3D Face Recognition Benchmarks on the Bosphorus Database with Focus on Facial Expressions

Neşe Alyüz[1], Berk Gökberk[2], Hamdi Dibeklioğlu[1], Arman Savran[3], Albert Ali Salah[4], Lale Akarun[1], Bülent Sankur[3]

[1] Boğaziçi University, Computer Engineering Department
nese.alyuz,hamdi.dibeklioglu,akarun@boun.edu.tr
[2] Philips Research, Eindhoven, The Netherlands
berk.gokberk@philips.com
[3] Boğaziçi University, Department of Electrical and Electronics Engineering
arman.savran,bulent.sankur@boun.edu.tr
[4] Centrum voor Wiskunde en Informatica, Amsterdam, The Netherlands
a.a.salah@cwi.nl

Abstract. This paper presents an evaluation of several 3D face recognizers on the Bosphorus database which was gathered for studies on expression and pose invariant face analysis. We provide identification results of three 3D face recognition algorithms, namely generic face template based ICP approach, one-to-all ICP approach, and depth image-based Principal Component Analysis (PCA) method. All of these techniques treat faces globally and are usually accepted as baseline approaches. In addition, 2D texture classifiers are also incorporated in a fusion setting. Experimental results reveal that even though global shape classifiers achieve almost perfect identification in neutral-to-neutral comparisons, they are sub-optimal under extreme expression variations. We show that it is possible to boost the identification accuracy by focusing on the rigid facial regions and by fusing complementary information coming from shape and texture modalities.

1 Introduction

3D human face analysis has gained importance as a research topic due to recent technological advances in 3D acquisition systems. With the availability of affordable 3D sensors, it is now possible to use three-dimensional face information in many areas such as biometrics, human-computer interaction and medical analysis. Especially, for automatic face recognition, expression understanding, and face/facial feature localization problems, three-dimensional facial data offers better alternatives over using 2D texture information alone [1]. The information loss when projecting the inherently 3D facial structure to a 2D image plane is the major factor that complicates the task of analyzing human faces. Problems arise especially when adverse situations such as head pose variations, changes in illumination conditions, or extreme facial expressions are present in the acquired

data. The initial motivation for the exploitation of 3D information was to overcome these problems in human facial analysis. However, most of the proposed solutions are still limited to controlled acquisition conditions and constrained to frontal and mostly neutral 3D faces. Although there are increasing number of studies that focus on pose and/or expression invariant face recognition, the databases upon which they are based have not been systematically constructed for the analysis of these variations or they remain limited in scope. For example, the most frequently used 3D face database, the Face Recognition Grand Challenge (FRGC) database [2], contains mostly frontal faces with slight arbitrary pose variations. In the FRGC database, there are several acquisitions for different expressions which are labeled according to the emotions such as sadness and happiness. Comparison of publicly available 3D face databases in terms of pose, expression and occlusion variations can be found in [3].

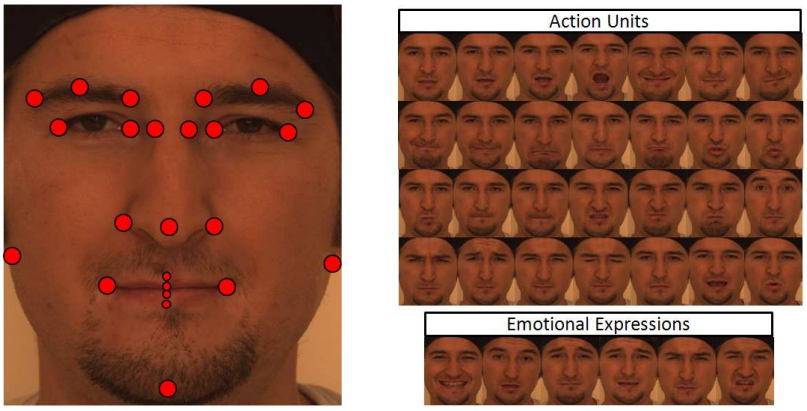
The desiderata of a 3D face database enabling a range of facial analysis tasks ranging from expression analysis to 3D recognition are the following: i) Action units (FACS) [4], both single and compound; ii) Emotional expressions; iii) Ground-truthed poses; iv) Occlusions originating from hair tassel and a gesticulating hand. Motivated by these exigencies, we set out to construct a multi-attribute database. In this paper, we present the characteristics of the database collected as well as preliminary results on face registration and recognition.

2 The Bosphorus 3D Face Database

The Bosphorus database is a multi-expression, multi-pose 3D face database enriched with realistic occlusions such as hair tassel, gesticulating hand and eyeglasses [5, 3]. The variety of expressions, poses and occlusions enables one to set up arbitrarily challenging test situations along the recognition axis or along the expression analysis axis. We want to point out the opportunities that the Bosphorus database provides for expression understanding. The Bosphorus database contains two different types of facial expressions: 1) expressions that are based on facial *action units* (AU) of the Facial Action Coding System (FACS) and 2) *emotional expressions* that are typically encountered in real life. In the first type, a subset of action units are selected. These action units are grouped into three sets: i) 20 lower face AUs, ii) five upper face AUs and iii) three AU combinations. In the second type, we consider the following six universal emotions: happiness, surprise, fear, sadness, anger and disgust. Figure 1(b) shows all different types of expressions. To the best of our knowledge, this is the first database where ground-truthed action units are available. In order to achieve more natural looking expressions, we have employed professional actors and actresses.

Facial data are acquired using Inspeck Mega Capturor II 3D, which is a commercial structured-light based 3D digitizer device [6]. The 3D sensor has about $x = 0.3mm$, $y = 0.3mm$ and $z = 0.4mm$ sensitivity in all dimensions and a typical pre-processed scan consists of approximately 35K points. The texture images are high resolution (1600×1200) with perfect illumination conditions. The locations of several fiducial points are determined manually on both 2D

and 3D images. On each face scan, 24 points are marked on the texture images provided that they are visible in the given scan. The landmark points are shown in Figure 1(a).



(a) (b)

Fig. 1. a) Manually located landmark points and b) expressions for the Bosphorus database.

The Bosphorus database contains 3396 facial scans of 81 subjects. There are 51 men and 30 women in the database. Majority of the subjects are Caucasian and aged between 25 and 35. The Bosphorus database has two parts: the first part, Bosphorus v.1, contains 34 subjects and each of these subjects has 31 scans: 10 types of expressions, 13 different poses, four occlusions, and four neutral/frontal scans. The second part, Bosphorus v.2, has more expression variations. In Bosphorus v.2, there are 47 subjects having 53 scans[1]. Each subject has 34 scans for different expressions, 13 scans for pose variations, four occlusions and one or two frontal/neutral face. 30 of these 47 subjects are professional actors/actresses.

3 Face Recognition Methodology

In this work, we apply commonly used techniques in face recognition to provide benchmarks for further studies. We have selected five face recognition approaches: three of them use shape information, and two use facial texture information. Two of the shape-based approaches are based on the Iterative Closest Point (ICP) algorithm, namely *one-to-all ICP* and average face model-based ICP (*AFM-based ICP*) [7]. The third one employs PCA coefficients obtained from 2D depth images. These techniques are explained in detail in Section 3.2. Texture-based approaches use either raw pixel information or PCA coefficients

[1] Some subjects have fewer than 53 scans due to acquisition errors

(eigenface technique). Before proceeding to identification methods, it is worthwhile to mention landmarking of faces because all these methods heavily rely on the quality of the initial alignment of facial surfaces.

3.1 Landmarking

Almost all 3D face recognition algorithms first need an accurate alignment between compared faces. There are various methods to align faces and most of them require several landmark locations that are easily and reliably detectable. ICP-based approaches which are explained later in this section, usually require these points at the initialization step. In our work, in addition to using 22 manually located landmark coordinates, we employ an automatic landmark localization method which estimates these points using the shape channel. The automatic landmarking algorithm consists of two phases [8]. In the first phase, a statistical generative model is used to describe patches around each landmark. During automatic localization, patches extracted from the facial surface are analyzed with these statistical models, and the region that produces the best likelihood value for each corresponding model is selected as the location of a landmark. A coarse-to-fine strategy is used to keep the search fast. We use inner and outer eye corners, nose tip and mouth corners, as these landmarks correspond to discriminative local structures. Figure 2(a) and 2(b) in Section 4 shows automatically found landmarks for a sample face image.

3.2 Shape-based Matchers

One-to-All ICP Algorithm: The 3D face recognition problem can be considered as a special case of a 3D object recognition problem. The similarity between two objects is inferred by features calculated from 3D models. Notice that most approaches require precise alignment (registration) of objects before similarity calculation and the performance depends heavily upon the success of registration [1].

The Iterative Closest Point (ICP) algorithm [9] has been one of the most popular registration techniques for 3D face recognition systems due to its simplicity The ICP algorithm basically finds the best rigid transformation (i.e., translation, scale, and rotation matrices) to align surface A to surface B. Traditionally, a probe face is registered to *every* gallery face and an estimate of the volumetric difference between aligned facial surfaces is used as a dissimilarity measure. Therefore, we call this method *one-to-all ICP*. If we assume 3D point cloud representations of faces, dissimilarity can be estimated by the sum of the distances between corresponding point pairs in given facial surface pair. Indeed, ICP uses this measure during its iterations and after convergence, it outputs this dissimilarity measure as the alignment error.

AFM-based ICP Algorithm: The one-to-all ICP approach requires as many alignments as the size of the gallery set, this easily becomes infeasible when the

gallery set size is large. An alternative approach would be to use a generic face model. All gallery faces are registered to this generic face model offline, before the identification phase [10], [11]. Thereby, only alignments between the probe faces and the generic face are needed to compute dissimilarities for the whole gallery set. This approach significantly shortens the identification delay by reducing the time complexity of the alignment phase. In the rest of the paper, we refer to this method as *AFM-based registration*.

Depth Image-based PCA Algorithm Most 3D sensors provide shape data in the form of 3D point clouds for the visible part of the object being scanned. For frontal facial 3D scans, the visible region usually contains the ear-to-ear frontal part of a human face. Therefore, there is at most one depth measurement, i.e., z coordinate, for any (x,y) coordinate pair. Due to this property, it is possible to project 2.5D data to an image plane where the pixels denote depth values. Images constructed in this way are called *depth images* or *range images*. 3D data should undergo post-processing stages during the conversion to depth images. Surface fitting is one of the important post-processing steps. A practical option for surface fitting is to obtain 3D triangulation of point cloud data and then to estimate the surface points inside the triangular patches by bilinear interpolation. Except for steep regions, such as the sides of the nose, information loss is minimal in depth image construction. Once 3D information is converted to 2D images, numerous approaches employed for 2D texture-based face recognition systems can be used for 3D face identification. Among them, using PCA coefficients as features is usually accepted as a baseline system for 3D depth image-based recognition. In our work, we perform whitening after computing PCA coefficients and use cosine distance for similarity calculation. As a pattern classifier, 1-nearest neighbor algorithm produces the estimated class label.

3.3 2D Texture Matchers

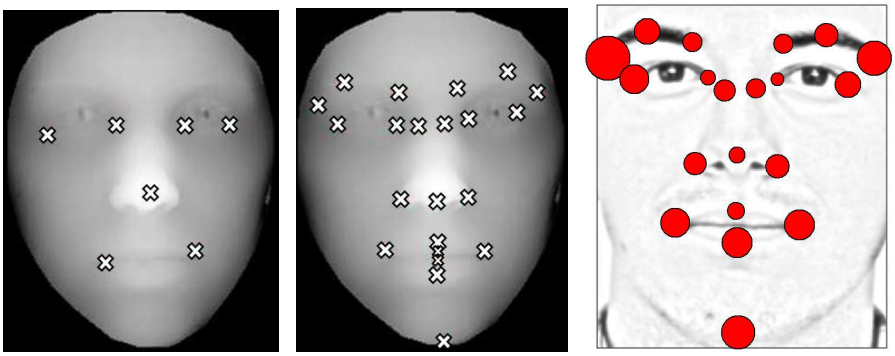
The Bosphorus database contains high quality texture information for each 3D facial model. In order to compare the performances of shape and texture channels we also implemented two 2D recognizers. The first, pixel-based method, simply uses gray-scale pixel information to represent a face. Texture images are normalized by scaling with respect to eye-to-eye distances. Illumination variations are handled by histogram equalization. In the pixel-based method, we use two regions: i) the whole face and ii) the upper facial region to test expression sensitivity. The second texture-based approach is the Eigenface technique where each face is transformed to a subspace by a PCA projection. As in the depth image method, we perform whitening and use 1-nearest neighbor classifier.

4 Experimental Results

We have performed recognition experiments on a subset of the Bosphorus database. The selected subset contains only neutral and expression-bearing images without

any pose variations or occlusions. Only one neutral image per person is used for enrollment, and the rest are used as the test set. First three rows of Table 1 show three experimental configurations. For the Bosphorus v.1, we have two experiments: one with the neutral probe set and the other with the non-neutral probe set. For v.2, there is only one experiment containing all non-neutral images of every subject in the probe set.

We have analyzed the effect of the number of landmarks and the effect of automatically detected landmarks in our tests. We use several subsets of landmarks that are presented in Figure 1(a). The performance of the automatic landmark detection module is summarized in Figure 2(c). We see that the most successful landmarks are the inner eye corners. In approximately 80% of the cases, they are found within tolerance, where the tolerance threshold is defined as 10% of the eye-to-eye distance. In general, inner eye corners and nose tip can be detected successfully, but outer eyebrows, and chin tip point usually can not be localized efficiently. The performance of the depth-image based automatic landmark detection is low. However, we include it here to test the performance of face recognizers with automatic landmarks.



(a) (b) (c)

Fig. 2. Automatically located landmarks: the locations of a) seven fiducial landmarks found by the first phase, b) all 22 landmarks after the second phase, and c) the performance of automatic landmarking. Circle size denotes average pixel distance error for each landmark location.

We have performed recognition experiments on the v.1 and v.2 expression subsets, as summarized in Table 1. The first experiment was the one-to-all ICP experiment (One-to-All ICP_{M22} method in Table 1): Although this takes a long time, we provide these results as a benchmark. In the ICP coarse alignment stage, we used the 22 manually detected landmarks. As observed in Table 1, one-to-all ICP yields 99.02% correct identification on the v.1 neutrals. However, the performance drops to 74.04% for v.1 non-neutral and to 72.41% for v.2. This performance drop is to be expected, since the gallery includes only one neutral face. Next, we compare the AFM approach with the one-to-all ICP. The

AFM approach is very fast since it performs only one match. The results of this approach with 22 manually detected landmarks is denoted as AFM_{M22} in Table 1. On the v.1 database, AFM based identification classifies every facial image in the neutral probe accurately. However, in the non-neutral v.1 probe set, the correct classification rate drops to 71.39%. For v.2 tests, only 67.67% of the probe set is identified correctly. On comparison with one-to-all results, we see that AFM performs better on neutral faces, but suffers a small drop in performance in faces with expressions. Since this drop is not very large, we use the AFM approach for the rest of the tests.

Table 1. Correct classification rates (%) of various methods on the Bosphorus database. Coarse alignment configurations used in these methods are denoted as subscripts: M and A is for manual and automatic landmarking, respectively. The numbers used in the subscripts denote the number of landmarks used; i.e., AFM_{M5} is the AFM method aligned with five manual landmark points.

Method	v.1 Neutral	v.1 Non-neutral	v.2 Non-neutral
Gallery Set Size	34	34	47
Probe Set Size	102	339	1508
AFM_{M5}	99.02	69.62	65.12
AFM_{M7}	100.00	73.75	68.83
AFM_{M8}	99.02	72.27	69.36
AFM_{M22}	100.00	71.39	67.67
AFM_{A7}	80.39	62.24	-
AFM_{A22}	81.37	62.24	-
One-to-All ICP_{M22}	99.02	74.04	72.41
$DI-PCA_{M22}$ (Whole face)	100.00	71.09	70.56
$DI-PCA_{M22}$ (Eye,Nose)	100.00	85.55	88.79
TEX-Pixel (Whole face)	97.06	93.51	92.52
TEX-Pixel (Upper face)	97.06	90.56	92.59
TEX-Eigenface (Whole Face)	97.06	87.61	89.25
Fusion of AFM_{M7} and TEX-Pixel (Whole face)	-	-	95.09
Fusion of $DI-PCA_{M22}$ (Eye,Nose) and and TEX-Pixel (Whole face)	-	-	98.01

The effect of facial landmarks on the identification rate is next analyzed. For this purpose, we look further into two quantities: 1) The subset of facial landmarks that should be used in coarse alignment and 2) The performance change caused by the use of automatic landmark localizer. For the first case, we formed three landmark subsets of size five, seven and eight. Landmark subset of size five only uses landmark points around the nose. The landmark set with seven landmarks contains eye corner points, nose tip, and mouth corners. The eight-point subset is the same as the seven-point set but with the added chin tip point. We see that using only seven landmarks leads to better performance than using all 22 landmarks. Accuracy in v.1 non-neutral set is 73.75% (see

Table 1, marked as AFM_{M7}) and in v.2, it is 68.83%. If faces are registered according to the nose region only (using five landmarks, AFM_{M5} in Table 1), we see degradation in accuracy. Adding chin tip to the previously selected seven landmarks does not change the identification rate significantly.

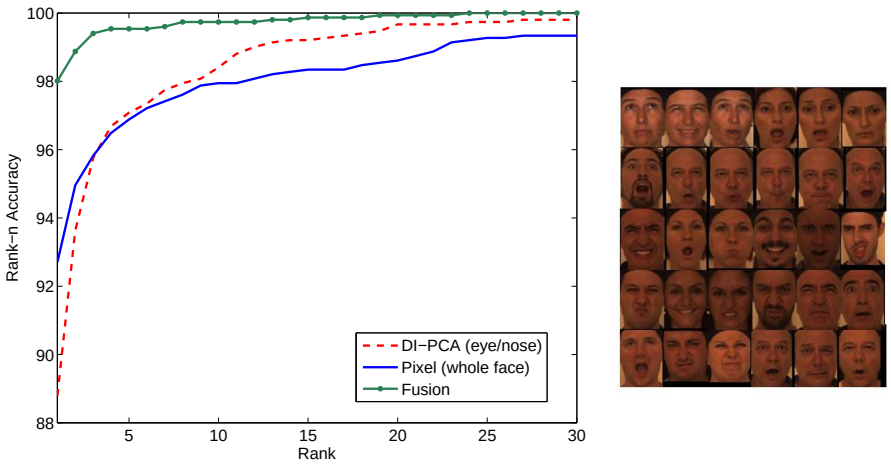
If we turn back to our second question about the effect of automatic landmarking on the identification rates, we see significant performance drop with automatic landmarking. Entries marked as AFM_{A7} and AFM_{A22} in Table 1 show that, irrespective of which landmark subset is used, there is approximately 20% and 10% accuracy decrease in neutral and non-neutral probe sets, respectively. This is mostly due to the localization errors in landmark detection.

Regarding all ICP-based experiments, we see that AFM_{M7} presents a good compromise in that: i) It is computationally much faster that one- to-all performance and performs only a little worse; and ii) It relies on only 7 landmarks, which are easier to find.

The next set of experiments are with the depth image PCA method ($DI-PCA_{M22}$ methods in Table 1). We have tried two versions: Using the whole face, and using only the eyes and the nose regions. Both perform perfectly with the neutral faces in v.1. In non-neutral v.1, and v.2, the performance of the whole face is 71.09% and 70.56%, respectively. When only the eye and nose regions are included, performance rises to 85.55% in v.1 non-neutrals and to 88.79% for v.2. Overall, we see that local PCA-based representation of eye and nose region is the best shape modality-based recognizer.

We have also used 2D textures to classify the faces. We have obtained very good identification performance with texture images. Note that the texture images are of very high quality, with perfect illumination and high resolution. The performance obtained with texture pixels is reported for i) the whole face and ii) the upper part (denoted as TEX-Pixel in Table 1). The eigenface technique is also applied (TEX-Eigenface). Identification performances of all three algorithms on the neutral v.1 are identical: 97.06%. On the non-neutral v.1, the three algorithms obtain 93.51%, 90.56%, and 89.25%, respectively. Recognition performance on the v.2 are unexpectedly higher: 92.52%, 92.59% and 89.25%, respectively. We note that the texture performances are higher than the shape performances. This is due to the perfect illumination conditions and the high resolution of the 2D images

And lastly, we fuse the results of the 3D and 2D classifiers. Using product rule to combine the dissimilarity scores of AFM-based ICP method and pixel-based textural classifier (See Table 1, Fusion of AFM_{M7} and TEX-Pixel), we achieve 95.09% correct identification rate in the v.2 experiment. If DI-PCA of the eye/nose region is used as a shape classifier in fusion, 98.01% accuracy is obtained (See Figure 3(b) for all 30 images misclassified in the v.2 set). Cumulative matching characteristic (CMC) curves of local DI-PCA and texture classifiers, together with their fusion performance, are shown in Figure 3(a). Notice that although rank-1 performance of the texture classifier is higher, shape classifier becomes superior after rank 3.



(a) (b)

Fig. 3. a) CMC curve for i) local PCA based depth image algorithm, ii) pixel-based texture algorithm and iii) their fusion, and b) Misclassified faces in the v.2 set by the fusion of DI-PCA and TEX-Pixel method.

5 Conclusion

In this work, benchmarking studies on a new challenging 3D face database are presented. We have used 3D recognition methods with proven performance: Two of these algorithms use ICP alignment for dissimilarity calculation. One is based on generic face template (AFM) for fast registration, and the other exhaustively searches the closest face model from the gallery set for a given probe image. In addition to ICP-based methods, depth images are also used where feature construction is handled via the PCA technique.

3D cameras almost always yield 2D texture images in addition to 3D data. At close range and under good illumination, the texture images turn out to be of high quality. In fact, texture images singly or in complementary role to 3D data can boost the performance. In our study, fusion of the shape and texture based methods has yielded recognition performances as high as 98.01%. The main conclusions of our work are as follows:

- The performance obtained with the one-to-all registration is comparable to that of AFM registration, both with neutral and expression faces. On the other hand, AFM method is orders of magnitude faster. Therefore AFM is preferable.
- The 3D recognition performance suffers heavily from inexactitude of landmarks. The present landmarking algorithm causes a heavy performance drop of 10-20% percentage points. Therefore real-time and reliable face landmarking remains still an open problem.

– Depth images with PCA form a viable competitor to the 3D point cloud feature set, and in fact outperform it. It remains to see if alternative feature sets, e.g., subspace methods or surface normals can bring improvements.
– The fusion of 2D texture and 3D shape information is presently the scheme with the highest performance.

The Bosphorus database is suitable for studies on 3D human face analysis under challenging situations such as in the presence of occlusion, facial expression, pose variations. The future work will consist of i) improving landmark localization performance, ii) testing the sensitivity of 3D face recognition algorithms under pose changes, and iii) employing different representation methods other than point clouds and depth images.

6 Acknowledgements

This work is supported by TÜBİTAK 104E080 and 103E038 grants.

References

1. Bowyer, K., Chang, K., Flynn, P.: A survey of approaches and challenges in 3D and multi-modal 3D + 2D face recognition. Computer Vision and Image Understanding **101** (2006) 1–15
2. Phillips, P., Flynn, P., Scruggs, T., Bowyer, K., Chang, J., Hoffman, K., Marques, J., Min, J., Worek, W.: Overview of the face recognition grand challenge. In: Proc. of. Computer Vision and Pattern Recognition. Volume 1. (2005) 947–954
3. Savran, A., Alyüz, N., Dibeklioğlu, H., Çeliktutan, O., Gökberk, B., Akarun, L., Sankur, B.: Bosphorus database for 3D face analysis. In: First European Workshop on Biometrics and Identity Management Workshop(BioID 2008)
4. Ekman, P., Friesen, W.: Facial action coding system: A technique for the measurement of facial movement. Consulting Psychologists Press (1978)
5. Savran, A., Çeliktutan, O., Akyol, A., Trojanova, J., Dibeklioğlu, H., Esenlik, S., Bozkurt, N., Demirkır, C., Akagündüz, E., Çalıskan, K., Alyüz, N., Sankur, B., Ulusoy, İ., Akarun, L., Sezgin, T.M.: 3D face recognition performance under adversarial conditions. In: Proc. eNTERFACE07 Workshop on Multimodal Interfaces. (2007)
6. Inspeck Mega Capturor II Digitizer: http://www.inspeck.com/
7. Gökberk, B., Dutağacı, H., Ulaş, A., Akarun, L., Sankur, B.: Representation plurality and fusion for 3D face recognition. IEEE Transactions on Systems Man and Cybernetics-Part B: Cybernetics **38**(1) (2008) 155–173
8. Salah, A.A., Akarun, L.: 3D facial feature localization for registration. In: Proc. Int. Workshop on Multimedia Content Representation, Classification and Security LNCS. Volume 4105/2006. (2006) 338–345
9. Besl, P., McKay, N.: A method for registration of 3-D shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence **14**(2) (1992) 239–256
10. İrfanoğlu, M., Gökberk, B., Akarun, L.: 3D shape-based face recognition using automatically registered facial surfaces. In: Proc. ICPR. Volume 4. (2004) 183–186
11. Gökberk, B., İrfanoğlu, M., Akarun, L.: 3D shape-based face representation and feature extraction for face recognition. Image and Vision Computing **24**(8) (2006) 857–869