Face Video Competition

Norman Poh[1], Chi Ho Chan[1], Josef Kittler[1], Sébastien Marcel[2], Christopher Mc Cool[2], Enrique Argones Rúa[3], José Luis Alba Castro[3], Mauricio Villegas[4], Roberto Paredes[4], Vitomir Štruc[5], Nikola Pavešić[5], Albert Ali Salah[6], Hui Fang[7], and Nicholas Costen[7]

[1] CVSSP, University of Surrey, Guildford, GU2 7XH, Surrey, UK
normanpoh@ieee.org, c.chan@surrey.ac.uk, j.kittler@surrey.ac.uk
[2] Idiap research institute, Marconi 19, Martigny, CH
[3] Signal Technologies Group, Signal Theory and Communications Dept., University of Vigo, 36310, Spain
[4] Universidad Politécnica de Valencia, Instituto Tecnológico de Informática, Camino de Vera s/n, 46022 Valencia, Spain
[5] Faculty of Electrical Engineering, University of Ljubljana, Tržaška 25, SI-1000 Ljubljana, Slovenia
[6] CWI, Science Park 123, 1098 XG Amsterdam
[7] Department of Computing and Mathematics, Manchester Metropolitan University, UK M1 5GD

Abstract. Person recognition using facial features, e.g., mug-shot images, has long been used in identity documents. However, due to the widespread use of web-cams and mobile devices embedded with a camera, it is now possible to realise facial video recognition, rather than resorting to just still images. In fact, facial video recognition offers many advantages over still image recognition; these include the potential of boosting the system accuracy and deterring spoof attacks. This paper presents the first known benchmarking effort of person identity verification using facial video data. The evaluation involves 18 systems submitted by seven academic institutes.

1 Introduction

With an increasing number of mobile devices with built-in web-cams, e.g., PDA, mobile phones and laptops, face is arguably the most widely accepted means of person verification. However, the biometric authentication task based on face images acquired by a mobile device in an uncontrolled environment is very challenging. One way to boost the face verification performance is to use multiple samples.

Previous attempts at assessing the performance of face verification algorithms have been restricted to matching still images, e.g., the three FERET evaluations[1] (1994, 1995 and 1996), the face recognition vendor tests (FRVTs 2000, 2002 and 2006)[2], and assessment on XM2VTS and BANCA databases [1,2]. The well known Face Recognition Grand Challenge [3] includes queries with multiple still images but this is far from the vast amount of data available in video matching.

[1] http://www.itl.nist.gov/iad/humanid/feret/feret_master.html
[2] http://www.frvt.org

M. Tistarelli and M.S. Nixon (Eds.): ICB 2009, LNCS 5558, pp. 715–724, 2009.

The evaluation exercise presented here is the first known effort in assessing *video-to-video* matching, i.e., in both enrolment and verification phases, the data captured is in the form of video sequence. This is different from still-image-to-video matching, one of the evaluation scenarios currently examined by the NIST Multiple Biometric Grand Challenge[3] (MBGC). Note that NIST MBGC aims at "portal application" where the task is to verify the identity of person as he/she walks through an access control check point. The video-to-video matching adopted here has a slightly different application, with a focus on mobile devices, where a sequence of unconstrained (talking) face images can be expected.

The video-to-video face verification assessment has several objectives, among which are:

- to promote the development of algorithms for analysing video sequences (e.g., exploring the talking face dynamics)
- to assess the merit of multi-template face representation
- to compare whether early integration (e.g., feature-level fusion) is better than late integration (e.g., decision-level fusion) in dealing with sequences of query images.

2 Database, Protocols, Facial Video Annotations

Towards the above goal, we have opted to use the publicly available BANCA database [4][4]. It has a collection of face and voice biometric traits of up to 260 persons in 5 different languages, but only the English subset is used here. It contains a total of 52 persons; 26 females and 26 males. The 52 persons are further divided into two sets of users, which are called g1 and g2, respectively. Each set (g1 or g2) is designed to be balanced in gender, i.e., having 13 males and 13 females. According to the experimental protocols reported in [4], when g1 is used as a development set (to build the user's template/model), g2 is used as an evaluation set. Their roles are then switched. This corresponds to a two-fold cross-validation procedure.

The BANCA database was designed to examine matching under the same recording conditions (as the enrolment session) and two different challenging conditions: recording under a noisy (adverse) environment and with a degraded device. In each of the three conditions, four recordings were performed. The clean conditions apply to sessions 1–4; adversed conditions to sessions 5–8; and degraded conditions to sessions 9–12. There are altogether seven experimental protocols specifying the sessions to be used for enrolment and for testing in an exhaustive manner. In this face video recognition evaluation, we focused on two protocols, namely the match controlled (Mc) and unmatched adversed (Ua) protocols. The first protocol was intended as a vehicle to design and tune their face verification systems. The second protocol aims at testing the systems under more realistic and challenging conditions.

In the Mc protocol, session 1 data is used for enrolment whereas the data from sessions 2–4 are reserved for testing. In the Ua protocol, the session 1 data again is used

[3] http://face.nist.gov/mbgc

[4] http://www.ee.surrey.ac.uk/CVSSP/banca

for enrolment but the test data is taken from session 5–8 (recorded under adversed conditions). The ICB2009 face video competition was thus naturally carried out in two rounds, with the first round focusing on the Mc protocol and the second round on the Ua protocol.

In order to be consistent with the previous BANCA evaluations [1,2], we also divided a query video sequence into 5 chunks, each containing 50 frames for convenience; the remaining frames were simply not used.

In order to standardise the evaluation, we provided a pair of eye coordinates, based on the face detector provided by the OmniPerception's SDK[5]. However, the participants could use their own face detectors. For each image in a video sequence, the SDK also annotated the following quality measurements. Note that the entire processes from detection to annotation were done automatically. No effort was made to fine tune the system parameters, and in consequence, some imperfectly cropped images were observed. The image quality measures assessed.

1. Overall reliability
2. Brightness
3. Contrast
4. Focus
5. Bit per pixel
6. Spatial resolution (between eyes)
7. Illumination
8. Background uniformity
9. Background brightness
10. Reflection
11. Presence of glasses
12. In-plane rotation
13. In-depth rotation
14. Frontalness

In the above list, "frontalness" quantifies the degree of similarity of a query image to a typical frontal (mug-shot) face image. The overall reliability is a compounded quality measure obtained by combining the remaining quality measures. Two categories of quality measures can be distinguished: face-specific or generic. The face-specific ones strongly depend on the result of face detection, i.e., frontalness, rotation, reflection, between-eyes spatial resolution in pixels, and the degree of background uniformity (calculated from the remaining area of a cropped face image). The generic ones are defined by the MPEG standards. All the annotation data (including eye coordinates and quality measures) has been published on the website "http://face.ee.surrey.ac".

A preliminary analysis shows that when the frontalness measure is 100%, the detected face is always frontal. On the other hand, any value less than 100% does indeed suggest an imperfect face detection, or else a non-ideal (non-frontal) pose.

3 Summary of Submitted Systems

The submitted face verification systems can be categorised according to whether they are image-set-based or frame-based (comparison) approach. In the image-set based approach, a video sequence is analysed and treated as a set of images. When comparing two video sequences, this approach, in essence, compares two *sets* of images. On the other hand, the frame-based approach directly establishes similarity between two images, each obtained from their respective video sequence. If there are P and Q images in both sequences, there will be at most PQ similarity scores. The frame-based approach

[5] http://www.omniperception.com

Table 1. Overview of the submitted face verification systems

	Systems	Pre-processing	Face rep.	Feature Extraction	Classifier	Quality measure used	Process all images
Holistic	idiap-pca-pearson	HEQ		PCA	Pearson	No	Yes
	idiap-pca-nc	HEQ		PCA	NC	No	Yes
	idiap-pca-cor	HEQ		PCA	StdCor	No	Yes
	idiap-lda-pearson	HEQ		PCAxLDA	Pearson	No	Yes
	idiap-lda-nc	HEQ		PCAxLDA	NC	No	Yes
	idiap-lda-cor	HEQ		PCAxLDA	StdCor	No	Yes
	mmu		AM	LDA	Avg(NC)	No	Yes
Local	idiap-dcthmmt-v1	HEQ		DCT	HMM	No	Yes
	idiap-dcthmmt-v2	HEQ		DCT	HMM	No	Yes
	idiap-dctgmm	HEQ		DCTmod2+xy	GMM	No	Yes
	idiap-LBP-dctgmm		LBP	DCTmod2+xy	GMM	No	Yes
	cwi-Cq			DCT	Max(NC)		Yes
	cwi-Eq			DCT	Max(NC)		Yes
	cwi-Cr			DCT	Max(NC)		No
	cwi-Er			DCT	Max(NC)		No
	upv	Local-HEQ	LF	PCA	Avg(KNN)	Yes	No
	uni-lj	ZMUV + HEQ	Gb2	KDA+PCA	WNC	Yes	No
	uvigo	Ani	Gb1		GMM	Yes	No

The following keys are used: AM = Appearance model, ZMUV = zero mean and unit-variance, Ani = Anisotropic+local mean subtraction, LF = Local feature Gb1 = Gabor(magnitude) Gb2 = Gabor(phase+magnitude, NC = Normalised correlation, WNC = Sum of whitened NC Note: OmniPerception's face detector was used by all systems.

would select, or otherwise combine these similarity scores to obtain a final similarity score. Among the systems, only the MMU system belongs to the image-set based approach, while the remaining systems are the frame-based approach.

Face verification systems can also be further distinguished by the way a face image is treated, i.e, either holistic or local (parts-based) appearance approach. In the former, the entire (often cropped) image is considered as input to the face classifier. In the latter, the face images are divided into (sometimes overlapping) parts which are then treated separately by a classifier.

Table 1 summarises the systems by this categorisation. Principal component analysis (PCA), or Eigenface, and local discriminant analysis (LDA), or Fisherface, are perhaps the most representative (and popular) examples of the holistic approach due to the pioneer work of Turk and Pentland [5]. Many of these systems were submitted by IDIAP as baseline systems, tested on the Mc protocol (and not the Ua protocol). Recent face verification research has been dominated by the local appearance approach, as exemplified by *most* of the submissions in this competition. The details of each system can be found in "http://face.ee.surrey.ac.uk/data/face_icb2009.pdf".

4 Evaluation Metrics

We use two types of curves in order to compare the performance: the Detection Error Trade-off (DET) curve [6] and the Expected Performance Curve (EPC) [7]. A DET

curve is actually a Receiver Operator Curve (ROC) curve plotted on a scale defined by the inverse of a cumulative Gaussian density function, but otherwise similar in all aspects. We have opted to use EPC because it has been pointed out in [7] that two DET curves resulting from two systems are not comparable. This is because such comparison does not take into account how the decision thresholds are selected. EPC turns out to be able to make such comparison possible. Furthermore, the performance across different data sets, resulting in several EPCs, can be merged into a single EPC [8]. Although reporting performance in EPC is more meaningful than DET as far as performance comparison is concerned, it is relatively new and has not gained a widespread acceptance in the biometric community. As such, we shall also report performance in DET curves, but using only a subset of operating points.

The EPC curve, however, is less convenient to use because it requires two sets of match scores, one used for tuning the threshold (for a given operating cost), and the other used for assessing the performance. In our context, with the two-fold cross-validation defined on the database (as determined by g1 and g2), these two match scores can be conveniently used.

According to [7], one possible, and often used criterion is the weighted error rate (WER), defined by:

$$\text{WER}(\beta, \Delta) = \beta\, \text{FAR}(\Delta) + (1 - \beta)\, \text{FRR}(\Delta), \tag{1}$$

where FAR is the false acceptance rate, FRR is the false rejection rate at a given threshold Δ and $\beta \in [0, 1]$ is a user-specified coefficient which balances FAR and FRR. The WER criterion generalises the criterion used in the annual NIST's speaker evaluation [9] as well as the three operating points used in the past face verification competitions on the BANCA database [1,2]. In particular the following three coefficients of β are used:

$$\beta = \frac{1}{1+R} \text{ for } R = \{0.1, 1, 10\}$$

which yields approximately $\beta = \{0.9, 0.5, 0.1\}$, respectively.

The procedure to calculate an EPC is as follows: Use g1 to generate the development match scores; and g2, the evaluation counterpart. For each chosen β, the development score set is used to minimise (1) in order to obtain an operational threshold. This threshold is then applied to the evaluation set in order to obtain the final pair of false acceptance rate (FAR) and false rejection rate (FRR). The EPC curve simply plots half total error rate (HTER) versus β, where HTER is the average of FAR and FRR. Alternatively, the generalisation performance can also be reported in WER (as done in the previous BANCA face competitions). To plot the corresponding DET curve, we use the pair of FAR and FRR of all the operating points, as determined by β. Note that this DET curve is a *subset* (in fact discrete version) of a conventional continuous DET curve because the latter is plotted from continuous empirical functions of FAR and FRR. By plotting the discrete version of the DET curve, we establish a *direct correspondence* between EPC and DET, satisfying both camps of biometric practitioners, while retaining the advantage of EPC which makes performance comparison between systems less biased.

5 Results

The DET curves of all submitted systems for the g1 and g2 data sets, as well as for the Mc and Ua protocols, are shown in Figure 1. By merging the results from g1 and g2, we plotted the EPCs for Mc and Ua in Figure 2 (plotting β versus HTER). To be consistent with the previous published BANCA evaluations [1,2], we also listed the individual g1 and g2 performance, in terms of WER, in Table 2 for the Mc protocol and in Table 3 for the Ua protocol.

The following observations can be made:

– **degradation of performance under adversed conditions:** It is obvious from Figure 2 that all systems systematically degrade in performance under adversed conditions.

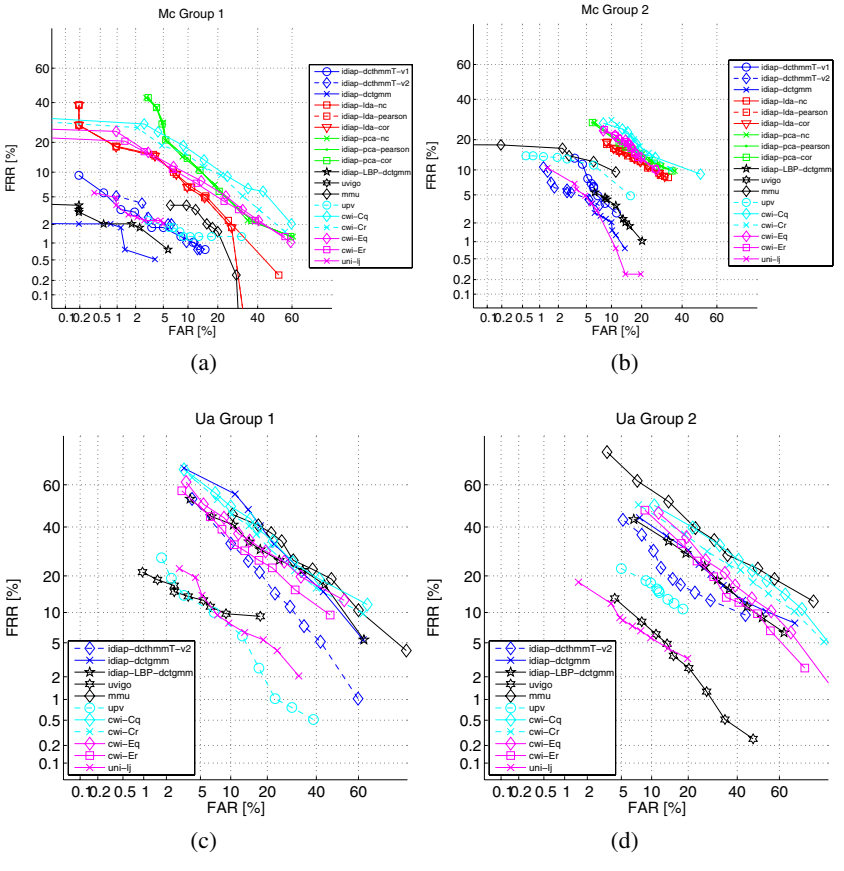
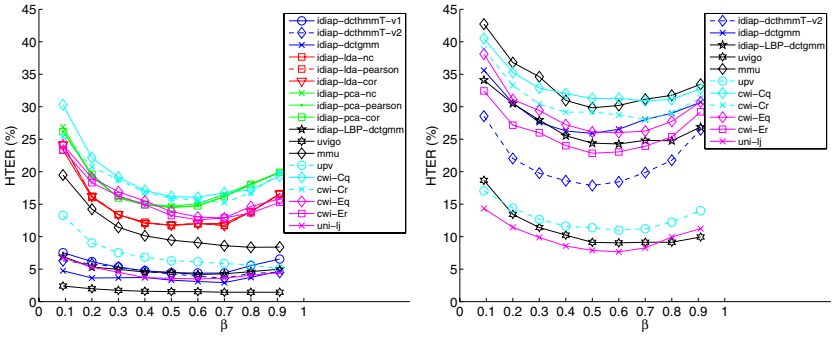


Fig. 1. DET curves of the submitted systems evaluated on the g2 (evaluation set) of the BANCA video based on the Mc protocol. Note that the uvigo system achieved zero EER on the Mc g2 datasets. As a result, its DET curve reduces to a single point at the origin ((∞, ∞) in the above normal inverse scales.



(a) Mc (b) Ua

Fig. 2. EPC curves of the submitted systems evaluated on the g2 (evaluation set) of the BANCA video based on the Mc protocol

Table 2. Performance of g1 and g2 based on the Mc protocol using video sequences

systems	WER (%)					
	$R = 0.1$		$R = 1$		$R = 10$	
	G1	G2	G1	G2	G1	G2
idiap-dcthmm†	7.52	4.90	5.45	0.64	2.56	0.12
idiap-dcthmm‡	7.78	3.76	5.13	2.08	1.17	2.74
idiap-dcthmmT-v2	1.34	2.03	4.20	4.29	1.92	3.93
idiap-dctgmm	0.82	5.14	1.12	5.48	0.82	1.96
idiap-LBP-dctgmm	0.75	6.26	1.63	7.37	1.22	2.77
uvigo	1.05	0.42	0.77	2.31	0.45	4.20
mmu	5.94	2.14	9.84	9.07	5.21	9.64
upv	3.01	1.81	5.06	7.50	4.00	5.86
cwi-Cq	3.80	9.84	14.20	18.14	7.28	12.76
cwi-Cr	3.66	11.72	13.14	18.69	6.49	12.40
cwi-Eq	2.84	9.51	10.90	16.83	6.32	11.49
cwi-Er	2.59	9.73	9.87	16.63	6.25	11.68
uni-lj	0.86	2.18	2.34	4.81	2.32	2.02

†: Experimental results on *still* images, taken from [1] with automatic localisation. ‡: Similar to †, except with manual localisation.

– **holistic vs. local appearance methods:** From Figure 1(a) and (b) as well as Figure 2(a), we observe that the performance of the holistic appearance methods (PCA and LDA) is worse than that of the local appearance methods, except for the CWI classifier (where photometric normalisation was not performed). Thus, we can expect that the performance of CWI to be similar to the performance of other local appearance methods in the raw image space, such as idiap-dctgmm, idiap-dcthmmt-v2 and upv if photometric normalisation were to be performed.
– **still vs. video comparison:** Among the submitted systems, only IDIAP's DCT-HMM system was involved in the previously reported results for the Mc protocol

Table 3. Performance of g1 and g2 based on the Ua protocol

systems	WER (%)					
	$R = 0.1$		$R = 1$		$R = 10$	
	G1	G2	G1	G2	G1	G2
idiap-dcthmmT-v2	8.52	8.66	18.65	17.08	6.37	12.61
idiap-dctgmm	9.10	11.03	27.31	24.49	10.54	13.31
idiap-LBP-dctgmm	8.34	10.08	23.85	24.94	10.58	11.47
uvigo	2.81	5.06	8.75	9.49	10.00	4.55
mmu	13.61	9.88	27.72	31.96	10.97	18.21
upv	4.00	6.60	9.29	13.46	3.98	11.45
cwi-Cq	9.06	14.18	28.08	34.46	16.54	11.19
cwi-Cr	9.43	11.41	26.60	31.79	14.50	11.79
cwi-Eq	8.72	14.73	24.23	27.98	16.50	8.48
cwi-Er	8.00	12.23	21.38	24.29	12.86	8.80
uni-lj	4.67	3.03	8.78	6.99	4.78	4.83

[1] which was based on 5 still images taken from a video sequence (as opposed to five video chunks as done here). The results for this classifier are shown in Table 2 (comparing rows 1-2 with row 3). In theory, one would expect the classifier tested on video sequence to be better than still images. Unfortunately, such conclusion cannot be made except for $R = 0.1$.

- **Pre-processing:** In dctgmm methods, the performance of applying HEQ is better than that of applying LBP as a pre-processing method for Mc protocol. However, the case is reversed for Ua protocol because HEQ enhances shadows while LBP features are invariant to such monotonic transformation (in relation to the neighbourhood pixels cast under shadows). In other words, the selection of the pre-processing methods should be dependent on the environmental conditions.
- **Sample size:** Cwi's submission has four variations: depending on the dichotomies: system complexity, i.e., Cheap (C) versus Exepnsive (E); and strategy for choosing the query samples, i.e, random (r) versus quality-based (q). Two observations can be noted: First, the performance of cwi-Eq and cwi-Er are better than that of cwi-Cq and cwi-Cr. Second, using *more* template and query features can improve the cwi system. A rigorous and systematic design of experiments is still needed to find out the usefulness of the provided quality measures, and more importantly, the most effective ways of using such auxilliary information. This is a challenging problem for two reasons. First, not all 14 quality measures provided are relevant to a face matching algorithm, e.g., an algorithm that is robust to illumination changes would, in principle, be invariant to some photometric measures used here (brightness, contrast, etc). This implies that a quality measure selection strategy is needed. Second, quality measures are themselves not discriminatory for distinguishing subjects but discriminatory in distinguishing environmental conditions.
- **Multi resolution Contrast Information:** The best algorithm of this competition for MC protocol is UVigo where the WER at R=1 is 0.77% for G1 and 2.31% for G2. For UA protocol, the best algorithm is uni-lj where WER at R=1 is 8.78% for G1 and 6.99% for G2. In fact, the performance of these two systems is very close

but uni-lj is slightly better overall as the average of WER at different R is 3.96% for G1 and 3.98% for G2, while the result of UVigo is 3.97% for G1 and 4.34% for G2. The success of these two algorithms derives from the use of multi resolution contrast information.

6 Discussion and Future Evaluation

Because the target application scenario of this assessment is on mobile devices, computational resources are crucial. For this reason, when benchmarking a face verification algorithm, the cost of computation has to be considered. For instance, a fast and light algorithm, capable of processing all images in a sequence, may be preferred over an extremely accurate algorithm only capable of processing a few selected images in a sequence. However, the former algorithm may be able to achieve better performance since it can process a much larger number of images within the same time limit and memory requirement. The above scenario highlights that the performance of two algorithms cannot be compared on equal grounds, unless both use comparable computation costs, taking the time, memory and computational resources into consideration.

The current evaluation has not taken this cost factor into consideration, but this will be carried out in future. The idea is to request each participant to run a benchmarking program, executable in any operating system. The time registered by the program will be used as a *standard unit time* for the participant's system. Thus the time to process a video file for a participant, for instance, will be reported in terms of multiples (or fractions) of the participant's standard unit time.

7 Conclusions

This paper presents a comparison of video face verification algorithms on BANCA database. Eighteen different video-based verification algorithms from a variety of academic institutions participated in this competition. The results show that the performance of the local appearance methods is better than that of the holistic appearance methods. Secondly, using more query and selected template features to measure similarity improve the system performance. Finally, the best algorithm in this competition clearly shows that multi resolution contrast information is important for face recognition.

Acknowledgement

The work of NPoh is supported by the advanced researcher fellowship PA0022_121477 of the Swiss NSF; NPoh, CHC and JK by the EU-funded Mobio project grant IST-214324; NPC and HF by the EPSRC grants EP/D056942 and EP/D054818; VS and NP by the Slovenian national research program P2-0250(C) Metrology and Biometric System, the COST Action 2101 and FP7-217762 HIDE; and, AAS by the Dutch BRICKS/BSIK project.

References

1. Messer, K., Kittler, J., Sadeghi, M., Hamouz, M., Kostyn, A., Marcel, S., Bengio, S., Cardinaux, F., Sanderson, C., Poh, N., Rodriguez, Y., Kryszczuk, K., Czyz, J., Vandendorpe, L., Ng, J., Cheung, H., Tang, B.: Face authentication competition on the BANCA database. In: Zhang, D., Jain, A.K. (eds.) ICBA 2004. LNCS, vol. 3072, pp. 8–15. Springer, Heidelberg (2004)
2. Messer, K., Kittler, J., Sadeghi, M., Hamouz, M., Kostin, A., Cardinaux, F., Marcel, S., Bengio, S., Sanderson, C., Poh, N., Rodriguez, Y., Czyz, J., Vandendorpe, L., McCool, C., Lowther, S., Sridharan, S., Chandran, V., Palacios, R.P., Vidal, E., Bai, L., Shen, L.-L., Wang, Y., Yueh-Hsuan, C., Liu, H.-C., Hung, Y.-P., Heinrichs, A., Muller, M., Tewes, A., vd Malsburg, C., Wurtz, R., Wang, Z., Xue, F., Ma, Y., Yang, Q., Fang, C., Ding, X., Lucey, S., Goss, R., Schneiderman, H.: Face authentication test on the BANCA database. In: Int'l. Conf. Pattern Recognition (ICPR), vol. 4, pp. 523–532 (2004)
3. Phillips, P.J., Flynn, P.J., Scruggs, T., Bowyer, K.W., Chang, J., Hoffman, K., Marques, J., Min, J., Worek, W.: Overview of the Face Recognition Grand Challenge. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 947–954 (2005)
4. Bailly-Baillière, E., Bengio, S., Bimbot, F., Hamouz, M., Kittler, J., Marithoz, J., Matas, J., Messer, K., Popovici, V., Porée, F., Ruiz, B., Thiran, J.-P.: The BANCA Database and Evaluation Protocol. In: Kittler, J., Nixon, M.S. (eds.) AVBPA 2003. LNCS, vol. 2688. Springer, Heidelberg (2003)
5. Turk, M., Pentland, A.: Eigenfaces for Recognition. Journal of Cognitive Neuroscience 3(1), 71–86 (1991)
6. Martin, A., Doddington, G., Kamm, T., Ordowsk, M., Przybocki, M.: The DET Curve in Assessment of Detection Task Performance. In: Proc. Eurospeech 1997, Rhodes, pp. 1895–1898 (1997)
7. Bengio, S., Marithoz, J.: The Expected Performance Curve: a New Assessment Measure for Person Authentication. In: The Speaker and Language Recognition Workshop (Odyssey), Toledo, pp. 279–284 (2004)
8. Poh, N., Bengio, S.: Database, Protocol and Tools for Evaluating Score-Level Fusion Algorithms in Biometric Authentication. Pattern Recognition 39(2), 223–233 (2005)
9. Martin, A., Przybocki, M., Campbell, J.P.: The NIST Speaker Recognition Evaluation Program, ch. 8. Springer, Heidelberg (2005)