XM2VTSDB: The Extended M2VTS Database

K. Messer, J. Matas, J. Kittler and K. Jonsson
University of Surrey, Guildford, Surrey, GU2 5XH, UK.

Abstract

In this paper we describe the acquisition and content of a large multi-modal database intended for training and testing of multi-modal verification systems. The XM2VTSDB database offers synchronised video and speech data as well as image sequences allowing multiple views of the face. It consists of digital video recordings taken of 295 hundred subjects at one month intervals taken over a period of five months. We also describe a protocol for evaluating verification algorithms on the database. The database has been made available to anyone on request to the University of Surrey through http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb.

1 Introduction

The use of biometric measurements in security applications is becoming common to a level where a dedicated journal [1] monitors the developments in the area. Extremely reliable methods of biometric personal identification exist, e.g. fingerprint analysis, retinal or iris scans. But most of these methods are considered unacceptable by users in all but high-security scenarios. Personal identification systems based the on analysis of speech, frontal or profile images of face are non-intrusive and therefore user-friendly. Moreover, personal identity can be often ascertained without client's assistance. However, the speech and image-based systems are less robust to imposter attack, especially if the imposter possesses information about a client, eg. a photograph or a recording of client's speech. Multi-modal personal verification is one of the most promising approaches to user-friendly (hence acceptable) highly secure personal verification systems [6].

Recognition and verification system need training; the larger the training set, the better the performance achieved [8]. The volume of data required for training a multi-modal system based on the analysis of video and audio signals is in the order of TBytes (1000 GBytes); technology allowing manipulation and effective use of such amounts of data has only recently become available in the form of digital video.

The M2VTS project was set up to address the problem of secured access to buildings or multi-media services by the use of automatic person verification based on such multi-modal strategies. In order for the project partners to reliably and robustly train, test and compare their algorithms a large multi-modal database was required. We are at present aware of only two publicly available medium or large scale multi-modal databases, the database collected within the M2VTS project, comprising 37 subjects [4] and the DAVID-BT database [2]. A survey of audiovisual databases prepared by Chibelushi et. al. list many others, but these are either mono-modal or small [7] . From the point of view of the database size DAVID-BT is comparable with the M2VTS database: 31 clients - 5 sessions. However, the speech part of DAVID-BT is significantly larger than that of M2VTSDB. On the other hand - the quality and reproducibility of the data available on an SVHS tape is low. It is for these reasons that it was decided to capture a large audio-visual database, the XM2VTSDB, using high quality digital video.

The paper is organised as follows. In the next section we define the database specification. The database acquisition system used is described in section 3. In Section 4 the content of the speech shot is described. Then the content of the head rotation shot is presented. Information about the XM2VTSDB protocol designed for training and testing personal identity verification algorithms is given in Section 6. In Section 7 we give information on how the database is distributed before reaching some conclusions.

2 Database Specification

The design of XM2VTSDB was based on the experience gained as a result of recording and experimenting with the M2VTS database. The database is primarily intended for research and development of personal identity verification systems where it is reasonable to assume that the client will be cooperative. The biometric data to be recorded is of the type that would normally be easily acquired during a normal access claim intercourse between an access point system and the client. The system may request from the client

some client specific information. Generally it will engage the user in a simple dialogue and request simple tasks to be performed which will introduce some subject dynamics into the intercourse session from which useful image sequence information can be extracted.

The scenario adopted reflected the above considerations. We assumed that a dialogue of some 30 secs duration would be perfectly acceptable in practical situations. The dialogue was simulated by asking the subjects to utter a predefined sentence. Each subject was also asked to move his/her head as might be necessary for instance to read some notice or instructions. Our objective was to induce the subjects to make extreme head rotation movements, so that we can also extract head side profile. In an operational scenario a side view camera could be used to capture this kind of biometric information instead.

In order to capture natural variability of clients caused by changes in physical condition, hair style, dress, and mood, subjects were recorded in four separate sessions uniformly distributed over a period of 5 months.

A continuous video recording was made of each subject, rather than a few snapshots from each recording session, as video data not only facilitates certain image processing tasks such as head segmentation, eye detection, but most importantly, it is a source of multiple biometric modalities. These include lip dynamics and face 3D surface modelling. Continuous video also supports verification of speech/lip shape correlation and speech/lip signal synchronisation.

The subjects were selected to include adults of both sexes and of different ages. As people wearing glasses may be interested in gaining access to services with glasses on or off, both instances would have to be present to develop robust algorithms.

A good quality consumer market digital camcorder was used to record the database. This particular choice had been made on the grounds that the camera system in a final product would have to be of low cost. The state of the art consumer products today will be low cost products tomorrow. With good quality recordings one can easily perform experiments on lower quality video which can be obtained by various processes of degradation (blurring, noise contamination, colour distortion, decrease in spatial and temporal resolution, reduced dynamic range and grey level resolution).

These design principles are similar to those adopted for the M2VTS database. The main difference between XM2VTSDB and M2VTS is in the size of the database and in the number of recordings taken for each subject during each session. The size of the M2VTS database (37 subjects) was reasonably representative of some application scenarios. However, even for client populations of such moderate size, any imposter tests should be carried out on a significantly larger database. There are also applications where the database of clients would be of the order of hundreds rather than tens. These considerations led to the conclusion that an order of magnitude increase in size of the M2VTS database was warranted. In view of the huge quantity of data, a target of some 300 subjects was thus aimed at.

Whereas M2VTS database comprised 5 different shots at distinct sessions recorder over a period of three months, the new database is constituted by 8 shots recorded in four distinct sessions. Thus each session contains two repetitions of the specified sentence. The main motivation for this recording policy was to increase the number of speech records for each subject to facilitate verification algorithm development and fusion. The design and training of an algorithm requires data from several records to encapsulate intra client variability. Further independent data may be required for feature selection and for the design of the supervisor in multiple expert fusion. It is assumed that the repetitions of the speech utterance in each session would be sufficiently different for them to be considered as independent records for experimental purposes. The difference would result not only from the natural variability that might be exhibited even under identical conditions, but also due to different emotional states of the subjects during the two consecutive attempts.

The database acquisition commenced with a population of 360 volunteers but through natural wastage only 295 completed the four sessions. 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 second part consisted of a head rotation shot. The data acquired during these two shots will be described in more detail in Sections 4 and 5 respectively.

3 The Database Acquisition System

The entire database was acquired using a Sony VX1000E digital cam-corder and DHR1000UX digital VCR. This captures video at a colour sampling resolution of 4:2:0 and 16bit audio at a frequency of 32kHz. The video data is compressed at the fixed ratio of 5:1 in the proprietary DV format. This format also defines a frame accurate timecode which is stored on the cassette along with the audio visual data.

This video hardware can be interfaced to a computer via a firewire(IEEE 1394) [3] port. We used an Intel based 586 PC running Windows 95 and connected it to the digital video equipment using an Adaptec AHA8940 firewire card. Software utilities were then written that enable a user to remotely control the VCR to frame accuracy searching through the stored timecodes on the cassettes. Routines were also written that allowed the capture of both video and audio data in real time to the computer hard disk.

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.

Before each video shot was recorded a short clipperboard sequence was taken that uniquely identified that shot. This clipperboard contained the subject unique identification number, the subjects name, shot type and session number. Also on the clipperboard was a colour test chart and resolution checker chart. This enables a check to be made that the quality of the recordings is consistent across the whole database and could help resolve any potential errors.

The raw database contains approximately 30 hours of digital video recordings. This has all been manually annotated. Every subject has an index file for each of the four recording sessions which contain the tape number and timecodes for selected key points in the speech and video data.

Using the information in these index files and the written software enable us to index into the database and automatically retrieve any subset of the database and enable us to automatically produce edited versions of the database.

4 The Speech Shot

After a short clipperboard sequence was recorded the subject was asked to sit in chair and a microphone was clipped onto their shirt. He/she was then asked to read three sentences which were written on a board positioned just below the camera. The subjects were asked to read at their normal pace, to pause briefly at the end of each sentence and to read through the three sentences twice. The three sentences remained the same throughout all four recording sessions and were

1. "0 1 2 3 4 5 6 7 8 9"
2. "5 0 6 9 2 8 1 3 7 4"
3. "Joe took fathers green shoe bench out"

The digits in the second sentence are in the same order as another large speech database whilst the third sentence was chosen because it is phonetically balanced.

Figures 2(a)-(d) show an image grabbed for a subject from each session. This image data can be used to train and test algorithms for frontal view authentication. Figures 2(e)-(h) show a sequence of images grabbed from the video taken at the first session during the speech shot. These sequences can be used to train and test lip-tracking systems. All the audio data from this shot have been grabbed and placed into audio files with each file containing a single sentence. This data can be used to train and test speaker verification and recognition algorithms.

5 The Head Rotation Shot

The next shot consisted of a sequence of rotating head movements. After the clipperboard shot the subject was asked to rotate his/her head from the centre to the left, to the right, then up, then down, finally returning it to the centre. They were told that a full side-profile was required and asked to run through the entire sequence twice.

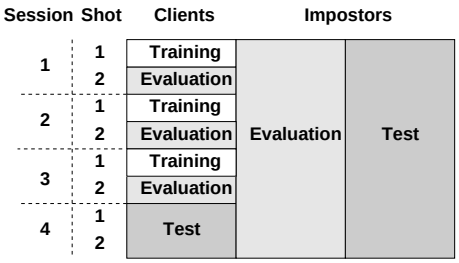
Figures 3(a)-(h) show selected frames from this sequence. This sequence was kept constant for all four sessions. These images can be used for profile or 3D based authentication.

Next, if the subject was wearing glasses he/she was asked to remove them and a short front profile video sequence was filmed. In total about 1.5 minutes of digital video was taken per subject, per session.

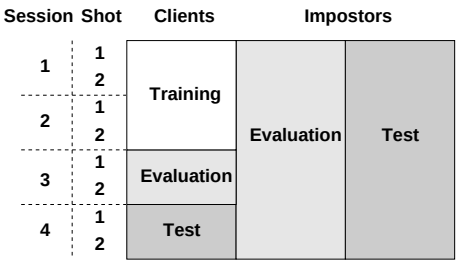
6 Evaluation Protocol

The evaluation protocol was specified to allow objective evaluation of vision- and speech-based person authentication systems on the extended M2VTS database. It is designed for the task of person verification (as opposed to recognition), where an individual asserts his identity. The verification system compares the features of the person with stored features corresponding to the claimed identity and computes their similarity, which is referred to as a score. Depending on the score, the system decides whether the identity claim is genuine or not. This authentication task corresponds to an "open-universe scenario" where persons unknown to the system may claim access. The subjects whose features are stored in the system's database are referred to as clients whereas persons claiming false identity are called impostors.

The database is divided into three sets: the training, evaluation and test sets (see Figure 1). The train-



(a) Configuration I



(b) Configuration II

Figure 1: The partitioning of the extended M2VTS database according to configuration (a) I and (b) II of the protocol.

ing set is used to build client models, the evaluation set to establish client-specific verification thresholds, and the test set to obtain estimates of the true verification rate on independent data. The decision whether a person should be accepted or rejected is determined from the input score and the verification threshold of the claimed identity. The threshold can be set to satisfy certain performance levels on the evaluation set. In the case of multi-modal classifiers, the evaluation set may also be used to optimally combine the outputs of several single-modal classifiers. As mentioned in the previous section, the database contains 295 subjects recorded in 4 different sessions and 2 shots (repetitions) per session. The database was randomly partitioned into 200 clients, 25 evaluation impostors and 70 test impostors Two different evaluation configurations were defined. They differ in the distribution of client training and evaluation data as can be seen in Figure 1.

6.1 Performance measures

The two error measures of a verification system are the false acceptance (FA) and false rejection (FR) rates. False acceptance is the case when an impostor, claiming the identity of a client, is accepted. In contrast, false rejection is the case when a client, claiming his true identity, is rejected. The FA and FR rates are given by:

$$\mathrm{FA} = n_i/m_i \qquad \mathrm{FR} = n_c/m_c \tag{1}$$

where n_i is the number of impostor acceptances, m_i the number of impostor claims, n_c the number of client rejections, and m_c the number of client claims. Both FA and FR are influenced by the verification thresholds. There is a trade-off between the two error rates, i.e. it is possible to reduce either of them with the risk of increasing the other one. For the test sets of both protocol configurations, m_i is 112000 (70 impostors $\times$ 4 sessions $\times$ 2 shots $\times$ 200 clients) and m_c is 400 (200 clients $\times$ 1 session $\times$ 2 shots).

The performance of a verification system is often given by the equal-error rate (EER). The EER can be obtained after a full authentication experiment has been performed. The true identities of the test subjects are then used to calculate the client thresholds for which the FA and FR are equal. Therefore, the EER does not correspond to a real authentication scenario and the predicition of the system performance may be inaccurate. In practical applications the thresholds need to be set a priori. An important measure of the performance of a system is therefore the deviation of the FA/FR test distribution from the corresponding distribution on the evaluation set. This is particularly the case for applications where the FA or FR are constrained to lay within certain limits. We therefore need to consider the distributions of FA and FR in addition to their (possibly weighted) sum.

Since we are interested in simulating real-world applications, we set the client thresholds on the evaluation data to obtain certain false acceptance (FAE) and false rejection (FRE) values. The same thresholds are then used on the test set. In a given application, there might be performance constraints which impose upper limits on the FA and FR rates. In the extended M2VTS protocol, it is suggested that the following three thresholds corresponding to FAE = 0, FRE = 0 and FAE = FRE should be used for evaluation:

$$\begin{aligned} \mathrm{T}_{\mathrm{FAE}=0} &= \arg\min_{\mathrm{T}}(\mathrm{FRE}|\mathrm{FAE}=0) \\ \mathrm{T}_{\mathrm{FAE}=\mathrm{FRE}} &= (\mathrm{T}|\mathrm{FAE}=\mathrm{FRE}) \\ \mathrm{T}_{\mathrm{FRE}=0} &= \arg\min_{\mathrm{T}}(\mathrm{FAE}|\mathrm{FRE}=0) \end{aligned} \tag{2}$$

Thus, there are 6 different scores associated with one experiment:

$$\begin{array}{ll} \mathrm{FA}_{\mathrm{FAE}=0} & \mathrm{FR}_{\mathrm{FAE}=0} \\ \mathrm{FA}_{\mathrm{FAE}=\mathrm{FRE}} & \mathrm{FR}_{\mathrm{FAE}=\mathrm{FRE}} \\ \mathrm{FA}_{\mathrm{FRE}=0} & \mathrm{FR}_{\mathrm{FRE}=0} \end{array} \tag{3}$$

For each threshold, the weighted error rate (WE) can be obtained as follows:

$$\begin{aligned} \mathrm{WE}_{\mathrm{FAE=0}} &= w_{\mathrm{FA}} \cdot \mathrm{FA}_{\mathrm{FAE=0}} + w_{\mathrm{FR}} \cdot \mathrm{FR}_{\mathrm{FAE=0}} \\ \mathrm{WE}_{\mathrm{FAE=FRE}} &= w_{\mathrm{FA}} \cdot \mathrm{FA}_{\mathrm{FAE=FRE}} + w_{\mathrm{FR}} \cdot \mathrm{FR}_{\mathrm{FAE=FRE}} \\ \mathrm{WE}_{\mathrm{FRE=0}} &= w_{\mathrm{FA}} \cdot \mathrm{FA}_{\mathrm{FRE=0}} + w_{\mathrm{FR}} \cdot \mathrm{FR}_{\mathrm{FRE=0}} \end{aligned} \tag{4}$$

The weights w_{FA} and w_{FR} are set depending on the relative importance of the false acceptance and rejection rates. If a general face verification and recognition system is used, without any specific application in mind we can weight the error rates equally, $w_{\mathrm{FA}} = w_{\mathrm{FR}} = 0.5$. Furthermore, if the error rates are obtained with the EER threshold $\mathrm{T}_{\mathrm{FAE=FRE}}$, i.e. $\mathrm{FA}_{\mathrm{FAE=FRE}}$, $\mathrm{FR}_{\mathrm{FAE=FRE}}$ and $\mathrm{WE}_{\mathrm{FAE=FRE}}$, then no assumptions about the constraints of any potential application are made.

7 Distribution

For ease of use, pre-selected image, audio and video subsets of the database are available on CDROM and DVD-RAM. The reader is pointed to [5] to find out about which subsets are currently being made available.

Although this large collection of speech and video data was commissioned in connection with biometric verification, many other uses are envisaged, eg. training of lip-tracking and lip-reading systems, face detection and animation.

8 Conclusions

The XM2VTSDB offers the research community the chance to test their multi-modal face verification algorithms on this high-quality large database. It is hoped that this database and protocol will become a standard enabling institutions to easily assess the performance of their own algorithms compared to others. In the future new sessions with the same people may be acquired in which the scale or illumination of the head will be changed.

In order to promote evaluation of pattern recognition algorithms using this publicly available dataset and the above standard performance assessment methodology, a competition for the best face authentication (verification) algorithm will take place in conjunction with the ICPR 2000 conference. More information about this competition can be found through [5].

Acknowledgements

This work has been performed within the framework of the M2VTS Project granted by the European ACTS programme. The authors would also like to thank the M2VTS partners for comments and suggestions for the design of the database and protocol.

M2VTS Partners

Matra Communication(France), Ibermatica SA (Spain), Cerberus AG (Switzerland), Aristotle University of Thessaloniki (Greece), Ecole Polytechnique Federale de Lausanne (Switzerland), Universite Catholique de Louvain (Belgium), University of Surrey (UK), University of Neuchatel (Switzerland), Renaissance (Belgium), Institut Dalle molle d'Intelligence Artificielle Perceptive (Switzerland), United Tecnica Auxiliar de la Policia (Spain), Compagnie Europeenne de Telesecurite (France), Banco Bilbao Vizcaya (Spain), Universidad Carlos III (Spain).

References

[1] *Biometric technology today. ISSN 0969-4765.*

[2] *BT-DAVID; http://faith.swan.ac.uk/SIPL/david.*

[3] *http://standards.ieee.org.*

[4] *The M2VTS database; http://www.tele.ucl.ac.be/M2VTS/m2fdb.html.*

[5] *The XM2VTSDB; http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/.*

[6] 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 *Multimedia Applications, Services and Techniques (ECMAST 96)*, Louvain-la-Neuve, 1996.

[7] C.C Chibelushi, F. Deravi, and J.S.D. Mason. Survey of audio visual speech databases. Technical report, University of Swansea.

[8] P Devijver and J Kittler. *Pattern Recognition: A Statistical Approach*. Prentice Hall, 1982.

(a) Session 1 (b) Session 2 (c) Session 3 (d) Session 4

(e) Speech 1 (f) Speech 2 (g) Speech 3 (h) Speech 4

Figure 2: (a)-(d): Selected front profile shots taken of a subject from each of the four sessions. (e)-(h): Selected images taken from one of the speech shots.

(a) First centre shot (b) Head left at 45 degrees (c) Left profile shot (d) Head right at 45 degrees

(e) Right profile shot (f) Looking Up (g) Looking Down (h) Second centre shot

Figure 3: Selected images from the rotating head sequence. The subject rotates head from left, to right, to up, to down then back to centre.