Realistic and Animatable Face Models for Expression Simulations in 3D

Nesli Erdogmus, Rémy Ethève, Jean-Luc Dugelay
Eurecom, Multimedia Communications Department
2229 Routes des Crêtes, 06904 Sophia Antipolis, France
{nesli.erdogmus, remy.etheve, jean-luc.dugelay}@eurecom.fr

ABSTRACT

In the face recognition problem, one of the most critical sources of variation is facial expression. This paper presents a system to overcome this issue by utilizing facial expression simulations on realistic and animatable face models that are in compliance with MPEG-4 specifications.

In our system, firstly, 3D frontal face scans of the users in neutral expression and with closed mouth are taken for one-time enrollment. Those rigid face models are then converted into animatable models by warping a generic animatable model using Thin Plate Spline method. The warping is based on the facial feature points and both 2D color and 3D shape data are exploited for the automation of their extraction. The obtained models of the users can be animated by using a facial animation engine. This new attribution helps us to bring our whole database in the same "expression state" detected in a test image for better recognition results, since the disadvantage of expression variations is eliminated.

Keywords: Facial expression animation, warping, 3D animation, face recognition

1. INTRODUCTION

Face recognition has been drawing an unabated interest in the research communities for decades; however despite the numerous proposed approaches, the solution for the intra-class variation problem introduced by facial expressions still remains unsolved. This paper proposes a method to eliminate the adverse effect of expression variations on face recognition and identification. Numerous solutions for the problem of face recognition in the presence of facial deformations, due to expressions, have been proposed which can be generally classified into three groups according to the data types used: 2D, 3D and 2D+3D approaches.

Among the approaches that utilize 2D facial images, subspace analysis method stands out. In [1], Principal Component Analysis (PCA) approach is extended by creating subsets of images through masking those regions where significant modifications are expected to occur and using them to build different face projection spaces. On the other hand, in [2], a local feature based method is proposed in which a set of feature points with highest deviations from the expectation is automatically extracted by using statistical Local Feature Analysis (LFA). Afterwards, each point is described by a sequence of local histograms captured from the Gabor responses at various frequencies and orientations around the feature point. Other than local approaches which divide the face into different parts, systems that analyze the whole face as described in [3], [4], and [5] exist. In [3], an extended eigenfaces method is proposed in which PCA is applied on motion vectors domain, caused by the motion of the facial features due to facial expressions and two spaces are used for the reconstruction of the test images. [4] combines Gabor wavelet transform and Nearest Neighbor Discriminant Analysis and [5] suggests using a subset of fractal codes of the whole face as feature for face recognition.

3D face recognition, relatively a younger research trend, has emerged especially with the acquisition devices becoming more accurate and less expensive. Similar to 2D case, local region analysis is commonly used. For instance, in [6], average region models, where local correspondences are inferred by the Iterative Closest Point algorithm are used. In a similar manner, in [7] a region-based face surface matching is applied where the similarity score is computed with higher importance given to more stable regions and in [8], again a multi-region approach is proposed which incorporates summation invariant features from those regions and optimal fusion by similarity scores between regions. On the other hand, different representations of the facial surface data is calculated for recognition, where points are referenced to the nose tip, which is highly robust to the expressional changes. In [17], the comparison between the faces is based on optimal deformations of the level curves, which are defined by a surface distance function, taking the nose tip as the reference point.

Lately, numerous techniques are proposed to combine both modalities in order to overcome the limitations they have separately. By assuming an isometric mapping between surfaces, [9] proposes a method in which 3D surface is used to define a polar geodesic coordinate space and the corresponding color image is embedded in this space to serve as the recognition data. On the other hand, in [10] the face recognition system is based on feature points which are described by Gabor filter responses in the 2D domain and Point Signature in the 3D domain.

In this paper, we applied the 2D+3D multimodality by inserting expressions that are similar to the detected ones on the test image (in 2D), to all enrolled models in the database (in 3D) by utilizing Thin Plate Splines (TPS) method, instead of trying to remove or avoid the deformation problem due to the facial expressions. Hence, by rendering images of enrolled faces with the detected expressions in the test images, we aim to improve the recognition rates. A detailed diagram of the proposed system is given in Fig. 1.

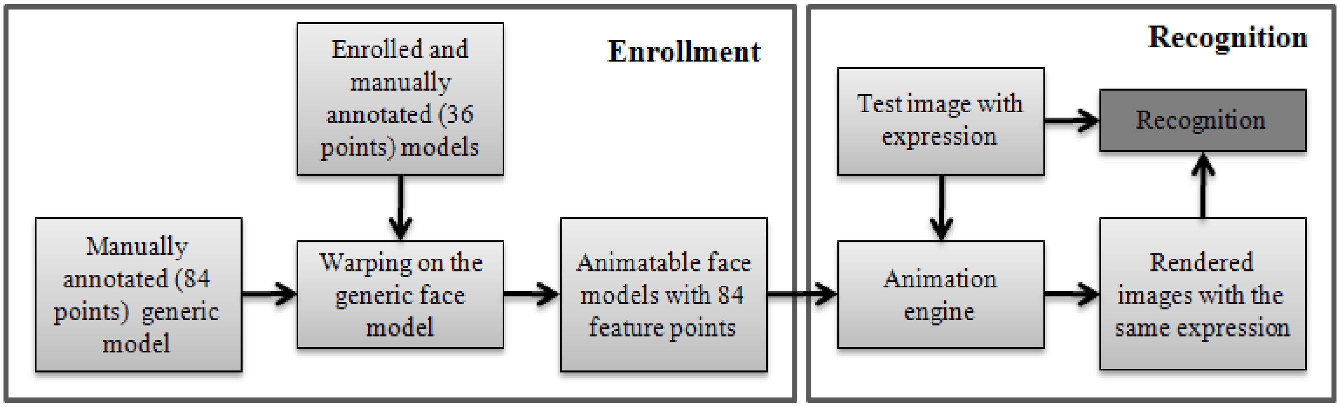
Enrollment
Enrolled and manually annotated (36 points) models
Manually annotated (84 points) generic model
Warping on the generic face model
Animatable face models with 84 feature points
Recognition
Test image with expression
Recognition
Animation engine
Rendered images with the same expression

Figure 1. Flow chart of the proposed system

The rest of this paper is organized as follows: In Section 2, a brief summary on MPEG-4 specifications and the facial animation object profiles is given. The detailed process for automatic extraction of facial feature points is explained in Section 3. In Section 4, for obtaining the realistic and animatable face models, the warping method based on TPS is explained. In Section 5, the resulting rendered images of expression-applied models will be provided along with the recognition results which are compared with the conventional approach. Finally in Section 6, some conclusion and future work are presented.

1.1. MPEG-4 Specifications and Facial Animation Object Profiles

MPEG-4 is an ISO/IEC standard developed by Moving Picture Experts Group which is a result of efforts of hundreds of researchers and engineers from all over the world. Mainly defining a system for decoding audiovisual objects, MPEG-4 also includes a definition for the coded representation of animatable synthetic heads. In other words, independent of the model, it enables coding of graphics models and compressed transmission of their animation parameters.

The facial animation object profiles defined under MPEG-4 often are classified under three groups [11] [12][18]:

- Simple facial animation object profile: The decoder receives only the animation information and the encoder has no knowledge of the model to be animated.
- Calibration facial animation object profile: The decoder also receives information on the shape of the face and calibrates proprietary model accordingly before animation.
- Predictable facial animation object profile: The full model description is transmitted. The encoder is capable of completely predicting the animation produced by the decoder.

The profile that is more conformable to our approach is the second one: calibration facial animation object profile, since we are aiming to calibrate the "generic" model according to the samples in our database. Actually, what we are doing is to build the part that differs between the first and the second profiles. Our system generates an animatable model by using 29 of 84 MPEG-4 specified face definition parameters (FDP) which are annotated manually. The rest of the points are only marked on the generic model. After the warping, those points are recalculated for animation.

In Fig. 2, the positions of the 84 feature points on the face are given. Most of these points are necessary to be supplied to an MPEG-4 compliant animation system, except for the points on the tongue, the teeth or the ears, depending on the animation tool structure.

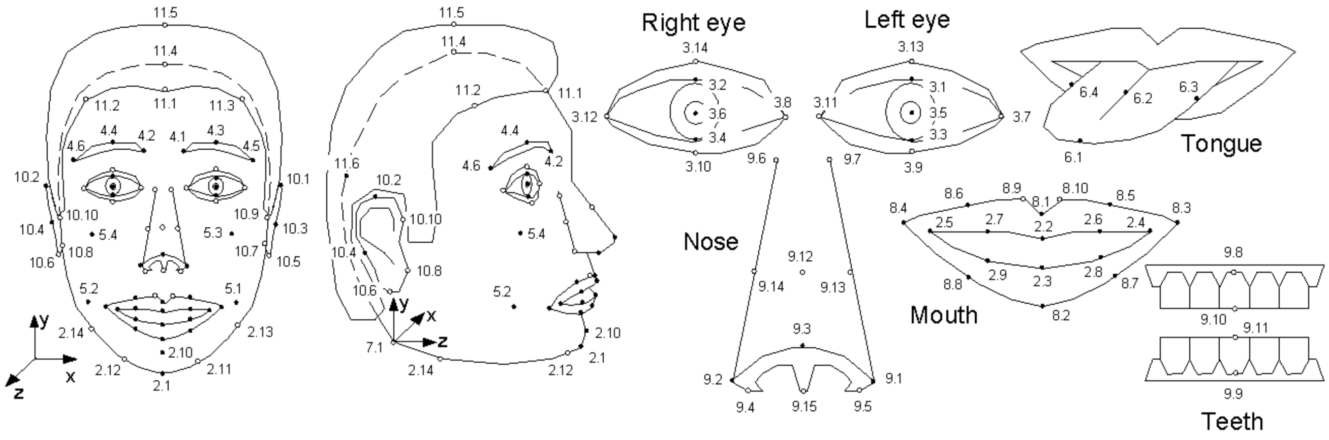

Figure 2. Feature points

2. PROPOSED METHOD

The proposed method is developed for the cases in which the enrollment is done with a closed mouth and a neutral face for each subject in 3D with texture. However, for the recognition, it is highly possible that the captured images will be under different facial expressions. In this method, the expressions on the captured images are assumed to be known and therefore, the generic model can be warped and animated accordingly for each model. The approach is described in more detail in the following subsections.

2.1. Feature Point Marking

As mentioned in the previous section, 84 feature points are necessary for MPEG-4 compliant animation engines. In our approach, we manually annotate 29 of those points which are utilized in warping process for the enrolled models. Those points are chosen to be in descriptive areas such as eyes (5x2), nose(6), mouth(7) and face borders(6). The rest of the feature points are only marked on the generic model and their positions are recalculated after warping.

2.2. Constructing the Animatable Face Models

In order to obtain an animatable 3D model for an enrolled user, the approach adopted in this paper is to warp an already animatable generic head model so that it transforms into the models in the database as correctly as possible. For this purpose, we choose to apply Thin Plate Spline (TPS) method for warping 3D faces. TPS method was made popular by Fred L. Bookstein in 1989 in the context of biomedical image analysis [13]. It is based on an analogy of the bending energy of thin metal plates under point constraints.

For the 3D surfaces S and T, and a set of corresponding points on each surface, P_i and M_i respectively, the TPS algorithm computes an interpolation function f(x,y) to compute T', which approximates T by warping S:

$$T' = \{(x',y',z') \text{ st. } \forall(x,y,z) \in S,\ x'=x,\ y'=y,\ z'=z+f(x,y)\} \qquad (1)$$

$$f(x,y) = a_1+a_xx+a_yy+\Sigma w_iU(|P_i-(x,y)|) \qquad (2)$$

with U(.), the kernel function, expressed as:

$$U(r) = r2ln(r),\ r = \surd x2+y2 \qquad (3)$$

In the interpolation function f(x,y), the w_i, $i\in\{1,2,...n\}$ are the weights. As given in (2), the interpolation function consists of two distinct parts. The affine part $(a_1+a_xx+a_yy)$ which accounts for the affine transformation necessary for the surface to match the constraint points and a warping part $(\Sigma w_iU(|P_i-(x,y)|))$.

The generic model, shown in Fig. 3, is a full head model with an open mouth and eye holes. Thereby, the mouth and eyes can be animated on the contrary of the static model. Due to the polynomial nature of the Thin-Plate Spline algorithm, regions far away from the constrained points "diverge". For this reason, additional constraint points are added on the back of the head model which will be used as the target points to preserve the human look of the model.

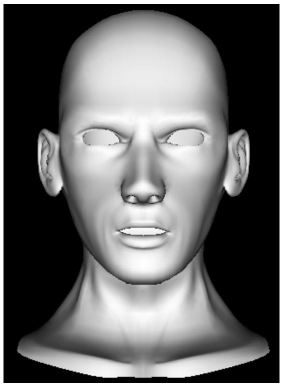

Figure 3. The generic head model

However, before applying the TPS method, we have to make sure that the target face and the generic face models are aligned correctly. Firstly, the generic model is roughly rescaled to match the target face dimensions. The rescaling coefficients are calculated simply by comparing x, y and z apertures. Afterwards, the feature point pairs are used to align the two models. These two sets of landmarks defines the transform, that gives the best fit mapping one onto the other, in a least squares sense.

After rescaling and alignment, a coarse warping is applied to the generic model by taking the feature point pairs as the source and the target landmarks. These point pairs describe a non-linear warp transform by which any point on the mesh close to a source landmark will be moved to a place close to the corresponding target landmark. The points in between are interpolated smoothly using Bookstein's Thin Plate Spline algorithm [13].

In the next step, by assuming for all points on the generic model, the corresponding pair on the target model is the one that is closest, fine warp transform is calculated. For each source landmark, the target landmark is found as the closest point on the other mesh. Hence, all points of the generic model become closest to the target model and maintain their smoothness. Afterwards, the center position and the radius for both eye spheres are calculated by using the four points annotated around the eyes.

Finally, the texture is copied onto the generic model and the new positions of the feature points are calculated. The obtained model is exported as a “vrml” file and the feature points are saved in an “fdp” file which will be explained in more detail in the next section.

2.3. Animating the warped models

In order to simulate the facial expressions on the obtained animatable model, the animation engine, called visage|life™ is utilized. visage|life™ allows user to prepare the face models that are produced in any 3D modeler (e.g. 3ds max™, Maya™, Softimage™ etc.) for animation using the unique Facial Motion Cloning (FMC) technology that automatically produces 86 standard morph targets for the new face model, which immediately becomes capable of full facial animation. [14] Additionally, it has an FDP editor to view and edit defined feature points with a user-friendly GUI. Face Definition Parameters are essential for animating the model. However, this property is not utilized since the FDP files needed for animation are generated automatically after warping.

3. EXPERIMENTS AND RESULTS

In order to apply and evaluate our approach, we worked with the Bosphorus 2D-3D face database [15] which was collected as a part of the Enterface'07 Workshop held in Bogazici University in Istanbul [16]. It includes a rich set of expressions, pose variations and different types of occlusions. Among the 105 enrolled subjects, and 10 types of expressions, we chose 73 models with multiple neutral scans.

In the preprocessing step, in order to remove the small protuberance, the 3D facial data is smoothed by applying bilateral filtering. Thus, the edges are preserved whereas the surface is denoised. In Fig 4, a sample model is shown before and after the smoothing.

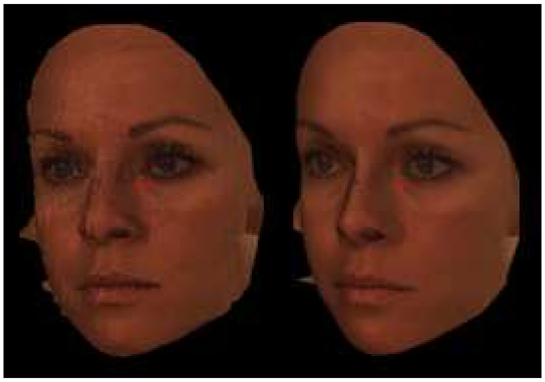

Figure 4. Facial surface before and after smoothing

After the preprocessing, the generic head model is processed as explained in detail in Section 4. The rescaling, alignment, coarse warping, fine warping and texturing steps are illustrated with an example in Fig. 5.

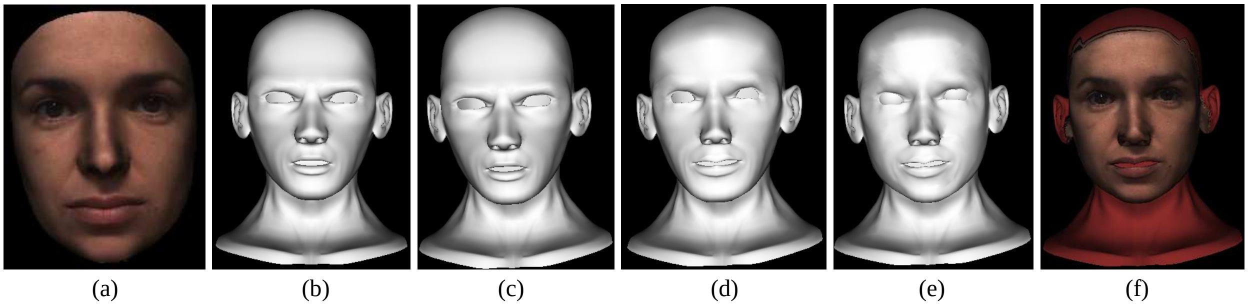

Figure 5. Generating the animatable model: (a) the target face, (b) the generic face rescaled, (c) aligned, (d) coarse warped, (e) fine warped, (f) textured

After the animatable model is obtained, the next step is animating the obtained face model in the visage|life™ animation tool. Since the FDP file which defines the feature points is automatically generated, we directly start facial motion cloning. In visage|life™, one can simply clone each morph target from a preexisting AFM rather than manually modeling them. After the Cloner is provided with two models, the target model and the source model, it copies morph targets from source to the target by utilizing the pre-defined feature points in order to find the correlation between the two faces. In the following image, you can see some of the resulting images.

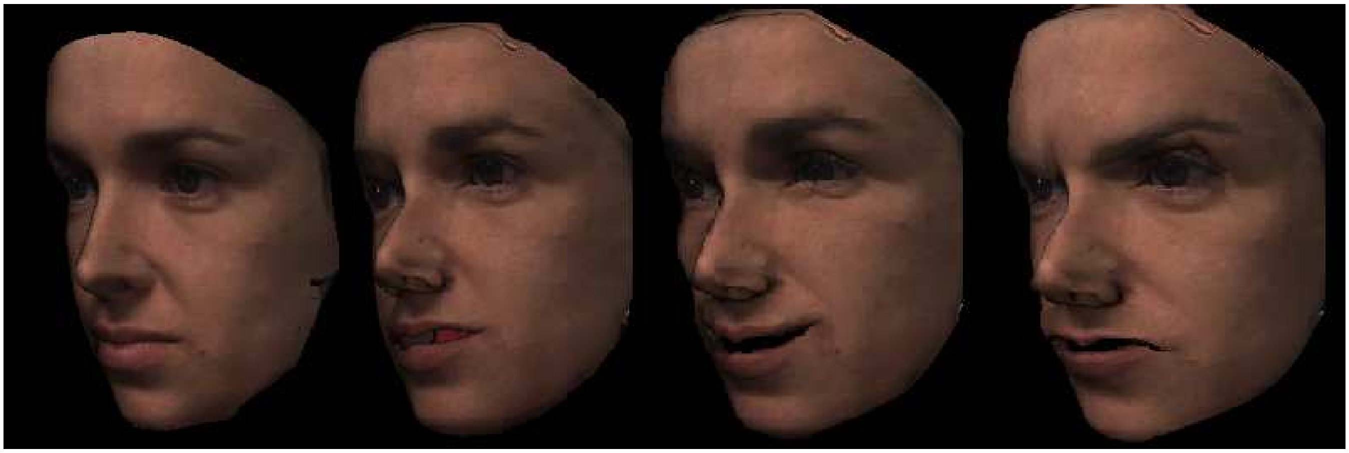

Figure 6. The target model, the resulting animatable model, the animated model with a smile and with a frown, respectively.

In order to observe the positive effect of the rendered images on face recognition problems while the presence of a facial expression, we experimented on the neutral and the "happy" faces of the Bosphorus database. We composed the gallery with one neutral image for each of the models which have multiple neutral shots in the database and its corresponding smiling image which is rendered using the animation engine. For testing, two probe sets are formed; one with another neutral image and one with a happy image of each subject.

For the identification purposes, simple Eigenfaces approach is adopted where decision-making is based on the Euclidian distance. In the first experiment, the identification is done on both the neutral and "happy" probe sets after the Eigenspace is constructed firstly by using the neutral images and secondly by using both neutral and rendered smiling faces in the gallery. In Fig. 7, the identification rates are given.

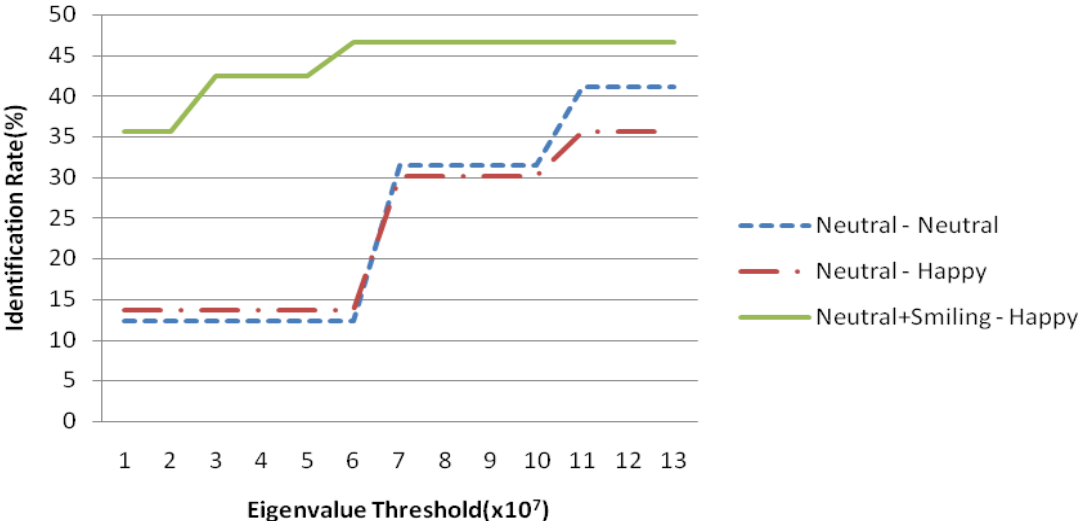

Figure 7. Identification rates

The second experiment is conducted for verification, where 36 people are removed from the gallery and treated as imposters. After constructing the Eigenspace with the optimum number of eigenvectors obtained from the first experiment, the verification tests are done in a similar manner on both the neutral and "happy" probe sets. Fig. 8 gives the ROC performances for each verification test. The increase in both identification and recognition rate indicates that our approach successfully eliminates the adverse effect of expression variations on face.

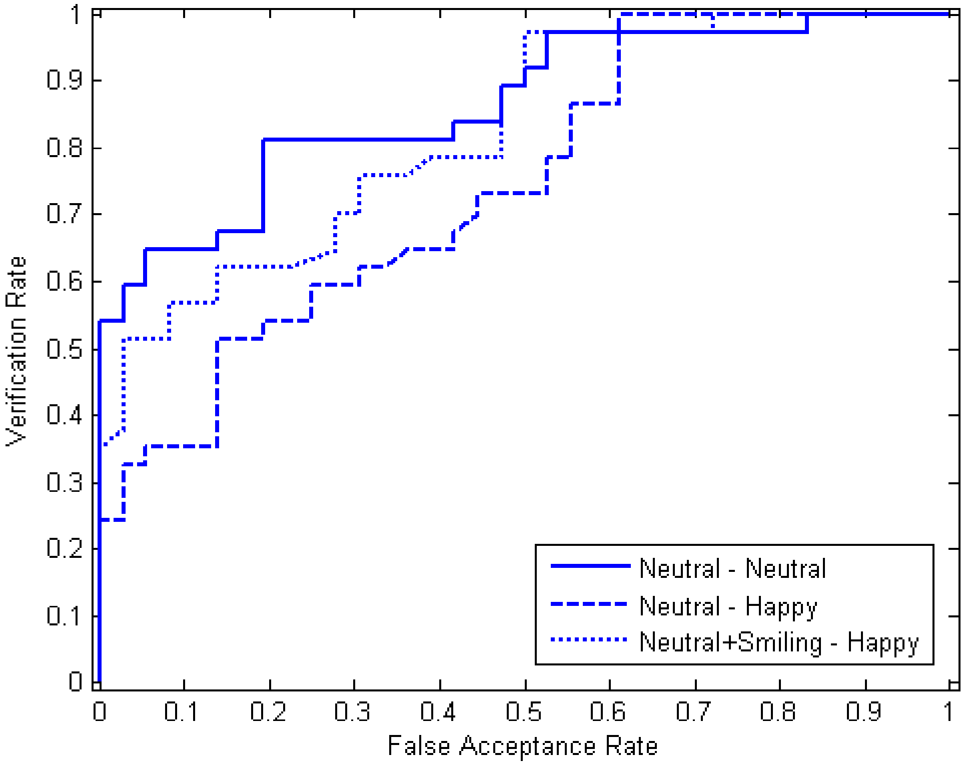

Figure 8. ROC performances for each verification test

4. CONCLUSION

In this paper, we propose a method in order to improve face recognition performances under the expression variations. By using both 2D and 3D data of the enrolled neutral faces, we constructed a system to obtain an animatable model by warping a generic one. Thereby, we are able to render images of the subjects with various facial expressions with the help of the visage|life™ animation tool. The results of the experiments conducted on a subset of the Bosphorus database are encouraging, however, there are still more possibility and necessity for further performance improvement, since several drawbacks remain, mainly due to manual land-marking of the face. Therefore, currently, we have been working on an automatic feature point detection system, which will allow us to fully automate the system while improving the recognition performance.

5. REFERENCES

[1] F.Tarrés, A. Rama, L. Torres, “A Novel Method for Face Recognition under Partial Occlusion or Facial Expression Variations”, 47th International Symposium ELMAR-2005, Multimedia Systems and Applications, 2005.

[2] E. F. Ersi, J. S. Zelek, “Local Feature Matching for Face Recognition”, Proceedings of the 3rd Canadian Conference on Computer and Robot Vision, p.4, 2006.

[3] Y. Bing, C. Ping, J. Lianfu, “Recognizing Faces with Expressions: Within-class Space and Between-class Space”, Proceedings of 16th International Conference on Pattern Recognition, vol. 1, p.139-142, 2005.

[4] K. Kirtac, O. Dolu, M. Gokmen, “Face Recognition by Combining Gabor Wavelets and Nearest Neighbor Discriminant Analysis”, 23rd International Symposium on Computer and Information Sciences, ISCIS '08, p.1-5, 2008.

[5] H.E. Komleh, V. Chandran, S. Sridharan, “Robustness to Expression Variations in Fractal-based Face Recognition”, Proceedings of 6th International Symposium on Signal Processing and its Applications, ISSPA’01, vol. 1, p.359-362, 2001.

[6] N. Alyuz, B. Gokberk, L. Akarun, “A 3D Face Recognition System for Expression and Occlusion Invariance”, 2nd IEEE International Conference on Biometrics: Theory, Applications and Systems, BTAS 2008, p.1-7, 2008.

[7] B. Ben Amor, M. Ardabilian, L. Chen, “Toward a Region-based 3D Face Recognition Approach”, IEEE International Conference on Multimedia and Expo, ICME’08, p.101-104, 2008.

[8] W. Lin, K. Wong, N. Boston, Y. Hu, “3D Face Recognition Under Expression Variations using Similarity Metrics Fusion”, IEEE International Conference on Multimedia and Expo, ICME’07, p.727-730, 2007.

[9] Mpiperis, S. Malassiotis, M. G. Strintzis, “Expression Compensation for Face Recognition Using a Polar Geodesic Representation”, 3rd International Symposium on 3D Data Processing, Visualization, and Transmission, 3DPVT'06, p.224-231, 2006.

[10] Y.J. Wang, C.-S. Chua, Y.-K. Ho, “Facial Feature Detection and Face Recognition from 2D and 3D Images”, Pattern Recognition Letters 23, p.1191-1202, 2002.

[11] F. Lavagetto, R. Pockaj, “The Facial Animation Engine: Toward a High-Level Interface for the Design of MPEG-4 Compliant Animated Faces”, IEEE Trans. Circuits and Systems for Video Technology, vol. 9, no. 2, pp. 277-289, 1999.

[12] V. Kuriakin, T. Firsova, E. Martinova, O. Mindlina, V. Zhislina, “MPEG-4 compliant 3D Face animation”, Proceedings of the 11th International Conference on Computer Graphics, GraphiCon’2001, p. 54-58, 2001.

[13] F. L. Bookstein, “Principal Warps: Thin-Plate Splines and the Decomposition of Deformations”, IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 11, no.6, 1989.

[14] Visage Technologies – The Character Animation Company
www.visagetechnologies.com/products_life.html

[15] Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, L. Akarun, “Bosphorus Database for 3D Face Analysis”, The First COST 2101 Workshop on Biometrics and Identity Management (BIOID 2008), Roskilde University, Denmark, May 2008.

[16] A. Savran, O. Çeliktutan, A. Akyol, J. Trojanova, H. Dibeklioğlu, S. Esenlik, N. Bozkurt, C. Demirkır, E. Akagündüz, K. Çalışkan, N.Alyüz, B.Sankur, İ. Ulusoy, L. Akarun, T. M. Sezgin, “3D Face Recognition Performance Under Adversorial Conditions”, in Proc. eNTERFACE’07 Workshop on Multimodal Interfaces, Istanbul, Turkey, July 2007.

[17] B. Ben Amor, H. Drira, B. Lahoucine, A. Srivastava, M. Daoudi, “An experimental illustration of 3D facial shape analysis under facial expressions”, Annals of Telecommunications journal, 64(5-6): 369-379 (2009)

[18] Igor S. Pandzic , Robert Forchheimer, “MPEG-4 Facial Animation: The Standard, Implementation and Applications”, John Wiley & Sons, Inc., New York, NY, 2003