Generic Face Alignment Using an Improved Active Shape Model

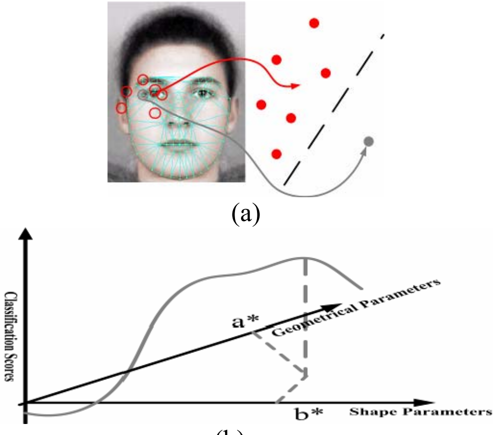
Liting Wang, Xiaoqing Ding, Chi Fang
Electronic Engineering Department, Tsinghua University, Beijing, China
{wanglt, dxq, fangchi} @ocrserv.ee.tsinghua.edu.cn

Abstract

Although conventional Active Shape Model (ASM) and Active Appearance Model (AAM) based approaches have achieved some success, however, evidence suggests that the performance of a person-specific face alignment which aligns the variation in appearance of a single person across pose, illumination, and expression is substantially better than the performance of generic face alignment which aligns the variation in appearance of many faces, including unseen faces not in the training set. This paper proposes a discriminative framework for generic face alignment. This technique is presented under the framework of conventional Active Shape Model (ASM) but has three improvements. First, random forest classifiers are trained to recognize local appearance around each landmark. This discriminative learning provides more robustness weight for the optimization fitting procedure. Second, to impose constrains, shape vectors are restricted to the vector space spanned by the training database. Third, data augment scheme is used for the benefit of a large training set. Experimental results show that this approach can achieve good performance on generic face alignment.

1. Introduction

Active Shape Model (ASM) [1] and Active Appearance Model (AAM) [2] are generative parametric models which are commonly used to align faces under various situations, such as pose, illumination, and expression changes. With the introduction of ASM and AAM by Cootes [1][2], face alignment becomes more popular in computer vision research area which allows rapidly location of the boundary of objects. By learning statistical distribution of shapes and textures from training database, a deformable shape model is built. The boundary of objects with similar shapes to those in the training set could be extracted by fitting this deformable model to images. Depending on the different tasks, ASM and AAM can be built in different ways. On one hand, we might construct a person specific ASM or AAM across pose, illumination, and expression. Such a person-specific model might be useful for interactive user interface applications including head pose estimation, gaze estimation etc. On the other hand, we might construct ASM or AAM to align any face, including faces unseen in training set. Evidence suggests that the performance of a person-specific face alignment is substantially better than the performance of generic face alignment. As indicated in [3], Gross's experimental results confirm that generic face alignment is far harder than person-specific face alignment and the performance degrades quickly when fitting to images which are unseen in the training set.



(a)

(b)

Figure 1. (a) Local appearance model: leaning a classifier for each landmark (In our model, 88 classifiers should be trained. We take left eye left corner point for example); (b) Parameter optimization: maximize the outputs of all 88 classifiers to get the best geometrical parameters a* and shape parameter b*

To remedy the generalization problems, this paper proposes an improved generic face alignment method under the framework of conventional ASM, but has three main improvements:

978-1-4244-1724-7/08/$25.00 ©2008 IEEE

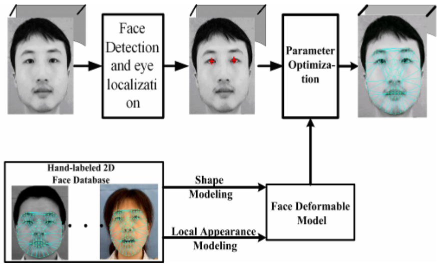
Firstly, as illustrated in Figure.1 (a), learning a classifier for each landmark simplifies the problem. We propose a novel discriminative method of local appearance modeling which distinguished correct point from incorrect point and such classification score could be used to parameter optimization. In our face model, 88 classifiers should be trained.

Secondly, after initializing the shape parameters, our optimization method iteratively updates the parameter such that the 88 classifiers outputs achieve the maximal scores and get the best shape parameter b* and geometrical parameter a* as illustrated in Figure.1 (b).

Thirdly, data augment scheme is used to improve the performance of generic face alignment via augmenting ground truth data. Although this scheme does not improve the misplaced labels, it significantly improves face alignment performance.

2. Face alignment framework overview

The proposed face alignment framework consists of training and aligning procedures, as illustrated in Figure.2.



Face Detection and eye localizati on
Parameter Optimiza-tion
Hand-labeled 2D Face Database
Shape Modeling
Local Appearance Modeling
Face Deformable Model

Figure 2. Face alignment framework

Training procedure is building a face deformable model via shape modeling and local appearance modeling. This procedure needs a great amount of hand labeled data. Aligning procedure consists of firstly face detection and eye localization and then parameter optimization based on the trained face deformable model.

In face detection and eye localization procedure, the state-of-the-art techniques are adopted. For face detection, a boosted cascade detector proposed by Viola and Jones [5] is used. For Eye localization, a robust and precise eye location method was found in [4], and we adopt this method in this paper to precisely locate the eye position. Both the two methods are real-time.

The training of generic face deformable model based on the hand labeled 2D face data and the parameter optimization are the main work of this paper. In the following paragraphs, they will be presented and discussed in detail.

3. Training face deformable model

Assume a training set of shape to be $\left\{X_i\right\}_{i=1}^{N}$. The shape $X_i = \left\{\left(x_j^i, y_j^i\right)\right\}_{j=1}^{K}$ is the sequence of hand-labeled K points in the image lattice. As illustrated in Figure.3, we manually label 88 points for each face image.

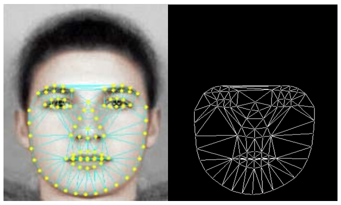


Figure 3. Manual labeled 88 points of training image

Training face deformable model consists of three steps: manually labeling faces, shape modeling and local appearance modeling. Shape modeling is the same as the conventional ASM [1]. Local appearance modeling was proposed in this paper using a novel discriminative method. We will briefly summarize shape modeling and discuss local appearance modeling in detail.

3.1. Shape modeling

Shape modeling is the same as the conventional ASM. It is represented as a vector *b* in the low dimensional shape eigenspace spanned by *k* principal modes (major eigenvectors) learned from the training shapes. A new shape X could be linearly obtained from shape eigenspace:

$$X \approx \bar{X} + Pb \tag{1}$$

where *P* is the matrix consisting of *k* principal modes of the covariance of $\{Xi\}$.

3.2. Local appearance modeling

In conventional ASM, the local appearance models, which describe local image feature around each landmark, are modeled as the first derivatives of the sampled profiles perpendicular to the landmark contour. However, this approach ignores the difference between landmarks and nearby backgrounds. This paper proposes to learn the local appearance classifier for each landmark. Several classification algorithms, such as SVM, or neural networks could have been chosen. Among those, Lepetit [7] has found random forest [6] to be eminently suitable because it is robust and fast, while remaining reasonably easy to train.

3.2.1. Random forest. In this section, we first describe them briefly in the context of our problem for the benefit of the unfamiliar reader. Figure.4 depicts a random forest. It consists of *N* decision trees. Each decision tree is trained by completely random approach. For each decision tree T_n, the samples are selected randomly from the training sample pool. It is a subset of all the training samples. After *N* trees are trained, the final decision combines all the outputs of T_1 T_2 $\ldots T_N$ by considering the average of all *N* outputs.

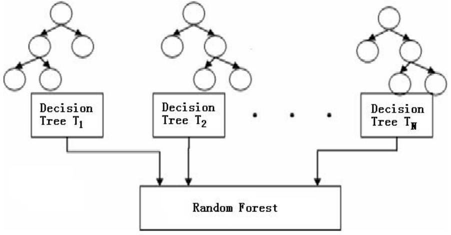


Figure 4. Random forest combines the outputs of all decision trees as a classifier fusion method.

3.2.2. Samples collection. For each landmark, we train one random forest. As illustrated in Figure.1 (a), we take an example of left eye left corner point. All the samples are cropped from faces (the distance between left eye center and right eye center is normalized into 60 pixels). Positive samples are the 32×32 image patches of all the training images with its center at the ground-truth landmark position. While negative samples are the 32×32 image patches of all the training images with its center inside 40×40 but outside 5×5 region from the ground-truth landmark position. Here, we take left mouth corner for example the positive samples and the negative samples are collected as Figure.5

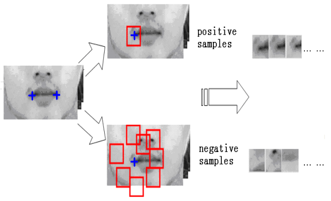


Figure 5. Positive and negative samples

3.2.3. Local appearance model. After training 88 random forest classifiers for 88 landmarks as illustrated in Figure.6, we could get 88 outputs. Each output of random forest classifier indicates the confidence a sample belongs to. The larger it is, the more probable it is a positive sample.

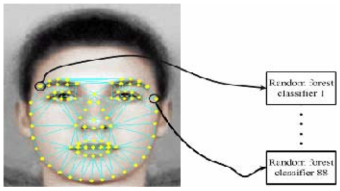


Figure 6. Learn a local appearance classifier for each landmark, in our face model, Totally train 88 random forest classifiers.

The position with the largest confidence is chosen to be the candidate position for the next ASM parameter optimization procedure.

4. Parameter optimization

Face alignment aims to find the best fit of the 88 points which defined in Figure.3. As illustrated in Figure.1 (b), this optimization procedure is to maximize the classification scores of all 88 random forest outputs. Finding the best shape parameter b* and geometrical parameter a*. The optimization procedure could be depicts as follows:

We use the eye location result [4] to initialize the deformable model. This optimization problem could be solved by two step procedure.

Step1 Relocating all the landmarks using the local appearance models, we obtain a new candidate shape Y and the weight matrix W from 88 random forest outputs.

Step2 After relocating all the landmarks using the local appearance models, we obtain a new candidate shape Y. In conventional ASM, The solution in shape eigenspace is derived by maximizing the likelihood:

$$\min_{a,b} \left\| Y - T_a(\bar{X} + Pb) \right\|_2 \qquad (2)$$

Where *a* represents the scale rotation and geometrical translation based on four parameters:

$$a = (X_t, Y_t, s, \theta)$$

$$T_{X_t,Y_t,s,\theta}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} X_t \\ Y_t \end{pmatrix} + \begin{pmatrix} s\cos\theta & -s\sin\theta \\ s\sin\theta & s\cos\theta \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} \qquad (3)$$

And *b* has been introduced in formula (1) which represents parameter of shape eigenvectors.

Along with the desired solution (2), two improvements are presented. First, it is essential to impose constraints on the formula (2). In our method, shape vector *b* is restricted to the vector space spanned by the database. Second, 88 facial points could be weighed by the outputs of random forest classifiers. We add the weight matrix W into the optimization. The optimization objective function is changed to (4):

$$\min_{a,b}(Y-T_a(\bar{X}+Pb))^T W(Y-T_a(\bar{X}+Pb))+\sum_{i=1}^{t} b_i^2/\sigma_i^2 \qquad (4)$$

After two step optimization procedure, the best shape parameter a* and b* are find. Face alignment result is $T_a(\bar{X}+Pb)$.

5. Data augment

To obtain a satisfactory performance, a large training data set is necessary. Obviously, manual labeling such large a training data set is time consuming and the quality of the labels is less than perfect. We use a data augment algorithm to augment the training set. First we manually label a subset of the training data and construct an original face deformable model. We then align the unlabeled images with the face model and modify the incorrect labeled points. After that, we construct another face deformable model with all labeled faces then align to the unlabeled and modify again until all the training data are labeled. It operates: (1) build ASM, (2) align unlabeled faces, (3)re-build ASM, (4) align unlabeled faces. Such data augment scheme enlarges our training data set which will significantly improve face alignment performance.

6. Experiments

In order to verify our algorithm, experiments have been conducted on a large data set consisting of 3,244 images from four databases as illustrated in Figure.7. We collect and construct the THFaceID database including totally 534 male and female aging from young to old with various facial expressions. Yale database [8], FRGC database [9] and JAFFE database [10] are all public available. Yale database includes illumination changes and facial expression changes; FRGC database also includes facial expression and illumination changes under controlled and uncontrolled situations; JAFFE database includes expression changes. All the 3,244 images are manually labeled 88 points as illustrated in Figure.3 by data augment procedure.

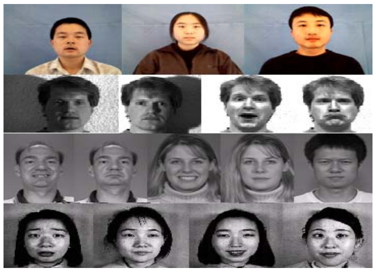


Figure 7. Images examples of four databases
THFaceID database (top line); Yale database (line 2); FRGC (line 3); JAFFE database (bottom line)

We divided all the images into three test sets. Set A is our training set; Set B is the test set which has the same subjects but different images from JAFFE; Set C is another test set which has unseen subjects from training set. Table 1 lists the images partition. FRGC for training only; Yale 15 subjects, 7 images per one subjects, totally 105 images for training; 15 subjects, 4 images per one subjects, totally 60 images for testing set B; JAFFE 10 subjects, 120 images for training and 93 images for testing set B; THFaceID 120 subjects for training but the other 414 subjects which are unseen subjects of training set for testing set C. Set A is manually labeled for training and Set B and Set C is manually labeled for ground truth data.

Table.1 Database description

	THFaceID	Yale	FRG	JAFFE
Images	1796	165	1070	213
Subjects	534	15	535	10
Set A	120(554)	15(105)	535	10(120)
Set B		15(60)		10(93)
Set C	414(1242)			

Given a dataset with ground truth landmarks, proposed face alignment algorithm automatically detects faces and locates eye positions. The eye localization is used as the initialization for parameter optimization procedure. After initialization, the faces are aligned by the generic face deformable model trained before.

The accuracy is measured by (5) calling the relative error *e*, which is the point to point error between the face alignment results P_a and manually labeled ground-truth P_m when the distance of left and right eye d_e is normalized to 60 pixels.

$$e=\sum_{i=1}^{88}\left\|\mathrm{P_a}-\mathrm{P_m}\right\|_2/(88\cdot d_e) \qquad (5)$$

Accuracy testing results are shown in Figure.8. It illustrates the distribution of the relative error on Test Set B and Set C. The mean relative error on Test Set B is ***3.60*** pixels and the mean relative error on Test Set C is ***4.86*** pixels. (distance of two eyes is normalized to 60 pixels). From the results, we could see the face alignment performance on unseen subjects is more difficult than on the seen subjects' unseen images.

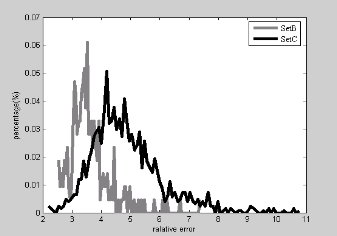


Figure 8. Accuracy performance tested on Set B and Set C

The X-axis is the relative error and the Y-axis is the number percentage of alignment results which shows the distribution of relative error on Set B and Set C.

Some results of face alignment under difficult illuminations, expressions and occlusions are shown in Figure.9.

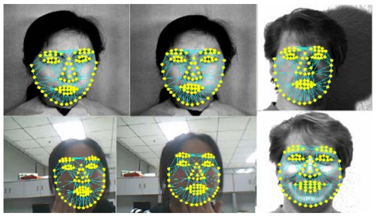


Figure 9. Some results of face alignment under difficult illuminations, expressions and occlusions

7. Conclusion

Generic face alignment on unseen subjects is a difficult task in face alignment research. In this paper, an improved generic face alignment method has been proposed. The main novelty is representing the local appearance via discriminative learning around each landmark. This discriminative learning provides more robustness weight for the parameter optimization procedure. Experimental results demonstrate its effectiveness on generic face alignment.

8. Acknowledgements

This work is supported by The National High Technology Research and Development Program of China (863 program) under contract No. 2006AA01Z115 & National Natural Science Foundation of China (project 60472002).

9. References

[1] T. Cootes, C. Taylor, D. Cooper and J. Graham，"Active shape models – their training and their applications," *Computer Vision and Image Understanding*, 61(1), pp. 38-59, January 1995.

[2] T. Cootes, G. Edwards, C. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23 (6) pp.681–685. 2001

[3] Gross, R, Matthews, I, Baker, S, "Generic vs. Person Specific Active Appearance Models," *Image and Vision Computing (23)*, No. 12, 1, pp. 1080-1093. November 2005

[4] Yong Ma., Xiaoqing Ding., Zhenger Wang., Ning Wang, "Robust precise eye location under probabilistic framework," *IEEE International Conference on Automatic Face and Gesture Recognition (FGR 2004)*, pp. 339-344, 2004

[5] Viola, P., Jones, M. "Rapid object detection using a boosted cascade of simple features," *In: CVPR*, vol.1, pp. 511–518, 2001

[6] Breiman L, "Random Forests," *Machine Learning*, 45, pp. 5–32, 2001

[7] V. Lepetit and P. Fua, "Keypoint Recognition Using Randomized Trees," *Digital Object Identifier 10.1109/TPAMI*, pp(s): 1465- 1479. September, 2006

[8] The Yale face database: http: // cvc.yale.edu/projects/yalefaces/yalefaces.html.

[9] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer,J. Chang, K. Hoffman, J. Marques, J. Min, and W.Worek， "Overview of the face recognition grand challenge," *In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[10] The JAFFE database: http:// www.mis.atr.co.jp/~mlyons/ jaffe.html