Performance Evaluation of Multibiometric Face Recognition Systems

Margarida Castro Neves

Security Technology Dept, Fraunhofer Institute for Computer Graphics Research IGD
Fraunhoferstr. 5, 64283 Darmstadt, Germany
margarida.castro-neves@igd.fraunhofer.de

Samuel Chindaro, Ming Ng, Ziheng Zhou, Farzin Deravi

Department of Electronics, University of Kent
Canterbury, CT2 7NT, United Kingdom
{s.chindaro|m.w.r.ng|z.zhou|f.deravi}@kent.ac.uk

Abstract: The *3D Face* project investigates the use of 3D face recognition technologies, and aims to improve their performance so that it would be possible to use such technologies in unsupervised access control scenarios in airports. During this project novel sensors, 2D, 3D, skin texture matchers and fusion algorithms have been developed. A technology performance test has been performed on all algorithms, in order to evaluate the technology improvements. This paper describes the independent test and evaluation activities for this project and gives an overview of the results obtained.

1 Introduction

The use of electronically stored biometric information in identity and especially travel documents has been rapidly increasing worldwide in recent years. According to the International Civil Aviation Organization (ICAO) recommendations, border control will be primarily based on 2D face technology [Ic04]. This technology has nevertheless some well known limitations. Acceptable performance can only be accomplished if certain conditions apply (e.g. frontal pose image, sufficient contrast, neutral face expression, etc.). In addition to this, effective liveness and fake detection, where the system detects that a stored image or photograph is presented to the camera to gain unauthorized access, may be difficult to achieve, potentially compromising the security when using this technology in un-supervised environments. The use of 3D face images could address some of these problems, thus increasing the level of system security and convenience for users [Bu08].

The *3D Face* project [3d06], is a European Integrated Project funded under the European Commission IST FP6 program. This three year R&D project is dedicated to advanced 3D facial recognition and aims to increase the performance of 3D-enabled face recognition technologies to a level allowing their fully operational implementation in airports, and to integrate privacy protection technologies that allow for a more trusted and secure usage of biometrics. During the project, several European partners, including industrial and research organizations, have developed *3D face* recognition technologies, including novel sensors, 2D, 3D and skin texture matchers, and fusion and template privacy protection algorithms in order to accomplish this goal. The challenge is to achieve a false acceptance rate (FAR) below 0.25% with a false rejection rate (FRR) below 2.5% in an operational environment. These are referred to as target values in this paper.

The algorithms developed for the project have to go through several phases of tests: During the development phase, internal tests are performed by the technology developing partners themselves, to check for problems and assure the quality of their products. In the technology testing phase, the produced software modules are then delivered to be tested by an independent test group, performing ISO/IEC 19795-1 [Is06] and ISO/IEC 19795-2 [Is07] compliant technology tests. These tests provide results on the performance of all submitted algorithms that can be compared. The best performing components will then be implemented into a prototype and a field trial will be performed. The goal of the field trial is the validation of the developed *3D face* recognition technology under operational conditions

This paper describes some of the performed technology tests. Conducting these tests presented new challenges given the number of components and the combinations in which they could be configured. The remainder of the paper is organized as follows. Section 2 describes the face recognition technologies tested. Section 3 gives an overview of the independent technology testing performed, its goals and activities. An overview of the results obtained will be shown in Section 4. Finally, in Section 5 some conclusions are presented.

2 Multibiometrics and Multimodality

In the *3D Face* project, 3D technologies are being investigated in a multibiometric framework together with 2D face recognition and the fusion of different 3D and 2D algorithms as well as novel template protection schemes are also part of this research programme. If the use of 3D shape can address some of the problems of 2D face recognition technologies, a combination of both aspects, texture as well as shape promises even better efficiency. The 2D texture information can be used, for example, for finding the eyes, and other relevant features, thus helping to find the angle of rotation of the 3D shape image. It is also possible to combine both methods in an algorithm, extracting both 2D and 3D features for comparison. This combination can also be done internally in the algorithm, where features are extracted using both methods. Score level fusion and also decision level fusion are also successful ways to combine multiple biometric technologies [Ve08].

3 Technology Testing

3.1 Test Database

A 3D face database of 600 individuals has been created from a population of volunteers including staff members and students from partner organizations. The individuals have been scanned in two sessions. It consists of 3D and 2D face images in different poses and different facial expressions, as shown in Figure 1.

Figure 1 Poses and expressions: Images from left to right; frontal, neutral expression without glasses (only for individuals usually wearing glasses), frontal-neutral expression, frontal,-smiling, frontal- talking, frontal-wearing a cap; head turned right, head turned left, head turned up, head turned down.

Three hundred individuals have been scanned using the ViSense scanner prototype, developed within the *3D Face* project. It produces 3D shape images with a resolution of 640x480 pixels, and 2D colour images with a resolution of 1280x960 pixels, using structured light method for scanning [Ne07][Po]. Another 300 individuals have been scanned using a 3DMD camera [3d]. As the field test prototype will use the *3D Face* scanner, the tests discussed in this paper have only used the images acquired by the ViSense prototype. Tests using the 3DMD images are still in progress.

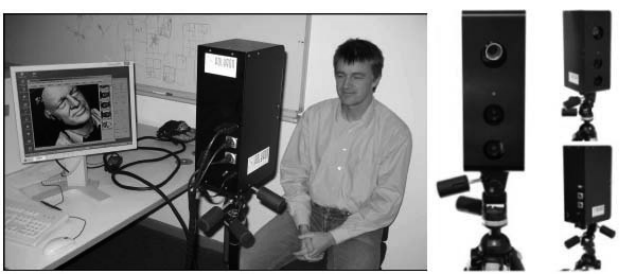

Figure 2 Acquisition set-up and scanner

3.2 Test Plan

A protocol has been developed that defines the kinds of test to be performed (conforming to [Is06] and [Is07]), the scenarios to be tested, and the data to be used at appropriate test phases. According to the objectives set for the independent technology testing, the performance of developed algorithms using the following technologies have been be tested and compared:

- 3D, 2D, inter-algorithm multimodal 3D2D and high-resolution skin texture algorithms
- Multibiometric score-level and decision-level fusion of different combinations of algorithms

Three different scenarios have been selected, as shown in Table 1. They differ on the level of difficulty of the query set (verification images) [Gr03]. The gallery/target set (enrollment images) is the same for all scenarios: one frontal pose and neutral expression image. The first scenario, S1, is a simple scenario, containing only neutral expression and frontal images, to test algorithm performance in ideal conditions. The scenario S2 also includes images with facial expressions and movements (where the test persons are smiling or talking), which is a more realistic scenario. Scenario S3 tests the robustness of the algorithms by allowing different head poses and head covering by a cap. Query set and target set contained images from two different sessions. In order to define overall performance criteria for choosing the best components and combinations, the performance of the algorithms were weighted as shown in Table 1.

Scenario ID	**Data Description (gallery v/s probe set)**	**Weight**
S1	Neutral/frontal images v/s Neutral/frontal images	50%
S2	Neutral/frontal images v/s frontal images including expressions (smiling, talking); no cap	35%
S3	Neutral/frontal images v/s all Images (frontal/non-frontal, with cap, neutral, talking and smiling)	15%

Table 1: Test scenarios and weighting scheme

The performance criteria was based on the following weighted FRR at FAR = 0.25% (referred to as FRR0025), indicating the relative importance associated with each test scenario according to operating conditions.

Weighted FRR0025 = 0.5 x FRR0025(S1) + 0.35 x FRR0025(S2) + 0.15 x FRR0025(S3)

The acquired database was intended not only for the technology test, but also to be used for internal testing and training by the developing partners. Therefore the database has been divided in two parts, one for each purpose. It was challenging to determine the best way of partitioning the database. If many images are made available for training and testing, then the algorithms get better, but there are not enough images to test the improvement with a good confidence interval. On the other hand, having more images for testing will make the results more exact, but the algorithms cannot be trained sufficiently, resulting in poorer performance.

The solution chosen for this trade-off was to have two phases of testing. In Phase 1, two thirds of the database has been used for testing, and one third for algorithm training and testing. After that, another third of the database has been released, so that in Phase 2 the new versions of algorithms have been trained/calibrated with 2/3, and tested with 1/3 of the database. With feedback from the test results, and more data for training, the algorithms could be improved and submitted for testing Phase 2.

3.3 Test System Implementation, Test Execution

Each partner provided feature extraction and authentication modules that used the *3D Face* API, which has been specified to make all modules from different vendors work together. The modules were integrated within a test system framework. Using this framework, images from the gallery/target set (enrolment images) and query set (verification images) have been read, and using the extraction modules, the templates for each image have been created. Then a cross comparison between all images from each set has been performed using the authentication modules, and the resulting scores have been written into similarity matrices. In fusion tests, the fusion module under test has been then applied to the similarity matrices generated by the individual module tests, resulting in a fused score matrix for each possible combination of modules.

4 Results

The results obtained provide an insight on how some of the state-of-the-art of 3D face recognition algorithms stand in comparison to 2D algorithms and to what extent multibiometrics and fusion techniques can enhance the overall performance. In all results henceforth, different modules are denoted by a letter prefix, followed by the type of modality. For example A3D2D and B3D2D denotes two different modules (A and B) using the internally fused modality 3D2D. High resolution texture modules are denoted by a suffix HR. Where this is not stated, the modules use low resolution texture images acquired by the *3D Face* Scanner.

4.1 Results for 3D, 2D, 3D2D and Skin Texture Algorithms

Phase 1

In Phase1, data from 196 volunteers has been used. For all scenarios, 196 enrolment images (gallery/target set) were compared with 1546 verification images (query set). A total of thirteen modules have been submitted for testing in this phase: Five of them based on internally fused 3D2D models, two based on low resolution texture (2D), two based on high resolution skin texture (2DHR), and four 3D modules. In S1, eleven out of thirteen modules reached or exceeded the target FRR0025. In S2, ten out of thirteen modules reached or exceeded the target. In S3, none of the modules reached or exceeded the target. For S1 and S2, all 2D and 3D2D modules reached or exceeded the target. One of the 3D algorithms exceeded the target for both S1 and S2 (see Figure 3 and Table 2).

Module	FRR0025 S1	FRR0025 S2	FRR0025 S3	Weighted FRR0025
A3D2D	0.0082	0.0087	0.0317	0.0119
B2D	0.0082	0.0087	0.0440	0.0137
C3D	0.0163	0.0175	0.0634	0.0238
A2DHR	0.0163	0.0192	0.4101	0.0764

Table 2 Performance of four types of modalities in individual algorithms Phase1, ranked by weighted FRR0025

It can be observed that both the 3D and 2D algorithms are capable of reaching the targets for the realistic scenario, with the internally fused 3D2D algorithm giving the best performance.

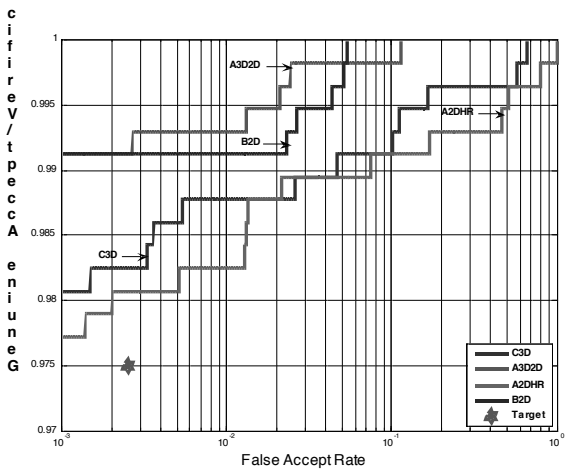

Figure 3 Performance of four types of modalities in individual algorithms Phase 1, S2.

Phase 2

Data from 99 volunteers has been used for Phase 2 test. For all scenarios, 99 enrolment images were compared with 772 verification images. Eighteen modules were tested in phase 2 and their performances were assessed using the same criteria as in Phase 1. Six 3D algorithms, two high resolution skin texture algorithms, three low resolution texture algorithms and seven internally fused 3D2D algorithms were tested.

In S1, fifteen out of eighteen modules reached or exceeded the target FRR. In S2, fourteen out of eighteen modules reached or exceeded the target. In S3, none of the eighteen modules reached or exceeded the target. Similarly to Figure 3 above, the highlights for the individual algorithms performance for S2, Phase 2 are shown in Figure 4 and Table 3 below.

Module	FRR0025 S1	FRR0025 S2	FRR0025 S3	Weighted FRR0025
A2DHR	0.0082	0.0070	0.0661	0.0165
A3D2D	0.0164	0.0140	0.0311	0.0178
B2D	0.0164	0.0140	0.0440	0.0197
C3D	0.0246	0.0210	0.0570	0.0282

Table 3 Performance of four types of modalities in individual algorithms Phase 2, ranked by weighted FRR0025

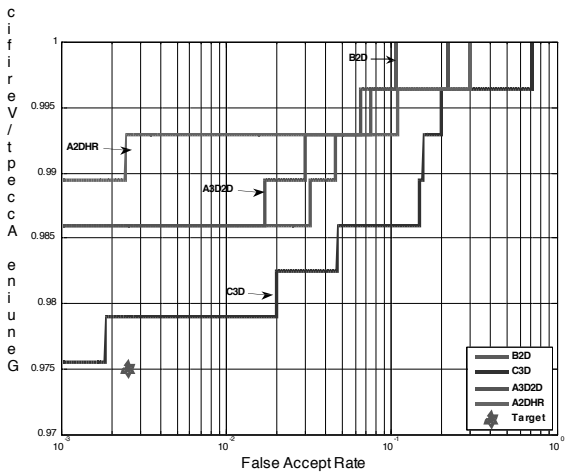

Figure 4 Performance of four types of modalities for individual algorithms Phase 2, S2.

All 2D and 3D2D algorithms reached or exceeded the target for S1 and S2. Two 3D algorithms reached the target for the same scenarios. As in Phase 1, none of the individual algorithms reached the target for S3. It can be observed by comparing Figure 3 and Figure 4, that there was a significant improvement in the performance of the 2D high resolution algorithm from Phase 1 to Phase 2, which can be partly attributed to the availability of more data for training in this phase, with the other algorithms maintaining their levels of performance, or achieving marginal improvements. In both phases, it can be observed that the state-of-the-art 3D algorithms can give performances which are as good as their 2D counterparts. This performance is seen to improve with algorithms which incorporate both 3D and 2D information in their internal processing.

4.2 Multibiometric Fusion

In Biometric applications, there are three levels of fusion which can be mainly employed. Feature Level Fusion, Score Level Fusion and Decision Level Fusion [Ro03]. With the participation of different matchers, it is easier and more feasible to access and combine scores generated by different matchers. Fusion experiments comprising of all possible permutations of 2D, 3D and 3D2D scores supplied by developers were carried out in this work. At score level, fusion can be done using two approaches; as a classification problem or as a combination problem.

In the combination approach, the individual matching scores are combined to generate a single scalar score which is then used to make the final decision. The Sum (SUM) and Weighted Sum (WSUM) combination methods were among the methods used for fusion in experiments described in this paper. For the Weighted Sum, the weights were calculated based on the equal error rates obtained from the fusion training set. In addition the LogFAR (LF) combination method which employs a simplified Neyman-Pearson fusion on FAR mapped scores was also employed, as was the recently developed Optimal OR (OPTOR) fusion method [Ta07].

A wide range of classification methods have been used for classification-based fusion at score level [Ro03]. This is due to the fact that scores can be treated as feature vectors for classification. Classification based fusion schemes combine the outputs of different matchers into single vector of scores which is then fed into a trained classifier. If scores are treated as such, they can then be exposed to a wide range of classification methods that are available for pattern recognition. There is no distinctive method of choice in this area; proposed methods are compared to differing sets of classifiers with differing results. As such, this approach is always open to further exploration. In this work the Logistic Regression Classifier (LOGC) [Ve99] and the Linear Discriminant Classifier (LDC) [Ma79][Mc92] were among the classifiers explored.

Score normalization refers to changing the location and scale parameters of the match score distributions at the outputs of different matchers, so that the match scores of different matchers are transformed into a common domain [Ro06]. Since different modalities and developers produce scores which are not heterogeneous, the scores have to be transformed into a common domain before combining them. There is a number of normalization methods investigated in literature [Ja05][Ro06]. After exploratory experiments, it was observed that no particular normalisation method resulted in any significant advantage over the others; the *z-score* normalisation method was then adopted because of its simplicity, robustness and efficiency. Z-score normalisation was applied to scores in implementing the Sum, Weighted Sum and Optimal OR fusion methods. No normalisation was applied for classifier based fusion methods.

Phase1 Fusion Testing

In Phase 1 testing, five fusion algorithms were tested. These were the SUM, WSUM, LOGC, LDC and LF. In this phase, all the top fusion combinations for each fusion method reached or exceeded the target for S1 and S2. None of the fusion algorithms reached the target for S3. These results are illustrated in Table 4.

Fusion Combination/Method	FRR0025 S1	FRR0025 S2	FRR0025 S3	Weighted FRR0025
A3D_A2D_B2D_C3D_D2DHR_LF	0.0082	0.0070	0.0259	0.0104
B2D_C3D_SUM	0.0082	0.0070	0.0310	0.0112
B2D_C3D_LDC	0.0082	0.0087	0.0317	0.0119
B2D_C3D_WSUM	0.0082	0.0087	0.0343	0.0123
C2D_A2D_B2D_C3D_LOGC	0.0122	0.0105	0.0608	0.0189

Table 4 Phase 1 Test: Best Combination and Fusion Result for Each Fusion Method

The best fusion method was the LF method, mainly based on its performance in S3; thus making it more robust than the other methods. However it required the most number of algorithms in its combination. The improvement over individual algorithms brought about by fusion is more significant in S3, the most difficult of the three scenarios (best individual FRR0025 was 0.0317 and for the best fusion method, 0.0259), thus highlighting the robustness introduced by fusion. Overall results achieved by the fusion methods bettered, or at least matched that achieved by the best individual algorithms (Table 3 and 4). With more training data in phase 2, the fusion results of the training-based fusion algorithms were expected to improve.

Phase 2 Fusion Testing

In Phase 2 six fusion algorithms were tested. In addition to the five from phase 1, the OPTOR method was also tested. In this phase, all the top fusion results for each method reached or exceeded the target for S1 and S2. The results for S3 showed a number of fusion algorithms and combinations reaching or exceeding the target. The table below shows the top fusion combinations for each fusion algorithm.

Fusion Combination/Method	FRR0025 S1	FRR0025 S2	FRR0025 S3	Weighted FRR0025
A2D_A2D3D_SUM	0.0082	0.0070	0.0259	0.0104
A2D_A3D_OPTOR	0.0082	0.0105	0.0313	0.0125
A2D_B2D_C3D_WSUM	0.0164	0.0105	0.0246	0.0156
A2D_A2D3D_LF	0.0082	0.0070	0.0661	0.0165
A2D_A3D_LDC	0.0082	0.0070	0.0687	0.0168
A2D3D_C3D_LOGC	0.0164	0.0140	0.0298	0.0176

Table 5 Phase 2 Test: Best Combination and Fusion Result for Each Fusion Method

Because of the low weight associated with S3, the table does not reflect the number of fusion algorithms which exceeded the target for S3. Six of the combinations reached or exceeded the target for S3 using the OPTOR method, highlighting the robustness of this fusion method in the presence of outliers as would exist in S3. One combination using WSUM exceeded the target for S3.These results are shown in Figure 5.

Figure 6 illustrates the improvement achieved by fusing 3D2D and 3D algorithms using the OPTOR method. It can be observed that even though the individual 2D and 3D2D algorithms failed to reach the target, by fusing them, the performance increases and the target is exceeded.

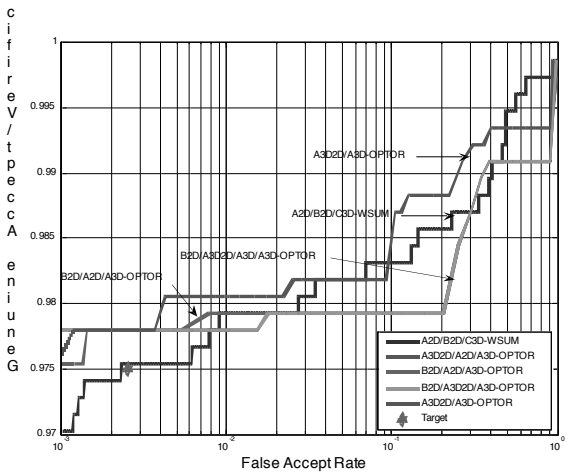

Figure 5 ROC for Top Fusions for Phase 2, S3

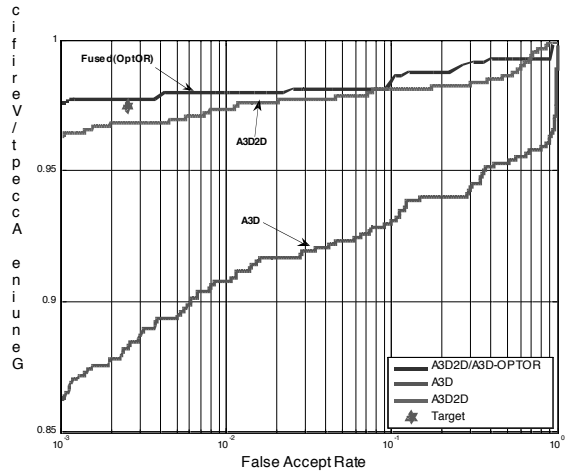

Figure 6 Modules and Fused (Optimal OR) Results: S3

4.3 Throughput and Template Size

The throughput was measured during the test on an Intel Core 2 Duo (2.66GHz), Windows XP professional edition, 3.23 GB RAM. The times for initialising and terminating modules are not included. Table 6 shows the maximum and minimum times achieved by the tested modules, for all tested modalities. They include enrolment (reading and feature extraction) and verification (reading, feature extraction and matching) times.

Modality and Phase	2D	3D	3D2D
Phase 1 maximum time	6.46 s	15.24 s	15.2 s
Phase 1 minimum time	1.76 s	6.4 s	6.46 s
Phase 2 maximum time	7.38 s	12.4 s	12.42 s
Phase 2 minimum time	3.41 s	3.74 s	3.79 s

Table 6 Enrolment and verification times

It appeared that the developers had to compromise on throughput time in order to achieve higher performances in some cases in Phase 2. For example A2DHR changed from being the best one in terms of throughput in Phase 1 to being the worst in Phase 2, but at the same time achieved the best accuracy for this phase.

The template sizes (in bytes) for the algorithms ranged between 192 and 210000. The best performing algorithms had a template sizes of 87075, 180 000, 16500 and 210000. The smallest template size was 192. It appeared again that the issue of accuracy and template sizes was a balancing act as in throughput, with the best performing algorithms providing relatively larger templates.

5 Conclusions and Outlook

The test results show that using both kinds of characteristics of an image - shape as well as texture - provides a better performance than any one of these sources of information alone. In addition, the fusion of some algorithm combinations (from different developers) is seen to result in significantly enhanced performance. By fusing different algorithms with different approaches the weaknesses of one algorithm appear to be compensated by the other. Fusion algorithms also resulted in the best performance in terms of robustness, and were the only algorithms to exceed the project target in the most challenging scenario considered during the tests. These technology tests remain to be verified by operational field tests which will show the performance in a real environment.

References

[3d06] 3D Face Consortium. 3D Face. Integrated Project funded by European Commission. http://www.3dface.org, 2006.

[Bu08] Busch et al. 3D Face Recognition for Unattended Border Control, Proceedings Net-ID 2008 Conference, (2008)

[Is06] ISO/IEC TC JTC1 SC37 Biometrics. ISO/IEC 19795-1:2006. Information technology – Biometric performance testing and reporting – Part 1: Principles and framework. International Organization for Standardization and International Electrotechnical Committee, 2006.

[Is07] ISO/IEC TC JTC1 SC37 Biometrics. ISO/IEC 19795-2:2007. Information

technology – Biometric performance testing and reporting – Part 2: Testing Methodologies for Technology and Scenario Evaluation. International Organization for Standardization and International Electrotechnical Committee, 2007.

[Ic04] International Civil Aviation Organization Technical Advisory Group 15 Machine Readable Travel Documents/New Technologies Working Group, Biometrics Deployment of Machine Readable Travel Documents, Version 2.0, May 2004.

[Ph04] P.J. Phillips, Face Recognition Grand Challenge, National Institute of Standards and Technology, USA, March 2004.

[Gr03] P. Grother, R. Micheals and P.J. Phillips. Face Recognition Vendor Test 2002 Performance Metrics. . Proceedings 4th International Conference on Audio Visual Based Person Authentication, 2003.

[Ne07] P. Neugebauer, Research on 3D Sensors - Minimizing the Environmental Impacting Factors,CAST Workshop Biometrics and eCards, (2007)

[Po] Webpage of the ViSense scanner; *http://www.polygon-technology.de/news.html*

[3d] Webpage of the 3DMD scanner*; http://www.3dmd.com*

[Ve08] R. Veldhius, F. Deravi, Q. Tao, Multibiometrics for Face Recognition, IT-Sicherheit & Datenschutz, 3, pp, 204-214 2008

[Ta07] Q. Tao and R. Veldhuis. Optimal decision fusion for a face verification system. In the 2nd International Conference on Biometrics, pp 958–967, Seoul, Korea, 2007.

[Ja05] A.K. Jain, K. Nandakumar and A. Ross, Score Normalisation in Multimodal Biometric Systems, Pattern Recognition, 38(12):2270-2285, 2005.

[Ro03] A. Ross, A. K. Jain, and J.-Z. Qian, Information Fusion in Biometrics, Pattern Recognition Letters, 24(13):2115-2125, 2003

[Ve99] P. Verlinde and G. Chollet. Comparing decision fusion paradigms using k-NN based classifiers, decision trees and logistic regression in a multi-modal identity verification application. In Second International Conference on Audio- and Videobased Biometric Person Authentication (AVBPA), Washington D. C., USA, March 1999.

[Ro06] Ross, A., K. Nandakumar, and A. Jain, Handbook of Multibiometrics. International Series on Biometrics. 2006: Springer

[Ma79] Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979), Multivariate Analysis, San Diego: Academic Press.

[Mc92] McLachlan, G. J. (1992), Discriminant Analysis and Statistical Pattern Recognition, New York: Wiley.