# Non intrusive multi-biometrics on a mobile device: a comparison of fusion techniques

Lorene Allano[1*1], Andrew C. Morris[2], Harin Sellahewa[3], Sonia Garcia-Salicetti[1], Jacques Koreman[2], Sabah Jassim[3], Bao Ly-Van[1], Dalei Wu[2], Bernadette Dorizzi[1]

[1] Institut National des Télécommunications, 9 rue Charles Fourier, 91011 Evry, France
[2] Saarland University, FR 4.7 Phonetics, Bldg. 17.2, rm 5.02 P.O. Box 151150, D-66041, Saarbruecken, Germany
[3] Department of Information Systems, University of Buckingham, Buckingham, UK

## ABSTRACT

In this article we test a number of score fusion methods for the purpose of multimodal biometric authentication. These tests were made for the SecurePhone project, whose aim is to develop a prototype mobile communication system enabling biometrically authenticated users to deal legally binding m-contracts during a mobile phone call on a PDA. The three biometrics of voice, face and signature were selected because they are all traditional non-intrusive and easy to use means of authentication which can readily be captured on a PDA. By combining multiple biometrics of relatively low security it may be possible to obtain a combined level of security which is at least as high as that provided by a PIN or handwritten signature, traditionally used for user authentication. As the relative success of different fusion methods depends on the database used and tests made, the database we used was recorded on a suitable PDA (the Qtek2020) and the test protocol was designed to reflect the intended application scenario, which is expected to use short text prompts. Not all of the fusion methods tested are original. They were selected for their suitability for implementation within the constraints imposed by the application. All of the methods tested are based on fusion of the match scores output by each modality. Though computationally simple, the methods tested have shown very promising results. All of the 4 fusion methods tested obtain a significant performance increase.

**Keywords**: multimodal biometrics, verification, score fusion, Gaussian Mixture Models, smartphone.

## 1 INTRODUCTION

Multibiometrics, i.e. the verification of a person's identity by more than one biometric trait, is expected to strongly enhance person authentication performance in real applications. But most of the presently available biometric databases have been acquired in more or less controlled environments, so that it is difficult to predict performance in real applications. The experiments presented here are performed on a database acquired on a personal mobile device (smartphone) as part of the SecurePhone project (IST-2002-506883 project "Secure Contracts Signed by Mobile Phone"). The project aim is to produce a prototype of a new mobile communication system (the "Securephone") enabling biometrically authenticated users to deal legally binding m-contracts during a mobile phone call in an easy yet highly dependable and secure way.

The use of signals recorded on a smartphone enables us to evaluate multibiometric person authentication with realistic signals under various degrees of degradation. The context of mobility generates degradations of input signals due to the variety of environments encountered (ambient noise, lighting variations, …), while the sensor's' lower quality further

[*] Lorene.allano@int-evry.fr ; phone +33 1 60 76 46 45; fax +33 1 60 76 42 84

contributes to decrease system performance. By fusing three different biometric traits, the effect on signal degradation on the system's performance can be counteracted.

Our aim in this work is to study the benefit of different fusion techniques for combining speech, face and handwritten signature on a smartphone. Many fusion techniques have so far been compared in the literature. Among them, score fusion techniques can be classified as score combination rules or as statistical learning techniques. Previous works have shown that one class of such techniques give better performance than the other depending on the experimental framework that is considered. Our aim in this work is to compare score combination rules and statistical learning techniques in a concrete application, performing non intrusive biometric verification on a smartphone.

## 2 BIOMETRIC DATA

In order to evaluate biometric systems in real application conditions, a database of voice, face and signature recordings [1] was captured on a mobile device. Examples of data are shown in Figure 1. We used the Qtek2020 device and the smartphone's own sensors (microphone, camera and touch-screen) that have limited sampling capabilities.

Audio-visual data for the smartphone database is in English. The database contains 60 speakers, 30 male and 30 female, of which 80% are native speakers. There are 3 age groups (< 30, 30-45, > 45) with 10 males and 10 females in each group. Three types of fixed prompt (5-digit, 10 digit and short phrase) were recorded, with 6 examples from each type. Each speaker is recorded in 2 recording sessions separated by at least one week, thus allowing for the evaluation of the influence of time variability on person authentication. Each session comprises 2 indoor and 2 outdoor recordings with variable environmental signal degradation. The 2 indoor recording conditions were respectively for voice/face: good/good and bad/bad. The 2 outdoor recordings conditions were respectively for voice/face: bad/good and bad/bad. The 4 recording conditions (2 indoors and 2 outdoors) reflect reasonable variation in noise and lighting conditions. The amount of data recorded per person is very limited, permitting only up to short 4 recordings for client model training. This reflects the practical need for fast client enrolment.

Signatures, which were always recorded under good but realistic recording conditions, were recorded from other subjects, because signers were recorded in a different place from the audio-visual subjects. Twenty signatures were recorded in one session from each of 30 male and 30 female subjects. The signers are balanced in the three age groups defined for audio-video persons. Twenty forgeries were made by impostors who were not clients of the signature (or audio-visual) database. Those forgeries were made after observing the genuine signatures together with their dynamic characteristics on the PDA's screen. The signature sampling rate was 100 Hz, with coordinates data but no pressure or pen inclination information.

Virtual SecurePhone clients were created by coupling subjects in the audio-visual and signature databases. This is possible because a client's signature can be assumed to be largely independent from his or her appearance and voice. Indeed, results from a previous work on signature and voice data from BIOMET [2] support the independence assumption of such modalities [3]. The data are coupled using gender-dependent virtual persons. Because both databases are balanced in age among the 3 age groups, virtual persons are also age-dependent. To be able to present results which are independent of any particular coupling between signatures and audio-visual data, 100 random couplings were made and the average over those 100 trials was computed.

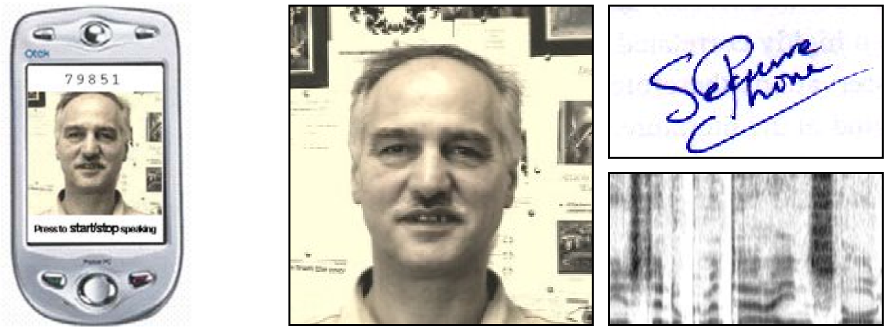

Figure 1: Examples of data of the PDA database

# 3 UNIMODAL PERSON AUTHENTICATION SYSTEMS

## 3.1 Signature verification

As described in [4] each writer's signature is modelled by a continuous left-to-right Hidden Markov Model (HMM), characterised by a given number of states with an associated set of transition probabilities among them, the data in each state being represented by a continuous density multivariate Gaussian mixture model (GMM). The HMM has a left-right topology, i.e. it only authorises transitions from each state to itself and to its immediate right-hand neighbour. An optimal number of states is estimated for each writer and a personalised feature normalisation (of 19 features) is carried out to improve the quality of the modelling. The system exploits a fusion strategy of two complementary sources of information provided by the same HMM. The first is the likelihood of the signature given the HMM. The second is the segmentation of the test signature when using the Viterbi algorithm. As shown in [4], the combination of these two sources of information results in a better separation of genuine and impostor distributions, thus significantly improving writer verification results.

## 3.2 Speaker verification

Both in terms of user enrolment as well as for actual verification, realistic applications impose strong restrictions on the amount of speech material that can be acquired from each client. To ensure optimal performance with small amounts of training data, the speaker verification system used here is text-dependent. We shall test only fixed digit sequences which allow us to keep the size of the acoustic model which needs to be stored on the device to a minimum. For a given fixed prompt, a GMM background model [5] is trained, from which individual speaker models are trained by maximum a-posteriori (MAP) adaptation to each speaker. Results are given for gender-dependent as well as gender-independent background models. Acoustic features use 19 Mel-frequency cepstral coefficients (MFCCs) together with cepstral mean subtraction and appended time difference features. GMMs have 128 Gaussian components and are trained by k-means clustering, followed by EM iteration. This is performed by the Torch machine learning API [6], using a variance threshold and minimum Gaussian weight determined on the basis of the development data set. The unimodal decision to accept or reject a speaker as a true client is based on a thresholding of the test score (the logarithm of the ratio of the speaker to background model likelihoods).

## 3.3 Face verification

Face verification is a challenging task due to varying conditions in the capturing process (variations in pose, facial expressions and illumination). In order to normalise variations in illumination conditions, we applied, as a pre-processing step, Histogram Equalization (HE) [7]. Our system, based on a wavelet-based verification scheme [8], uses the coefficients in the wavelet transformed LL-subbands at depth 3 or more as a feature vector of a given face image. The LL-subband corresponds to the low-pass filtering, which captures the scaled energy of the image. The final classification decision of a test video is based on verifying each of 10 frames selected from the full video sequence. The verification of a frame is performed by computing the City-Block distance between the feature vector of a test frame and the feature vectors of enrolled frames. For each test frame, the match-score is the minimum distance to one of the enrolled frames. The match score for a test video is the minimum of the 10 frame match scores.

# 4 FUSION METHODS

Two types of score fusion methods are presented. The first type is based on the Arithmetic Mean Rule after a previous normalization of each score separately. The second type is based on a 3D density estimation followed by class posterior probabilities computation.

### 4.1 Fusion by AMR with associated normalisations

The 3 unimodal scores are combined by means of a simple Arithmetic Mean Rule (AMR) after performing a normalisation of these scores. Two types of normalisation are studied: the first one is based on the Min-Max normalisation [9] and the second uses a posteriori class probabilities.

The "Min-Max" normalisation of score s of one unimodal expert is defined as *n=(s-m)/(M-m)* where *M* is the maximum and *m* is the minimum. We consider the mean (μ) and standard deviations (σ) of both the client and impostor distributions in the training database, and set*:* $m= \mu_{imp}-2\ \sigma_{imp}$ and $M= \mu_{cl}+2\ \sigma_{cl}$. Indeed, assuming that genuine and impostor scores follow Gaussian distributions, 95% of the values lie in the *[μ-2σ, μ+2σ]* interval; following this model, our choice of *m* and *M* permits to cover most of the scores. Values higher than *M* or lower than *m* are thresholded. This linear normalisation maps the score in the [0,1] interval.

Bayes normalisation uses the a-posteriori client class probability *P(C|s)* given score *s*, as a normalised score. A-posteriori probabilities are obtained using Bayes' rule:

$$P(C/s) = \frac{p(s/C)P(C)}{p(s/C)P(C) + p(s/I)P(I)} \tag{1}$$

where *P(C)* and *P(I)* are the client and impostor priors, set to 0.5 because we have no prior knowledge, and *p(s|C)* and *p(s|I)* are the client and impostor likelihoods. Conditional probability densities are computed from Gaussian score distributions whose parameters are estimated on the training database. Assuming independence between the 3 scores *s1, s2* and *s3,* and following [10], we compute the arithmetic mean of *P(C|s1), P(C|s2)* and *P(C|s3).*

### 4.2 Fusion by 3D density estimation

#### 4.2.1 3D Gaussian density estimation

In this case, instead of estimating the conditional score densities of each modality separately, the 3D Gaussian density is estimated. This is done using a Gaussian assumption on the 3D class conditional densities. The difference with the previous case is that there is no independence assumption. We estimate p(s1, s2, s3|C) and p(s1, s2, s3|I) and compute the a posteriori class probability p(C|s1, s2, s3), as:

$$p(C/s_1, s_2, s_3) = \frac{P(C)p(s_1, s_2, s_3/C)}{P(C)p(s_1, s_2, s_3/C) + P(I)p(s_1, s_2, s_3/I)} \tag{2}$$

*P(C)* and *P(I)* are the client and impostor priors, set to 0.5 because we have no prior knowledge.
For the 3D Gaussian density, we are considering a diagonal covariance matrix, assuming that the 3 scores are uncorrelated. This is done for computational reasons, and also considering that there is not enough data to estimate cross-correlation.

#### 4.2.2 Estimation with a Gaussian Mixture Model (GMM)

Fusion using Gaussian Mixture Models (GMM) [5] follows the same idea as AMR using 3D Gaussian density estimates: instead of considering that the conditional 3D density is Gaussian, we estimate it as a mixture of Gaussian densities (a sum of *N (here N=3)* 3D Gaussian components of a Gaussian Mixture Model). The i^th 3D Gaussian component is represented by $\mu_i$, $\Sigma_i$ and $\alpha_i$, the mean vector, covariance matrix and weight of the i^th Gaussian component in the sum. Therefore, with $s=(s_1, s_2, s_3)$, the joint density is:

$$P(s/C) = \sum_{i=1}^{N} \alpha_i \frac{1}{(2\pi)^{n/2}\sqrt{\det(\Sigma_i)}} \exp\left(-\frac{1}{2}(s-\mu_i)^t \Sigma_i^{-1}(s-\mu_i)\right) \tag{3}$$

For impostors' conditional density, other mean vectors, covariance matrix and weights are estimated and, to obtain the final score, we compute the a posteriori class probability $p(C|s_1, s_2, s_3)$ using equation *(2).*

As before, since cross-correlations cannot be estimated both for computational reasons and because of data sparsity, Σ is considered diagonal for both client and impostor densities, assuming that the 3 scores are uncorrelated.

## 5 EXPERIMENTS

Results for fusion are obtained on the PDA database [1] of virtual subjects built by coupling audio-video data from a given person to signatures of another person. Virtual persons are gender and age dependent. We perform 100 random couplings and compute the average performance over those 100 trials. The three types of prompt are considered: type 1 (T1) of 5 digits, type 2 (T2) of 10 digits and type 3 (T3) of phrases (of approximately the same duration as the 5-digit sequences). Each type is represented by 6 prompt (P1, P2, P3, P4, P5 and P6) examples. Results reported in Table 1 and Table 2 are also averaged over these 6 examples per type.

Scores used for fusion are obtained from the three single modality systems using specific protocols. Signature scores are generated considering 5 randomly chosen signatures for model training and the remaining 15 signatures for test purposes. For voice and face, the 4 sequences (for a given prompt example) of session 1 are used for training and the 4 sequences of session 2 are used for test purposes. This means that models are trained on both indoor and outdoor data and evaluated on both too.

The protocol is the following:

- 24 persons (balanced in gender and age group) out of 60 available in the PDA database are reserved for the construction of the Universal Model Background (UBM) for the speech modality.
- The remaining 36 persons are split in two sets g1 and g2 of 18 persons each (3 for each age group of both gender). Each person has 4 associated client accesses and 20 impostor accesses.
- The fusion system is trained on g1 and tested on g2, and their roles are interchanged. The resulting two EERs are averaged to obtain the final error rate which is given in Table 2.

Table 1: Single modalities results (EER) for the 6 examples (P1 to P6) of each prompt type.

| **Prompt type** | T1 (5-digits) | | | | | | T2 (10-digits) | | | | | | T3 (phrases) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Prompt examples** | **P1** | **P2** | **P3** | **P4** | **P5** | **P6** | **P1** | **P2** | **P3** | **P4** | **P5** | **P6** | **P1** | **P2** | **P3** | **P4** | **P5** | **P6** |
| **voice** | 7.21 % | | | | | | 3.24 % | | | | | | 5.5 % | | | | | |
| | 9.51 | 7.22 | 8.40 | 5.69 | 6.87 | 5.55 | 4.03 | 4.03 | 2.91 | 2.64 | 2.22 | 3.61 | 4.23 | 6.66 | 6.94 | 5.62 | 6.25 | 3.543.54 |
| **face** | 28.40 | | | | | | 27.55 % | | | | | | 28.33 % | | | | | |
| | 27.43 | 26.46 | 27.08 | 29.16 | 28.40 | 31.87 | 27.84 | 27.84 | 26.73 | 28.33 | 27.78 | 26.73 | 30.55 | 26.87 | 24.86 | 28.54 | 27.98 | 31.18 |
| **signature** | 8.01 % | | | | | | | | | | | | | | | | | |

Table 2: Fusion results (EER) compared to single modalities in the PDA database

| | | T1 (5-digits) | T2 (10-digits) | T3 (phrases) |
|---|---|---|---|---|
| Single Modalities | Voice | 7.21 % | 3.24 % | 5.54 % |
| | Face | 28.40 % | 27.55 % | 28.33 % |
| | Signature | 8.01 % | | |
| Fusion Methods | AMR with Min-Max | 4.09 % | 3.16 % | 3.85 % |
| | AMR of posterior probabilities | 2.41 % | 1.67 % | 2.30 % |
| | 3D Gaussian density estimation | 2.67 % | 1.93 % | 2.52 % |
| | GMM with 3 Gaussian components | 2.56 % | 1.66 % | 2.68 % |
| | **MinMax + GMM** | **2.39 %** | **1.54 %** | **2.30 %** |

Compared with the baseline results for the single modalities, a strong improvement in user authentication is found for most fusion methods. GMM after a Min-Max normalisation of scores leads to the lowest EERs. For this method, and for each prompt type, the 6 values of the EER corresponding to each occurrence are reported in Table 3. We also indicate their associated standard deviation for 100 random couplings of virtual subjects.

A comparative study of the fusion methods described in Section 4 show that training-based methods, here based on GMMs, seem to be more suited to degraded test conditions, as already shown in [11].

Table 3: ERRs for each example of the 3 prompt types and their standard deviation for the best fusion method

| | T1 (5-digits) | | | | | | T2 (10-digits) | | | | | | T3 (phrases) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Averaged EER | 2.39 % | | | | | | 1.54 % | | | | | | 2.30 % | | | | | |
| ERRs | 2.51 | 1.99 | 2.89 | 2.58 | 2.39 | 1.96 | 1.85 | 1.54 | 1.43 | 1.43 | 1.74 | 1.27 | 2.06 | 2.81 | 3.04 | 1.91 | 2.08 | 1.90 |
| Standard Dev. | 1.01 | 0.87 | 1.07 | 1.10 | 0.87 | 0.83 | 0.91 | 0.87 | 0.81 | 0.76 | 0.83 | 0.78 | 1.04 | 2.98 | 3.07 | 1.97 | 1.16 | 0.88 |
| Averaged St.Dev. | 0.96 % | | | | | | 0.83 % | | | | | | 1.85 % | | | | | |

In order to visualize performance for GMM fusion using Min-Max normalised scores from the individual modalities compared to single experts for every value of the decision threshold, DET curves [12] for T1, T2 and T3 are shown in Figure 2.

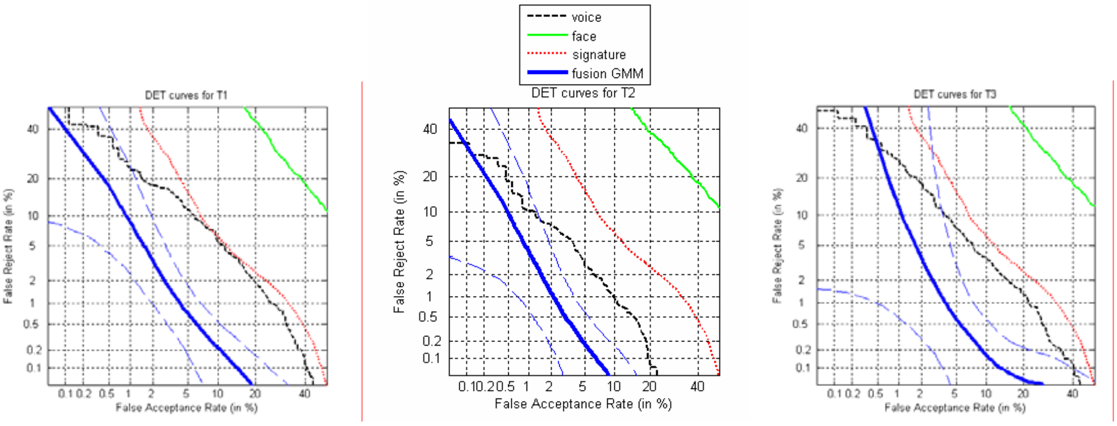

Figure 2: DET curves for signature, voice, face and best fusion method, for T1 (left), T2 (centre) and T3 (right)

Figure 2 shows that fusion of three modalities by a GMM improves system performance compared to single modalities for the 3 prompt types. The standard deviation of the error rates is also shown for fusion (dashed lines) For most values of decision threshold, the upper bound of the interval of variation of error rates for 100 databases of virtual subjects is still better than the best modality.

Other specific functioning points, namely the WER (Weighted Error Rate) for 3 different weights: R=1, R=0.1 and R=10 and their corresponding FAR (False Acceptance Rate) and FRR (False Rejection Rate) are reported in Table 3. WER is defined as:

$$WER(R) = \frac{FRR + R * FAR}{1 + R}$$

The threshold corresponding to the minimum of WER on the development group g1 is used to report the error rate on the evaluation group g2 (a priori threshold), and vice versa. A-priori error rates are then averaged.

Table 3: fusion results (WER) for the best fusion method (MinMax + GMM).

| | | T1 (5-digits) | T2 (10-digits) | T3 (phrases) |
|---|---|---|---|---|
| EER | | 2.39 % | 1.54 % | 2.30 % |
| FAR / FRR<br>WER | R=1 | 1.57 %/3.24 %<br>2.40 % | 0.89 %/3.32 %<br>1.60 % | 1.61 %/3.14 %<br>2.37 % |
| | R=0.1 | 4.97 %/1.56 %<br>1.87 % | 3.05 %/1.20 %<br>1.37 % | 4.54 %/1.78 %<br>2.03 % |
| | R=10 | 0.43 %/6.95 %<br>1.02 % | 0.25 %/4.37 %<br>0.63 % | 0.38 %/6.34 %<br>0.92 % |

Table 3 reports weighted error rates (WERs) for 3 different R values: 1, 0.1 and 10, together with their corresponding FAR and FRR. Results show that low error rates can be obtained for a wide range of cost ratios. Results for 10-digit prompts are better than for either 5-digit prompts or phrases, which show similar performance.

## 6 CONCLUSIONS

The aim of this work is to study the benefit of multimodal fusion on single non intrusive modalities acquired on a PDA. To that aim, a virtual subjects database was built from audio-video data and signatures from different subjects, captured on the PDA. Single modality experts show in general higher error rates on such database compared to state-of-the-art results on standard databases. This is due to the fact that mobility conditions are well reflected in our acquisition protocol. This remark emphasizes the interest of multimodal fusion in mobile applications, which is confirmed by our study: fusion by a training-based method, based on GMMs, proves to be robust in the variety of acquisition conditions considered in this work. Moreover, considering the variance of the fusion system performance due to random couplings of virtual subjects, the upper bound of the error rate still does better than the best unimodal system, for most values of the decision threshold. Tests have shown that the non-intrusive biometrics used in the SecurePhone project can be used to achieve a level of authentication accuracy which would be sufficient for a wide range of applications.

Further work will focus on a more difficult protocol, relying on training unimodal experts on controlled reference data and testing the fusion system on variable acquisition conditions (degraded and adverse). This would indeed simplify the enrolment phase for the user. Also this work exploits a small database in number of persons and client accesses; we look forward extending our database for further tests. Another possibility is to consider a text-independent speaker verification system, which would permit to consider the 18 available prompt examples as being of the same type. This way the fusion system could be trained on enough data, and particularly density estimation could be done with full covariance matrices.

## 7 ACKNOWLEDGEMENTS

This work was supported by the EC SecurePhone project IST-2002-506883.

## 8 REFERENCES

[1] A.C. Morris, H. Sellahewa, L. Allano, “The SecurePhone PDA database and automatic test procedure for multimodal user verification”, Tech Report, Jan. 2006.
http://www.coli.uni-saarland.de/SecurePhone/documents/PDA_database_and_test_protocol.pdf

[2] Garcia-Salicetti, S., Beumier, C., Chollet, G., Dorizzi, B., Leroux-Les Jardins, J., Lunter, J., Ni, Y. & Petrovska-Delacretaz, D., "BIOMET: a Multimodal Person Authentication Database Including Face, Voice, Fingerprint, Hand and Signature Modalities", Proc. of 4th International Conference on Audio and Video-Based Biometric Person Authentication, pp. 845-853, Guildford, UK, July 2003.

[3] S. Garcia-Salicetti, M.A.Mellakh, L. Allano, B. Dorizzi, “Multimodal Biometric Score Fusion: the Mean rule vs. Vector Support Classifiers”, in Proc. of EUSIPCO 2005, Antalya, Turkey, September 4-8, 2005.

[4] B. Ly Van, S. Garcia-Salicetti, B. Dorizzi, “Fusion of HMM’s Likelihood and Viterbi Path for On-line Signature Verification”, Biometric Authentication Workshop (BioAW), Lecture Notes in Computer Science (LNCS) 3087, pp. 318-331, Prague, Czech Republic, May 2004.

[5] Reynolds, D.A.: Speaker identification and verification using Gaussian mixture speaker models. Speech Comm. 17, 1995, pp.91-108.

[6] Collobert, R., Bengio, S. & Mariéthoz, J.: Torch: a modular machine learning software library. Technical Report IDIAP-RR 02-46, 2002.

[7] R. Beveridge, D. Bolme, M. Teixeira and B. Draper “The CSU Face Identification Evaluation System User’s Guide: Version 5.0”, Computer Science Department, Colorado State University, May 1, 2003.
http://www.cs.colostate.edu/evalfacerec/index.html (10/01/06)

[8] H. Sellahewa and S. Jassim, “Wavelet-based Face Verification for constrained platforms”, Proc. SPIE on Biometric Technology for Human Identification II, Florida 2005, Vol. 5779, pp 173-183, March 2005.

[9] M. Indovina, U. Uludag, R. Snelick, A. Mink, A. Jain, “Multimodal Biometric Authentication Methods : A COTS Approach”, in Proc. MMUA 2003, Santa Barbara, California, USA, Dec. 2003, pp. 99-106.

[10] J. Kittler, M. Hatef, R.P.W. Duin, J. Matas, “On Combining Classifiers”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, N°3, pp. 226-239, March 1998.

[11] S. Garcia-Salicetti, A. Mellakh, L. Allano, B. Dorizzi, “Multimodal Biometric Score Fusion: the Mean Rule vs. Support Vector Classifiers”, in Proc. of EUSIPCO’05, Antalya, Turkey, 4-8 September 2005.

[12] A. Martin, G. Doddington, T. Kamm, M. Ordowski, M. Przybocki, “The DET Curve in Assessment of Detection Task Performance”, in Proc. EUROSPEECH’97, Vol. 4, pp. 1895-1898, Rhodes Greece, 1997.