Comparison of face verification results on the XM2VTS database

J. Matas[1,2], M. Hamouz[1], K. Jonsson[1], J. Kittler[1], Y. Li[1], C. Kotropoulos[3], A. Tefas[3], I. Pitas[3], Teewoon Tan[5], Hong Yan[5], F. Smeraldi[7], J. Bigun[4], N. Capdevielle[7], W. Gerstner[7], S. Ben-Yacoub[8], Y. Abdeljaoued[7], E. Mayoraz[6]

[1]Centre for Vision Speech and Signal Processing
University of Surrey, Guildford, GU2 7XH, UK
{g.matas, m.hamouz, k. jonsson, j.kittler, y.li}@eim.surrey.ac.uk

[2]Center for Machine Perception
CTU Prague, Karlovo nám. 13, 121 35
Czech Republic

[3]Department of Informatics
Aristotle University of Thessaloniki
Box 451, Thessaloniki 540 06, Greece
{costas, tefas, pitas}@zeus.csd.auth.gr

[4]Halmstadt University
Box 823, S-30115 Halmstad, Sweden
josef.bigun@ide.hh.se

[5]University of Sydney
NSW 2006 Australia
{teewoon,yan}@ee.usyd.edu.au

[6]Motorola Inc.
Eddy.Mayoraz@lexicus.mot.com

[7]Swiss Federal Institute of Technology
DI, 1015 Lausanne, Switzerland
{nathalie.capdevielle, wulfram.gerstner, yousri.abdeljaoued}@epfl.ch

[8]Swisscom AG
Souheil.BenYacoub@swisscom.com

Abstract

The paper presents results of the face verification contest that was organized in conjunction with International Conference on Pattern Recognition 2000 [14]. Participants had to use identical data sets from a large, publicly available multimodal database XM2VTSDB. Training and evaluation was carried out according to an a priori known protocol ([7]). Verification results of all tested algorithms have been collected and made public on the XM2VTSDB website [15], facilitating large scale experiments on classifier combination and fusion. Tested methods included, among others, representatives of the most common approaches to face verification – elastic graph matching, Fisher's linear discriminant and Support vector machines.

1 Introduction

Hundreds of papers have been published on the face verification and recognition problem [11, 2]. Direct comparison of the reported methods is typically rather difficult, because tests are performed on different data, with large variations in test and model database sizes, viewing conditions, background etc. Standard face databases are publicly available, e.g. Yale [21], Harvard [16] , Olivetti [19], M2VTS [18] to name a few commonly used (see the face recognition homepage [17] for a longer list. Even if the same database is used, it may be split differently into test and training sets. Moreover, results are often evaluated using different methodologies.

For face recognition, the FERET test [10, 9] provided a comparison of a number of algorithms. However, a similar test was missing for the *face verification* (or authentication) task. Verification and recognition differ in at least three fundamental aspects. Firstly, a client – an authorized user of a personal identification system – is assumed to be cooperative and makes an identity claim. Computationally this means that it is not necessary to consult the complete set of models (reference images in our case) in order to verify a claim. A test image is thus compared to a small number of reference images of the person whose identity is claimed and not, as in the recognition scenario, with every image (or some descriptor of an image) in a potentially large database. Secondly, an automatic authentication system must operate in near-real time to be acceptable to users. And finally, in recognition experiments only images of people from the training database are presented to the system, whereas the case of an impostor (most likely a previously unseen person) is of utmost importance for authentication.

In order to collect face verification results on an identical, publicly available, data set using a standard performance assessment methodology a contest was organized in conjunction with the ICPR 2000. Besides assessing the quality of various face verification methods, the contest's secondary objective was to make the results of different methods on particular algorithms available to the research community. Placing the results, in a predefined format, on a publicly accessible web site [15] enables large scale experiments on classifier combination and fusion as well as the study of dependencies of errors of a wide range of face verification methods.

The results published are based completely on self-assessment of the research groups providing the error rates. In the original call for participation [14], a second part of the test on sequestered data was mentioned, but we were not able to carry it out in time to meet the publication deadline. Unlike in the FERET test, where each research group obtained a different subset of the database, all research groups have identical data sets and therefore can assess their performance at any time. We believe that this open approach, trusting the published results, will increase in the long term the number of algorithms that will be tested on the XM2VTSDB subset.

The rest of the paper is structured as follows. In section 2, the image dataset and the evaluation protocol is described. In section 3 results evaluated according to the Lausanne protocol are presented. Section 4 introduces other results, that are not exactly according to the Lausanne protocol.

2 XM2VTS database and Lausanne protocol

The XM2VTS database [8] is a multimodal database consisting of face images, video sequences and speech recordings taken of 295 subjects at one month intervals. This database is available at the cost of distribution from the University of Surrey (see [20] for details). 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. Since the data acquisition was distributed over a long period of time, significant variability of appearance of clients, e.g. changes of hair style, facial hair, shape and presence or absence of glasses, is present in the recordings - see figure 1.

The subjects were volunteers, mainly employees and PhD students at the University of Surrey of both sexes and many ethnical origins. The XM2VTS database contains 4 sessions. During each session two head rotation and "speaking" shots were taken. From the "speaking" shot, where subjects are looking just below the camera while reading a phonetically balanced sentence, a single image with a closed mouth was chosen. Two shots at each session, with and without glasses, were acquired for people regularly wearing glasses.

Figure 1. Sample images from XM2VTS database

For the task of personal verification, a standard protocol for performance assessment has been defined. The so called Lausanne protocol splits randomly all subjects into a client and impostor groups. The client group contains 200 subjects, the impostor group is divided into 25 evaluation impostors and 70 test impostors. Eight images from 4 sessions are used.

From these sets consisting of face images, training set, evaluation set and test set is built. There exist two configurations that differ by a selection of particular shots of people into the training, evaluation and test set. The training set is used to construct client models. The evaluation set is selected to produce client and impostor access scores, which are used to find a threshold that determines if a person is accepted or not (it can be a client-specific threshold or global threshold). According to the Lausanne protocol the threshold is set to satisfy certain performance levels (error rates) on the evaluation set. Finally the test set is selected to simulate realistic authentication tests where impostor's identity is unknown to the system. The evaluation set is also used in fusion experiments (classifier combination) for training, but this is not relevant in the context of this paper.

The performance measures of a verification system are the False Acceptance rate (FA) and the False Rejection rate (FR). False acceptance is the case where an impostor, claiming the identity of a client, is accepted. False rejection is the case where a client, claiming his true identity, is rejected. FA and FR are given by:

$$FA = EI/I * 100\% \qquad FR = EC/C * 100\% \qquad (1)$$

where EC is the number of impostor acceptances, I is the number of impostor claims, EC the number of client rejections, and C the number of client claims. Both FA and an FR are influenced by an acceptance threshold. To simulate real application the threshold is set on the data from evaluation set to obtain certain false acceptance on the evaluation set (FAE) and false rejection error (FRE). The same threshold is afterwards applied to the test data and FA and FR on the test set are computed. Three thresholds are defined on the evaluation set:

$$\begin{array}{l} T_{FAE=0} = \arg\min_T(FRE|FAE = 0) \\ T_{FAE=FRE} = (T|FAE = FRE) \\ T_{FRE=0} = \arg\min_T(FAE|FRE = 0) \end{array} \quad (2)$$

Consequently, performance on the test set is characterised by six error rates:

$$\begin{array}{ll} FA_{FAE=0} & FR_{FAE=0} \\ FA_{FAE=FRE} & FR_{FAE=FRE} \\ FA_{FRE=0} & FR_{FRE=0} \end{array} \quad (3)$$

3 Results on XM2VTS database evaluated according to the Lausanne protocol

This section describes results of face verification methods that either participated in the contest or had been tested according to the Lausanne protocol. In all cases, files storing verification results have been made public.

3.1 Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP)

The Elastic Graph Matching introduces a specific face representation. Each face is represented by a set of feature vectors positioned on nodes of a coarse, rectangular grid placed on the image. Moduli of complex Gabor responses from filters with 6 orientations and 3 resolutions are used as features. The matching consists of two consecutive steps: rigid matching and deformable matching. Advantages of the elastic graph matching are the robustness against variation in face position, and expression. This owes to the Gabor features, the rigid matching stage, and the deformable matching stage [1]. Results can be found in tables 1 and 2.

3.2 Aristotle University of Thessaloniki

The approach of C. Kotropoulos, A. Tefas and I. Pitas [5] also falls into the Elastic Graph Matching category. However, novel features based on multiscale dilation and erosion operations are computed at each node of the grid. Verification score is a function of the matching energy which in turn is a complex function of the grid deformation and the difference of response of the morphological operators obtained at node locations. Results of the method are labelled **AUT** in tables 1 and 2.

3.3 University of Surrey

University of Surrey provided several results (see tables 1 and 2).

In [6] the problem of face verification using linear discriminant analysis was addressed and the issue of matching score investigated. The improved understanding about the role of metric led to a novel way of measuring the distance between probe image and a model. The effect of various photometric normalizations on the matching scores was also investigated. In tables 1 and 2 the group of results are referred to as **UniS–X–X–NC** which stands for normalized correlation used as distance measure and as **UniS–X–X–SM** which stands for the novel proposed metric. The experiments were conducted using both types of thresholding - a client-specific thresholds and global thresholds. Also both types of registration were used — fully-automatic registration (based on robust correlation described in [4] — see next paragraph) and semi-automatic registration, where the eyes of people were located manually. Results with the automatic registration are available only in Configuration I.

Another verification approach is based on the use of robust correlation and Support vector machines [4]. The problem of the registration was treated as an optimization task via estimating the optimal transformation parameters by maximizing a similarity function. The influence of signal noise, occluding objects and suboptimalities in the transformation models were reduced by applying robust estimation techniques. The classification of face patterns was carried out by using a support vector machine. The registered and photometrically normalized images were projected into a subspace optimized for representation or discrimination. The client–specific thresholding was used. The results are referred to as **UniS–SVM** and are available only for configuration I.

3.4 University of Sydney

Fractal image coding was applied to the task of face verification. Two subsystems constituted this face verification system, namely the face detection and the face verification components. Central to both systems is the notion of the Fractal Neighbor Distance (FND). The detection system firstly performed a rough location of the head, based on the assumption of a blue background. A search was then performed in the reduced region. This involved the use of a generic face template, the fractal code of which had been generated. Afterwards the Fractal Neighbor Distances between localized head images and the images stored in the database were computed. The minimal FND was taken as a score [13]. See tables 1 and 2 for results.

4 Other results

In this section results that have not been obtained according to the protocol are presented.

4.1 Swiss Federal Institute of Technology (EPFL)

In the approach of Smeraldi et al. [12] a concept of Retinal vision was introduced. The raw visual input was analyzed by means of a log-polar retinotopic sensor, whose receptive fields consisted of a vector of modified Gabor filters designed in the log-polar frequency plane. The Gabor responses extracted by placing the sensor over the corresponding facial regions were then used to perform authentication. The implementation of knowledge representation using Support vector machine classifier was used. Since the training and test sets were used exactly according to the Lausanne protocol (although evaluation set was not used at all), it is still possible to compare the results. The aposteriori equal error rate on the test set was about 0.50%.

5 Conclusion and Future Work

This paper presents a comparison of face verification algorithms that was organized in conjunction with International Conference on Pattern Recognition 2000. Fourteen face verification methods were tested using identical data sets from a large, publicly available multimodal database XM2VTSDB. Training and evaluation was carried out according to an a priori known protocol. Verification results of all tested algorithms have been collected and made public on the internet [15], facilitating large scale experiments on classifier combination and fusion.

References

[1] S. Ben-Yacoub, Y. Abdeljaoued, and E. Mayoraz. Fusion of face and speech data for person identity verification. *IEEE Transactions on Neural Networks*, 10(05):1065–1074, 1999.

[2] R. Chellappa, C. L. Wilson, and S. Sirohey. Human and machine recognition of faces: A survey. *Proceedings of the IEEE*, 83(5):704–740, May 1995.

[3] K. Jonsson. *Robust Correlation and Support Vector Machines for Face Identification*. PhD thesis, University of Surrey, 2000.

[4] K. Jonsson, J. Kittler, Y. P. Li, and J. Matas. Support Vector Machines for Face Authentication. In T. Pridmore and D. Elliman, editors, *British Machine Vision Conference*, pages 543–553, 1999.

[5] C. Koutropoulos, A. Tefas, and I. Pitas. Morphological elastic graph matching applied to frontal face authentication under well-controlled and real conditions. Technical report, Aristotle university of Thessaloniki, 1999.

[6] Y. Li, J. Kittler, and J. Matas. On Matching Scores of LDA-based Face Verification. In T. Pridmore and D. Elliman, editors, *Proc British Machine Vision Conference BMVC2000*, page submitted, London, UK, September 2000. University of Bristol, British Machine Vision Association.

[7] J. Luettin and G. Maître. Evaluation Protocol for the extended M2VTS Database (XM2VTSDB). IDIAP-COM 05, IDIAP, 1998.

[8] 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.

[9] P. Phillips, H. Wechsler, J.Huang, and P. Rauss. The FERET database and evaluation procedure for face-recognition algorithm. *Image and Vision Computing*, 16:295–306, 1998.

[10] P. J. Phillips, H. Moon, P. Rauss, and S. A. Rizvi. The FERET evaluation methodology for face-recognition algorithms. In *Proceedings of CVPR97*, pages 137–143, 1997.

[11] A. Samal and P. Iyengar. Automatic Recognition and Analysis of Human Faces and Facial Expressions: A Survey . *Pattern Recognition*, 25:65–77, 1992.

[12] F. Smeraldi, N. Capdevielle, and J. Bigun. Face Authentication by retinotopic sampling of the Gabor decomposition and Support Vector Machines. In *Proceedings of the 2nd International Conference on Audio and Video Based Biometric Person Authentication (AVBPA'99), Washington DC (USA)*, 1999.

[13] T. Tan and H. Yan. Face recognition by fractal transformations. In *Proc. IEEE ICASSP*, pages 3537–3540, 1999.

[14] http://xm2vtsdb.ee.surrey.ac.uk/face-icpr2000/index.html.

[15] http://xm2vtsdb.ee.surrey.ac.uk/results/face/verification_LP/.

[16] ftp://hrl.harvard.edu/pub/faces.

[17] http://www.cs.rug.nl/~peterkr/FACE/face.html.

[18] http://ns1.tele.ucl.ac.be/M2VTS/.

[19] http://www.cam-orl.co.uk/facedatabase.html.

[20] http://xm2vtsdb.ee.surrey.ac.uk/.

[21] http://cvc.yale.edu/projects/yalefaces/yalefaces.html.

EXPERIMENT	Configuration I								
	Evaluation set			Test set					
	[1]FAE = [2]FRE	FAE (FRE=0)	FRE (FAE=0)	FAE=FRE		FRE=0		FAE=0	
				FA	FR	FA	FR	FA	FR
[3]**AUT**	8.1	48.4	19.0	8.2	6.0	46.6	0.8	0.5	20.0
[4]**IDIAP**	8.0	54.9	16.0	8.1	8.5	54.5	0.5	0.5	20.5
[5]**Sydney**	12.9	94.4	70.5	13.6	12.3	94.0	0.0	0.0	81.3
[6]**UniS–A–G–NC**	5.7	96.4	26.7	7.6	6.8	96.5	0.3	0.0	27.5
[7]**UniS–A–S–NC**	5.3	99.3	25.3	7.4	6.8	99.4	0.0	0.0	24.3
[8]**UniS–S–G–NC**	3.5	81.1	16.2	3.5	2.8	81.2	0.0	0.0	14.5
[9]**UniS–S–S–NC**	3.3	92.9	15.7	3.3	3.0	93.5	0.0	0.0	14.0
[10]**UniS–A–G–SM**	10.0	99.8	93.8	9.8	8.8	100.0	0.0	0.0	97.3
[11]**UniS–A–S–SM**	7.0	97.3	63.7	5.8	7.3	99.6	0.0	0.0	58.5
[12]**UniS–S–G–SM**	6.5	93.9	40.0	6.5	5.3	94.1	0.0	0.0	37.5
[13]**UniS–S–S–SM**	2.5	84.2	25.7	2.3	2.5	85.0	0.3	0.0	24.3
[14]**UniS–SVM**	6.9	—	—	7.7	6.3	—	—	—	—

Table 1. Error rates according to the Lausanne protocol for configuration I

[1]**FAE** stands for false acceptance error rate on evaluation set
[2]**FRE** stands for false rejection error rate on evaluation set
[3]**AUT** Aristotle University of Thessaloniki
[4]**IDIAP** Dalle Molle Institute for Perceptual Artificial Intelligence
[5]**Sydney** University of Sydney
[6]**UniS–A–G–NC** University of Surrey, full automatic registration, global threshold, normalized correlation
[7]**UniS–A–S–NC** University of Surrey, full automatic registration, client-specific threshold, normalized correlation
[8]**UniS–S–G–NC** University of Surrey, semi-automatic registration, global threshold, normalized correlation
[9]**UniS–S–S–NC** University of Surrey, semi-automatic registration, client-specific threshold, normalized correlation
[10]**UniS–A–G–SM** University of Surrey, full automatic registration, global threshold, special metric
[11]**UniS–A–S–SM** University of Surrey, full automatic registration, client-specific threshold, special metric
[12]**UniS–S–G–SM** University of Surrey, semi-automatic registration, global threshold, special metric
[13]**UniS–S–S–SM** University of Surrey, semi-automatic registration, client-specific threshold, special metric
[14]**UniS–SVM** University of Surrey, fully-automatic registration, Support vector machine

EXPERIMENT	Configuration II								
	Evaluation set			Test set					
	[1]FAE = [2]FRE	FAE (FRE=0)	FRE (FAE=0)	FAE=FRE		FRE=0		FAE=0	
				FA	FR	FA	FR	FA	FR
[3]**AUT**	6.5	36.9	18.8	6.2	3.5	34.7	0.8	0.5	16.3
[4]**IDIAP**	7.0	59.4	19.8	7.7	7.3	0.3	54.2	1.0	18.0
[5]**Sydney**	14.1	98.4	80.8	13.0	12.3	98.1	0.0	0.0	84.8
[6]**UniS–S–G–NC**	1.3	43.5	9.3	1.3	1.8	44.2	0.3	0.0	9.0
[7]**UniS–S–S–NC**	1.3	55.4	8.5	1.2	1.5	55.6	0.3	0.0	8.5
[8]**UniS–S–G–SM**	3.5	42.0	18.8	3.5	3.8	42.1	0.5	0.0	19.5
[9]**UniS–S–S–SM**	1.3	23.1	18.8	1.2	1.0	22.6	0.3	0.0	20.5

Table 2. Error rates according to the Lausanne protocol for configuration II

[1]**FAE** stands for false acceptance error rate in evaluation set
[2]**FRE** stands for false rejection error rate in evaluation set
[3]**AUT** Aristotle University of Thessaloniki
[4]**IDIAP** Dalle Molle Institute for Perceptual Artificial Intelligence
[5]**Sydney** University of Sydney
[6]**UniS–S–G–NC** University of Surrey, semi-automatic registration, global threshold, normalized correlation
[7]**UniS–S–S–NC** University of Surrey, semi-automatic registration, client-specific threshold, normalized correlation
[8]**UniS–S–G–SM** University of Surrey, semi-automatic registration, global threshold, special metric
[9]**UniS–S–S–SM** University of Surrey, semi-automatic registration, client-specific threshold, special metric