A Multi-Modal Recognition System Using Face and Speech

Samir Akrouf, Yahia Belayadi, Messaoud Mostefai, Youssef Chahir

A Multi-Modal Recognition System Using Face and Speech

Samir Akrouf[1], Belayadi Yahia[2], Mostefai Messaoud[2] and Youssef chahir[3]

**[1]Department of Computer Science, University of Bordj Bou Arréridj, algeria
El Anasser, 34030, BBA Algeria**

**[2]Department of Computer Science, University of Bordj Bou Arréridj, algeria
El Anasser, 34030, BBA Algeria**

**[2]Department of Computer Science, University of Bordj Bou Arréridj, algeria
El Anasser, 34030, BBA Algeria**

**[3]Department of Computer Science, University of Caen Lower Normandie, France
Caen, State ZIP/Zone, France**

Abstract

Nowadays Person Recognition has got more and more interest especially for security reasons. The recognition performed by a biometric system using a single modality tends to be less performing due to sensor data, restricted degrees of freedom and unacceptable error rates. To alleviate some of these problems we use multimodal biometric systems which provide better recognition results. By combining different modalities, such us speech, face, fingerprint, etc., we increase the performance of recognition systems.

In this paper, we study the fusion of speech and face in a recognition system for taking a final decision (i.e., accept or reject identity claim). We evaluate the performance of each system differently then we fuse the results and compare the performances.

Keywords: Biometrics, data fusion, face recognition, automatic speaker recognition, data processing, decision fusion.

1. Introduction

Identity recognition is becoming more and more used in the last years. Demand is increasing for reliable automatic user identification systems in order to secure accesses to lots of services or buildings. Biometric Identification [1] is the area related to person recognition by means of physiological features (fingerprints, iris, voice, face, etc.).

A biometric person recognition system can be used for person identification or verification. For the verification, a user claims a certain identity ("I am X"). The system accepts or rejects this claim (deciding if really the user is who he claims to be). For identification, there is no identity claim. The system decides who the user is. In this paper we use two the biometrics which appears to be the most popular ones and are less restricting for person identification (voice and face). The major strength of voice and face biometrics is their high acceptance by the society.

This multiple sensors capture different biometric traits. Such systems, known as multi-modal biometric systems [2], are more reliable due to the presence of multiple pieces of evidence. These systems are able to meet the stringent performance requirements imposed by various applications. Moreover, it will be extremely difficult for an intruder to violate the integrity of a system requiring multiple biometric traits.

In the literature we find that combining different biometric modalities enables to achieve better performances than techniques based on single modalities [3]–[10]. Combining different modalities allows to outcome problems due to single modalities. The f*usion* algorithm, which combines the different modalities, is a very critical part of the recognition system. So before the fusion one would ask what strategy do we have to adopt in order to make the final decision?

The sensed data (face and speech) are processed by different recognition systems: a face identification system and a speaker identification system. Each system, given the sensed data, will deliver a matching score in the range between zero (reject) and one (accept). The fusion module will combine the opinions of the different systems and give a binary decision: accept or reject the claim.

An identification scenario involving two modalities is shown in Fig. 1. The paper will address the issue of which binary classifier to use for the fusion of different expert "opinions."

The face recognition system will be presented in paragraph 2. The speaker recognition system based on text-dependent approach is discussed in paragraph 3.

The fusion [2]-[4] of different modalities is described in paragraph 5.

Finally we present the evaluation results and the main conclusions.

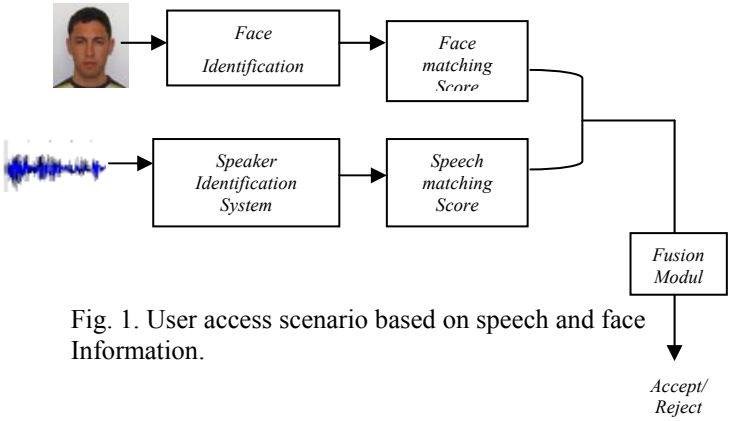


Fig. 1. User access scenario based on speech and face Information.

2. Face Recognition

This paper uses a hybrid method combining principal components analysis (PCA) [11] and the discrete cosine transform (DCT) [12] for face identification [13].

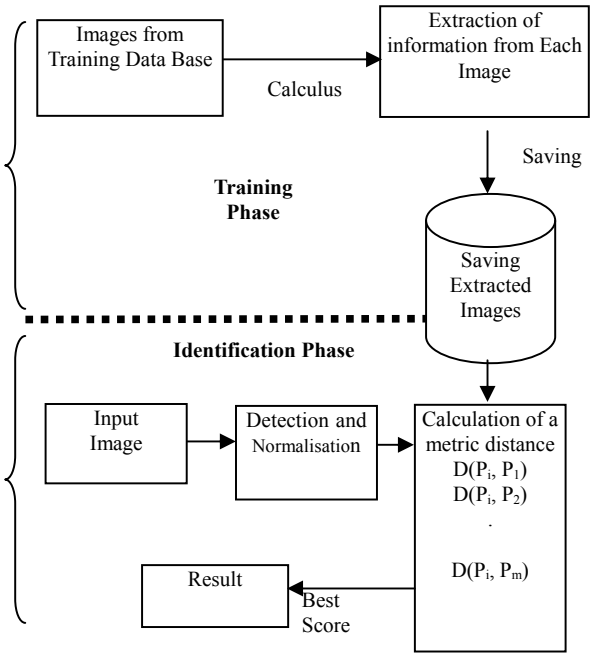


Fig. 2. Recognition Algorithm Stages.

2.1 Presentation of the Hybrid Method

PCA and DCT have certain mathematical similarities since that they both aim to reduce the dimensions of data. The use of a hybrid method combining these two techniques gave performances slightly higher than those obtained with only one method (experiments being made on three different image data bases). Its principle is very simple: each image is transformed into a coefficient vector (in the training and recognition phase). We first use the DCT method which produces a result used as entry for the PCA method. We use PCA with coefficients vectors instead of pixels vectors. We notice that this technique requires more time than PCA (because of the calculation of the coefficients) in particular with data bases of average or reduced size but it should be noted that it requires less memory what makes its use advantageous with bases of significant size.

2.2 Experimental Results

The tests were performed by using the image data bases ORL, Yale Faces and BBAFaces. The latter was created at the University Center of Bordj Bou Arreridj in 2008. It is composed by 23 people with 12 images for each one of them (for the majority of the people, the images were taken during various sessions). The images reflect various facial expressions with different intensity variations and different light sources. To facilitate the tests, the faces were selected thereafter manually in order to get images of 124 X 92 pixels, we then convert them into gray levels and store them with JPG format. Fig. 3. represents a typical example of the data. It should be noted that certain categories of this data are not retained for the tests.

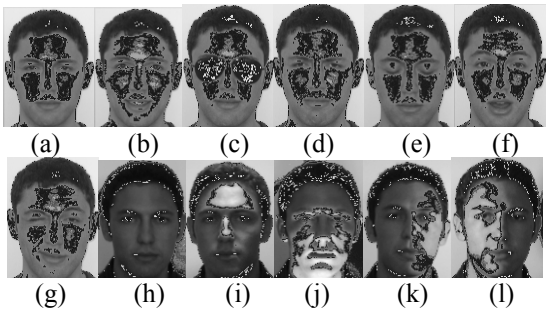


Fig. 3. Example from BBAFaces. (a): normal, (b): happy, (c): glasses, (d): sad, (e): sleepy, (f): surprised, (g): wink, (h): dark, (i): top light, (j): bottom light, (k): left light, (l): right light.

In the following we will expose the results obtained for the tests realized with Yale Faces and BBA Faces.

Table 1: Rates of Recognition

Data Base	***PCA***	***PCA + DCT***
BBA Faces	57.06 %	66.30 %
Yale Faces	62 %	72.77 %
ORL Base	71.38 %	72.77 %

Finally we conclude that the combination of PCA with DCT offers higher rates of recognition than those obtained with only one method which justifies our choice for the algorithm used in our system.

3. Speaker Recognition System

Nowadays The Automatic Treatment of speech is progressing, in particular in the fields of Automatic Speech Recognition "ASR" and Speech Synthesis.
The automatic speaker recognition is represented like a particular pattern recognition task. It associates the problems relating to the speaker identification or verification using information found in the acoustic signal: we have to recognize a person by using his voice. ASR is used in many fields, like domestic, military or jurisprudence applications.
In this work we use an automatic speaker recognition system presented an earlier paper [15]. We will use speaker recognition in text independent mode since we dispose of very few training data. We have to estimate with few data a robust speaker model to allow the recognition of the speaker.

3.1 Basic System

A speaker recognition system comprises 4 principal elements:

1. An acquisition and parameterization module of the signal: to represent the message in an exploitable form by the system.

2. A training module: who is charged to create a vocal reference of the speaker starting from a sample of his voice «GMM Gaussian Mixture Models».

3. A resemblance calculus module: who calculates the resemblance between a sample signal and a given reference corresponding to a person.

4. A decision module: based on a strategy of decision.

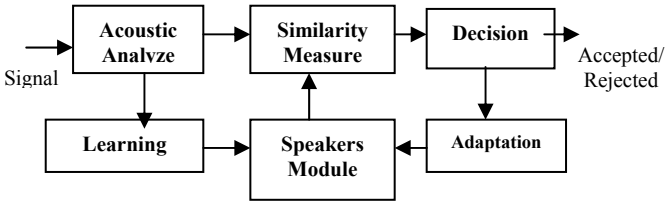


Fig. 4. Typical diagram of a checking speaker system

3.2 Speaker Identification "SI"

The speaker identification consists in recognizing a person among many speakers by comparing his vocal expression with known references. From a diagrammatic point of view "see figure 4", a sequence of word is given in entry of the ASR system. For each known speaker, the sequence of word is compared with a characteristic reference of the speaker. The identity of the speaker whose reference is the nearest to the sequence of word will be the output datum of the system (ASR). Two modes of identification are possible: identification in a closed unit for which the speaker is identified among a known number of speakers or identification in an open unit for which the speaker to be identified does not belong inevitably to this unit [16].

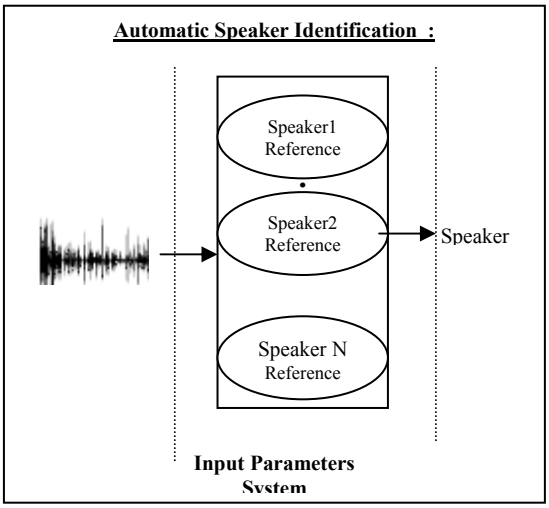


Fig.5. Automatic Speaker Identification

3.4 Speaker Verification "SV"

The checking "or authentification" of the speaker consists in, after the speaker declines his identity, checking the adequacy of its vocal message with the acoustic reference of the speaker who it claims to be. A measurement of similarity is calculated between this reference and the vocal message then compared with a threshold. In the case the measurement of similarity is higher than the threshold, the speaker is accepted. Otherwise, the speaker is considered as an impostor and is rejected [16].

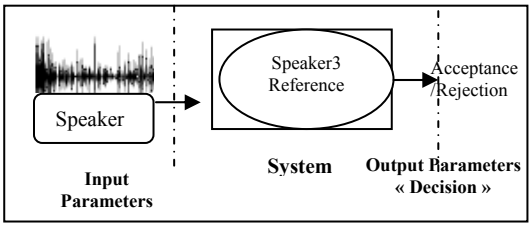


Fig. 6. Automatic Speaker Verification

3.5 Text Dependent and independent mode

We distinguish between the speaker recognition independently of the contents of the sentence pronounced "text independent mode" and the speaker recognition with the pronunciation of a sentence containing a key word "text dependent mode". The levels of dependence to the text are classified according to the applications:

- Systems with free text "or *free-text"*: the speaker is free to pronounce what he wants. In this mode, the sentences of training and test are different.
- Systems with suggested text "or *text-prompted"*: a text, different on each session and for each person, is imposed to the speaker and is determined by the machine. The sentences of training and test can be different.
- Systems dependent on the vocabulary "or *vocabulary-dependent"*: the speaker pronounces a sequence of words resulting from a limited vocabulary. In this mode, the training and the test are carried out on texts made up and starting from the same vocabulary.
- Personalized systems dependent on the text (or *to use-specific text dependent)*: each speaker has his own password. In this mode, the training and the test are carried out on the same text.

The vocal message makes the task of ASR systems easier and the performances are better. The recognition in text mode independent requires more time than the text mode dependent [17].

3.6 Speaker Modeling

Here we briefly introduce the most usually used techniques in the speaker recognition. Here the problem (speaker recognition) can be formulated as a classification problem. Various approaches were developed; nevertheless we can classify them in four great families:

1. Vectorial approach*:* the speaker is represented by a set of parameter vectors in the acoustic space. The principal is he recognition containing "*Dynamic Time Warping"* DTW and by vectorial quantification.
2. Statistical approach*: it* consists in representing each speaker by a probabilistic density in the acoustic space parameters. It covers the techniques of modeling by the Markov hidden models, the Gaussian mixtures and statistical measurements of the second order.
3. The connexionnist approach*:* mainly consists in modeling the speakers by neuron networks.
4. Relative approach*: here* we model a speaker relatively with other reference speakers which models are well learned.

Finally we say that the automatic speaker recognition is probably the most ergonomic method to solve the access problems. However, the voice cannot be regarded as a biometric characteristic of a person taking into account intra-speaker variability. A speaker recognition system generally proceeds in three stages: acoustic analysis of the speech signal, speaker modeling and finally taking the decision. In acoustic analysis, the MFCC are the most used acoustic coefficients. As for the modeling, GMM constitutes the state of the art in ASR. The decision of an automatic speaker recognition system is based on the two processes of speaker identification and/or checking whatever the application or the task is concerned with.

4. Performance of Biometric Systems

The most significant and decisive argument which makes the difference between a biometric system and another is its error rate, a system is considered ideal if its:

False Rejection Rate= False Acceptance Rate= 0;

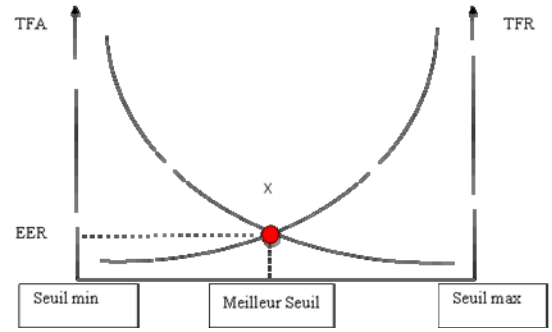


Fig. 7. Illustration of typical errors in a biometric system.

Consequently it is necessary to find a compromise between the two rates which are the junction of the curves (point X) where couple (TFR, TFA) is minimal.

5. Fusion by Decision Methods

Among the fusion of decision methods the most used one quotes:

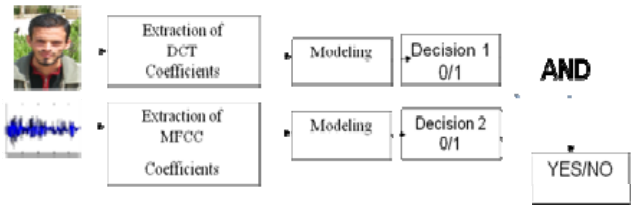
5.1 Fusion by the AND operator:

If all the systems decided 1 then the final decision is **YES** with the operator AND, a false acceptance occurs only if the result of each test is a false acceptance. The probability of false acceptance is thus the product of the probabilities obtained for each test.

$$P(FA) = P1(FA).P2(FA)$$

But in a symmetrical way, the probability of false rejections becomes:

$$P(FR) = P1(FR) + P2(FR) - P1(FR).P2(FR)$$



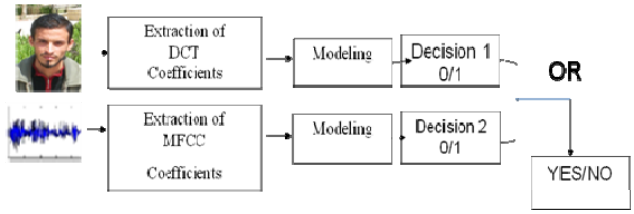
5.2 Fusion by OR operator

If one of the systems decided 1 then the final decision is **YES**. The user is accepted so at least one of the two tests is positive. In this configuration, a false rejection can exist only if the two tests produce a false rejection. The final probability of false rejection P (FR) is the product of the two probabilities of false rejection

P (FR) = P1 (FR)*P2 (FR)

The probability of false final acceptance is described by:

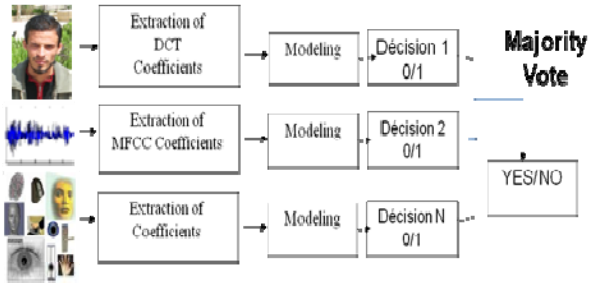
P (FA) = P1 (FA) + P2 (FA) - P1 (FA)*P2 (FA)



5.3 Fusion by the majority vote:

If the majority of the systems decided 1 then the final decision is **YES**.

Majority Vote is a simple method to combine the exits of multiple sources and use a voting process. In this case, each source must provide a decision of its choice and the final decision is based on a majority rule.



5.4 Experimental Results

In order to test our system we used ORL and TIMIT bases. We used 30 customers and 30 impostors with a base containing 100 elements. The face recognition system generated 13 false rejections and 6 false acceptances in an average time equal to 5.6 seconds whereas the speaker recognition system produced 7 false rejections and 12 false acceptances in an average time equal to 6.1 seconds.

- In the face recognition system we obtained:
  - ✓ $P(FA_1)=0.1$.
  - ✓ $P(FR_1)=0.6$.
- In the speaker recognition system we obtained:
  - ✓ $P(FA_2)=0.3$.
  - ✓ $P(FR_2)=0.2$.

When applying fusion operators AND and OR we obtain:

- **AND Operator:**

P (FR) = 0.12
P (FA) = 0.37

- **AND Operator :**

P (FA) = 0.03
P (FR) = 0.68

The tests carried out confirm not only the importance of biometric fusion but also the robustness and the effectiveness of the new system which makes its appearance much more through the real tests where the one modal systems had a fall of performances.

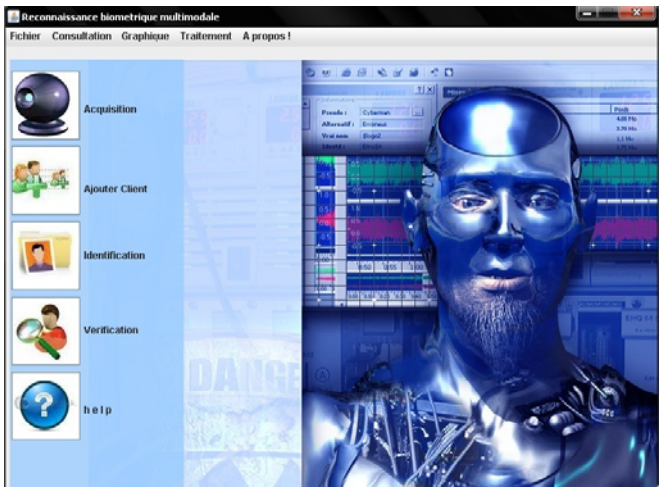
We noticed that Fusion give better results than those obtained by the first system.

We also noticed that the performances are closely related to the number of coefficients taken and the number of GMM. Finally we could say that the significant factor is the size of the base.

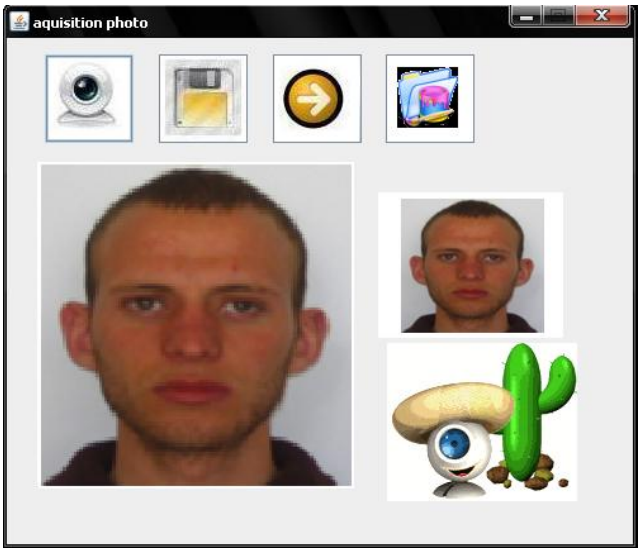
6. Demonstration System

In the following we present some interfaces of our Multi-Modal Recognition system which was developed using a Pentium IV cadenced at 2 Ghz and using 1 Giga bytes of RAM. It was running under Windows XP professional edition and using Java 1.6 as programming language.

1. Main Interface

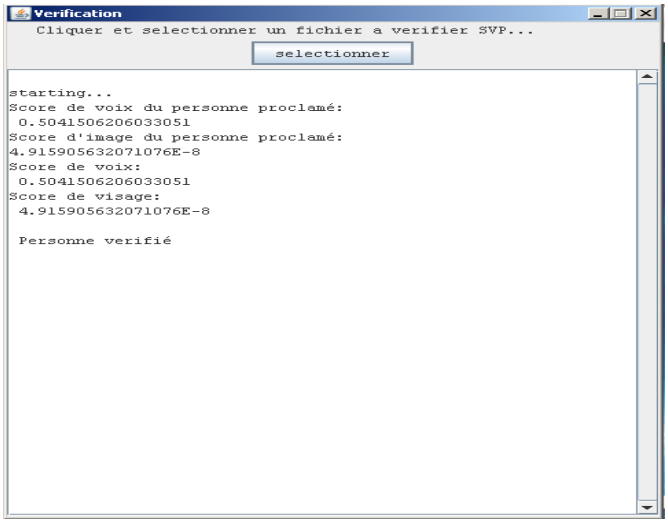


2. Acquisition Module for Face

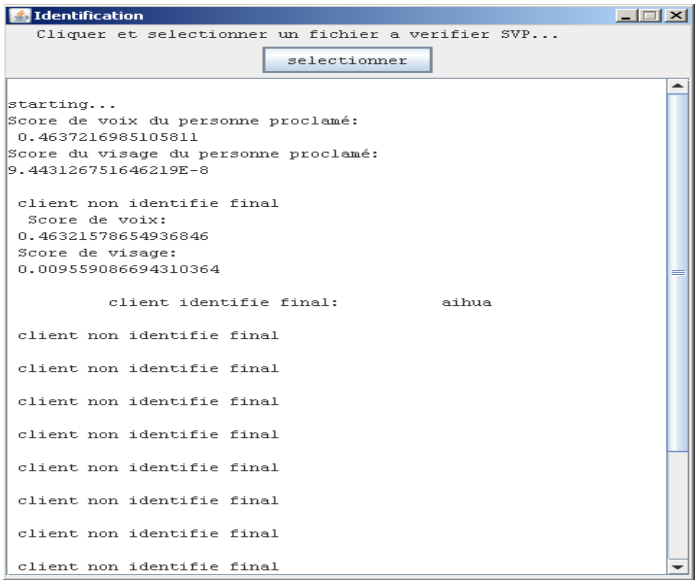


3. Acquisition Module for Speaker

4. Verification Process



5. **Identification Process**



7. Conclusions

This paper provides results obtained on a multi-modal biometric system that uses face and voice features for recognition purposes. We used fusion at the decision level with OR and AND operators. We showed that the resulting system (multi-modal) considered here provide better performance than the individual biometrics. For the near future we are collecting data corresponding to three

biometric indicators - fingerprint, face and voice in order to conceive a better multi-modal recognition system.

Acknowledgments

Special thanks to Benterki Mebarka and Bechane Louiza for their contribution to this project.
Samir Akrouf thanks the Ministry of Higher Education for the financial support of this project (project code: B*0330090009) .

References

[1] A. K. Jain, R. Bolle, and S. Pankanti, Biometrics: Personal Identification in Networked Society. Boston, MA: Kluwer, 1998.

[2] A. K. Jain, S. Prabhakar, and S. Chen, \Combining multiple matchers for a high security fingerprint verification system," Pattern Recognition Letters, vol. 20, pp. 1371-1379, 1999.

[3] R. Brunelli and D. Falavigna, "Person identification using Multiple cues," *IEEE Trans. Pattern Anal. Machine Intell.,* vol. 17, pp. 955–966, Oct. 1995.

[4] B. Duc, G. Maitre, S. Fischer, and J. Big¨un, "Person Authentication by fusing face and speech information," in *1st Int. Conf. Audio- Video- Based Biometric Person Authentication AVBPA'97,* J. Big¨un, G. Chollet, and G. Borgefors, Eds. Berlin, Germany: Springer-Verlag, Mar. 12–14, 1997, vol. 1206 of Lecture Notes in Computer Science, pp. 311–318.

[5] E. Big¨un, J. Big¨un, B. Duc, and S. Fischer, "Expert conciliation for multi modal person authentication systems by Bayesian statistics," in *Proc. 1st Int. Conf. Audio-Video-Based Biometric Person Authentication AVBPA'97.* Berlin, Germany: Springer-Verlag, Lecture Notes in Computer Science, 1997, pp. 291–300.

[6] L. Hong and A. K. Jain, "Integrating faces and fingerprint for Personal identification," *IEEE Trans. Pattern Anal. Machine Intell.,* vol. 20, 1997.

[7] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On Combining classifiers," *IEEE Trans. Pattern Anal. Machine Intell.,* vol. 20, pp. 226–239, 1998.

[8] A. K. Jain, L. Hong, and Y. Kulkarni, "A multimodal biometric system using fingerprints, face and speech," in *Proc. 2nd Int. Conf. Audio-Video Based Biometric Person Authentication,* Washington, D.C., Mar. 22–23, 1999, pp. 182–187.

[9] T. Choudhury, B. Clarkson, T. Jebara, and A. Pentland, "Multimodal person recognition using unconstrained audio and video," in *Proc. 2nd Int. Conf. Audio-Video Based Person Authentication,* Washington, D.C., Mar. 22–23, 1999, pp. 176–180.

[10] S. Ben-Yacoub, "Multimodal data fusion for person authentication using SVM," in *Proc. 2nd Int. Conf. Audio-Video Based Biometric Person Authentication,* Washington, D.C., Mar. 22–23, 1999, pp. 25–30.

[11] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Science*, pages 71–86, 1991.

[12] Ronny Tjahyadi, Wanquan Liu, Svetha Venkatesh. Application of the DCT Energy Histogram for Face Recognition. 2nd International Conference on Information Technology for Application (ICITA 2004) PP 305-310

[13] Samir Akrouf, Sehili Med Amine, Chakhchoukh Abdesslam, Messaoud Mostefai and Youssef Chahir 2009 Fifth International Conference on Mems Nano and Smart Systems 28-30 December 2009 Dubai UAE.

[14] N Morizet, Thomas Ea, Florence Rossant, Frédéric Amiel Et Amara Amara, Revue des algorithmes PCA, LDA et EBGM utilisés en reconnaissance 2D du visage pour la biométrie, Tutoriel Reconnaissance d'images, MajecStic 2006 Institut Supérieur d'Electronique de Paris (ISEP).

[15] Akrouf Samir, Mehamel Abbas, Benhamouda Nacéra, Messaoud Mostefai An Automatic Speaker Recognition System, 2009 the 2nd International Conference on Advanced Computer Theory Engineering (ICACTE 2009) Cairo, Egypt September 25-27 2009

[16] Approche Statistique pour la Reconnaissance Automatique du Locuteur : Informations Dynamiques et Normalisation Bayesienne des Vraisemblances ″, October, 2000.

[17] Yacine Mami "Reconnaissance de locuteurs par localisation dans un espace de locuteurs de référence" Thèse de doctorat, soutenue le 21 octobre 2003.

Samir Akrouf was born in Bordj Bou Arréridj, Algeria in 1960. He received his Engineer degree from Constantine University, Algeria in 1984. He received his Master's degree from University of Minnesota, USA in 1988. Currently; he is an assistant professor at the Computer department of Bordj Bou Arréridj University, Algeria. He is an IACSIT member and is a member of LMSE laboratory (a research laboratory in Bordj Bou Arréridj University). He is also the director of Mathematics and Computer Science Institute of Bordj Bou Arréridj University. His main research interests are focused on Biometric Identification, Computer Vision and Computer Networks.

Yahia Belayadi was born in Bordj Bou Arréridj, Algeria in 1961. He received his Engineer degree from Setif University Algeria in 1987. He received his magister from Setif University Algeria in 1991. Currently; he is an assistant professor at the Computer department of Bordj Bou Arréridj University, Algeria. He also is the director of University Center of Continuous Education in Bordj Bou Arreridj.

Messaoud Mostefai was born in Bordj Bou Arréridj, Algeria in 1967. He received his Engineer degree from Algiers University, Algeria in 1990. He received a DEA degree en Automatique et Traitement Numérique du Signal (Reims - France) in 1992. He received his doctorate degree en Automatique et Traitement Numérique du Signal (Reims - France) in 1995. He got his HDR Habilitation Universitaire **:** Theme : « Adéquation Algorithme /Architecture en traitement d'images » in (UFAS Algeria) in 2006. Currently; he is a professor at the Computer department of Bordj Bou Arréridj University, Algeria. He is a member of LMSE laboratory (a research laboratory in Bordj Bou Arréridj University). His main research interests are focused on classification and Biometric Identification, Computer Vision and Computer Networks.

Youssef Chahir is an Associate Professor *(since '00) at* GREYC Laboratory CNRS UMR 6072, Department of Computer Science, University of Caen Lower-Normandy France.