The BANCA Database and Evaluation Protocol

Enrique Bailly-Baillire[4], Samy Bengio[1], Frédéric Bimbot[2], Miroslav Hamouz[5], Josef Kittler[5], Johnny Mariéthoz[1], Jiri Matas[5], Kieron Messer[5], Vlad Popovici[3], Fabienne Porée[2], Belen Ruiz[4], and Jean-Philippe Thiran[3]

[1] IDIAP, CP 592, rue du Simplon 4, 1920 Martigny, Switzerland
[2] IRISA (CNRS & INRIA) / METISS, Campus de Beaulieu, 35042 Rennes, France
[3] EPFL, STI-ITS, 1015 Lausanne, Switzerland
[4] University Carlos III,Madrid, Spain
[5] University of Surrey, Guildford, Surrey, GU2 7XH, UK.
K.Messer@ee.surrey.ac.uk

Abstract. In this paper we describe the acquistion and content of a new large, realistic and challenging multi-modal database intended for training and testing multi-modal verification systems. The BANCA database was captured in four European languages in two modalities (face and voice). For recording, both high and low quality microphones and cameras were used. The subjects were recorded in three different scenarios, controlled, degraded and adverse over a period of three months. In total 208 people were captured, half men and half women. In this paper we also describe a protocol for evaluating verification algorithms on the database. The database will be made available to the research community through http://banca.ee.surrey.ac.uk.

1 Introduction

In recent years the cost and size of biometric sensors and processing engines has fallen, a growing trend towards e-commerce, teleworking and e-banking has emerged and people's attitude to security since September 11th has shifted. For these reasons there has been a rapid increase in the use of biometric technology in a range of different applications. For example, at Schipol Airport in Amsterdam, frequent flyers are able to use their iris scans to check-in for flights. The same technology is also used to grant access to airport personnel in secure areas. In Spain, fingerprint scans are used on social security cards and in the U.S.A the Federal Bureau of Prisons uses hand geometry to track the movements of its prisoners, staff and visitors within its prisons.

However, even though these are all fairly reliable methods of biometric personal authentication they are unacceptable by users in all but these high-security situations. They require both user co-operation and are considered intrusive. In contrast, personal identification systems based on the analysis of speech and face images are non-intrusive and more user-friendly. Moreover, personal identity can be often ascertained without the client's assistance. However, speech and image-based systems are more susceptible 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 [5].

BANCA is a European project whose aim is to develop and implement a secure system with enhanced identification, authentication and access control schemes for applications over the Internet such as tele-working and Web - or remote - banking services. One of the major innovations targeted by this project is to obtain an enhanced security system by combining classical security protocols with robust multi-modal verification schemes based on speech and face images.

In order to build reliable recognition and verification systems they need training and in general the larger the training set, the better the performance achieved [8]. However, 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). It is only recently that technology allowing manipulation and effective use of such amounts of data has become available.

For the BANCA project there was a need for a multi-modal database that would include various but realistic recording scenarios, using different kinds of material and in different European languages. We are at present aware of only three publicly available medium or large scale multi-modal databases, the database collected within the M2VTS project, comprising 37 subjects [3], the DAVID-BT database [2] and the database collected within the Extended M2VTS EU project [10]. A survey of audio visual databases prepared by Chibelushi et al [7], lists many others, but these are either mono-modal or small e.g. FERET [12], Yale [15], Harvard [13] and Olivetti [14].

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. The XM2VTS database, together with the Lausanne protocol, [10] contains 295 subjects recorded over 4 sessions. However, it was not possible to use it as the controlled recording environment was not realistic enough compared to the real world situations when one makes a transaction at home through a consumer web cam or through an ATM in a variety of surroundings. Therefore it was decided that a new database for the project would be recorded and a new experimental protocol using the database defined [6]. Typically, an evaluation protocol defines a set of data, how it should be used by a system to perform a set of experiments and how the performance of the system should be computed [11]. Hopefully, the protocol should of been designed in such a manner that there no bias in the performance is introduced, e.g. the training data is not used for testing. It should also represent a realistic operating scenario. Performing experiments according to a defined protocol allows different institutions to easily asses their results when compared to others. The purpose of this paper is to present the BANCA database and its associated protocol.

The rest of this paper is organised as follows. In the next section we define the task of identity verification. In section 3 the BANCA database specification is detailed. Information about the BANCA protocol designed for training and testing personal verification algorithms is given in section 4 before some conclusions are drawn.

2 Identity Verification

Identity Verification (IV) can be defined as the task that consists in verifying the identity X claimed (explicitly or implicitly) by a person U, using a *sample* y from this person, for instance an image of the face of U, a speech signal produced by U, etc... By comparing the sample to some *template* (or *model*) of the claimed identity X, the IV system outputs a decision of *acceptance* or *rejection*. The process can be viewed as a hypothesis testing scheme, where the system has to decide within the following alternative:

- U is the *true client* (acceptance, denoted $\hat{X}$),
- U is an *impostor* (rejection, denoted $\hat{\bar{X}}$).

In practice, an IV system can produce 2 types of errors:

- False Acceptance (FA) if the system has wrongly accepted an impostor,
- False Rejection (FR) if a true client has been rejected by the system.

In practical applications, these 2 types of error have an associated cost, which will be denoted as C_{FA} and C_{FR} respectively. Moreover, in order to measure the quality of the system independently of the the distribution of the accesses, we define the following quantities:

- the False Acceptance Rate (P_{FA}) is the ratio between the number of FA and the number of impostor accesses,
- the False Rejection Rate (P_{FR}) is the ratio between the number of FR and the number of client accesses.

IV approaches are usually based on the characterization of hypotheses X and $\bar{X}$ by a client template and a non-client template respectively, which are learned during a *training* (or enrollment) phase (the non-client model may even be trained during a preliminary phase, also called *installation* phase, and is often the same for every client, in which case it is called the *world model*). Once the template for client X has been created, the system becomes operational for verifying identity claims on X. In the context of performance evaluation, this is referred to as the *test* phase. Conventionally, the procedure used by an IV system during the test phase can be decomposed as follows:

- *feature* extraction, i.e. transformation of the raw sample into a (usually) more compact representation,
- *score* computation, i.e. output of a numerical value $S_X(y)$ based on a (normalized) distance between y and the templates for X (and $\bar{X}$),
- *decision* by comparing the score $S_X(y)$ to a threshold Θ, independent of X.

3 The BANCA Database

3.1 The Acquisition System

To record the database, two different cameras were used; a cheap analogue web cam and a high quality digital camera. For the duration of the recordings the cameras were left in automatic mode. In parallel, two microphones, a poor quality one and a good quality one were used. The database was recorded onto a PAL DV system. PAL DV is a proprietary format which captures video at a colour sampling resolution of 4:2:0. The audio was captured in both 16bit and 12bit audio frequency at 32kHz. The video data is lossy compressed at the fixed ratio of 5:1. The audio data remains uncompressed. This format also defines a frame accurate timecode which is stored on the cassette along with the audio-visual data.

This video hardware can easily be interfaced to a computer allowing frame accurate retrieval of the data in the database onto the computer disk.

3.2 The Specification

The BANCA database was designed in order to test multi-modal IV with various acquisition devices (2 cameras and 2 microphones) and under several scenarios (controlled, degraded and adverse). For 4 different languages (English, French, Italian and Spanish), video and speech data were collected for 52 subjects (26 males and 26 females), i.e. a total of 208 subjects. Each language - and gender - specific population was itself subdivided into 2 groups of 13 subjects, denoted in the following $g1$ and $g2$.

Each subject recorded 12 sessions, each of these sessions containing 2 recordings: 1 true *client access* and 1 informed (the actual subject knew the text that the claimed identity subject was supposed to utter) *impostor attack*. The 12 sessions were separated into 3 different scenarios:

– *controlled* (c) for sessions 1-4,
– *degraded* (d) for sessions 5-8,
– *adverse* (a) for sessions 9-12.

The web cam was used in the degraded scenario, while the expensive camera was used in the controlled and adverse scenarios. The two microphones were used simultaneously in each of the three scenarios with each output being recorded onto a separate track of the DV tape.

During each recording, the subject was prompted to say a random 12 digit number, his/her name, their address and date of birth. Each recording took an average of twenty seconds. Table 1 gives an example of the speech a subject would be expected to utter at a single session (i.e. two recordings). For each session the true client information remained the same. For different sessions the impostor attack information changed to another person in their group.

More formally, in a given session, the impostor accesses by subject X were successively made with a claimed identity corresponding to each other subject

True Client	Impostor Attack
0 3 8 9 2 1 6 7 4 5 0 1	8 5 7 9 0 1 3 2 4 6 0 2
Annie Other	Gertrude Smith
9 St Peters Street	12 Church Road
Guildford	Portsmouth
Surrey	Hampshire
GU2 4TH	PO1 3EF
20.02.1971	12.02.1976

Table 1. Example of the speech uttered by a subject at one of the twelve Banca sessions.

from the *same group* (as X). In other words, all the subjects in group g recorded one (and only one) impostor attempt against each other subject in g and each subject in group g was attacked once (and only once) by each other subject in g. Moreover, the sequence of impostor attacks was designed so as to make sure that each identity was attacked exactly 4 times in the 3 different conditions (hence 12 attacks in total).

In the rest of this paper the following notation will be used:

X_i^g : subject i in group g $\quad g \in \{g1, g2\}, \quad i \in [1, 13]$

$y_k(X)$: true client record from session k by subject X $\quad k \in [1, 12]$

$z_l(X)$: impostor record (from a subject X') claiming identity X during a session l (with $X' \neq X$) $\quad l \in [1, 12]$

For each language, an additional set of 30 other subjects, 15 males and 15 females, recorded one session (audio and video). This set of data is referred to as *world data*. These individuals claimed two different identities, recorded by both microphones. Finally, any data outside the BANCA database will be referred to as *external data*.

Figure 1 shows a few examples of the face data from the English part of the database, whilst figure 2 shows a few examples of face data from the French part.

4 Experimental Protocol

In verification, two types of protocols exist; closed-set and open-set. In closed-set verification the population of clients is fixed. This means that the system design can be tuned to the clients in the set. Thus both, the adopted representation (features) and the verification algorithm applied in the feature space are based on some training data collected for this set of clients. Anyone who is not in the training set is considered an impostor. The XM2VTS protocol is an example of this type of verification problem formulation.

In open-set verification we wish to add new clients to the list without having to redesign the verification system. In particular, we want to use the same feature space and the same design parameters such as thresholds. In such a scenario the feature space and the verification system parameters must be trained using

Fig. 1. Examples of the BANCA database images taken from the English part of the database *Up:* Controlled, *Middle:* Degraded and *Down:* Adverse scenarios.

completely independent data from that used for specifying client models. The Banca protocol is an example of an open-set verification protocol.

In this paper, we present a configuration of the Banca protocol using only one language [6]; other protocols, taking into account all 5 languages, will be presented later.

4.1 A Monolingual Protocol

In order to define an experimental protocol, it is necessary to define a set of evaluation data (or *evaluation set*), and to specify, within this set, which are to be used for the training phase (enrollment) and which are to be used for the test phase (test accesses).

Moreover, before becoming operational, the development of an IV system requires usually the adjustment of a number of configuration parameters (model size, normalization parameters, decision thresholds, etc.). It is therefore necessary to define a *development set*, on which the system can be calibrated and

Fig. 2. Examples of the BANCA database images taken from the French part of the database *Up:* Controlled, *Middle:* Degraded and *Down:* Adverse scenarios.

adjusted, and for which it is permitted to use the knowledge of the actual subject identity during the test phase. Once the development phase is finished, the system performance can then be assessed on the evaluation set (without using the knowledge of the actual subject identity during the test phase).

To avoid any methodological flaw, it is essential that the development set is composed of a distinct subject population as the one of the evaluation set. In order to carry realistic (and unbiased experiments), it is necessary to use different populations and data sets for development and for evaluation. We distinguish further between 2 circumstances: single-modality evaluation experiments and multi-modality evaluation experiments. In the case of single-modality experiments, we need to distinguish only between two data sets: the development set, and the evaluation set. In that case, $g1$ and $g2$ are used alternatively as development set and evaluation set (when $g1$ is used as development set, $g2$ is used as evaluation set, and vice versa).

In the case of multi-modality experiments, it is necessary to introduce a third set of data: the *(fusion) tuning set* used for tuning the fusion parameters, i.e. the way to combine the outputs of each modality. If the tuning set is identical to the development set, this may introduce a pessimistic bias in the estimation of the tuning parameters (*biased* case). Another solution is to use three distinct sets for development, tuning and evaluation (*unbiased* case). In that case, we expect the experimenters to use data from the other languages as development set, while $g1$ and $g2$ are used alternatively for tuning and evaluation.

In the BANCA protocol, seven distinct experimental configurations have been specified which identify which material can be used for training and which for testing. In all configurations, the true client records for the first session of each condition is reserved as training material, i.e. the true client record from sessions 1, 5 and 9. In all experiments, the client model training (or template learning) is done on at most these 3 records.

The seven configurations are Matched Controlled (MC), Matched Degraded (MD), Matched Adverse (MA), Unmatched Degraded (UD), Unmatched Adverse (UA), Pooled test (P) and Grand test (G). Table 2 describes the usage of the different sessions in each configuration. “TT” refers to the client training and impostor test session, and “T” depicts clients and impostor test sessions. A more detailed description of the seven configurations can be found in Annex A.

For example, for configuration MC the true client data from session 1 is used for training and the true client data from sessions 2, 3 and 4 are used for client testing. All the impostor attack data from sessions 1,2,3 and 4 is used for impostor testing.

Session	MC	MD	MA	UD	UA	P	G
1	TT			TT	TT	TT	TT
2	T					T	T
3	T					T	T
4	T					T	T
5		TT					TT
6		T		T		T	T
7		T		T		T	T
8		T		T		T	T
9			TT				TT
10			T		T	T	T
11			T		T	T	T
12			T		T	T	T

Table 2. The usage of the different sessions in the seven BANCA experimental configurations (“TT”: clients training and impostor test, “T”: clients and impostor test)

From analysing the performance results on all seven configurations it is possible to measure:

- the intrinsic performance in a given condition,
- the degradation from a mismatch between controlled training and uncontrolled test,

– the performance in varied conditions with only one (controlled) training session,
– the potential gain that can be expected from more representative training conditions.

It is also important to note that for the purpose of the protocol, 5 frontal images have been extracted from each video recording to be used as true client and impostor attack images.

4.2 Performance Measure

In order to visualize the performance of the system, irrespective of its operating condition, we use the conventional DET curve [9], which plots on a log-deviate scale the *False Rejection Rate* P_{FR} as a function of the *False Acceptance Rate* P_{FA}. Traditionally, the point on the DET curve corresponding to $P_{FR} = P_{FA}$ is called EER (Equal Error Rate) and is used to measure the closeness of the DET curve to the origin. The EER value of an experiment is reported on the DET curve, to comply with this tradition. Figure 3 shows an example DET curve. We also recommend to measure the performance of the system for 3 specific operating conditions, corresponding to 3 different values of the Cost Ratio $R = C_{FA}/C_{FR}$, namely $R = 0.1, R = 1, R = 10$. Assuming equal *a priori* probabilities of genuine clients and impostor, these situations correspond to 3 quite distinct cases:

$R = 0.1$ $\rightarrow$ a FA is an order of magnitude less harmful than a FR,
$R = 1$ $\rightarrow$ a FA and a FR are equally harmful,
$R = 10$ $\rightarrow$ a FA is an order of magnitude more harmful than a FR.

When R is fixed and when P_{FR} and P_{FA} are given, we define the Weighted Error Rate (WER) as:

$$WER(R) = \frac{P_{FR} + R\,P_{FA}}{1 + R}\ . \tag{1}$$

P_{FR} and P_{FA} (and thus WER) vary with the value of the decision threshold Θ, and Θ is usually optimized so as to minimize WER on the development set D:

$$\hat{\Theta}_R = \arg\min_{\Theta_R}\ WER(R)\ . \tag{2}$$

The *a priori threshold* thus obtained is always less efficient than the *a posteriori threshold* that optimizes WER on the evaluation set E itself:

$$\Theta^*_R = \arg\min_{\Theta_R}\ WER(R)\ . \tag{3}$$

The latter case does not correspond to a realistic situation, as the system is being optimized with the knowledge of the actual test subject identities on the evaluation set. However, it is interesting to compare the performance obtained with *a priori* and *a posteriori* thresholds in order to assess the reliability of the threshold setting procedure.

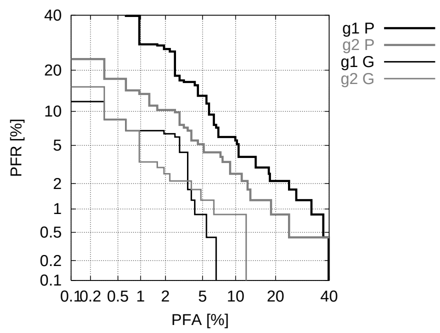

Fig. 3. An example DET used to help measure performance. Curves are for the groups 1 and 2 using protocols P and G.

5 Distribution

It is intended to make the database available to the research community to train and test their verification algorithms. The reader is pointed to [1] to find out which parts of the database are currently available. Already, part of the database has been made available for use for standards work for MPEG-7 [4]. To date, 14 copies of this bench mark MPEG-7 test set have been distributed.

Although this data was captured in connection with biometric verification many other uses are envisaged such as animation and lip-tracking.

6 Conclusion

In this paper, a new multi-modal database and its associated protocol have been presented which can be used for realistic identity verification tasks using up to two modalities.

The BANCA database offers the research community the opportunity to test their multi-modal verification algorithms on a large, realistic and challenging database. It is hoped that this database and protocol will become a standard, like the XM2VTS database, which enables institutions to easily compare the performance of their own algorithms to others.

7 BANCA Partners

EPFL(Switzerland); IRISA(France); University Carlos III (Spain); Ibermatica (Spain); Thales (France); BBVA (Spain); Oberthur(France); Institut Dalle Molle d'Intelligence Artificielle Perceptive (IDIAP) (Switzerland); Université de Louvain(France); University of Surrey(UK);

8 Acknowledgments

This research has been carried out in the framework of the European BANCA project, IST-1999-11169. For IDIAP this work was funded by the Swiss OFES project number 99-0563-1 and for EPFL by the Swiss OFES project number 99-0563-2.

A Detailed Description of the Different Protocols

Tables 3, 4 and 5 describe more formally the 7 training-test configurations.

		Mc (use only controlled data)	Md (use only degraded data)	Ma (use only adverse data)
Client training	$\forall\ X_i \quad i = 1, ..., 13$ train with:	$y_k(X_i), \quad k = 1$	$y_k(X_i), \quad k = 5$	$y_k(X_i), \quad k = 9$
Non client training	All World data + any external data (it is forbidden to use other client data from the same group)			
Client testing	$\forall\ X_i \quad i = 1, ..., 13$ test with:	$y_k(X_i), \quad k = 2, 3, 4$	$y_k(X_i), \quad k = 6, 7, 8$	$y_k(X_i), \quad k = 10, 11, 12$
Impostor testing	$\forall\ X_i \quad i = 1, ..., 13$ test with:	$z_l(X_i), \quad l \in \{1, 2, 3, 4\}$	$z_l(X_i), \quad l \in \{5, 6, 7, 8\}$	$z_l(X_i), \quad l \in \{9, 10, 11, 12\}$
Number of tests per experiment		client: $13 \times 3 = 39$ impostor: $13 \times 4 = 52$	client: 39 impostor: 52	client: 39 impostor: 52
Total number of image tests		client: $2 \times 5 \times 2 \times 39 = 780$ impostor: $2 \times 5 \times 2 \times 52 = 1040$	client: 780 impostor: 1040	client: 780 impostor: 1040

Table 3. Description of protocols Mc, Md and Ma.

		Ud (use controlled data for training and degraded data for testing)	Ua (use controlled data for training and adverse data for testing)
Client training	$\forall\ X_i \quad i = 1, ..., 13$ train with:	$y_k(X_i), \quad k = 1$	$y_k(X_i), \quad k = 1$
Non client training	All World data + any external data (it is forbidden to use other client data from the same group)		
Client testing	$\forall\ X_i \quad i = 1, ..., 13$ test with:	$y_k(X_i), \quad k = 6, 7, 8$	$y_k(X_i), \quad k = 10, 11, 12$
Impostor testing	$\forall\ X_i \quad i = 1, ..., 13$ test with:	$z_l(X_i), \quad l \in \{5, 6, 7, 8\}$	$z_l(X_i), \quad l \in \{9, 10, 11, 12\}$
Number of tests per experiment		client: $13 \times 3 = 39$ impostor: $13 \times 4 = 52$	client: 39 impostor: 52
Total number of image tests		client: $2 \times 5 \times 2 \times 39 = 780$ impostor: $2 \times 5 \times 2 \times 52 = 1040$	client: 780 impostor: 1040

Table 4. Description of protocols Ud and Ua.

		P (use controlled data for training and all data for testing)	G (use all data for training and all data for testing)
Client training	$\forall\ X_i \quad i = 1, ..., 13$ train with:	$y_k(X_i), \quad k = 1$	$y_k(X_i), \quad k = 1, 5, 9$
Non client training	All World data + any external data (it is forbidden to use other client data from the same group)		
Client testing	$\forall\ X_i \quad i = 1, ..., 13$ test with:	$y_k(X_i), \quad k = 2, 3, 4, 6, 7, 8, 10, 11, 12$	$y_k(X_i), \quad k = 2, 3, 4, 6, 7, 8, 10, 11, 12$
Impostor testing	$\forall\ X_i \quad i = 1, ..., 13$ test with:	$z_l(X_i), \quad l \in \{1, ..., 12\}$	$z_l(X_i), \quad l \in \{1, ..., 12\}$
Number of tests per experiment		client: $13 \times 9 = 117$ impostor: $13 \times 12 = 156$	client: 117 impostor: 156
Total number of image tests		client: $2 \times 5 \times 2 \times 117 = 2340$ impostor: $2 \times 5 \times 2 \times 156 = 3120$	client: 2340 impostor: 3120

Table 5. Description of protocols P and G.

References

1. *The BANCA Database - English part; http://banca.ee.surrey.ac.uk/.*
2. *BT-DAVID; http://faith.swan.ac.uk/SIPL/david.*
3. *The M2VTS database; http://www.tele.ucl.ac.be/ M2VTS/m2fdb.html.*
4. *MPEG-7 Overview; http://mpeg.telecomitalialab.com/standards/mpeg-7/mpeg-7.htm.*
5. 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.
6. S. Bengio, F. Bimbot, J. Mariéthoz, V. Popovici, F. Porée, E. Bailly-Baillière, G. Matas, and B. Ruiz. Experimental protocol on the BANCA database. Technical Report IDIAP-RR 02-05, IDIAP, 2002.
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.
9. A. Martin et al. The DET curve in assessment of detection task performance. In *Eurospeech'97*, volume 4, pages 1895–1898, 1997.
10. K Messer, J Matas, J Kittler, J Luettin, and G Maitre. Xm2vtsdb: The extended m2vts database. In *Second International Conference on Audio and Video-based Biometric Person Authentication*, March 1999.
11. P. J. Phillips, A. Martin, C. L. Wilson, and M. Przybocki. An introduction to evaluating biometric systems. *IEEE Computer*, pages 56–63, February 2000.
12. P.J. Phillips, H. Wechsler, J.Huang, and P.J. Rauss. The FERET database and evaluation procedure for face-recognition algorithm. *Image and Vision Computing*, 16:295–306, 1998.
13. ftp://hrl.harvard.edu/pub/faces.
14. http://www.cam-orl.co.uk/facedatabase.html.
15. http://cvc.yale.edu/projects/yalefaces/yalefaces.html.