Novel Example-Based Shape Learning For Fast Face Alignment

Xiujuan Chai[1, 2], Shiguang Shan[2], Wen Gao[1, 2], Bo Cao[2]

[1]Vilab, Computer College, Harbin Institute of Technology, Harbin, China, 150001
[2]ICT-YCNC FRJDL, Institute of Computing Technology, CAS, Beijing, China, 100080
xjchai@vilab.hit.edu.cn, {sgshan, wgao, bcao}@jdl.ac.cn

ABSTRACT

In this paper, a novel Example-based Shape Learning (ESL) strategy has been proposed for facial feature alignment. The method is motivated by an intuitive and experimental observation that there exists an approximate linearity relationship between the image difference and the shape difference, that is, similar face images imply similar face shapes. Therefore, given a learning set of face images with their corresponding face landmarks labeled, the shape of any novel face image can be learned by estimating its similarities to the training images in the learning set and applying these similarities to the shape reconstruction of the novel face image. Concretely, if the novel face image is expressed by an optimal linear combination of the training images, the same linear combination coefficients can be directly applied to the linear combination of the training shapes to construct the optimal shape for the novel face image. Our experiments have convincingly shown the effectiveness and efficiency of the proposed approach in both speed and accuracy performance compared with other methods.

Keywords: Face recognition, Face alignment, Example-based shape learning (ESL)

1. INTRODUCTION

Face recognition has a variety of potential applications in multi-modal interface, commerce and law enforcement, such as face screen saver, face logon interface, mug-shot database matching, identity authentication, access control, information security, and surveillance. Related research activities have significantly increased during the past decades [1, 2]. Though much progress has been made in the past few years and several successful commercial face recognition systems have emerged in the Biometrics market, the FERET [3] and FRVT2000 [4] show that practical face recognition system is still a great challenge both academically and industrially. The bottlenecks in a practical face recognition system mainly include the misalignment, pose and illumination problem [3, 4]. And first of all, accurate face alignment has been recognized as the most important task.

Face alignment can be generally classified into three levels: affine transformation, sparse feature correspondence and pixel-wise dense correspondence. Different feature correspondence is related to different feature definitions.

Affine alignment is to normalize face images by fixing the distance between two eyes and the vertical distance between eye and mouth, or only normalize the former. Most current face recognition systems are based on this level of correspondence. However, it is believed to be too simple for complex face analysis tasks. Therefore, sparse feature correspondence has currently attracted much attention in computer vision community. Active Shape Models (ASMs) and Active Appearance Models (AAMs) [5, 6] are among the dominant methods for sparse feature alignment. ASMs combine the local texture matching and the global shape subspace constraint. By alternating local search and global constraint, an iterative procedure is expected to converge to an optimal solution. AAMs build up an appearance model combining shape and texture. Alignment is achieved by optimizing the parameters of the appearance model. In spite of their accuracy for face alignment, however, due to the optimization procedure, ASM and AAM have been proved to be time-consuming and prone to trapping into local minimum.

Based on optical flow technology, Beymer etc. [7, 9] proposed an algorithm named *vectorizer*, to compute the dense correspondence between two different face images. In *vertorizer*, the 2D shape is represented as the optical flow field between the probe image and a reference image. However, such an algorithm is unavoidably constrained by the accuracy of optical flow computation, which is obviously a challenging problem in such an application.

In this paper, we investigate the relationship between a face image and its facial shape. An approximate linearity relationship is observed between the image difference and shape difference, that is, similar images imply similar shapes. According to such an observation, we propose a simple Example-based Shape Learning (ESL) strategy to solve the face alignment problem. In our method, given a learning set of face images with their corresponding face shapes labeled, the shape of any novel face image can be extracted by estimating its similarity to the training images

in learning set and applying the similarity to shape reconstruction.

The remaining part of the paper is organized as follows. In Section 2, we explain our preliminary experiment on the relationship between face image and face shape. In Section 3 we describe our example-based shape-learning algorithm in detail. Experimental results for shape extraction are given in Section 4. Finally, a short conclusive remark is presented in Section 5.

2. OBSERVATION ON THE RELATIONSHIP BETWEEN IMAGE AND SHAPE

In our approach, face shape is sparsely represented the same as in ASMs and AAMs, that is, a vector S of length $2n$ concatenated by the x and y coordinates of the pre-defined landmarks on the face:

$$S = (x_1, y_1, ..., x_n, y_n)^T,$$

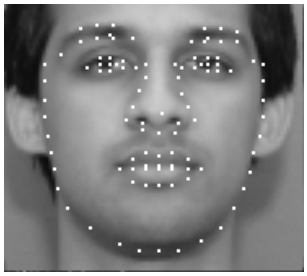

Fig.1 Example face with manually labeled landmarks

In our system, 103 landmarks on the face are chosen, as illustrated in Figure 1. And the face image is directly represented by the concatenated intensity values pixel by pixel in the image.

To find the relationship between a face image and its shape, we design the following experiments: Select some face images and label shape feature points manually. Therefore, we acquire the training data: m face images $I_1, I_2, \cdots, I_m$ and their corresponding shapes $S_1, S_2, \cdots, S_m$. And all of these images are aligned to the same scale by simple affine transformation. We choose one image randomly as the reference image, and then calculate the Euclidean distances between the images in training set and the reference image:

$$\Delta I_j = \sqrt{\sum_{i=0}^{n-1} (I_j(x_i, y_i) - I^*(x_i, y_i))^2}, \quad j = 0,1,\cdots,m-1,$$

where n is the total number of pixels in one image, and m is the number of images in our training set, I^* is the reference image. Then, we compute the corresponding Euclidean distance between S_j and S^*.

$$\Delta S_j = \sqrt{\sum_{i=0}^{N-1} (S_j(x_i, y_i) - S^*(x_i, y_i))^2}, \quad j = 0,1,\cdots,m-1,$$

where N is the number of shape feature points labeled manually. S^* is the shape vector of the reference image I^*. For any fixed image, we can get a set of ΔI_j and ΔS_j. When their distributions are displayed in a two-dimension coordinate system, we find that there is an approximate linearity relationship between ΔI_j and ΔS_j for the reference image. According to these data, a straight line can be fitted. Figure 2 illustrates some results of these data distribution and corresponding fitted lines, where the horizontal axis represents the image difference ΔI, and the vertical axis denotes the shape difference ΔS.

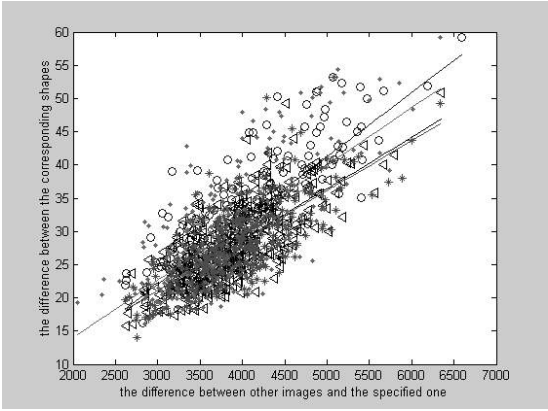

Fig.2 Approximate linearity relationship between the image difference and the shape difference

Consequently, a conclusion can be drawn from this experiment: there exists an approximate linearity relationship between the image difference and the shape difference. It suggests that similar face images imply similar face shapes, based on which, we propose our example-based shape learning algorithms for face alignment.

3. EXAMPLE-BASED FACE SHAPE LEARNING ALGORITHM

Since similar face images imply similar face shape as concluded in last section, we derive the following idea to extract face shapes: given a novel image, we first seek for the linear combination coefficients of the example face images that minimize the residue between the reconstructed image and the novel image, and then apply these optimal coefficients to the linear combination of the corresponding example shapes of these training images to reconstruct the facial shape of the novel face image.

To make the linear combination of the face images reasonable, all of the facial images used in this proposed method have been normalized to the same size by fixing the locations of two eyes at the same position. As above-mentioned, in our system, 103 landmarks on the face are chosen, as illustrated in Fig.1. And the face images are described by concatenating the intensity values pixel by

pixel in the image. Also, the intensity values of face images are all normalized to zero mean and one variance to reducing the influence of the lighting variance. For any given probe face image, the two eyes are first located by using some iris locating method such as in [11, 12].

By labeling some facial images manually, we can get a set of learning facial images $\{I_1, I_2, \cdots, I_N\}$ and the set of their corresponding facial shapes $\{S_1, S_2, \cdots, S_N\}$. Linear object class is a good representation approach widely used by researchers [8, 10]. In our work, similar ideas are utilized to represent images and shapes as follows:

$$I = \sum_{j=1}^{N} \omega_j I_j ,$$

$$S = \sum_{j=1}^{N} \omega_j S_j ,$$

for a given image *I* and its shape *S* respectively, that is, once we have obtain the linear combination coefficients for the image, we can get its face shape. Thereby, the pivotal problem is to compute the appropriate coefficients that can well reconstruct the input image linearly. It is formulated as follows:

For a given novel image, normalizing it in scale and in grey level, we gain image *I* which consists of *m* pixels, it can be denoted as:

$$I = (b_1, b_2, ..., b_m)^T .$$

And each normalized training image in the learning set is denoted as:

$$I_j = (a_{j1}, a_{j2}, ..., a_{jm})^T .$$

Therefore, we need to solve the following linear equation group:

$$\begin{cases} a_{11}\omega_1 + a_{21}\omega_2 + \cdots + a_{N1}\omega_N = b_1 \\ a_{12}\omega_1 + a_{22}\omega_2 + \cdots + a_{N2}\omega_N = b_2 \\ \cdots \\ a_{1m}\omega_1 + a_{2m}\omega_2 + \cdots + a_{Nm}\omega_N = b_m \end{cases} \quad (1)$$

Generally, it is an over-determined linear system, since *m* is far greater than *n*. To cope with this problem, the least square is commonly exploited. Let A be the training images matrix, $x = (\omega_1, \omega_2, \cdots, \omega_N)^T$ is the coefficients vector, then:

$$Y = \sum_{i=0}^{N} \omega_i I_i = Ax$$

is the linear approximation of the given unfamiliar image I, called reconstructed image, with an error term between Y and I:

$$E = \|I - Y\|^2 = \|I - Ax\|^2 = \left\| I - \sum_{i=1}^{N} \omega_i I_i \right\|^2$$

Therefore, to solve (1) is equivalent to solve the following minimization problem:

$$x^* = \min_x E \quad (2)$$

Suppose the space spanned by the vectors $I_1, I_2, \cdots, I_N$ is L. Then, to solve (2) is equivalent to find a vector *Y* in *L* that is closest to the input vector *I*. Suppose $Y = Ax$ is just the appropriate vector. Then $C = I - Y = I - Ax$ must be orthogonal to L space, i.e., satisfying:

$$(C, I_1) = (C, I_2) = \cdots = (C, I_N) = 0 .$$

Thus, we get:

$$I_1' C = 0, I_2' C = 0, \cdots, I_N' C = 0 .$$

It can be rewritten as matrix form as follows:

$$A'(I - Ax) = 0 .$$

Such a problem can be easily solved by:

$$x = A^{\perp} I. \quad (3)$$

Where $A^{\perp}$ is the pseudo-inverse matrix of A, given by:

$$A^{\perp} = (A'A)^{-1} A' .$$

Having determined the coefficients x, the shape *S* for the novel normalized image *I* can be computed relatively straightforward:

$$S = (S_1, S_2, \cdots, S_N)x \quad (4)$$

Finally, we can get the shape of the original input image by transforming the normalized shape feature points coordinate into the original image coordinate.

4. EXPERIMENTS

Experiments are carried out on a face database containing 300 near-frontal face images of 240x320, most of which are faces with neutral expression. Before the experiment, all of the 300 test images are labeled manually. To evaluate our method, the *leave-one-out* strategy is used to separate the training and the testing images. Therefore, the shape vector of any image selected randomly is calculated with the training set composed by the remaining 299 images through our shape learning method. Finally, we acquire all the shape vectors of the 300 images by our ESL method. So we can compute the error between the resulting shape of the proposed method and the ground truth shape vector by Euclidian distance. An average error of **1.95** pixels has been achieved, which demonstrates the effectiveness of the ESL.

In addition, since the pseudo-inverse matrix $A^{\perp}$ can be calculated offline, our method can work very fast, especially when compared with iterative methods such as ASMs or AAMs. Table 1 shows the performance

comparison between our ESL method and the ASM method. The ASM is also initialized according to the same positions of the two eyes. Our ESL based methods performs 7 times faster than the ASM, while with similar accuracy. Therefore, we can conclude that our shape learning method is an effective approach to extract facial shape.

Table 1. Comparison with other methods

Method	Time (ms/frame)	Average Error(pixels)	Error Variance
Mean Shape	/	2.48	3.45
ASM	70	2.14	2.23
ESL	10	1.95	1.85

Some experimental results of our method are illustrates in Fig.3, from which good performance of the proposed method can be seen. Moreover, ESL has the potentialities to extend to faces from different races, under various lighting conditions, and even for occlusion situations.

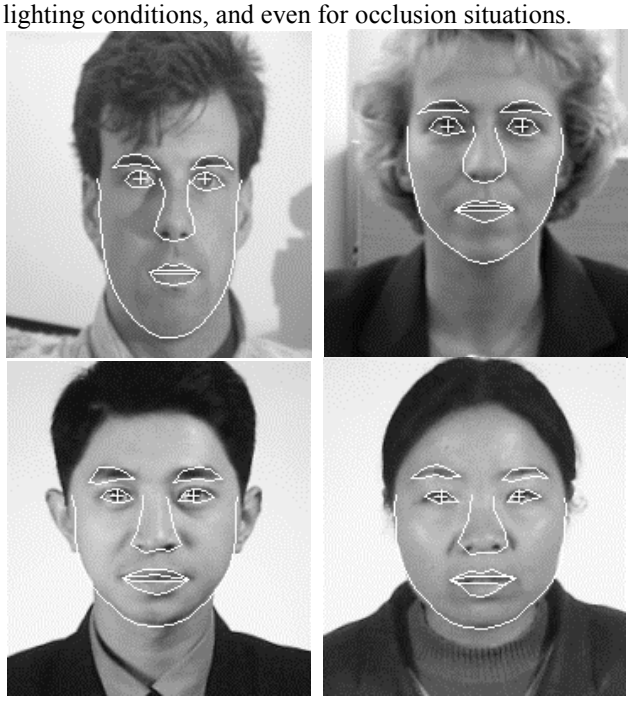

Fig.3. Results of the proposed ESL method.

5. CONCLUSION

In this paper we discuss the problem of face alignment in terms of the 2D face shape. An example-based learning strategy is proposed to derive shape information directly from the novel face image. It is based on the observation that similar face image imply similar face shape. Hence, the shape can be computed by the linear combination of training shapes by using the same weights as the corresponding linear combination to reconstruct the original face image. Our experiments have demonstrated the good performance of such a learning-based approach.

It is exciting to introduce successfully learning strategy to face alignment, though we just investigated a simple linear least square approach. More sophisticated learning algorithms are expected to produce much better performance, which would be one of our future works.

As an example-based learning method, sufficient examples are needed in the bootstrap set, or the results will not be precise enough. Illumination, expression and other sources of variation will also be considered carefully in the future.

ACKNOWLEDGEMENTS

This research is sponsored partly by National Hi-Tech Program of China (No.2001AA114160), SiChuan Chengdu YinChen Net. Co. (YCNC) and 100 Talents Foundation of Chinese Academy of Sciences.

REFERENCES

[1] R.Brunelli and T.Poggio, "Face Recognition: Features versus Template", *TPAMI*, 15(10), pp1042-1052, 1993

[2] R.Chellappa, C.L.Wilson ect. "Human and Machine Recognition of Faces: A survey", *Proc. of the IEEE*, 83(5), pp705-740, 1995.5

[3] P.J.Phillips, H.Moon, etc. "The FERET Evaluation Methodology for Face-Recognition Algorithms," *IEEE TPAMI*, Vol.22, No.10, pp1090-1104, 2000

[4] D.M.Blackburn, M.Bone, P.J.Phillips, Facial Recognition Vendor Test 2000: evaluation Report, Feb.2001, htt://www.frvt.org/frvt2000/

[5] T.F.Cootes, C.J.Taylor,D.Cooper, and J.Graham. "Active Shape Models--Their Training and Application," *Computer vision and image understanding*, 61(1): pp38-59, 1995.

[6] T.F.Cootes, G.J.Edwards and C.J.Taylor, "Active Appearance Models," Proc. European Conf. Computer Vision, vol. 2, pp. 484-498, 1998.

[7] D.Beymer and T.Poggio, "Face Recognition from One Example View," Proc. Int'l Conf. Computer Vision, pp. 500-507, 1995.

[8] D.Beymer, T.Poggio, "Image Representations for Visual Learning, Science," Vol.272, pp1905-1909, 1995.

[9] D.Beymer, "Vectorizing Face Images by Interleaving Shape and Texture Computation," A.I.Memo No. 1537, 1995.9

[10] T.Vetter, T.Poggio, "Linear Object Classes and Image Synthesis from a Single Example Image," IEEE Trans. On PAMI, Vol.19, No. 7, pp733-742, 1997

[11] J.Miao, B.C.Yin, K.Q.Wang, et al, "A Hierarchical Multiscale and Multiangle System for Human Face Detection in a Complex Background using Gravity-Center Template," Pattern Recognition, 32(7), 1999.

[12] B.Cao, S.Shan, W.Gao, "Localizing the Iris Center by Region Growing Search," Proceeding of the *ICME2002*.