text-independent methods. We propose two different algorithms: a text-independent method based on the sphericity measure [3] and a text-dependent technique using hidden Markov models (HMM) [19].

3.1 Text-independent Speaker Verification

The first processing step aims to remove silent parts from the raw audio signal as these parts do not convey speaker dependent information. We use the speech activity detector proposed by Reynolds *et al.* [18] on the 16 kHz sub-sampled audio signal.

The cleaned audio signal is converted to linear prediction cepstral coefficients (LPCC) [1] using the autocorrelation method. We use a pre-emphasis factor of 0.94, a Hamming window of length 25 ms, a frame interval of 10 ms, and an analysis order of 12. We have applied cepstral mean subtraction (CMS), where the mean cepstral parameter is estimated across each speech file and subtracted from each frame. The energy is normalized by mapping it to the interval [0, 1] using the tangent hyperbolic function. The normalized energy is included in the feature vector, leading to 13-dimensional vectors. A client model is represented by the covariance matrix $\mathbf{X}$, computed over the feature vectors of the client's training data. Similarly, an accessing person is represented by the covariance matrix $\mathbf{Y}$, computed over that person's speech data. We use the arithmetic-harmonic sphericity measure $D_{SPH}(\mathbf{X}, \mathbf{Y})$ [3] as similarity measure between the client and the accessing person:

$$D_{SPH}(\mathbf{X},\mathbf{Y}) = \log\left[\frac{tr(\mathbf{Y}\mathbf{X}^{-1})tr(\mathbf{X}\mathbf{Y}^{-1})}{m^2}\right], \quad (4)$$

where m denotes the dimension of the feature vector and $tr(\mathbf{x})$ the trace of $\mathbf{x}$. The similarity values were mapped to the interval $[0,1]$ with a sigmoid function $f(D_{SPH}) = (1 + exp(-(D_{SPH} - t)))^{-1}$ where $f(t) = 0.5$. A claimed speaker is rejected if $S_{SPH} < 0.5$, otherwise she/he is accepted. We have used person-dependent thresholds t which were estimated on the evaluation set. The processing time, on an Sun UltraSparc 30, required by the speech verification module is $\frac{1}{20}$ the time of the utterance duration.

3.2 Text-dependent Speaker Verification

Hidden Markov models (HMMs) represent a very efficient approach to model the statistical variations of speech in both the spectral domain and in the temporal domain. Our HMM-based verification technique makes use of 3 HMM sets: client models, world models, and silence models. Utterances of a client are represented by client HMMs. The world models serve as speaker-independent models to represent speech of an average person. They are trained on the POLYCOST[3] database, which represents a distinct set of speakers, that neither includes clients nor impostors of the XM2VTS database. Finally, three silence HMMs are used to model the silent parts of the signal.

The same feature extraction as in the previous section is performed. In addition, the first and second order temporal derivatives were included, leading to 42-dimensional feature vectors. All models were trained based on the maximum likelihood criterion using the Baum-Welch (EM) algorithm. The world models were trained on the segmented words of the POLYCOST database, where one HMM per word was trained.

For both training and verification the sentences of the XM2VTSDB are first segmented into words and silence using the world and silence models. This consists in computing the best path between the sentence and the sequence of known HMMs using the Viterbi algorithm. To do this we used an HMM network that allowed optional silence at the beginning of a sentence, between words, and at the end of a sentence. The client models could then be trained on the segmented training words. For verification, the Viterbi algorithm is used to calculate the likelihood $p(X_j|\mathcal{M}_{ij})$, where X_j represents the observation of the segmented word j; $\mathcal{M}_{ij}$ represents the model of subject M_i and word j. We normalize the log-likelihood of word j by the numbers of frames N_j and sum them over all words W, which leads to the following measure:

$$\log p(X|M_i) = \frac{1}{W}\sum_{j=1}^{W}\frac{\log p(X_j|\mathcal{M}_{ij})}{N_j} \quad (5)$$

This measure is calculated for the models $\mathcal{M}_c$ of a given client M_c and for the world models $\mathcal{M}_w$. The following similarity:

$$D_{HMM} = \log p(X|\mathcal{M}_c) - \log p(X|\mathcal{M}_w) \quad (6)$$

is computed and compared to a threshold t. The claiming subject is rejected if $D_{HMM} < t$, otherwise she/he is accepted. The quantities D_{HMM} were mapped to the interval $[0,1]$ as described in Section 3.1. The processing time is half the time of the utterance duration.

4 Multi-Modal Data Fusion

Combining different experts results in a system which can outperform the experts when taken individually [15, 10]. This is especially true if the different experts are not correlated. We expect from the fusion of vision and speech to achieve better results. In the next section, we compare the Support Vector Machine (SVM) with tradition fusion methods to combine different modalities. The use of SVM is motivated by the fact that verification is basically a binary classification problem (i.e. accept or reject user) [2].

4.1 SVM

The Support Vector Machine is based on the principle of *Structural Risk Minimization* [20]. Classical learning approaches are designed to minimize the empirical risk (i.e error on a training set) and therefore follow the *Empirical Risk Minimization* principle. The SRM principle states that better generalization capabilities are achieved through a minimization of the bound on the generalization error.

We assume that we have a data set $\mathcal{D}$ of M points in a n dimensional space belonging to two different classes **+1** and **-1**:

$$\mathcal{D} = \{(X_i, y_i)|i \in \{1..M\}, X_i \in \mathbb{R}^n, y_i \in \{+1,-1\}\}$$

A binary classifier should find a function f that maps the points from their data space to their label space.

It has been shown [20] that the optimal separating hyperplane is expressed as:

$$f(x) \;=\; sign(\sum_i \alpha_i y_i K(X_i, x) + b) \quad (7)$$

where K(x,y) is a positive definite symmetric function, b is a bias estimated on the training set, α_i are

[3] For more informations see http://circwww.epfl.ch/polycost

the solutions of the following Quadratic Programming (QP) problem:

$$\begin{cases} \min_{\mathcal{A}} W(\mathcal{A}) = -\mathcal{A}^t I + \frac{1}{2}\mathcal{A}^t D \mathcal{A} \\ \text{with the constraints:} \\ \sum_i \alpha_i y_i = 0 \text{ and } \alpha_i \geq 0 \\ \\ \text{where:} \\ (i,j) \in [1..M] \times [1..M] \\ (\mathcal{A})_i = \alpha_i \\ (I)_i = 1 \\ (D)_{ij} = y_i y_j K(X_i, X_j) \end{cases}$$

The kernel functions $K(x,y)$ define the nature of the decision surface that will separate the data. They satisfy some constraints in order to be applicable (Mercer's conditions, see [20]). Some possible kernel functions have been already identified (we assume (x,y) $\in \mathbb{R}^n \times \mathbb{R}^n$) :

- $K(x,y) = (x^t y + 1)^d$ with $d \in \mathbb{N}$, this defines a polynomial decision surface of degree d.
- $K(x,y) = e^{-g\|x-y\|^2}$ is equivalent to one RBF classifier.

The computational complexity of the SVM during the training depends on the number of data points rather than on their dimensionality. The number of computation steps is $O(n^3)$ where n is the number of data points. At run time the classification step of SVM is a simple weighted sum. The classification of 112400 claims requires 5.6sec on an Ultra-Sparc 30.

5 The XM2VTS database

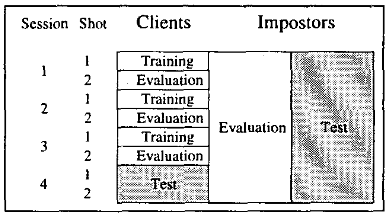
The XM2VTSDB database contains synchronized image and speech data as well as sequences with views of rotating heads. The database includes four recordings of 295 subjects taken at one month intervals. On each visit (session) two recordings were made: a speech shot and head rotation shot. The speech shot consisted of frontal face recording of each subject during the dialogue.

The database was acquired using a Sony VX1000E digital cam-corder and DHR1000UX digital VCR. Video is captured at a color sampling resolution of 4:2:0 and 16bit audio at a frequency of 32kHz. The video data is compressed at a fixed ratio of 5:1 in the proprietary DV format. In total the database contains approximately 4 TBytes (4000 Gbytes) of data.

When capturing the database the camera settings were kept constant across all four sessions. The head was illuminated from both left and right sides with diffusion gel sheets being used to keep this illumination as uniform as possible. A blue background was used to allow the head to be easily segmented out using a technique such as chromakey. A high-quality clip-on microphone was used to record the speech. The speech sequence consisted in uttered digits from 0 to 9.

5.1 Evaluation Protocol

The database was divided into three sets: training set, evaluation set, and test set (see Fig. 1). The training set is used to build client models. The evaluation set is selected to produce client and impostor access scores which are used to estimate parameters (i.e. thresholds). The estimated threshold is then used on the test set. The test set is selected to simulate real authentication tests. The three sets can also be classified with respect to subject identities into client set, impostor evaluation set, and impostor test set. For this description, each subject appears only in one set. This ensures realistic evaluation of imposter claims whose identity is unknown to the system. The protocol is based on 295 subjects, 4 recording sessions, and two shots (repetitions) per recording sessions. The database was randomly divided into 200 clients, 25 evaluation impostors, and 70 test impostors (See [16] for the subjects' IDs of the three groups).



Session	Shot	Clients	Impostors	
1	1	Training	Evaluation	Test
	2	Evaluation		
2	1	Training		
	2	Evaluation		
3	1	Training		
	2	Evaluation		
4	1	Test		
	2			

Figure 1: Diagram showing the partitioning of the XM2VTSDB according to protocol Configuration I.

5.2 Performance Measures

Two error measures of a verification system are the *False Acceptance rate* (FA) and the *False Rejection rate* (FR). False acceptance is the case where an impostor, claiming the identity of a client, is accepted. False rejection is the case where a client, claiming his true identity, is rejected. FA and FR are given by $FA = EI/I * 100\%$ and $FR = EC/C * 100\%$, where EI is the number of impostor acceptances, I the number of impostor claims, EC the number of client rejections, and C the number of client claims. A trade-off between FA and FR can be controled by a threshold. For the protocol configurations, I is $112,000$ (70 impostors × 8 shots × 200 clients) and C is 400 (200 clients × 2 shots).

6 Experiments and Results

The video and audio stream of each user are processed by the different verification modules. Three different modalities are considered: Face verification (Section 2), Sphericity-based speaker verification (Section 3.1) and HMM-based speaker verification (Section 3.2). Data generated by the verification modules and processed by the fusion algorithms are publicly available on the XM2VTS ftp server [4]. This will enable people from the community to compare their methods and results.

The performance of each modality is displayed in Table 1.

Modality	FA (%)	FR (%)
Face	7.76	7.25
Voice (Sphericity)	1.6	5.00
Voice (HMM)	0.00	1.48

Table 1: Performance of Modalities on Test Set

We performed a series of experiments to evaluate different configuration sets of modalities. The sets are defined as follows:

- C1: Face and HMM.
- C2: Face, Sphericity and HMM.
- C3: HMM and Sphericity.
- C4: Face and Sphericity.

For the SVM-based fusion, we used polynomial and gaussian kernels in our experiments. The training set was used as an evaluation set to see how performance changes with different kernel parameters. The main conclusion is that the performance does not change significantly with different polynomial. The conclusion is also valid for the gaussian kernel. We chose to run the experiments with the following configurations:

- Linear: $K(x, y) = x^t y$
- Polynomial: $K(x, y) = (x^t y + 1)^3$
- Gaussian: $K(x, y) = exp(-4||x - y||^2)$

The dimensionality of the data corresponds to the number of modalities to combine. Moreover, SVM computes only dot products with the data and therefore the complexity of SVM is independent from the number of modalities to combine. As a baseline fusion experiment we combined the output of the HMM, Sphericity and face expert using simple combination rules: maximum, minimum, median, average score and a product of scores. These methods do not require an independent evaluation set for training of the fusion algorithm (and such set is often not available). Conditions under which such schemes perform well are theoretically understood and have been shown to hold in applications [14]. However, in the case of high performance speech verification modules and a medium performance vision module the conditions are violated and none of the above-mentioned fusion scheme performed better than the best individual expert (the HMM).

Modalities	weights		FA (%)	FR (%)
HMM and Face	0.9	0.1	0.86	0.25
Spher. and Face	0.95	0.05	1.37	2.5
HMM and Spher.	0.84	0.16	0.64	0.25

Table 2: Performance on the test set of the linear weighted fusion.

We then considered linear weighted combination rules (also used in [12]). Optimal weights and acceptance threshold were chosen using the evaluation set. The performance of the scheme on the test set is summarized in Table 2. The results show that the trained linear classifier outperforms the linear SVM. This is not unexpected since SVMs minimize maximum distance from decision boundaries whereas the training of the linear classifier minimizes error rate (over training is not a problem for a simple 1-parameter linear classifier). Surprisingly, the linear classifier compares well even with non-linear SVMs. One more interesting observation can be made. A posteriori, a threshold (point on the ROC curve) can be found for the HMM where this expert outperforms the face and HMM combination. However, at the threshold *predicted* from training and evaluation data the weighted sum of Face and HMM expert has a lower error. This suggest that more stable prediction of the operating point can be made for the fused data.

Kernel	Polynomial		Gaussian		Linear	
Set	FA	FR	FA	FR	FA	FR
C1	1.07	0.25	1.18	0	1.47	0
C2	0.34	0.50	0.78	0	1.47	0
C3	0.39	0.50	0.38	0.50	1.47	0
C4	0.13	10.0	1.18	0	1.23	1.25

Table 3: SVM Fusion Performance

[4] http://www.ee.surrey.ac.uk/Research/VSSP/xm2vts

7 Conclusion

We have described a complete multi-modal person identity verification system with very low error rates (less than 1% total error rate). It was evaluated and tested on a large database (295 people) with a published protocol. Combining different modalities increases the performance of the system and yields better results than individual modalities. One of the major problems is how to combine modalities with different skills. We compared two approaches: a linear weighted classifier and SVM. The linear classifier performed well and even better than linear SVM in combining two modalities (face/speech). SVM has the advantage of combining any number of modalities at the same computational cost with very good fusion results.

References

[1] B.S. Atal. Effectivness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *JASA*, 55(6):1304–1312, 1974.

[2] S. Ben-Yacoub. Multi-Modal Data Fusion for Person Authentication using SVM. In *Proc. of AVBPA'99, Washington DC*, pages 25–30, 1999.

[3] F. Bimbot, I. Magrin-Chagnolleau, and L. Mathan. Second-order statistical measure for text-independent speaker identification. *Speech Communication*, 17(1-2):177–192, 1995.

[4] M. Bober and J. Kittler. Robust motion analysis. In *CVPR'94*, pages 947–952, Washington, DC., Jun 1994. Computer Society Press.

[5] R. Brunelli and D. Falavigna. Person identification using multiple cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(10):955–966, October 1995.

[6] T. Choudhury, B. Clarkson, T. Jebara, and A. Pentland. Multimodal Person Recognition using Unconstrained Audio and Video. In *Proc. of AVBPA'99, Washington DC*, pages 176–180, 1999.

[7] Benoît Duc, Elizabeth Saers Bigün, Josef Bigün, Gilbert Maître, and Stefan Fischer. Fusion of audio and video information for multi modal person authentication. *Pattern Recognition Letters*, 18(9):835–843, 1997.

[8] D. Gibbon, R. Moore, and R. Winski, editors. *Handbook of Standards and Resources for Spoken Language Systems.* Mouton de Gruyter, 1997.

[9] F. R. Hampel, E. M. Ronchetti, P.J. Rouseseeuw, and W.A. Stahel. *Robust Statistics.* John Wiley, 1986.

[10] J.Kittler and A Hojjatoleslami. A weighted combination of classifiers employing shared and distinct representations. In *Proc. Conference on CVPR*, pages 924-929, 1998.

[11] K. Jonsson, J. Matas, and J. Kittler. Fast face localisation and verification by optimised robust correlation. Technical report, U. of Surrey, Guildford, Surrey, United Kingdom, 1997.

[12] P. Jourlin, J. Luettin, D. Genoud, and H. Wassner. Acoustic-labial speaker verification. *Pattern Recognition Letters*, 18(9):853–858, 1997.

[13] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, May 1983.

[14] J. Kittler, M. Hatef, R.P.W Duin, and J. Matas. On Combining Classifiers. *IEEE PAMI*, 20(3):226–239, 1998.

[15] J. Kittler. Combining classifiers: A theoretical framework. *Pattern Analysis and Applications*, 1:18–27, 1998.

[16] J. Luettin and G. Maître. Evaluation protocol for the extended m2vts database (xm2vtsdb). Technical Report IDIAP-COM 98-05, IDIAP, 1998.

[17] J. Matas, K. Jonsson, and J. Kittler. Fast face localisation and verification. In A. Clark, editor, *British Machine Vision Conference*, pages 152–161. BMVA Press, 1997.

[18] D.A. Reynolds, R.C. Rose, and M.J.T. Smith. Pc-based tms320c30 implementation of the gaussian mixture model text-independent speaker recognition system. In *ICSPAT, DSP Associates*, pages 967-973, 1992.

[19] A. E. Rosenberg, C. H. Lee, and S. Gokoen. Connected word talker verification using whole word hidden markov model. In *ICASSP-91*, pages 381–384, 1991.

[20] V. Vapnik. *Statistical Learning Theory.* Wiley Inter-Science, 1998.

[21] J. Zhang, Y. Yan, and M. Lades. Face recognition: Eigenfaces, elastic matching, and neural nets. *Proceedings of IEEE*, 85:1422–1435, 1997.