Performance Characterisation of Face Recognition Algorithms and Their Sensitivity to Severe Illumination Changes

Kieron Messer[1], Josef Kittler[1], James Short[1], G. Heusch[2], Fabien Cardinaux[2], Sebastien Marcel[2], Yann Rodriguez[2], Shiguang Shan[3], Y. Su[3], Wen Gao[3], and X. Chen[3]

[1] University of Surrey, Guildford, Surrey, GU2 7XH, UK
[2] Dalle Molle Institute for Perceptual Artificial Intelligence, CP 592, rue du Simplon 4, 1920 Martigny, Switzerland
[3] Institute of Computing Technology, Chinese Academy of Sciences, China

Abstract. This paper details the results of a face verification competition [2] held in conjunction with the Second International Conference on Biometric Authentication. The contest was held on the publically available XM2VTS database [4] according to a defined protocol [15]. The aim of the competition was to assess the advances made in face recognition since 2003 and to measure the sensitivity of the tested algorithms to severe changes in illumination conditions. In total, more than 10 algorithms submitted by three groups were compared[1]. The results show that the relative performance of some algorithms is dependent on training conditions (data, protocol) as well as environmental changes.

1 Introduction

Over the last decade the development of biometric technologies has been greatly promoted by an important research technique instrument, namely comparative algorithm performance characterisation via competitions. Typical examples are the NIST evaluation campaign in voice based speaker recognition from telephone speech recordings, finger print competition, and face recognition and verification competitions. The main benefit of such competitions is that they allow different algorithms to be evaluated on the same data, using the same protocol. This makes the results comparable to much greater extent than in the case of an unorchestrated algorithm evaluation designed by individual researchers, using their own protocols and data, where direct comparison of the reported methods can be difficult because tests are performed on different data with large variations in test and model database sizes, sensors, viewing conditions, illumination and background. Typically, it is unclear which methods are the best and for which scenarios they should be used. The use of common datasets along with evaluation protocols can help alleviate this problem.

[1] This project was supported by EU Network of Excellence Biosecure.

D. Zhang and A.K. Jain (Eds.): ICB 2006, LNCS 3832, pp. 1–11, 2005.

In face recognition, the two main series of competitions has been run by NIST and the University of Surrey [13, 8, 14] respectively. For the purpose of the excersise, NIST collected a face database, known as FERET. A protocol for face identification and face verification [17] has been defined for the FERET database. However, only a development set of images from the database are released to researchers. The remaining are sequestered by the organisers to allow independent testing of the algorithms. To date three evaluations have taken place, the last one in the year 2000, and an account of these, together with the main findings can be found in [16].

More recently, two Face Recognition Vendor Tests [3] have been carried out, the first in 2000 and the second in 2002. The tests are done under supervision and have time restrictions placed on how quickly the algorithms should compute the results. They are aimed more at independently testing the performance of commercially available systems, however academic institutions are also able to take part. In the more recent test 10 commercial systems were evaluated. The FERET and FRVT have recently evolved in a new initiative known as Face Recognition Grand Challenge which is promoting research activities both in 2D and 3D face recognition.

The series of competitions organised by the University of Surrey commenced in the year 2000. It was initiated by the European Project M2VTS which focused on the development of multimodal biometric personal identity authentication systems. As part of the project a large database of talking faces was recorded. For a subset of the data, referred to as the XM2VTS database, two experimental protocols, known as Lausanne Protocol I and II, were defined to enable a cooperative development of face and speaker verification algorithms by the consortium of research teams involved in the project. The idea was to open this joint development and evaluation of biometric algorithms to wider participation.

In the year 2000 a competition on the XM2VTS database using the Lausanne protocol [15] was organised [13]. As part of AVBPA 2003 a second competition on exactly the same data and testing protocol was organised [8]. All the data from the XM2VTS database is available from [4]. We believe that this open approach increases, in the long term, the number of algorithms that will be tested on the XM2VTS database. Each research institution is able to assess their algorithmic performance at any time.

The competition was subsequently extended to a new database, known as the BANCA database [5] which was recorded as part of a follow up EU project, BANCA. The database was captured under 3 different realistic and challenging operating scenarios. Several protocols have also been defined which specify which data should be used for training and testing. Again this database is being made available to the research community through [1]. The first competition on the BANCA database was held in 2004 and the results reported in [14].

In this paper, the competition focuses once again on XM2VTS data with two objectives. First of all it is of interest to measure the progress made in face verification since 2003. The other was to gauge the sensitivity of face verification algorithms to severe changes to illumination conditions. This test was carried

out on a section of the XM2VTS database containing face images acquired in side lighting. As with the previous competition, the current event was held under the auspices of EU Project Biosecure.

The rest of this paper is organised as follows. In the next section the competition rules and performance criterion are described. Section 3 gives an overview of each algorithm which entered the competition and in the following section the results are detailed. Finally, some conclusions are drawn in Section 4.

2 The Competition

All experiments were carried out using images acquired from the XM2VTS database on the standard and darkened image sets. The XM2VTS database can be aquired through the web-page given by [4].

There were two separate parts to the competition.

Part I: Standard Test. The XM2VTS database contains images of 295 subjects, captured over 4 sessions in a controlled environment. The database uses a standard protocol. The Lausanne protocol splits the database randomly into training, evaluation and test groups [15]. The training group contains 200 subjects as clients, the evaluation group additional 25 subjects as impostors and the testing group another 70 subjects as impostors.

There are two testing configurations of the XM2VTS database. In the first configuration, the client images for training and evaluation, were collected from each of the first three sessions. In the second configuration, the client images for training were collected from the first two sessions and the client images for evaluation from the third.

Part II: Darkened Images. In addition to the controlled images, the XM2VTS database contains a set of images with varying illumination. Each subject has four more images with lighting predominantly from one side; two have been lit from the left and two from the right.

To assess the algorithmic performance the *False Rejection Rate* P_{FR} and *False Acceptance Rate* P_{FA} are typically used. These two measures are directly related, i.e. decreasing the false rejection rate will increase the number of false acceptances. The point at which $P_{FR} = P_{FA}$ is known as the EER (Equal Error Rate).

Fig. 1. Example images from XM2VTS database

Fig. 2. Example images from dark set of XM2VTS database

3 Overview of Algorithms

In this section the algorithms that participated in the contest are summarised.

3.1 Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP)

IDIAP proposed three different classifiers, used with two distinct preprocessing steps, resulting in a total of six complete face authentication systems. The preprocessing steps aims to enhance the image or to reduce the effect of illumination changes. The first preprocessing step we used is simple histogram equalization whereas the second one is the illumination normalization model first proposed by Gross & Brajovic [10] and described in details in [9].

The first two classification systems (called GMM and HMM) are based on local features and statistical models while the third one (called PCA-LDA) uses discriminant holistic features with a distance metric.

IDIAP-GMM. The GMM based system uses DCT-mod2 features [18] and models faces using Gaussian Mixture Models (GMMs) [6]. In DCTmod2 feature extraction, each given face is analyzed on a block by block basis: from each block a subset of Discrete Cosine Transform (DCT) coefficients is obtained; coefficients which are most affected by illumination direction changes are replaced with their respective horizontal and vertical deltas, computed as differences between coefficients from neighboring blocks. A GMM is trained for each client in the database. To circumvent the problem of small amount of client training data, parameters are obtained via Maximum a Posteriori (MAP) adaptation of a generic face GMM: the generic face GMM is trained using Maximum Likelihood training with faces from all clients. A score for a given face is found by taking the difference between the log-likelihood of the face belonging to the claimed identity (estimated with the client specific GMM) and log-likelihood of the face belonging to an impostor (estimated with the generic face GMM). A global threshold is used in making the final verification decision.

IDIAP-HMM. The HMM based system uses DCT features and models faces using Hidden Markov Models (HMMs). Here, we use a simple DCT Feature extraction: each given face is analyzed on a block by block basis; from each block, DCT coefficients are obtained; the first fifeteen coefficients compose the feature

vector corresponding to the block. A special topology of HMM is used to model the client faces which allows the use of local features. The HMM represents a face as a sequence of horizontal strips from the forehead to the chin. The emission probabilities of the HMM are estimated with mixture of gaussians modeling the set of blocks that composes a strip. A further description of this model is given in [7]. A HMM is trained for each client in the database using MAP adaptation. A score for a given face is found by taking the difference between the log-likelihood of the face belonging to the claimed identity (estimated with the client specific GMM) and log-likelihood of the face belonging to an impostor (estimated with the generic face GMM). A global threshold is used in making the final verification decision.

IDIAP-PCA/LDA. Principal component analysis (PCA) is first applied on the data so as to achieve decorrelation and dimensionality reduction. The projected face images into the coordinate system of eigenvectors (Eigenfaces) are then used as features to derive the optimal projection in the Fisher's linear discriminant sense (LDA) [12]. Considering a set of N images $\{x_1, x_2, ..., x_N\}$, an image x_k is linearly projected to obtain the feature vector y_k:

$$y_k = W^T x_k \qquad k = 1, 2, ..., N$$

where $W^T = W_{lda}^T W_{pca}^T$. Finally, classification is performed using a metric: considering two feature vectors, a template y_t and a sample y_s, their correlation is computed according to:

$$1 - \frac{< y_t, y_s >}{\|y_t\| \cdot \|y_s\|}$$

3.2 Chinese Academy of Sciences

The adopted method, Gabor Feature based Multiple Classifier Combination (CAS-GFMCC), is an ensemble learning classifier based on the manipulation of Gabor features with multiple scales and orientations. The basic procedure of CAS-GFMCC is described as follows: First, face images are aligned geometrically and normalized photometrically by using region-based histogram equalization. Then, Gabor filters with 5 scales and 8 orientations are convolved with the normalized image and the magnitude of the transform results are kept for further processing. These high dimensional Gabor features, with a dimension of 40 times of the original normalized face images, are then adaptively divided into multiple groups. For each feature group, one classifier is learnt through Fisher discriminant analysis, which will result in an ensemble of classifier. These classifiers are then combined using a fusion strategy. In addition, face image re-lighting techniques are exploited to adapt the method for more robustness to the face images with complex illumination (named by CAS-GFMCC-L). For automatic evaluation case, AdaBoost-based methods are exploited for both the localization of the face and facial landmarks (the two eyes). Please refer to http://www.jdl.ac.cn/project/faceId/index_en.htm for more details of our methods.

3.3 University of Surrey (UniS)

Two algorithms have been tested using the competition protocol. The first algorithm (Unis-Components) applies client-specific linear discriminant analysis to a number of components of the face image. Firstly, twelve sub-images are obtained. The images are found relative to the eye positions, so that no further landmarking is necessary. These images are of the face, both eyes, nose and mouth and of the left and right halves of each, respectively. All twelve images have the same number of pixels, so that the images of smaller components will effectively be of higher resolution. These components are then normalised using histogram equalisation. Client-specific linear discriminant analysis [11] is applied to these sub-images separately. The resulting scores for each of the components are fused using the sum rule.

The second algorithm (UniS-Lda) is based on the standard LDA. Each image is first photometrically normalised using filtering and histogram equalisation. The corrected images are then projected into an LDA space which has been designed by first reducing the dimensionality of the image representation space using PCA. The similarity of probe and template images is measured in the LDA space using normalised correlation. In contrast to the results reported in the AVBPA2003 competition, here the decision threshold is globally optimal rather than client specific. For the automatic registration of the probe images, an SVM based face detection and eye localisation algorithm was used. Exactly the same system was used in Part II of the competition, without any adjustmenst of the system parameters, including the decision threshold.

4 Results and Discussion

4.1 Part I

Most of the algorithm entries provide results for Part I of the competition with manually registered images, which is aimed at establishing a bench mark for the other parts. As there were so few entrants, the competition was used as a framework for comparative evaluation of different algorithms from two of the groups, rather than just the best performing entry. This offered an interesting insight into the effectiveness of different decision making schemes under the same photometric normalisation conditions, and the dependence of each decision making scheme on different photometric normalisation methods. Interestingly, the best combination of preprocessing and decision making methods investigated by IDIAP differed from one evaluation protocol to another.

In general the performance of the algorithms achieved under Protocol II was better. This is probably the consequence of more data being available for training and the evaluation data available for setting the operational thresholds being more representative, as it was recorded in a completely different session. The best performing algorithm was CAS, which also achieved the best results on the BANCA database in the previous competition. The CAS algorithm outperformed the winning algorithm on the XM2VTS database at the AVBPA03 competition [8].

Table 1. Error rates according to Lausanne protocol for configuration I with manual registration

	Evaluation Set			Test Set		
Method	FA	FR	TER	FA	FR	TER
ICPR2000-Best	-	-	5.00	2.30	2.50	4.80
AVBPA03-Best	1.16	1.05	2.21	0.97	0.50	1.47
IDIAP-HE/GMM	2.16	2.16	4.32	2.00	1.50	3.50
IDIAP-HE/HMM	2.48	2.50	4.98	2.57	1.50	4.07
IDIAP-HE/PCA/LDA	3.16	3.33	6.49	3.72	2.00	5.72
IDIAP-GROSS/GMM	2.20	2.17	4.37	2.32	2.00	4.32
IDIAP-GROSS/HMM	6.00	6.00	12.0	6.31	4.75	11.06
IDIAP-GROSS/PCA/LDA	5.96	6.00	11.96	7.04	4.50	11.54
UNIS-Components	5.50	5.50	11.00	4.44	3.50	7.94
UNIS-Lda	1.66	1.67	3.33	1.66	1.25	2.91
CAS	0.80	0.80	1.63	0.96	0.00	0.96

Table 2. Error rates according to Lausanne protocol for configuration II with manual registration

	Evaluation Set			Test Set		
Method	FA	FR	TER	FA	FR	TER
AVBPA03-Best	0.33	0.75	1.08	0.25	0.50	0.75
IDIAP-HE/GMM	1.00	1.00	2.00	0.04	4.75	4.79
IDIAP-HE/HMM	1.75	1.75	3.50	1.80	1.25	3.05
IDIAP-HE/PCA/LDA	1.64	1.75	3.39	1.86	3.25	5.11
IDIAP-GROSS/GMM	1.00	1.00	2.00	1.15	1.00	2.15
IDIAP-GROSS/HMM	5.25	5.25	10.50	5.13	3.25	8.38
IDIAP-GROSS/PCA/LDA	3.25	3.25	6.50	4.01	5.75	9.76
UNIS-Components	2.64	2.75	5.39	1.99	1.75	3.74
UNIS-Lda	1.00	1.00	2.00	1.26	0.00	1.26
CAS	0.24	0.25	0.49	0.26	0.25	0.51

Table 3. Error rates according to Lausanne protocol for configuration I with automatic registration in test phase

	Evaluation Set			Test Set		
Method	FA	FR	TER	FA	FR	TER
ICPR2000-Best	-	-	14.0	5.80	7.30	13.10
AVBPA03-Best	0.82	4.16	4.98	1.36	2.50	3.86
CAS	1.00	1.00	2.00	0.57	1.57	1.57

Only one of the algorithms, CAS, was also subjected to the test on automatically registered images. The automatic registration was accomplished with a CAS in house face detection and localisation method. By default, CAS is the winning entry. However, the achievement of the CAS method should not be

Table 4. Error rates according to Lausanne protocol for configuration II with auto registration in test phase

	Evaluation Set			Test Set		
Method	FA	FR	TER	FA	FR	TER
AVBPA03-Best	0.63	2.25	2.88	1.36	2.00	3.36
CAS	0.49	0.50	0.99	0.28	0.50	0.78

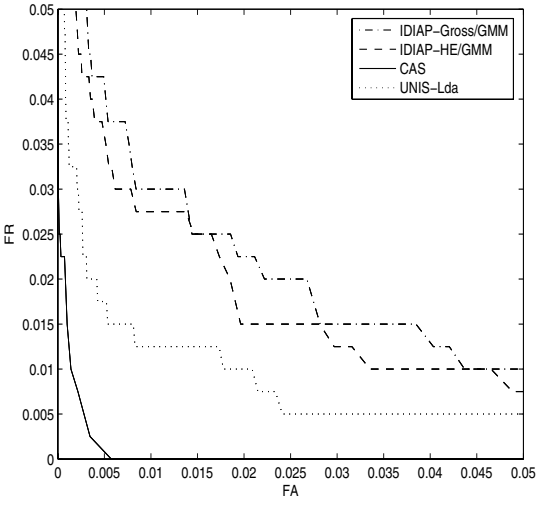

Fig. 3. ROC curves for configuration I with manual registration

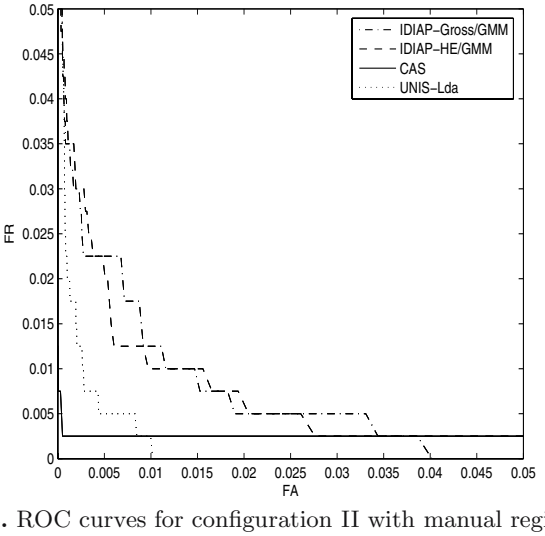

Fig. 4. ROC curves for configuration II with manual registration

underrated, as the overall performance shown in Table 3 and Table 4 is very impressive. The results show only a slight degradation, in comparison with the

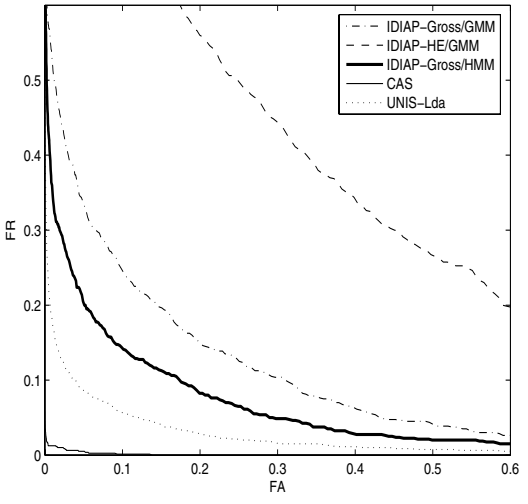

Fig. 5. ROC curves for the dark test set with manual registration

manually registered figures. More over, the results are a significant improvement over the previously best reported results.

Figures 3, 4, 5 provide the ROC curves for the better performing methods. It is interesting to note that if the operating points were selected aposteriori, then the performance of the algorithms would be even better. This suggests that if the evaluation data set was more extensive and therefore fully representative, the error rates could be reduced even further.

4.2 Part II

This part of the competition provided a useful insight into the sensitivity of the tested algorithms to severe changes in subject illumination. In some cases the performance degraded by an order of magnitude. Surprisingly, the error rates of some of the lower ranking methods, such as the Unis-Components and IDIAP LDA based procedures, deteriorated only by a factor of two. Again, the CAS approach achieved the best performance, which was an order of magnitude better than the second best algorithm. The comparability of the results was somewhat affected by the interesting idea of CAS to relight the training and evaluation set data to simulate the illumination conditions of the test set. This has no doubt limited the degree of degradation from good conditions to side lighting. However, it would have been interesting to see how well the system would perform on the original frontal lighting data sets. This would better indicate the algorithm sensitivity to changes in lighting conditions.

The CAS algorithm was the only entry in Part II, automatic registration category. Again the reported results are consistently excellent, demonstrating a high degree of robustness of the CAS system and the overall high level of performance.

Table 5. Darkened image set with manual registration

	Evaluation Set			Test Set		
Method	FA	FR	TER	FA	FR	TER
IDIAP-HE/GMM				6.20	77.37	88.68
IDIAP-HE/HMM				12.78	60.75	73.53
IDIAP-HE/PCA/LDA				2.41	29.50	31.91
IDIAP-GROSS/GMM				10.54	23.75	34.29
IDIAP-GROSS/HMM				8.14	15.86	24.00
IDIAP-GROSS/PCA/LDA				6.49	18.75	25.24
UNIS-Components				4.01	17.38	21.39
UNIS-Lda				17.88	0.98	18.86
CAS	1.18	1.17	2.35	0.77	1.25	2.02

Table 6. Darkened image set with automatic registration

	Evaluation Set			Test Set		
Method	FA	FR	TER	FA	FR	TER
CAS	1.18	1.17	2.35	1.25	1.63	2.88

5 Conclusions

The results of a face verification competition [2] held in conjunction with the Second International Conference on Biometric Authentication have been presented. The contest was held on the publically available XM2VTS database [4] according to a defined protocol [15]. The aim of the competition was to assess the advances made in face recognition since 2003 and to measure the sensitivity of the tested algorithms to severe changes in illumination conditions. In total, more than 10 algorithms submitted by three groups were compared. The results showed that the relative performance of some algorithms is dependent on training conditions(data, protocol). All algorithms were affected by environmental changes. The performance degraded by a factor of two or more.

References

1. *BANCA; http://www.ee.surrey.ac.uk/banca/.*
2. *BANCA; http://www.ee.surrey.ac.uk/banca/icba2004.*
3. *Face Recognition Vendor Tests; http://www.frvt.org.*
4. *The XM2VTSDB; http://www.ee.surrey.ac.uk/ Research/VSSP/xm2vtsdb/.*
5. E. Bailly-Bailliere, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariethoz, J. Matas, K. Messer, V. Popovici, F. Poree, B. Ruiz, and J. P. Thiran. The BANCA database and evaluation protocol. In *Audio- and Video-Based Biometric Person Authentication: Proceedings of the 4th International Conference, AVBPA 2003*, volume 2688 of *Lecture Notes in Computer Science*, pages 625–638, Berlin, Germany, June 2003. Springer-Verlag.

6. F. Cardinaux, C. Sanderson, and S. Bengio. User authentication via adapted statistical models of face images. *To appear in IEEE Transactions on Signal Processing*, 2005.
7. Fabien Cardinaux. Local features and 1D-HMMs for fast and robust face authentication. Technical report, 2005.
8. Kieron Messer et al. Face verification competition on the xm2vts database. In *4th International Conference on Audio and Video Based Biometric Person Authentication*, pages 964–974, June 2003.
9. F. Cardinaux G. Heusch and S. Marcel. Efficient diffusion-based illumination normalization for face verification. Technical report, 2005.
10. R. Gross and V. Brajovic. An Image Preprocessing Algorithm for Illumination Invariant Face Recognition. In *International Conference on Audio- and Video-Based Biometric Person Authentication*, 2003.
11. J. Kittler, Y. P. Li, and J. Matas. Face verification using client specific fisher faces. *The Statistics of Directions, Shapes and Images (2000) 63–66.*
12. S. Marcel. A symmetric transformation for lda-based face verification. In *Proc. Int. Conf. Automatic Face and Gesture Recognition (AFGR), Seoul, Korea*, 2004.
13. J Matas, M Hamouz, K Jonsson, J Kittler, Y P Li, C Kotropoulos, A Tefas, I Pitas, T Tan, H Yan, F Smeraldi, J Bigun, N Capdevielle, W Gerstner, S Ben-Yacoub, Y Abdeljaoued, and E Mayoraz. Comparison and face verification results on the xm2vts database. In A Sanfeliu, J J Villanueva, M Vanrell, R Alquezar, J Crowley, and Y Shirai, editors, *Proceedings of International Conference on Pattern Recognition, Volume 4*, pages 858–863, 2000.
14. K Messer, J Kittler, M Sadeghi, M Hamouz, A Kostin, and et al. Face authentication test on the banca database. In J.Kittler, M Petrou, and M Nixon, editors, *Proc. 17th Intern. Conf. on Pattern Recognition*, volume IV, pages 523–529, Los Alamitos, CA, USA, August 2004. IEEE Computer Society Press.
15. 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.
16. P. J. Phillips, H. Moon, P. Rauss, and S. A. Rizvi. The feret evaluation methodology for face-recognition algorithms. volume 22, pages 1090–1104, October 2000.
17. 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.
18. C. Sanderson and K.K. Paliwal. Fast features for face authentication under illumination direction changes. *Pattern Recognition Letters*, 24(14):2409–2419, 2003.