The M2VTS Multimodal Face Database (Release 1.00)

Stéphane Pigeon - Luc Vandendorpe
UCL - Laboratoire de Télécommunications et Télédétection
Place du Levant, 2 – B-1348 Louvain-La-Neuve – Belgium
E-Mail : pigeon@tele.ucl.ac.be

Abstract

The primary goal of the M2VTS project is to address the issue of secured access to buildings or multi-media services by the use of automatic verification based on multimodal strategies (secured access based on speech, face images and other information). This paper presents an overview of the multimodal face database recorded at UCL premises for the purpose of research applications inside the M2VTS project. This database offers synchronized image and speech data as well as video sequences allowing to access multiple views of a face. This material should permit the design and the testing of identification strategies based on speech and/or labial analysis, frontal and/or profile face analysis as well as 3-D analysis thanks to the multiple views. The M2VTS Database is available to any non-commercial user on request to the European Language Resource Agency. *Keywords : face, database, multimodal, M2VTS.*

1 Introduction

Among the different European ACTS projects, the M2VTS project (Multi Modal Verification for Telesevices and Security applications) deals with access control by the use of multimodal identification of human faces. The goal of using a multimodal recognition scheme is to improve the recognition efficiency by combining single modalities, namely face and voice features [1]. This requires the development of fusion methods, i.e. the merge of individual results given by separate analysis of different modalities (e.g. voice texture and face features) or, more efficiently, the direct analysis of a combination of different modalities (e.g. study of speech/lip synchronization). Due to the relative novelty of multi-modal identification, our own test material had to be recorded since no existing database could meet our requirements of offering all modalities needed by the multiple recognition tasks. These requirements are the presence of synchronized speech and image and the opportunity to extract 3-D face features from the database.

The constitution this multimodal database required many efforts and a lot of time to be grabbed and edited. As the material being recorded couldn't be found in any other public database, it has been decided to distribute the M2VTS database to any non-commercial user that could take benefit from such a product. This distribution has been taken in charge by the *European Language Resource Agency* (see section 4).

2 Overview of the M2VTS Database

Our current database is made up from 37 different faces and provides 5 shots for each person. These shots were taken at one week intervals or when drastic face changes occurred in the meantime. During each shot, people have been asked to count from '0' to '9' in their native language (most of the people are French speaking), rotate the head from 0 to -90 degrees, again to 0, then to +90 and back to 0 degrees. Also, they have been asked to rotate the head once again without glasses if they wear any. From this whole sequence, 3 parts have been extracted : the *voice* sequence, the *motion* sequence and the *glasses off* motion sequence (if any). The first sequence can be used for speech verification, 2-D dynamic face verification (choosing the most appropriated picture out of the sequence) and for speech/lips correlation analysis. The other two sequences are meant for face recognition purposes only and provide information

about the 3-D face features thanks to the motion. They may be used to implement and compare other techniques like identification from 2-D facial pictures, profile view or multiple views. For each person belonging to the database, the most difficult shot to recognize is labeled as the 5th shot. These shots mainly differ from the others because of face variations (head tilted, eyes closed, different hairstyle, presence of a hat/scarf...), voice variations or shot imperfections (poor focus, different zoom factor, poor voice SNR...).

It was decided to use good quality material for the recording, leaving space in the future to degrade quality in order to simulate low-cost acquisition systems. A Hi8 video camera (576x720, 50Hz-interlaced, 4:2:2) was chosen for the shooting and a D1 digital recorder for the recording and editing. In order to reduce the storage requirement, television sequences are down-converted into CIF (288x360 pixels, 25Hz-Progressive, 4:2:2). This conversion removes one field out of two and performs an horizontal down-sampling in the remaining frame with respect to the MPEG-2 TM5 specification. By keeping active pixels only, the final resolution for the database images is 286x350 pixels. Concerning voice acquisition, the sound track is digitally recorded using a 48kHz sampling frequency and 16 bit linear encoding.

But for the particular case of the 5th shot, the database can be considered as having been produced under "ideal" shooting conditions (good picture quality, indoor shooting, nearly constant lightening, uniform grey background) and within a highly co-operative scenario (as much as they could, people followed the instructions they were given). Nevertheless, we can notice some impairments with respect to the theoretical case :

- some people do no rotate their head properly (horizontal translation of the head in the direction of the rotation, vertical tilt depending on the rotation angle, no full covering of the 180 frontal degrees...)
- some people might have their mouth open during one rotation of the head, closed during the other, ending up on different shapes in the profile view
- some people close their eyes while moving the head
- the direction of starting the rotation of the head is not fixed over the different shots
- some people are speaking very low, resulting in a poor sound SNR
- some people can not keep from smiling during the shot
- rotation speed can be highly variable between different shots, but also within the same shots.
- reflections on eyes and glasses
- blurry images during fast head rotation, due to limited shutter speed

However, similar imperfections – combined with other as well – will appear when implementing a practical recognition scheme. Moreover people will expect the recognition algorithms to be able to deal with such imperfections. From this point of view, the M2VTS Database can be seen as a good material to test the robustness of the recognition algorithms with regards to common problems. Assuming an algorithm would not overcome the imperfections encountered here, it would be difficult for this algorithm to overcome those associated with true operational conditions.

Further information can be found in [4]. Images of the M2VTS Database can be viewed from [5].

3 Experimentation procedure as defined within the M2VTS Project

Up to now, only four shots of our database have been used during our current experimentations [2, 3], leaving the 5th shot for future experiments (as reminder, the 5th shot is made up of the most difficult cases to recognize due to particular shooting conditions).

The procedure for experimentation chosen inside M2VTS follows the "leave one out" principle. One *experiment session* will use a *training* and a *test* database. The *training* database is built of 3 shots (4 are

available) of 36 persons (37 available). The *test* database is built of the last (4th) shot of the remaining person and the left-out shots of the 36 persons not present in the training database.

The training database is used to build a *reference* model for each client. The performance of the identification algorithms is evaluated by matching the 37 candidate persons (36 clients and 1 imposter) from the test database with the 36 reference clients. Such an experiment session provides 36 *authentic* and 36 *imposture* tests. An authentic test consists of candidate claims which are true. An imposture test consists of candidate claims which are false.

There are 4*37 experiment sessions by leaving out one person and one shot. The order of the experiment sessions is determined by the left-outs, and will be as follows: first, the shot number will be varied in decreasing order, starting with the last one, and then the person in decreasing alphabetic order of the shot label, starting with the last one.

4 How to acquire a copy of the M2VTS Database?

A copy of the M2VTS Database (release 1.00) can be requested through the *European Language Resource Agency* :

European Language Resource Agency - Distribution Agency (ELDA)
87, Avenue d'Italie — F-75013 Paris — FRANCE
Phone : +33 1 45 86 53 00 Fax : +33 1 45 86 44 88
E-mail : *elra@calvanet.calvacom.fr*
WWW : *http://www.icp.grenet.fr/ELRA/home.html*

Please note that, although we were intending to distribute this database for free, a distribution fee is asked by ELRA in order to compensate for the distribution costs. As this cost is a cost price only - no benefits are expected from the distribution - we kindly ask the end-users to acknowledge the M2VTS project whenever its database is used (see section 5).

5 End-user agreement

The use of the M2VTS is restricted to non-commercial applications only. Commercial applications including - but not restricted to - the distribution of the M2VTS Database as part of reseller's own product or the marketing the M2VTS Database in a modified form, are strictly forbidden.

On the other hand, authors are encouraged to use the M2VTS database in their contributions in the field of face, speech or multimodal person authentication. However, we kindly ask the end-user to acknowledge the M2VTS project whenever the M2VTS database is used inside a publication. Also, a clear link to the M2VTS Database CEC Deliverable [4] or the M2VTS Database WWW Site [5] should be put in the references of the paper, allowing interested people to access further information about the M2VTS Database.

Acknowledgments

This work has been performed within the framework of the M2VTS Project granted by the European ACTS programme.

M2VTS Partners are : Matra Communication (France), Ibermatica SA (Spain), Cerberus AG (Switzerland), Aristotle University of Thessaloniki (Greece), Ecole Polytechnique Fédérale de Lausanne (Switzerland), Université Catholique de Louvain (Belgium), University of Surrey (UK), University of Neuchatel

(Switzerland), Renaissance (Belgium), Institut Dalle molle d'Intelligence Artificielle Perceptive (Switzerland), Unidad Tecnica Auxiliar de la Policia (Spain), Compagnie Européenne de Télésécurité (France), Banco Bilbao Vizcaya (Spain), Universidad Carlos III (Spain).

References

[1] M. Acheroy, C. Beumier, J. Bigün, G. Chollet, B. Duc, S. Fischer, D. Genoud, P. Lockwood, G. Maitre, S. Pigeon, I. Pitas, K. Sobottka and L. Vandendorpe. "Multi-Modal Person Verification Tools using Speech and Images". In *Proceedings of the European Conference on Multimedia Applications, Services and Techniques (ECMAST '96)*, pp. 747-761, Louvain-La-Neuve, Belgium, May 28-30 1996.

[2] P. Jourlin, J. Luettin, D. Genoud and H. Wassner. "Acoustic-Labial Speaker Verification". Submitted to *The First International Conference on Audio- and Video-based Biometric Person Authentication (AVBPA '97)*, Crans-Montana, Switzerland, March 12-14, 1997.

[3] S. Pigeon and L. Vandendorpe. "Profile Authentication Using a Chamfer Matching Algorithm". Submitted to *The First International Conference on Audio- and Video-based Biometric Person Authentication (AVBPA '97)*, Crans-Montana, Switzerland, March 12-14, 1997.

[4] S. Pigeon, "The M2VTS Multimodal Face Database (Release 1.00)", *CEC ACTS/M2VTS Deliverable no. AC102/UCL/WP1/DS/P/161*, October 1996. A postscript version of this document can be downloaded from [5].

[5] The M2VTS Database WWW site : *<http://www.tele.ucl.ac.be/M2VTS/>*