SIC_DB: Multi-Modal Database for Person Authentication

Charles Beumier, Marc Acheroy

Abstract

This paper presents a multi-modal database intended for person authentication from multiple cues. It currently contains three sessions of the same 120 individuals, providing profile and frontal color images, 3-D facial representations and many french and some english speech utterances. People were selected for their availability so that new sessions will be easily acquired. Individual recognition performances for speech, profile and 3D facial surface modalities are presented. The combination of these experts is the subject of current research.

1 Introduction

The *SIC_DB* database has been acquired at the SIC (*Signal & Image Center*) of the Royal Military Academy of Belgium. This center is dealing with image and speech processing activities, and more particularly, with person authentication from visual and acoustical clues. For this specific activity, three persons are currently working in face recognition, speaker verification and fusion (combination of different verification experts).

In 1991, a first activity in face recognition was started by considering the grey-level analysis of frontal images. Automatization, robustness and speed were criteria which were difficult to be satisfied as lighting conditions, pose, mood and hair may induce large changes. In 1993, we turned to profile analysis [2], taking advantage of the 1-Dimensional nature of the external contour and the rather high rigidity of the parts involved (forehead, nose, chin). The success of the profile approach led us to consider a more complete geometrical representation of a face obtained from a 3D acquisition of the facial surface. This can easily solve the problem of pose, scale and face detection but at the cost of a 3D acquisition system. In 1995, we joined the starting M2VTS [1] consortium about Multimodal Verification for Teleservices and Security Applications. This European project of the ACTS programme aimed at developing, integrating and fusing biometric experts (using speech, face, profile, 3D) for security in building or teleservice applications. Within this project, we first built a prototype for automatic profile identification in real-time [3]. Then we concentrated on the 3D comparison of facial surfaces. We also participated to the important effort devoted to expert fusion [7].

In 1992, a text-dependent speaker recognition method based on HMM (Hidden Markov Model) was implemented and later improved by introducing the fundamental frequency (pitch). Since 1994, we have been studying the possibilities of speaker recognition on vocoder links. We developed a new text-independent speaker recognition method based on histogram classifiers. Simple and quick, this method gives very promising performances. It was used as a new objective test to measure the quality of a vocoder.

For speech as well as for image analysis, we oriented our research towards the development of a system with practical constraints, such as a limited response time, a normal user presentation and a maintainable software with sufficient recognition performance. The multi-modality concept, by combining several sources of information, meets most of these requirements. It first allows to keep each modality simple for maintainance and cost reasons. It also enables better performances using the additional information. Computation time might not increase much, as modalities can be kept simpler, can be processed in parallel or can be used sequentially, stopping as soon as the decision becomes clear enough.

The problem of a multi-modal approach is the need for the manifold sources of information required by the modalities. The fusion engine needs multi-modal data of each person for learning. Several practical considerations led us to create our own multi-modal database. First, there was no available multi-modal database including 3D data. Secondly, we had to test the system we developed for quick 3D acquisition. Thirdly, we wanted to have the possibility to acquire new sessions rapidly.

Another multi-modal database is now available. XM2VTSDB [6], by-product of the M2VTS consortium, contains 4 sessions of 2 shots of 295 people, including voice, frontal and profile. A high quality 3D shot has been acquired during one of the sessions.

2 Database content

2.1 Database description

To be worth the effort, we imposed a minimum of 100 people to be present in the database. On the one hand, we decided to go beyond the common size (30-40) of our previous experiments. On the other hand, we tried to keep a reasonable size to be able to have many sessions. Indeed, we selected people among the academic population likely to stay in our reach for several years. This will allow to make long term studies about facial and voice characteristics.

This database is intended to be used for experiments in cooperative situations. The person wants to be recognised and obeys to given guidelines. Pictures are taken in the sitting attitude to reduce the height and position variation and the microphone distance. People are asked to gaze in some direction.

Two sessions were first acquired: one in November 97 and the second in January 98. They both contain the same 120 individuals. Only a few of them were not correctly represented in some shots. 80 persons are students from the Military Academy. The remaining 40 people come from the teaching staff. From this population, only 14 women were present. Hair is most of the time short, there are few black and no asian people. A third shot was captured in May 99, consisting of 100 individuals from the 120 original ones.

2.2 Speech

The speech part of the database is sampled at 8 kHz from a high-quality microphone (ref SENNHEISER MD441). The samples are encoded on 16-bit. The recording environment is a working room with some background noise (computers and projector ventilators). One purpose of this speech database is to support all kinds of speaker verification methods (text-dependent, prompted-text, text-independent, verbal information). We intend to study the long term statistics of speech parameters and the influence of noise.

The 110 French and 10 Dutch native speakers were asked to pronounce a series of sample utterances in French and English.

The nine utterances in French are:

- the full name and address.
- the birthday date.
- the phone number.
- the digit sequence: 0-1-2-3-4-5-6-7-8-9.
- a session and speaker dependent sequence of digits.
- the sentence "Le menuisier a scié une planche et l'a rabotée"
- A session independent, speaker dependent sentence.
- A session and speaker dependent sentence.
- A (free text) description of a speaker independent, session dependent picture.

The three utterances in English are:

- The digit sequence: 0-1-2-3-4-5-6-7-8-9.
- The digit sequence: 5-0-6-9-2-8-1-3-7-4.
- The sentence "Joe took father's green shoe bench out"

A file attached to each speaker contains the literal transcription of all sample utterances pronounced during the session. The phone number is limited to 9 digits. Its pronunciation is imposed digit per digit. The birthday date is spoken freely (for instance, 04-07-69 or Seven April 69). The variable sentences have been chosen in order to have the same duration (+/- 2 second) and to present a lot of different voiced phonemes. The English part offers the possibility to study the influence of the language on the speaker recognition methods.

2.3 Frontal Face

Frontal face images were captured by a color camera (JVC, model KY-F55BE) and digitized by the Matrox Meteor frame grabber in full (768x576) resolution. People were asked to sit on a chair, look at the camera and center themselves. After the first shot, people were asked to move a little and if relevant, to put their glasses on. A second shot was then taken.

We selected for a blue background to ease face detection. This rather uniform background also allowed to save disk space by storing images in the 'gif' format (file size are typically 300 Kb instead of 1.3 Mb).

The camera was rotated 90^o to get the maximal resolution. First, the head fits better in a portrait mode, allowing for a larger scale. Secondly, important details about the mouth and eye regions are made of horizontal edges. These edges are better captured if they are vertical in the image, because the horizontal bandwidth of the camera/digitizer pair is higher.

The quality of the images is not optimal. Indeed, balanced lighting without reflexion is not easy so we preferred to use the ambient light. The gain of the camera had to be very high, resulting in high noise. However, the quality is sufficient to get most of the facial information.

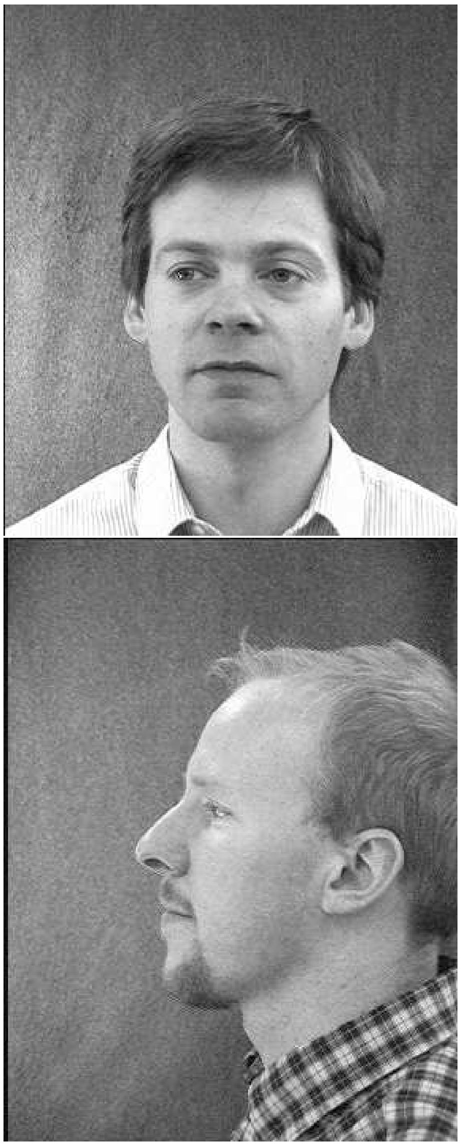

Figure 1. a) Frontal image (here in grey) b) Profile image (here in grey)

2.4 Profile

Profile acquisition followed the same guidelines as frontal image capture. The same camera/digitizer pair was used, resulting in the same resolution. People were asked to present their profile by looking towards a reference direction. One profile image was shot, and a second for people wearing glasses.

The blue background was here more critical as profile contours are to be extracted automatically from the images. Files were also stored in the 'gif' format.

The camera was rotated 90^o to benefit from the higher resolution of portrait images for the head. We kept the same camera and light settings as for frontal images, so that the noise is important but the external contour of the face can be easily delineated from the hue information.

2.5 3D Facial surface

Much more than for frontal and profile images, 3D acquisition and 3D analysis needed their own database. It was the opportunity to test the hardware prototype we developed to capture 3D facial surface very quickly (in less than 1 second). It is based on the projection of a known pattern of light ("structured light"). The 3D database also allowed to perform recognition tests at a sufficiently large scale.

In accordance with conditions of use of the 3D acquisition system [4], people were asked to sit on a chair positioned at correct distance to ensure sharp images. The black and white camera (Panasonic WV-BL600) coupled with the Matrox Meteor frame grabber delivered 768x576 grey images. For each person, the first image was shot without stripe projection, to give a texture image of a nearly frontal pose. This image was rather dark as ambient light was used to have balanced illumination. The second image was taken directly after, with the projector switched on. The second and third striped images (with projection) were then shot, with limited (up to 15^o) left/right and up/down rotations. Each session has thus three striped shots and one grey shot in near correspondence with the first striped shot. Mention that these images were taken with a 40^o inclination of the camera/projector head to ease stripe detection in eyebrows, mouth and nose (see Fig. 2).

From those striped images, we extracted 3D information which consist of a set of 3D coordinates of points along the projected stripes. The coding limits the resolution to 0.1 mm, but the real quality is lower than this. Since the position of the face in the space is of little concern, the calibration procedure optimised relative distances so that the resolution is given in terms of relative distance. For an object of 15 cm, the error in measured distances after calibration is up to 5 mm, depending on its localisation in the field of view. A mean error for an object of this size centered on the

Figure 2. Samples from the database

image is 2 mm.

Because the stripe visibility depends on the underlying texture, 3D extraction is disturbed by beards, moustaches, opened mouth, eyes and hair. This may result in no 3D data or important noise.

3 Applications

3.1 Speech

To assess the quality of the speech database and the performances of the developed method, identification tests were carried out. The retained method was based on histogram classifiers [5]. The free speech data was used to train the system (about 15 second) and the three sentences (about 9 second) were used as test data.

The first session gave very poor results. We found a very important source of noise in the preamplifier used for the microphone. Solving the problem for the second session, identification results (Equal Error Rate: 6 %) were in accordance with expected results.

Figure 3. 3D reconstructions from the top left image of Fig. 2

3.2 Frontal Face

Frontal images of the databases have not been used yet.

3.3 Profile

The quality of the profile images has been studied by first detecting the profile contour (automatically). Based on color segmentation, thresholding on the blue component, a perfect binarization between the head (on the profile side) and the background was obtained for all the people of the first two sessions. The identification program ([3]), comparing session 2 profiles with session 1 ones, achieved an EER of 10.5 %, what is very good considering that no tuning was performed relative to the new database, all the persons were considered, only one image was used as reference and the two sessions were acquired two months apart.

3.4 3D Facial Surface

Two very important roles of the SIC_DB have been the validation of the automatic 3D acquisition system and the development of 3D comparison approaches.

The automatic 3D extraction procedure from striped images was validated on 720 images, performing good in situations where there is no bushy beard or moustache and no spectacles with thick frames. The precision of the data seems sufficient as far as recognition performances are concerned.

3D comparison experiments were driven on two similar approaches: the comparison of the full facial surfaces by the matching of parallel profiles extracted from the surfaces and the comparison of the central and lateral profiles. To perform full tests in different conditions, only 30 persons were considered from the database. Intra-session comparisons results in EER of about 8 % while inter-session give 10 %, for both methods. Errors come from acquisition noise and bad matching due to local minima. Manual refinement of the 3D extraction and of the matching procedure give EER of about 3 to 4 %.

3.5 Intended Fusion

One important goal of the development of this database is to dispose of sufficient data for fusion experiments. On the one hand, all the modalities need their own database. On the other hand, the different modalities must exist for each person for fusion purposes.

Expert combination will of course bring robustness. It will also enable us to see how modalities are correlated. For instance, the profile information should more or less be contained in the 3D modality. We can check how far this is true, remembering that the acquisition of those modalities is not the same.

Ideally, 3D will also integrate grey information, possibly overpassing frontal analysis, since pose and light normalization could be helped by 3D. The difference in authentication results of the frontal analysis and the grey analysis from 3D is a way to see if the grey analysis brings as much as it could.

4 Conclusions

The presented database was mainly developed for the analysis of authentication methods based on 3D and speech. Profile and frontal images are also provided but their quality is not perfect and their use in our lab is limited.

New sessions are expected to be acquired regularly, to get more data for tests, to see the influence of noise or microphone on the speech data and to evaluate possible improvements of the 3D acquisition system.

5 Acknowledgements

The author would like to thank Frédéric Jauquet for taking care of all speech aspects of the database. Many thanks to Alain Roufosse who designed the interface for speech prompting and to everybody who participated to the sessions. The work on the 3D part of the database has been supported by the project M2VTS of the European ACTS programme.

References

[1] Acheroy, M., Beumier, C., Bigün, J., Chollet, G., Duc, B., Fischer, S., Genoud, D., Lockwood, P., Maitre, G., Pigeon, S., Pitas, I., Sobatta, K., Vandendorpe, L., Multimodal person verification tools using speech and images, Proceedings of the European Conference on Multimedia Applications, Services and Techniques (ECMAST '96)(1996), 747-761.

[2] C. Beumier, M.P. Acheroy, “Automatic Face Identification”, In *Applications of Digital Image Processing XVIII*, SPIE, vol. 2564, July 1995, pp. 311-323.

[3] C. Beumier, M.P. Acheroy, “Automatic Profile Identification”, In *Audio- and Video-based Biometric Person Authentication*, Crans-Montana, Switzerland, 12-14 March 1997.

[4] C. Beumier, M.P. Acheroy, “Automatic Face Authentication from 3D Surface”, In *British Machine Vision Conference BMVC 98*, University of Southampton UK, 14-17 Sep, 1998, pp. 449-458.

[5] F. Jauquet, P. Verlinde, and C. Vloeberghs. “Histogram classifiers using vocal tract and pitch information for text-independent speaker identification”, In *Proc. ProRISC'97 "Circuits, Systems and Signal Processing"*, November 1997.

[6] K. Messer, J. Matas, and J. Kittler, “Acquisition of a large database for biometric identity verification”, in Proc. of BioSig 98, June 1998.

[7] P. Verlinde and G. Chollet, “Comparing decision fusion paradigms using k-NN based classifiers, decision trees and logistic regression in a multi-modal identity verification application”, Accepted for presentation in the *Second International Conference on Audio- and Video-based Biometric Person Authentication (AVBPA), Washington D.C.*, March 1999.