Face Recognition with 3D Model-Based Synthesis

Xiaoguang Lu[1], Rein-Lien Hsu[1], Anil K. Jain[1], Behrooz Kamgar-Parsi[2], and Behzad Kamgar-Parsi[2]

[1] Michigan State University, East Lansing, MI 48824.
{lvxiaogu, hsureinl, jain}@cse.msu.edu
[2] Office of Naval Research, 800 N. Quincy St., Arlington, VA 22217.

Abstract. Current appearance-based face recognition system encounters the difficulty to recognize faces with appearance variations, while only a small number of training images are available. We present a scheme based on the analysis by synthesis framework. A 3D generic face model is aligned onto a given frontal face image. A number of synthetic face images are generated with appearance variations from the aligned 3D face model. These synthesized images are used to construct an affine subspace for each subject. Training and test images for each subject are represented in the same way in such a subspace. Face recognition is achieved by minimizing the distance between the subspace of a test subject and that of each subject in the database. Only a single face image of each subject is available for training in our experiments. Preliminary experimental results are promising.

1 Introduction

After decades of research [1], face recognition is still a very challenging topic. Current systems can achieve a good performance when the test image is taken under similar conditions as the training images. However, in real applications, a face recognition system may encounter difficulties with intra-subject facial variations due to varying lighting conditions, different head poses and facial expressions. Most of the face recognition methods are appearance-based [2–6] which require that several training samples be available under different conditions for each subject. However, only a small number of training images, are generally available for a subject in real applications, which can not capture all the facial variations.

A human face is a 3D elastic surface, so the 2D image projection of a face is very sensitive to the changes in head pose, illumination, and facial expression. Utilizing 3D facial information is a promising way to deal with these variations [5–12]. Adopting Waters' animation model [9] as our generic face model, we propose a face recognition system that synthesizes various facial variations to augment the given training set which contains only a single frontal face image for each subject. Both the training and test images are subjected to the model adaptation and synthesis in the same way. We use the synthetic variations to

construct an affine subspace for each subject. The recognition is achieved by matching the subspace of the test image with that for each of the subjects in the training database. Yamaguchi et al. [13] used the minimal principal angle between two subspaces, which are generated from two sets of images, to measure the dissimilarity between them. However, the mean of the set of images is not taken into account. We use an alternative distance metric to measure the dissimilarity between two affine subspaces.

2 Face Synthesis

2.1 Face Alignment

The face alignment module adapts a 3D generic face model [10, 9] onto a face image to extract facial shape and texture information. Waters' animation model contains 256 vertices and 441 triangular polygons for one half of the face. The other half of the face can be generated using the symmetry assumption of the human face. The face alignment is based on two processes: labeling and adaptation. In *labeling*, feature vertices are chosen from the 3D model. Currently, the 2D projected positions of these feature vertices (totally 115 feature vertices in our experiments) are labeled manually in the given intensity image. Figures 1(a)(b) illustrates these feature vertices. In *adaptation*, vertices in the 3D model, other than the feature vertices, are adjusted iteratively based on the propagation of feature vertex displacement [14].

Since there is no depth information available in the given 2D intensity image, the depth information (Z coordinate in Fig. 1(c)) of each model vertex is adapted by scaling the original generic model based on the average value of the global scaling factors (in the X and Y coordinates) of the generic model. Although the reconstructed face is not very realistic (the model is not dense enough), Fig. 1(d) shows that the face identity is preserved in the frontal view.

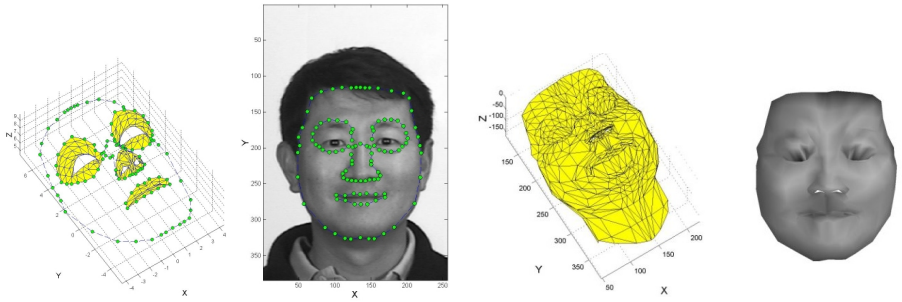
After face alignment, each model vertex is mapped with an intensity value at its corresponding coordinate on the face image. Texture values over non-vertex regions of each triangular facet on the 3D face model are interpolated by the intensity values of facet's vertices. See Fig. 1(d) for an example.

2.2 Eye Augmentation

There is no mesh for the eye regions in Water's model, so we create an additional mesh in the eye region to augment the original 3D mesh model. In each eye region, based on the vertices on its boundary, a grid is generated. The augmented eye mesh is obtained by using the Delaunay triangulation on these vertices. These vertices are adjusted according to the adaptation of the boundary vertices (see Fig. 3(b) for an example of the eye augmented reconstruction result).

2.3 Facial Variation Synthesis

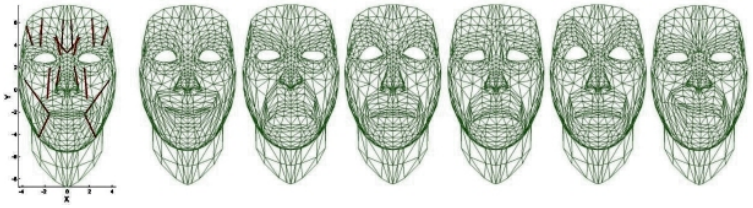
The face synthesis module synthesizes variations in head pose, illumination, and facial expression as follows. Rotating the face model and projecting it to



(a) (b) (c) (d)

Fig. 1. Face alignment: (a) feature vertices shown as "beads" on the 3D generic face model; (b) overlaid on a given intensity face image; (c) adapted 3D face model; (d) reconstructed images using the model shown in (c) with texture mapping.

the image plane can generate different pose variations. Lighting is simulated by adding a virtual light source around the reconstructed face surface. Phong shading technique is employed to render lighting effects on the face surface [15]. We use Terzopoulos and Waters [9, 10] approach of physics-based synthetic facial tissue and a set of anatomically motivated facial muscle actuators to synthesize facial expressions, see figure 2. Figure 3 shows the given intensity face image and



(a) (b) (c) (d) (e) (f) (g)

Fig. 2. Expression synthesis through 18 muscle contractions. The generic face mesh is: (a) shown in neutral expression (the dark bars represent 18 muscle vectors); distorted with six facial expressions (b) happiness ; (c) anger; (d) fear; (e) sadness; (f) surprise; (g) disgust.

several synthesis results from the adapted 3D model. Different types of synthesis (pose, lighting and expression) are done independently, so the combination of different types of synthesis is seamless.

3 Face Matching

3.1 Subspace Construction

For each subject, besides the input training image, a number of synthetic images are available after face synthesis. An affine subspace is constructed based on the original and synthetic images of this subject for face representation.

In the classical subspace analysis, an image is represented as a high-dimensional vector by concatenating each row (or column) of the image. Given a set of linearly

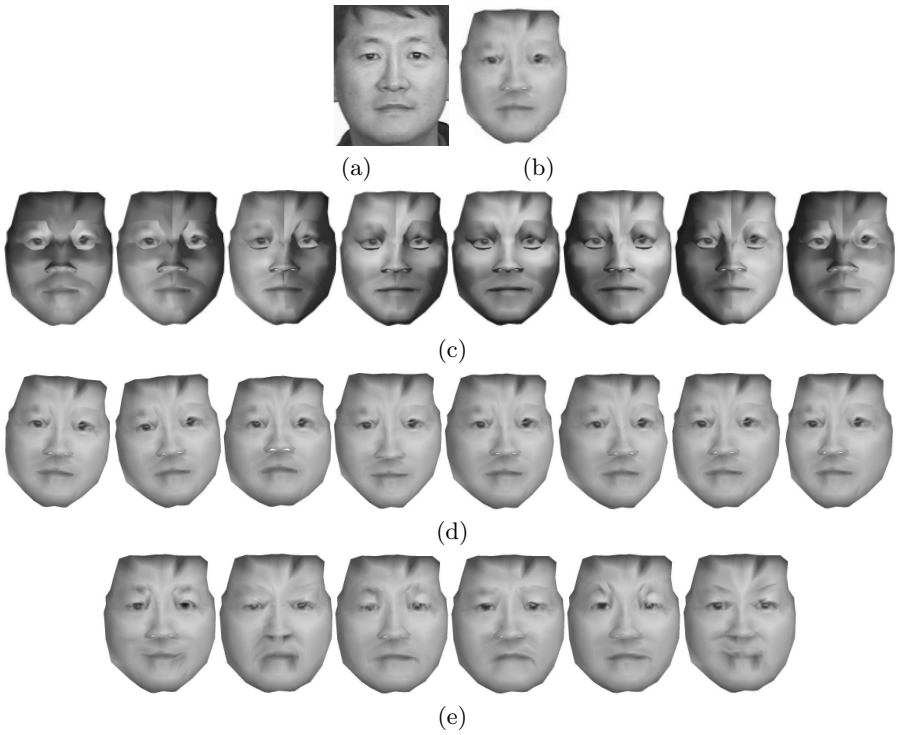


Fig. 3. Synthesis results: (a) input intensity image; (b) the reconstructed image; (c) image variations under 8 different lighting conditions; (d) 8 different pose variants; (e) synthesis results for 6 different expressions.

independent vectors, Gram-Schmidt orthogonalization can be used to obtain a set of orthogonal basis vectors to span the subspace. An affine subspace [16] M is of the form $M = M_0 + L$, where M_0 is a fixed vector and L is a subspace. Here, we call M_0 as the center of the affine subspace. Therefore, the sample mean of each set of images is taken into account in the following matching. Let $X_i (i = 1, \ldots, N)$ be the set of image vectors (original image and synthesized image vectors) for one subject, M be the affine subspace of this subject. The center M_0 of M is calculated as $M_0 = \frac{1}{N} \sum_{i=1}^{N} X_i$. Then image vectors $\tilde{X}_i = X_i - M_0$, are used to generate a set of basis vectors to span the subspace L, i.e., $L = \sum_{i=1}^{N} w_i \tilde{X}_i$, where w_i is a weight on $\tilde{X}_i$. Thus any vector X in this affine subspace can be represented as

$$X = M_0 + \sum_{i=1}^{N} w_i (X_i - M_0). \tag{1}$$

3.2 Subspace Matching

The distance between two affine subspaces (DAS) is defined as the minimum Euclidean distance between any two vectors from two different affine subspaces.

$$DAS = \min \|P - Q\|, \tag{2}$$

where P and Q are any vectors coming from two different affine subspaces. Finding this minimum distance is an optimization problem, whose close-form solution is derived in Appendix A.

4 Experiments and Discussion

For each of the ten subjects, a frontal image is captured once a week over a period of five weeks. As a result, 5 images for each subject are captured, resulting in a total of 50 images. The size of face area in the images is approximately 128×128. Figure 4 shows 20 cropped images in our database, two samples for each subject. One image from the 5 images of each subject is randomly selected to construct

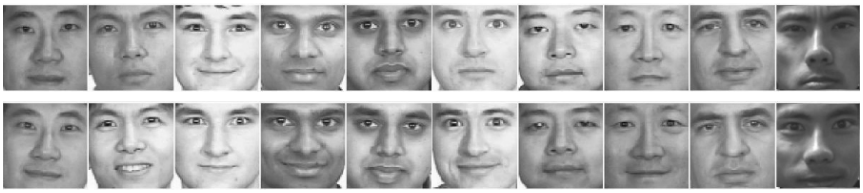


Fig. 4. Sample face images in the data set. All cropped images are aligned using the two eye centers.

the training set. Use the rest of the data set as the test set. This process is repeated 50 times. The same adaptation, synthesis and affine subspace analysis procedures are applied to each image.

In our experiments, a total of 8 images per subject with different poses are synthesized. They correspond to the in-plane-rotation to the left and right by 5 degrees (2 images are synthesized); tilt up and down by 10 degrees (2 synthetic images); pan to left and right by 5 and 10 degrees (4 synthetic images). The light source is placed above the face and 8 different angles are chosen to render the simulated lighting effect. A total of 8 synthetic images per subject are generated. Using the approach and parameters proposed in [9], we synthesize 6 different expressions per subject. Thus, a total of 22 synthetic images for each subject are obtained. Twenty two images based on one input image per subject are illustrated in Fig. 3. The images are closely cropped[4]. Cropped images are aligned by the centers of the two eyes and normalized to 64×64. All image vectors are normalized to be of unit length. For comparison purposes, the classical eigenface framework [2] (PCA) is applied to the data set without any synthesis augmentation. We captured 5 additional images for each subject. This additional data set is used for augmenting the original data to construct the eigenspace.

The matching results are illustrated in Fig. 5. On this data set our proposed method outperforms the PCA-based (without synthesis) method, indicating that the proposed synthesis for recognition scheme is promising.

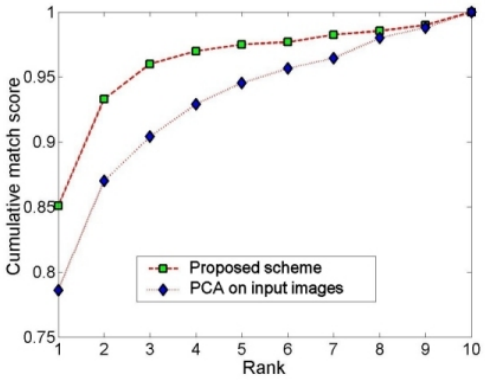


Fig. 5. Performance of proposed method and classical PCA based (without data augmentation by synthesis) method.

5 Conclusions and Future Work

We have proposed a human face recognition framework with 3D model based synthesis scheme to augment data set for face recognition when only a single training sample for each subject is available. A novel distance metric are utilized to measure the dissimilarity between two affine subspaces, which are constructed by the synthetic set of images. We are currently investigating the automatic alignment of a 3D generic model onto an intensity image. Experiments on a large public domain face database are also being carried out.

Appendix A: Distance between Two Affine Subspaces

The distance between two affine subspaces (denoted as DAS) is defined as the minimum distance between any vectors in the two affine subspaces. Let M_u and M_v be two affine subspaces. Any vectors u (in M_u) and v (in M_v) can be represented as $u = \mu_u + Ut_u, v = \mu_v + Vt_v$, where μ_u and μ_v are the centers of M_u and M_v, U and V are the basis matrixes spanning the subspace L of M_u and M_v, t_u and t_v are the coefficients vectors, respectively.

$$H = (u - v)^T(u - v) = (\mu_{uv} + Ut_u - Vt_v)^T(\mu_{uv} + Ut_u - Vt_v), \tag{3}$$

where H is the square of the DAS, . Taking the derivative of H, find the u and v (i.e. t_u and t_v) that minimizes H.

$$\begin{cases} \frac{\partial H}{\partial t} = (\mu_{uv} + Ut_u - Vt_v)^T U \\ \frac{\partial H}{\partial t} = (\mu_{uv} + Ut_u - Vt_v)^T V \end{cases} \tag{4}$$

$$(\mu_{uv}^T \ \mu_{uv}^T) = (t_u^T \ t_u^T) \begin{pmatrix} -U^T U & -U^T V \\ V^T U & V^T V \end{pmatrix}. \quad (5)$$

If the rightmost matrix in Eq. 5 is not singular, the projection coefficients can be derived as follows:

$$(t_u^T \ t_u^T) = (\mu_{uv}^T U \ \mu_{uv}^T V) \begin{pmatrix} -U^T U & -U^T V \\ V^T U & V^T V \end{pmatrix}^{-1}. \quad (6)$$

Acknowledgements

This research was supported by the Office of Naval Research, contract No. N00014-01-1-0266.

References

1. Zhao, W., Chellappa, R., Rosenfeld, A., Phillips, P.: Face recognition: A literature survey. CVL Technical Report, University of Maryland (2000), ftp://ftp.cfar.umd.edu/TRs/CVL-Reports-2000/TR4167-zhao.ps.gz.
2. Turk, M., Pentland, A.: Eigenfaces for recognition. Journal of Cognitive Neuroscience **3** (1991) 71-86
3. Lee, K.C., Ho, J., Kriegman, D.J.: Nine points of light: acquiring subspaces for face recognition under variable lighting. In: Proc. CVPR. (2001) 519-526
4. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. IEEE Trans. PAMI **19** (1997) 711-720
5. Zhao, W., Chellappa, R.: Face recognition using symmetric shape from shading. In: Proc. CVPR. (2000) 286-293
6. Sim, T., Kanade, T.: Combining models and exemplars for face recognition: An illuminating example. In: Proc. CVPR 2001 Workshop on Models versus Exemplars in Computer Vision. (2001)
7. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Proc. ACM SIGGRAPH. (1999) 187-194
8. Romdhani, S., Blanz, V., Vetter, T.: Face identification by matching a 3d morphable model using linear shape and texture error functions. In: Proc. ECCV. Volume 4. (2002) 3-19
9. Parke, F.I., Waters, K.: Computer Facial Animation. A. K. Peters Ltd. (1996), http://crl.research.compaq.com/publications/books/waters/Appendix1/appendix1.html.
10. Terzopoulos, D., Waters, K.: Analysis and synthesis of facial image sequences using physical and anatomical models. IEEE Trans. PAMI **15** (1993) 569-579
11. Essa, I., Pentland, A.: Coding, analysis, interpretation, and recognition of facial expressions. IEEE Trans. PAMI **19** (1997) 757-763
12. Guenter, B., Grimm, C., Wood, D., Malvar, H., Pighin, F.: Making faces. In: Proc. ACM SIGGRAPH. (1998) 55-66
13. Yamaguchi, O., Fukui, K., ichi Maeda, K.: Face recognition using temporal image sequence. In: Proc. IEEE FG'98, Nara, Japan (1998) 318-323
14. Hsu, R., Jain, A.: Face modeling for recognition. In: Proc. IEEE ICIP. Volume 2. (2001) 693-696
15. Foley, J., van Dam, A., Feiner, S., Hughes, J.: Computer Graphics: Principles and Practice. 2nd ed. Addison-Wesley, New York (1996)
16. Oja, E.: Subspace Methods of Pattern Recognition. Research Studies Press (1983)