Database Size Effects on Performance on a Smart Card Face Verification System

Thirimachos Bourlai Josef Kittler Kieron Messer

Centre of Vision, Speech and Signal Processing, School of Electronics and Physical Sciences
University of Surrey, Guildford GU2 7XH, UK
E-mail: {t.bourlai, j.kittler, k.messer} @ surrey.ac.uk

Abstract

We study the effect of development set size on system performance, as measured by verification error. The study was performed using the FERET and FRGC2 databases to construct development training sets of varying size, while XM2VTS was used to test the system. Surprisingly, the achievable performance levels off relatively quickly. Increasing the size of the development set does not bring any benefit. On the contrary it may result in performance degradation. This finding appears to be development set independent. However, the choice of the development set size is protocol dependent.

1. Introduction

In this paper we revisit the *client specific linear discriminant analysis* (CSLDA) technique [5] used on the novel distributed architecture proposed in [1], where the decision making is carried out on the smart card itself. No user data ever leaves the card for verification, making the system more secure and user friendly. There are many severe engineering constraints and limitations imposed by small computing platforms. Thus special considerations have to be given to the system design issues in order to improve performance while increasing system speed as reported in [2]. However, an important aspect of the approach which was not explored was its scalability, which describes the performance as the database size increases.

A certain number of experiments were performed in the field (FERET and FRVT evaluations) where performance variation was checked either against multiple galleries of the same size [7] or against multiple gallery/probe sets of variable sizes generated by different databases. In the experiments reported over this problem, the identification ("who am I?") task gained more attention [11, 10, 9]. The watch list ("Are you looking for me?") task was also investigated as a function of gallery size [9].

In the verification task ("Am who I say I am?") that we are interested, performance was estimated against the size of the probe set in [11]. The effect of the size of the training set on verification performance was studied in [8]. However, in this case the size of the set was varied up to 8.192 images and false accept rate was fixed to 0.1%. Moreover, only the FRGC set was used for training and testing and there in no report on the way the training sets were selected, on the number of training sets per size or on the faces/clients ratio per set.

Interesting results on scalability can be found in [3] and most importantly in the FERET and FRVT evaluations. The FERET evaluation methodology [10] was developed to test face recognition algorithms under different scenarios (identification/verification) and categories of images (lighting change and the time between the acquisition date of the gallery and probe images). Over the course of the FERET evaluations from 1994-97, performance was increased as the size of the database increased and it was dependent on the gallery sets used. In FRVT 2002, new statistical methods were developed to estimate the variation in performance over the multiple galleries that alter the underlying classification problem. Different commercially available and mature prototype face recognition systems were evaluated. For the best system, for every doubling of the database size, identification performance decreased by two to three overall percentage points. A similar effect was observed for the watch list tasks.

To identify the trade off between the development set size and system performance (as measured by the verification error) in our system, a number of experiments were performed in order to answer the following main questions:

- How does the training set size affect the performance of the verification system?
- How does the system behave when going from a small (FERET) to a much larger (FRGC2) dataset?
- How many eigenfaces are needed as the size of the training set increases?
- How does performance vary when the number of Faces/number of Clients ratio changes as the size of the training set increases?

In order to model a system that might be installed at different locations where possibly different cameras, background

Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition (FGR'06)
0-7695-2503-2/06 $20.00 © 2006 IEEE

and illumination conditions were used, a number of development sets are randomly generated from the FERET and FRGC2 datasets. These sets are used to generate the initial statistical model, while the XM2VTS (C1/C2) dataset is used for testing our system.

In this paper the size of the development set is increased up to 44278 images and it is demonstrated that the use of a relatively low number of face images (in the range of 500-3000) to generate the development set can achieve maximum system performance. However, the optimum development set size in terms of performance depends on the testing configuration used to evaluate the system. As the database size increases, the verification performance decreases monotonically. From then on the performance appears to saturate. The experimental results suggest that the preservation of the necessary informational content to achieve high performance is ensured even when the number of different individuals used to create the initial statistical model is considerably reduced (see Table 3).

The rest of the paper is organised as follows. In the next section the basic face verification process will be reviewed. In Section 3, the database and the protocols used in our experiments will be presented. Section 4 covers the system evaluation process used, whilst Section 5 presents the experimental setup. In Section 6 the scalability effects on system performance will be analysed before finally, some conclusions are made.

2. Face Verification System

The face verification method adopted for the implementation on a smart card is the CSLDA technique [5]. It combines face representation and decision making into a single step, requiring a template of the size of the input image. The overall face verification system involves face detection, photometric normalisation and finally the verification test. All but the last processing step are carried out in the host system. The photometrically normalised image is then transmitted to the smart card where the user biometric template is stored, the verification score computed and the final decision made.

The experiments were performed with a relatively low resolution face images, namely 55x51. This resolution was initially used as a reference for our study. After detecting the face and localising the eye centres, a simple rotation and scaling of the eye positions onto two fixed points allowed geometric transformation. The photometric normalisation was achieved by a homomorphic filter and histogram equalisation.

For the feature extraction stage, a *PCA* model is built to achieve dimensionality reduction and then an *LDA* model is produced to get the overall client i specific linear discriminant transformation a_i, which defines the client specific fisher face for testing the claimed identity. The decision making stage produces a score, which defines how close the probe of the claimed identity is to the class of impostors. The thresholds in this stage have been determined based on the EER criterion.

3. Face Datasets

For the purpose of this study, XM2VTS, FERET and FRGC2 face databases were used in the experiments. **XM2VTS** [6] is a multi-modal database consisting of face images, video sequences and speech recordings taken from 295 subjects at one month intervals. In this database, the data acquisition was under controlled lighting and distributed over a long period of time that resulted in significant variability of the appearance of clients. The original size of the colour images is 720x576 pixels. This database contains 4 sessions. During each session two head rotation and "speaking" shots were taken. Eight images from 4 sessions are used. 200 subjects are used for training, that results in a total of 600/800 face images for *configuration* CI/CII. In *CI*, both the client validation and training sets contain very similar data, which is likely to lead to an optimistic threshold choice. In contrast, the test client images are made up of the last session that has quite different from the training images. In *CII* the amount of data available for creating user templates (4 images) is higher and only two images per system user are available for setting the threshold. Since the two validation images per user are coming from different sessions, the threshold is likely to be more accurate so that there should be a better balance between training and validation data in this case.

The XM2VTS protocol is an example of a closed-set verification protocol where the population of clients is fixed to a relatively large size (295 persons and 2360 images in total) and the system design can be tuned to the clients in the set. For the purpose of personal verification the Lausanne protocol[4] has been defined, which randomly splits all subjects into a client group (containing 200 subjects) and impostor group (divided into 25 evaluation and 70 test impostors). Eight images from 4 sessions are used. 200 subjects are used for training, that results in a total of 600/800 face images for configuration CI/CII.

The **FERET** image database [10] was assembled to support government monitored testing and evaluation of face recognition algorithms using standardised tests and procedures. The final database consists of 14051 eight-bit grey scale images of human heads with views ranging from frontal to left and right profiles. The images were collected over 15 session and were acquired in a semi-controlled environment. Images of an individual were acquired in sets of 5 to 11 images. Two frontal views (labelled **fa** and **fb**), where taken for different facial expressions. For 200 sets of im-

ages, a third frontal image was taken (labelled as **fc**) using a different camera and different lighting. The rest of the images were collected at various aspects between the right and left profiles. Simple variations to the database were added by taking a second set of images for which the subjects were asked to put on their glasses and/or pull their hair back. In some cases a second set of images was taken on a later date (*duplicate set*) and included variations in pose, scale, illumination and expression of the face. The 1201 clients resulted in a total of 3570 face frontal images (used for training).

The Face Recognition Grand Challenge (**FRGC**) is designed to achieve an increase in performance of the latest face recognition techniques [8]. The FRGC ver2.0 distribution contains high resolution still images taken under controlled lighting conditions and with unstructured illumination, 3D scans, and still images collected during the same period of time. It consists of three parts, the FRGC ver2.0 data set, the FRGC BEE, which includes all the data sets for performing and scoring the six ver2.0 experiments and a set of baseline algorithms for experiments 1 through 4. The data for FRGC ver2.0 consists of 50,000 recordings divided into training and validation partitions. The validation partition consists of data from 4,003 subject sessions. A subject session is the set of all images of a person taken each time a person's biometric data is collected and consists of four controlled still images, two uncontrolled still images, and one three-dimensional image. The controlled images, taken in a studio setting, are full frontal facial images taken under two lighting conditions and with two facial expressions (smiling and neutral). The uncontrolled images were taken in varying illumination conditions. Each set of uncontrolled images contains two expressions, smiling and neutral.

4. System Evaluation

Our face verification system has been evaluated via a set of experiments using both the XM2VTS[6], FERET[10] and Face Recognition Grand Challenge 2 (FRGC2)[8] data sets in a total of four different testing configurations:

- FERET-XM2VTS CI/CII: Configuration I/II for the XM2VTS database but with FERET data set used as the development set to generate the initial statistical model.
- FRGC2-XM2VTS CI/CII: Configuration I/II for the XM2VTS database but with FRGC2 data set used as the development set to generate the initial statistical model.

All cases represent an open set protocol where clients are not known to the system prior to enrolment. The system performance levels were measured in terms of half-total error rate (HTER). This was done on the test set of each protocol obtained using the EER threshold determined from the ROC curve (computed on an independent evaluation set). Both ROC curves on the evaluation as well as on the test set produce additional information about the system behaviour.

5. Experimental Setup

An aspect of our face verification system which has not been explored in our previous work is its performance as the face database increases in size. With the current system, the FERET and FRGC2 databases were used to generate the initial statistical model. Then the system was tested on the two protocols of the XM2VTS database. Examples of the (cropped) face images used in our experiments are provided in *Figure 1*.

Figure 1. Examples of the cropped face images used in our experiments. The first two are taken from the FERET database and the remaining two from the FRGC2.

The experimental setup of our system consists of the following stages. Initially, **n** face images, corresponding to the development set size, were randomly selected (based on a uniform distribution) from the complete gallery set of each database. By doing this we constructed the necessary xml files for the training set, which hold all the information about each face image of the dataset, such as its location in the system, face/subject ID, feature set (eyes/mouth/nose coordinates), pose (i.e.frontal), wearing glasses or not etc. In the case of FERET database (see Table 1) where the total number of face images is 3570 (of 1201 clients), **n** face images were randomly selected , where $n = [10, 20, 50, 100, ..., 3000]$, and therefore created an independent development set for each value of **n**. In the case of FRGC2 where the total number of face images (of 568 clients) of the complete set used was 44278, **n** was selected accordingly ($n = [10, 20, 50, 100, ..., 30000, 35000, 40000]$) as Table 2 shows.

The process was repeated 10 times (for each value of **n**) to generate 10 different xml files, which allowed for 10 independent experiments to be performed. This way 110/160 development sets were produced from the FERET/FRGC2 databases and a total of 270 experiments were performed.

For each experiment the normalisation process discussed in Section 2 was followed. Initially, face detection was carried out and then the eye centres were localised and the image geometrically normalised. For the relatively low resolution of 55×51, binomial filtering was then applied by

using an 1×11 mask, followed by histogram equalisation and finally homomorphic filtering. 95% of variance was used to define the number of principal components to be retained from PCA and the 236/292-dimensional subspace (for FERET/FRGC2 respectively) was used as the input for the client specific LDA algorithm. Verification results in terms of *half total error rate* (HTER), information about the number of clients used, PCA components needed and the ROC curve generated for each trial were recorded for all experiments.

6. Development Set Size Effects on System Performance

In this set of experiments we examine how verification performance varies with the use of different development set sizes that are selected from different datasets. This complicates the problem since different cameras and quality images are used to generate the two databases from which the development sets are created. Such an experiment models the performance of a system that might be installed at different locations where different cameras, background and illumination conditions may be used. Note also that the maximum number of faces/clients ratio is almost 3/1 in the case of FERET and more than 70/1 when FRGC2 is used.

After performing all of the experiments for all testing configurations, HTER, the number of clients (CL), and the number of PCA components used to create the statistical models based on the 10 trials per **n** face images, are averaged. Standard deviation of HTER per **n** is also measured from the errors across the 10 random samples. The overall results for each database are presented in *Figures 2* and 3 and Tables 1, 2 and 3. In all tables, the coloumn of *faces* represents the development set size that is varied.

In all testing configurations it was found that as the number of faces increased, performance increased as well up to a point where is started to saturate. In all cases this point was reached when only a relatively small number of faces was used to generate the development set. After that no real contribution was observed, just noise and perturbation. In practice, a development set of 500 images proved to be the operating point from which the performance started to saturate (see Tables 1, 2). In the case of FERET, the results showed that for C1/C2 the mean HTER (from 10 trials per **n**) becomes maximum when 2000/2500 face images were used as the development set. However, it is very interesting to note that in C1, one sample of 500 images proved to achieve the maximum performance(see Table 3). In *Figure 2*(i)/(ii) we see the performance variance for each experiment in C1/C2 protocols.

Similarly, when FRGC2 was used to randomly generate the development sets, the performance started to saturate already after 500 images and no more than 1000 images were required to achieve the minimum mean HTER. Again, in Table 3 we can see the sample cases where maximisation of performance is achieved. Even though we are dealing with a much larger and different dataset sizes, in C2 one sample of 500 images proved to achieve the best performance.

It is important to note that a higher HTER variation than expected is obtained in some experiments i.e. in the middle cases of FRGC2 where we have a gallery of i.e. 500,1000,...,4000 images. This is because the number of images per client for training is not fixed and we may have cases of one image/client that just add up an extra dimension to the problem that contributes negatively in the variance fluctuation of HTER. Better models and therefore less variance is achieved when the gallery set consists of more than 4000 images and enough faces/client are included.

Table 1. The overall results for FERET (FE) database. The "FACES" column represents the development set size. The HTER, number of clients (CL), number of PCA components and standard deviation (STD) are produced from the 10 trials per samples image in both Configurations I and II of the XM2VTS (XM) dataset, and using FERET to generate the initial statistical model. The last row represents the results on the full set.

FERET							
DB	FACES	CL	PCA	STDC1	STDC2	HTERC1	HTERC2
FE-XM	10	10.0	8.0	0.02126	0.02084	0.24746	0.27049
FE-XM	20	19.9	15.9	0.01961	0.02548	0.16108	0.16970
FE-XM	50	49.3	35.2	0.01457	0.01743	0.11948	0.10999
FE-XM	100	95.3	60.7	0.00908	0.00935	0.09963	0.07665
FE-XM	200	184.0	95.9	0.00742	0.00514	0.08050	0.05731
FE-XM	500	408.7	152.5	0.00687	0.00447	0.06735	0.04506
FE-XM	1000	689.5	192.2	0.00441	0.00250	0.06525	0.04067
FE-XM	1500	897.0	209.4	0.00301	0.00150	0.06495	0.04016
FE-XM	2000	1031.4	220.3	0.00311	0.00158	0.06445	0.03896
FE-XM	2500	1125.9	227.1	0.00216	0.00135	0.06511	0.03853
FE-XM	3000	1181.8	232.3	0.00194	0.00110	0.06657	0.03870
FE-XM	3570	1201.0	236.0	-	-	0.06816	0.04028

Table 2. Similarly, the overall results for FRGC2 (FR) database.

FRGC2							
DB	FACES	CL	PCA	STDC1	STDC2	HTERC1	HTERC2
FR-XM	10	9.9	8.0	0.01390	0.0096	0.23753	0.23895
FR-XM	20	19.9	15.7	0.01777	0.01539	0.16511	0.14797
FR-XM	50	47.5	34.2	0.00843	0.00789	0.10274	0.08206
FR-XM	100	87.1	59.3	0.00545	0.00222	0.08364	0.06186
FR-XM	200	157.0	96.1	0.00509	0.00221	0.06985	0.05186
FR-XM	500	286.8	157.0	0.00182	0.00323	0.05546	0.04212
FR-XM	1000	391.9	202.8	0.00335	0.00195	0.05315	0.04018
FR-XM	2000	466.3	239.6	0.00325	0.00213	0.05561	0.04060
FR-XM	4000	516.7	262.9	0.00203	0.00257	0.05875	0.04169
FR-XM	7500	546.2	277.8	0.00172	0.00159	0.05977	0.04148
FR-XM	15000	564.7	285.8	0.00093	0.00105	0.05973	0.04112
FR-XM	20000	567.4	288.2	0.00084	0.00101	0.05991	0.04168
FR-XM	30000	567.9	289.9	0.00047	0.00091	0.05934	0.04146
FR-XM	35000	567.8	290.9	0.00101	0.00063	0.05991	0.04089
FR-XM	40000	568.0	291.1	0.00092	0.00079	0.05987	0.04108
FR-XM	44278	568.0	292.0	-	-	0.05989	0.04048

7. Conclusions

In this paper the development set scalability was investigated in the context of a smart card face verification system.

Table 3. Summary of results obtained with FERET/FRGC2 development sets.

DB	Case	FACES	CL	PCA	HTERC1	HTERC2
FE-XM	Best C1	500	415.0	152.0	0.05587	-
FE-XM	Aver C1	2000	1031.4	220.3	0.06445	-
FE-XM	Best C2	2500	1142.0	226.0	-	0.03623
FE-XM	Aver C2	2500	1125.9	227.1	-	0.03853
FE-XM	Full Set	3570	1201.0	236.0	0.06816	0.04028
FR-XM	Best C1	2000	466.0	242.0	0.04864	-
FR-XM	Aver C1	1000	391.9	202.8	0.05315	-
FR-XM	Best C2	500	277.0	156.0	-	0.03647
FR-XM	Aver C2	1000	391.9	202.8	-	0.04018
FR-XM	Full Set	44278	568.0	292.0	0.05989	0.04048

Development sets of different sizes were created by sampling randomly the complete FERET and FRGC2 databases. The system was tested on the XM2VTS database using two different protocol configurations. It transpired that the use of a relatively low number of face images in the development set can achieve the best system performance. The number of images needed is variable and depends on the testing configuration used to evaluate the system. However, the performance results indicate that the development set can be limited to the range of 500-3000 images. The verification performance monotonically decreases until the above range is reached. Depending on the testing configuration, when performance achieves to a point, it becomes stable whether the training set increases or not (the slight changes of the error rates are not statistically significant). The experimental results suggest that the necessary informational content to achieve high performance was ensured even when the number of clients involved in creating that initial statistical model was about one half of the set size. 287/568 (Tables 1, 2) clients and 157/292 PCA components was good enough to achieve comparable performance results to those when the complete FERET/FRGC2 database respectively was used. Note also that the results show that as the database size grows, the number of eigenfaces needed to represent the ensemble of faces grows at much smaller rate[13, 12].

A typical behaviour is illustrated by the results of testing on XM2VTS (CII) shown in *Table 3*. In this case, there was a sample of 2500/500 images randomly selected from FERET/FRGC2 dataset that achieved a 10% improvement over the system performance obtained when the complete FERET/FRGC2 dataset was used. The average gain computed across all 10 trials was 4%/1%.

The effect of the development database size on a face verification system is interesting. However, the conclusions of this work were drawn in the context of the CSLDA algorithm [5]. For future work we plan to conduct the same evaluation on other typical face verification algorithms to check the general validity of the results.

References

[1] T. Bourlai, K. Messer, and J. Kittler, 'Face verification system architecture using smart cards', *ICPR 2004*, **1**, 793–796, (23-26 August 2004).

[2] T. Bourlai, K. Messer, and J. Kittler, 'Scenario based performance optimisation in face verification using smart cards', *AVBPA 2005*, (22-25 July 2005).

[3] J. Heo, B. Abidi, J. Paik, and M. A. Abidi, 'Face recognition: Evaluation report for faceit identification and surveillance', *Proc. Of SPIE 6th International Conference on QCAV03, Gatlinburg, TN, USA*, (May 2003).

[4] J. Leuttin and G. Maître, 'Evaluation protocol for the extended m2vts database (xm2vts)', *IDIAP*, (July 1998).

[5] Y.P. Li, J. Kittler, and J. Matas, 'Face verification using client specific fisher faces', *In J. T. Kent and R. G. Aykroyd, editors, Proc. Int. conf. on The Statistics of Directions, Shapes and Images*, 63–66, (September 2000).

[6] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, 'Xm2vtsdb: The extended m2vts database', *AVBRA*, 72–77, (March 1999).

[7] H. Moon and P.J. Phillips, 'Computational and performance aspects of pca-based face-recognition algorithms', *Perception*, **30**, 303–321, (2001).

[8] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, 'Overview of the face recognition grand challenge', *CVPR2005*, (June 2005).

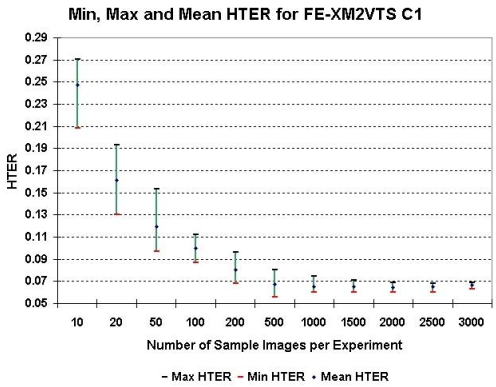
[9] P. J. Phillips, P. Grother, R.J. Michaels, D. M. Blackburn, and M. Bone, 'Face recognition vendor test 2002, http://www.frvt.org', *Evaluation Report*, (2003).

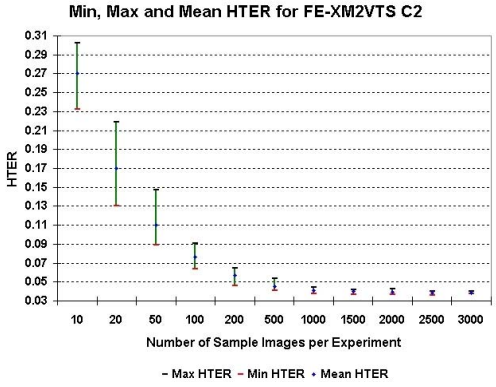
[10] P. J. Phillips, H. J. Moon, S. A. Rizvi, and P. J. Rauss, 'The feret evaluation methodology for face-recognition algorithms', *PAMI*, **22**(10), 1090–1104, (October 2000).

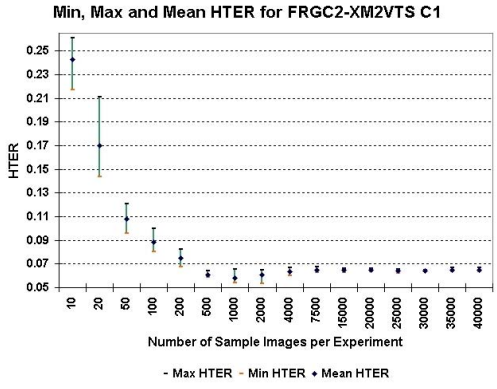
[11] S.A. Risvi, P. J. Phillips, and H. J. Moon, 'The feret verification testing protocol for face recognition algorithms', *IEEE AFGR'98*, 45–53, (April 1998).

[12] L. Sirovich and M. Kirby, 'Application of the karhunen-loeve procedure for the characterization of human faces', *PAMI*, **12**(1), 103–108, (Jan. 1990).

[13] L. Sirovich and M. Kirby, 'Low-dimensional procedure for the characterization of human faces', *J. Opt. Soc. Am*, **4**(3), 519–524, (March 1987).







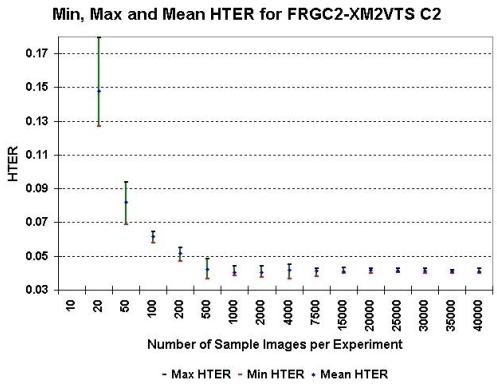
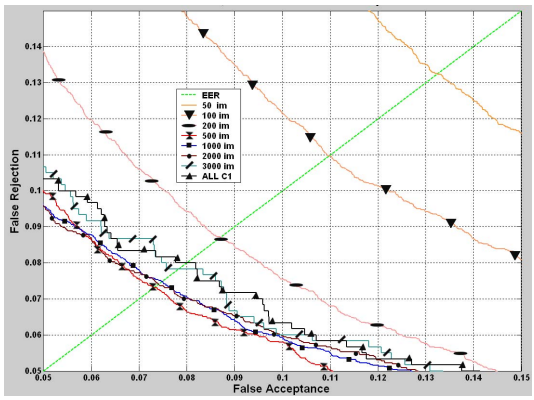
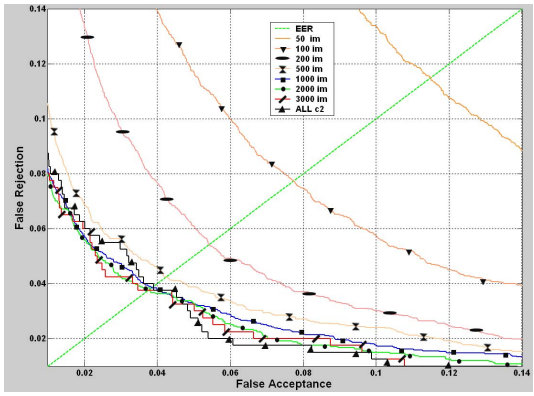
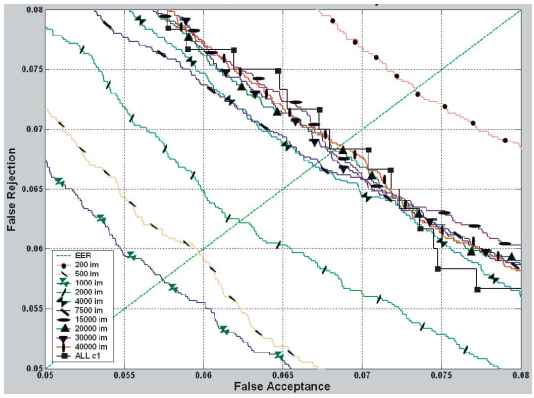


Figure 2. Performance/Variance for each protocol of the XM2VTS database, when FERET (i),(ii) and FRGC2(iii),(iv) were used to generate the initial statistical model.







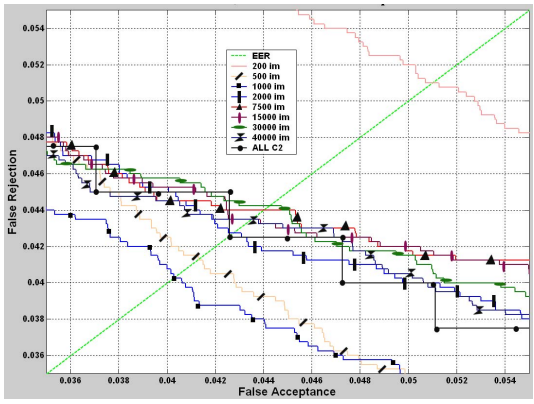


Figure 3. ROC curves (zoom in) for each protocol of the XM2VTS database. Figures (i)/(ii) for FERET-XM2VTS C1/C2 and (iii)/(iv) for FRGC2-XM2VTS C1/C2.