Chapter 12
Facial Landmark Localization

Xiaoqing Ding and Liting Wang

12.1 Introduction

Face detection and recognition is a vibrant area of biometrics with active research and commercial efforts over the last 20 years. The task of face detection is to search faces in images, reporting their positions by a bounding box. Recent studies [19, 31] have shown that face detection has already been a state-of-the-art technology in both accuracy and speed. However, face detection is not sufficient to acquire facial landmarks, for example, eye contours, mouth corners, nose, eyebrows, etc. This is therefore the task of facial landmark localization which aims to find the accurate positions of the facial feature points as illustrated in Fig. 12.1. It is a fundamental and significant work in face-related areas, for example, face recognition, face cartoon/sketch, face pose estimate, model-based face tracking, eye/mouth motion analysis, 3D face reconstruction, etc.

There is a wide variety of works related to facial landmark localization. The early researches extract facial landmarks without a global model. Facial landmarks, such as the eye corners and centers, the mouth corners and center, the nose corners, chin and cheek borders are located based on geometrical knowledge. The first step consists of the establishment of a rectangular search region for the mouth and a rectangular search region for the eyes. The borders are extracted by applying corner detection algorithm such as SUSAN border extraction algorithm [17]. Such methods are fast, however, they could not deal with faces of large variation in appearance due to pose, rotation, illumination and background changes.

X. Ding (✉) · L. Wang
State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
e-mail: dingxq@tsinghua.edu.cn

L. Wang
e-mail: wangltmail@tsinghua.edu.cn

S.Z. Li, A.K. Jain (eds.), *Handbook of Face Recognition*,
DOI 10.1007/978-0-85729-932-1_12,

Fig. 12.1 Facial landmark localization

Different with the earlier model-independent algorithm, some researches focus on model-dependent algorithm. Hsu and Jain [18] propose an approach which represents human faces semantically via facial components such as eyes, mouth, face outline, and the hair outline. Each facial component is encoded by a closed (or open) snake that is drawn from a 3D generic face model. The face shape model here is not based on statistical learning and it still could not deal with faces of large variation in appearance due to pose, rotation, illumination and background changes.

With the prominent successful research of Active Shape Model (ASM) [7, 9, 3, 8] and Active Appearance Model (AAM) [4–6, 10–13], face shape is well modeled as a linear combination of principal modes (major eigenvectors) learned from the training face shapes. 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 fit any face, including faces unseen in training set. Evidence suggests that the performance of the person-specific facial landmark localization is substantially better than the performance of generic facial landmark localization. As indicated in [15], Gross's experimental results confirm that generic facial landmark localization is far harder than person-specific facial landmark localization and the performance degrades quickly when fitting to images which are unseen in the training set.

In recent years, there are several improved research works based on the framework of AAM. Papandreou and Maragos [27] introduce two enhancements to inverse-compositional AAM matching algorithms in order to overcome the limitation when inverse-compositional AAM matching algorithms are used in conjunction with models exhibiting significant appearance variation, such as AAMs trained on multiple-subject human face images. Liu Xiaoming [25, 26] proposes a discriminative framework to greatly improve the robustness, accuracy and efficiency of face alignment for unseen data. Liebelt et al. [24] develop an iterative multi-level algorithm that combines AAM fitting and robust 3D shape alignment. Xiao et al. [33] also develop the research work of combining 2D AAM and 3D Morphable Model (3DMM). Hamsici and Martinez [16] derive a new approach carries the advantages of AAM and 3DMM that can model nonlinear changes in examples without the

need of a pre-alignment step. Lee and Kim [21] propose a tensor-based AAM that can handle a variety of subjects, poses, expressions, and illuminations in the tensor algebra framework. They reported Tensor-based AAM reduced the fitting error of the conventional AAM by about two pixels and the computation time by about 0.6 second.

There are also several improved research works based on the framework of ASM. Tu et al. [30] propose a hierarchical CONDENSATION framework to estimate the face configuration parameter under the framework of ASM. Jiao et al. [20] present a W-ASM, in which Gabor wavelet features are used for modeling local image structure. Zhang and Ai [34] propose an Adaboost discriminative framework which improves the accuracy, efficiency, and robustness of ASM. The same research works are also carried on by Li and Ito [23] who describe a modeling method by using AdaBoosted histogram classifiers. Brunet et al. [2] define a new criterion to select landmarks that have good generalization properties. Vogler et al. [32] combine the ASM with 3D deformable model which governs the overall shape, orientation and location.

In the following, we will introduce a coarse-to-fine facial landmark localization algorithm which uses discriminant learning to remedy the generalization problems based on the framework of Active Shape Model.

12.2 Framework for Landmark Localization

This facial landmark localization framework consists of training and locating procedures, as illustrated in Fig. 12.2.

The 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. The locating procedure consists of firstly the face detection, the eye localization and then the facial landmark localization based on the face deformable model. In the eye localization procedure, we will introduce a robust and precise eye localization method, and then adopt this method to precisely locate the eye position. The eye localization method is real-time. In the facial landmark localization procedure, a random forest embedded active shape model is adopted. In the following paragraphs, they will be presented and discussed in detail.

12.3 Eye Localization

The eye localization is a crucial step towards automatic face recognition and facial landmark localization due to the fact that these face related applications need to normalize faces, measure the relative positions or extract features according to eye positions. Like other problems of object detection under complex scene such as face detection, car detection, eye patterns also have large variation in appearance due to various factors, such as size, pose, rotation, the closure and opening of eyes,

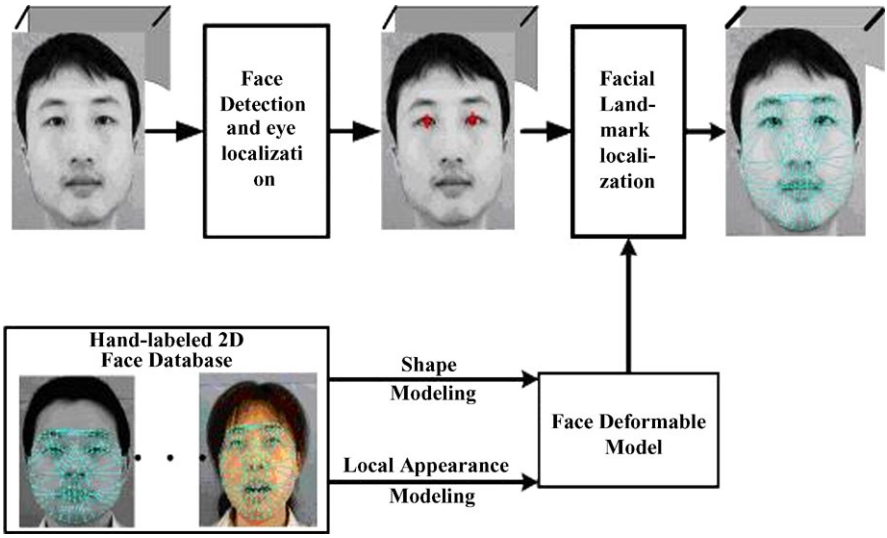

Fig. 12.2 Facial landmark localization processing framework

illumination conditions, the reflection of glasses and the occlusion by hairs etc. Even having found the positions of faces grossly, robustly and precisely locating the eye's center is still a challenging task. A variety of eye detection and tracking algorithms have been proposed in recent years, but most of them can only deal with part of these variations or be feasible under some constraints. We have devised a novel approach for precisely locating eyes in face areas under a probabilistic framework. The experimental results demonstrate that our eye localization method can robustly cope with different eye variations and achieve higher detection rate on diverse test sets.

The block diagram of the proposed method is shown in Fig. 12.3. When a rough face region is presented to the system, mean projection function and variance projection function [14] are adopted for determining the midline between the left and right eye. Then in the two areas, the appearance-based eye detector is used to find eye candidates separately. All the eye candidates are subsampled according to their probabilities. The remaining left and right eye candidates are paired. All the possible eye pairs are classified by an appearance-based eye-pair classifier. The most probable eye pairs are taken as the locations of left and right eyes.

12.3.1 Midline of Eyes

For an upright frontal face, the vertical midline between left and right eye is near the bridge of nose. According to the observations that the change of gray intensity on eye area is more obvious than bridge of nose and the eye area is often darker than the bridge of nose, vertical mean and variance projection function [14] are

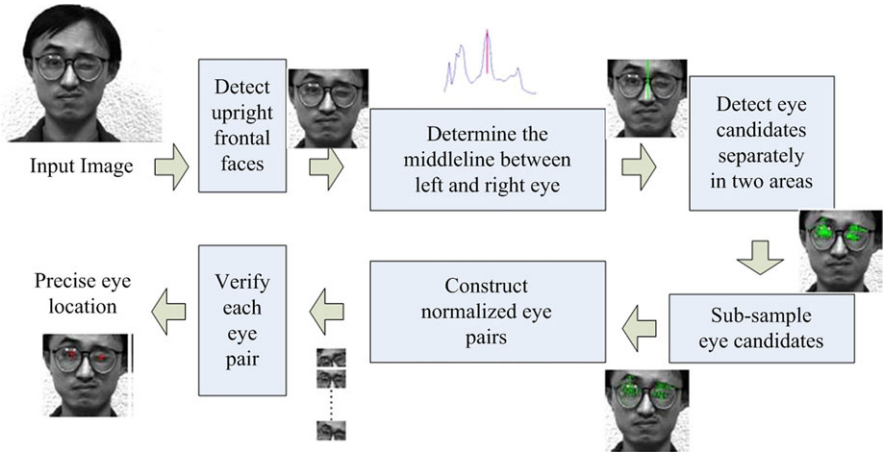

Fig. 12.3 Flowchart of the precise eye localization method under probabilistic framework

used. Suppose $I(x, y)$ is the intensity of a pixel at location (x, y), the vertical mean projection function $\mathrm{MPF}_v(x)$ and vertical variance projection function $\mathrm{VPF}_v(x)$ of $I(x, y)$ in intervals $[y_1, y_2]$ can be defined respectively, as:

$$\mathrm{MPF}_v(x) = \frac{1}{y_2 - y_1} \sum_{y=y_1}^{y_2} I(x, y) \tag{12.1}$$

$$\mathrm{VPF}_v(x) = \sqrt{\frac{1}{y_2 - y_1} \sum_{y=y_1}^{y_2} \left[I(x, y) - \mathrm{MPF}_v(x) \right]^2} \tag{12.2}$$

Applying the two functions to upper half of a face region, an obvious response around the bridge of nose will be obtained (Fig. 12.3b). So the position of vertical midline separating the left eye from right eye can be estimated. An appearance-based eye detector will be applied in the two areas separately.

12.3.2 Eye Candidate Detection

We used standard AdaBoost training methods combined with Viola and Jones's cascade approach to build appearance based eye detector. The cascade structure enables the detector to rule out most of the face areas as eye with a few tests and allows computational resources to be concentrated on the more challenging parts of the images. The features used in AdaBoost training process are Haar basis vectors [31] as elementary features. For an eye sample with size of 24×12, there are about 40 000 features in total. There are in total 6800 eye samples in the positive training set, some of which can be seen in Fig. 12.4. All the eye samples are cropped from faces

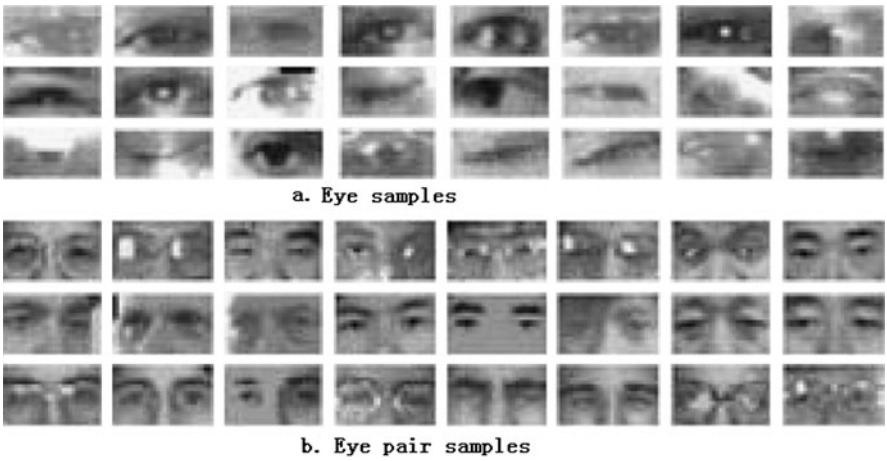

Fig. 12.4 Positive training examples for AdaBoost

with the eye center being the center of the example. The negative examples are obtained by a bootstrap process [28]. All the samples are processed with gray scale normalization and size normalization to 24×12 pixels. In this step, we avoid making premature decisions about the precise location of an eye. By contrast, we just exclude most background and give all the candidates with the probabilities at the expense of some false positives. The face regions most easily confused with eyes are eyebrows, thick frames of glasses, etc. In Fig. 12.3c, the center of every detected candidate is denoted by a dot in the face area.

12.3.3 Eye Candidate Subsampling

Because the local appearance of eyes is not nearly as distinctive as that of the whole face, some spurious eyes such as eyebrows, spectacle frames would be found, and true eyes would be located in different scales and near positions (in Fig. 12.3d). If all the candidates were considered in the next step, the processing time would be too long (e.g., for 40 left eye candidates and 40 right eye candidates, we have 1600 eye pairs in next step). To merge the candidates in a neighborhood, we subsample candidates with a factor of N in horizontal and vertical direction according to the probabilities (as shown in Fig. 12.3d). N is adjusted according to the face width. After the subsampling step, the number of eye pairs is reduced to $1/3$ or less of the original amount.

12.3.4 Eye-Pair Classification

To exclude spurious and inaccurate eye candidates, we build an eye-pair classifier in a similar way constructing the eye detector described above. Each eye-pair sample includes the bounding rectangle around the left and right eyes with a small amount of space above and below the eyes, some of which can be seen in Fig. 12.4.

Negative eye-pair examples are collected also using bootstrap method. All the samples are normalized to 25×15 pixels. In the test stage, for every pair of candidates in our list we figure out all possible pairings such that a priori information on inter ocular distances is satisfied. Then we use the affine warp to normalize the pair's region so that its left and right eye center positions line up with the left and right eye center positions of training data. The probability of the pairing constituting a true eye-pair is estimated. The average position of the 3 most probable eye-pairs' eye-center is considered as the precise position of the eye center of the face.

12.4 Random Forest Embedded ASM

With the prominent successful research of ASM and AAM, face shape is well modeled as a linear combination of principal modes (major eigenvectors) learned from the training face shapes. Here, we define shape as a series of coordinates of facial feature points. Facial landmark localization, thus, can be solved under the framework of ASM. Both the methods consist of three steps, shape modeling, distance measurement and global optimization. The facial landmark localization method is described as fitting the 2D face model to the novel face image. 2D face model is a deformable model based on random forest embedded key point recognition under the framework of ASM. We name our 2D face model as Random Forest Embedded Active Shape Model (RFE-ASM). The novelty is that this method embeds the discriminant learning into ASM. In our method, the 2D face model is represented by 88 landmarks; therefore, it can describe eyes, eyebrows, nose, mouth and cheek. Each landmark is accurately recognized by a fast classifier, which is trained from the appearance around this landmark. The proposed 2D face model embedding discriminant learning is illustrated in Fig. 12.5. Our facial landmark localization using RFE-ASM is presented in the following. Firstly, face shape is modeled and the fitting problem is defined as an optimization problem; then, distance between shape fitting results and the novel face image should be measured; finally, best fit should be optimized and facial landmark localization is performed by optimizing all the defined 88 facial feature points.

12.4.1 Shape Modeling

We define shape as a series of coordinates of facial feature points as:

$$X_i = [x_{i1},\ y_{i1},\ x_{i2},\ y_{i2} \cdots x_{ij},\ y_{ij} \cdots x_{i88},\ y_{i88}]^{\mathrm{T}}. \tag{12.3}$$

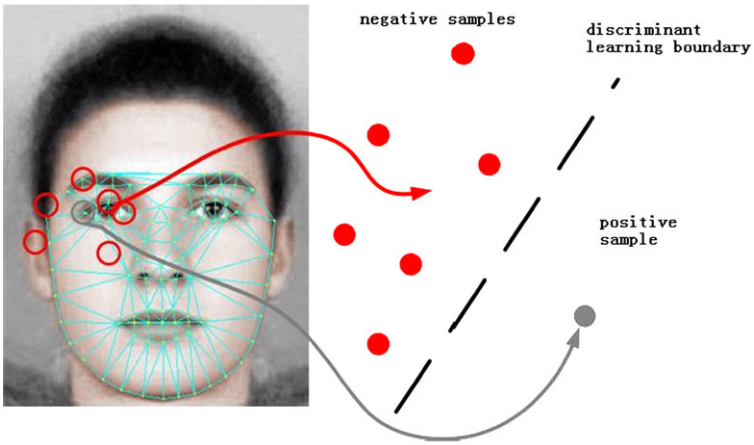

Fig. 12.5 2D face model embedding discriminant learning

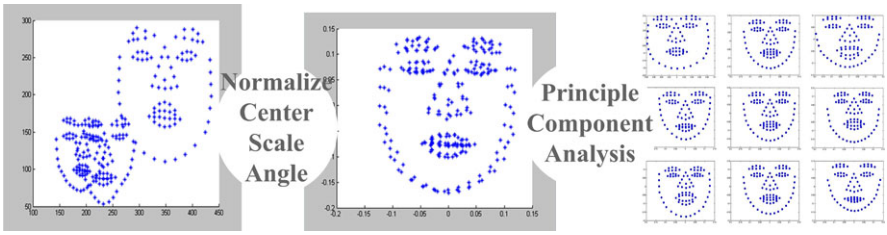

Fig. 12.6 2D shape modeling

It is the sequence of hand-labeled 88 points in the image lattice. We manually label 88 points for each face image in the training set. The manually labeled face images are used to train the face model. With the trained model, we can automatically locate 88 facial feature points of the face images which are unseen in the training set. 2D face shape is firstly normalized (center, scale, angle) and then well modeled as a linear combination of principal modes (major eigenvectors) learned from the training face shapes as illustrated in Fig. 12.6.

Principal component analysis (PCA) is used to represent the normalized shape 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 P with shape parameter vector b, and then transformed by center, scale and angle, presented by geometry parameter a as shown in:

$$X = T_a(\bar{X} + Pb), \tag{12.4}$$

$$a = (X_t, Y_t, s, \theta), \tag{12.5}$$

$$T_a \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} X_t \\ Y_t \end{pmatrix} + \begin{pmatrix} s\cos\theta & -s\sin\theta \\ s\sin\theta & s\cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}. \tag{12.6}$$

s is scale factor, θ is angle factor, and X_t, Y_t are horizontal and vertical shift variables. 2D face modeling builds 2D face deformable model based on a large training set. Such 2D face model needs two parameters (geometry parameter a and shape parameter b) to present a face shape. Facial landmark localization algorithm is thus defined as the method to find the best geometry parameter a and shape parameter b for a novel face image.

12.4.2 Distance Measurement

In conventional ASM, local image features 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 their nearby backgrounds. This study proposes to add key point recognition into ASM by embedding discriminant learning as illustrated in Fig. 12.5. Each landmark is accurately recognized by a fast classifier, which is trained from the appearance around this landmark. Several classification algorithms, such as SVM or neural networks, could have been chosen. Among those, Lepetit and Fua [22] have found random forest to be eminently suitable because it is robust and fast, while remaining reasonably easy to train. The proposed method is under the framework of ASM with embedded random forest learning, so called RFE-ASM.

Random forest classifier is trained to recognize each landmark. The samples are collected on a large training set. All the samples are cropped from faces (the distance between the center of the left eye and the center of the right eye is normalized into 60 pixels). Positive samples are the 32×32 image patches of all the training images with the center at the ground-truth landmark position. While negative samples are the 32×32 image patches of all the training images with the center inside the 40×40, but outside the 5×5 region from the ground-truth landmark position. As illustrated in Fig. 12.7, we take an example of the left mouth corner point. To find the left mouth corner point accurately, we train one random forest classifier for this landmark. All the samples are cropped from face images.

Random forest is a classifier combination method. A random forest consists of N binary trees. Each node of a binary tree is a weak classifier. The structure of random forest combines all the weak classifiers into a strong classifier. The output of random forest classifier is the voting of each binary tree. Figure 12.8 depicts a random forest. It consists of N decision trees. Each decision tree is trained by a 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 12.9 depicts a generic tree. Each node contains a simple comparison of the intensity in a pair of points that split the space of image patches to be classified. The training step aims to get an estimate

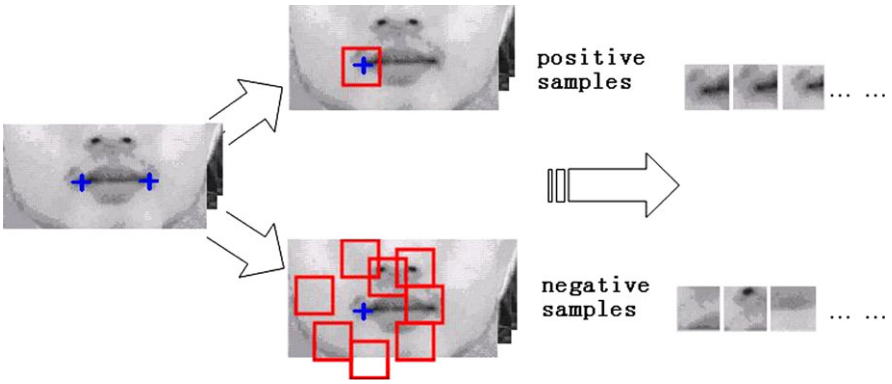

Fig. 12.7 To train one random forest for left mouth corner point, this figure shows an example of positive and negative sample collection

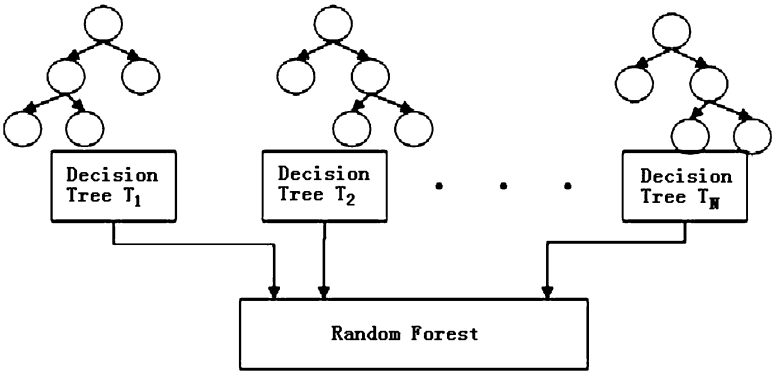

Fig. 12.8 Random forest combines the outputs of all decision trees as a classifier fusion method

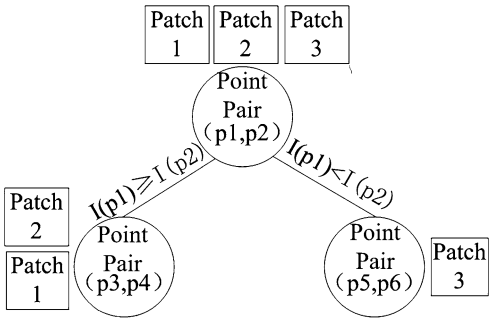

Fig. 12.9 A generic binary tree in random forest: each node contains a simple comparison of the intensity in a pair of points that split the space of image patches to be classified

based on training data of the posterior distribution over the classes in each leaf (the end node of a binary tree, which does not have child nodes).

In this training case, a random forest consists of multiple binary trees so that each tree yields a different partition of the space of image patches. Each node of a binary tree stores the best point pair, which is the weak classifier. The weak classifier is the comparison of intensity in a pair of points as in:

$$\mathrm{h} = \begin{cases} 1 & \text{if } I(p_1) \geq I(p_2), \\ 0 & \text{otherwise.} \end{cases} \tag{12.7}$$

Each node chooses the best point pair as the best weak classifier and the random forest combines the results of each weak classifier to a strong classifier as

$$\hat{F}(p) = \arg\max_{c} p_c(p) \tag{12.8}$$

$$= \arg\max_{c} (1/N) \sum_{n=1\ldots N} p_{n,p}\big(f(p) = c\big) \tag{12.9}$$

where N is the total number of binary trees and n specifies one binary tree; p is the image patch to be classified; c is the label of class such that when $c = 0$, the image patch does not belong to the landmark and when $c = 1$, the image patch belongs to the landmark; $p_{n,p}$ is the probability classified by the nth binary tree that the image patch p belongs to the landmark.

By dropping the image patch down the tree and performing a determined point pair comparison at each node, the image patch is sent to one side or the other (each node of a binary tree has two splits). When it reaches a leaf, it is assigned probabilities of belonging to a class depending on the distribution stored in the leaf. Responses of all the binary trees are combined during classification to achieve a better recognition rate than a single tree could. The distance of the novel face image and the 2D face model is thus measured by the output of random forest classifiers. The point that is more similar to the landmark will get a bigger random forest output probability.

12.4.3 Global Optimization

Facial landmark localization aims to find the best fit of the 88 points in the novel face image with the 2D face model. A new shape X could be obtained with geometry parameter a and shape parameter vector b. For each landmark, a random forest classifier gives the result measuring the distance of one point belonging to the landmark. 2D face shape consists of 88 facial feature points; therefore, all random forest results should be embedded into the global optimization. The global optimization objective function proposed is:

$$\big(\hat{a}, \hat{b}\big) = \arg\min_{a,b} \left(\big(Y - T_a(\bar{X} + Pb)\big)^{\mathrm{T}} W \big(Y - T_a(\bar{X} + Pb)\big) + k \sum_{j=1}^{t} b_j^2 / \sigma_j^2 \right) \tag{12.10}$$

where W is the output of random forest classifier and is embedded into the global optimization objective function to weigh 88 facial points. The shape parameter vector b is restricted to the vector space spanned by the training database.

The optimization includes the following steps:

1. Initialization: a is initialized according to face detection bounding box and two eyes positions. PCA shape parameter b is initialized to 0.
2. Finding new shape candidate: $Y \leftarrow \hat{Y}$ Random Forest output in the nearby location of the last Y.
3. $a \leftarrow \hat{a}$. $\hat{a} = \min_a((Y - T_a(\bar{X} + Pb))^{\mathrm{T}} W(Y - T_a(\bar{X} + Pb)))$.
4. $b \leftarrow \hat{b}$. $\hat{b} = \min_b((Y - T_a(\bar{X} + Pb))^{\mathrm{T}} W(Y - T_a(\bar{X} + Pb))) + k \cdot \sum_{j=1}^{t} b_j^2/\sigma_j^2$.
5. If $\|\hat{a} - a\| + \|\hat{b} - b\| < \epsilon$, stop. else, go to 2.

12.5 Experiments

12.5.1 Eye Localization

The training set is drawn from FERET, ARData, Bern, BioID, ORL, OCRFace database, and a total of 6800 eyes and 18 000 eye-pairs are cropped and normalized for training. The experimental test set consists of Yale (15 persons, 165 images), AeroInfo Face (165 persons, 3740 images), Police Face (30 persons, 448 images), and a total of 4353 faces are involved in the evaluation of localization performance and the influence of eye locations on face recognition. In the training databases, FERET, ARData, BioID, Bern, ORL are open databases, and OCRFace, built by our lab, consists of 1448 face images with different views, expressions and glasses. Among the test databases, Yale is an open database, which features extremely unbalanced lightening and thick glasses; Police Face, provided by the First Research Institute of Ministry of Public Security of China, features strong glaring of glasses and large pose variation; AeroInfo, provided by the Aerospace Information Co. Ltd. of China, features a large variety of illumination, expression, pose, face size, and complex background. The three test sets are from diverse sources to cover different eye variations in view angles, sizes, illumination, and glasses. Experiments based on such diverse sets should be able to test the generalization performance of our eye localization algorithm.

To evaluate the precision of eye localization, a scale independent localization criterion [29] is used. This relative error measure compares the automatic localization result with the manually marked locations of each eye. Let C_l and C_r be the manually extracted left and right eye positions, C_l' and C_r' be the detected positions, dl be the Euclidean distance between C_l' and C_l, d_r be the Euclidean distance between C_r' and C_r, d_{lr} be the Euclidean distance between the ground truth eye centers. Then the relative error of this detection is defined as follows:

$$\mathrm{err} = \frac{\max(d_l, d_r)}{d_{lr}}. \tag{12.11}$$

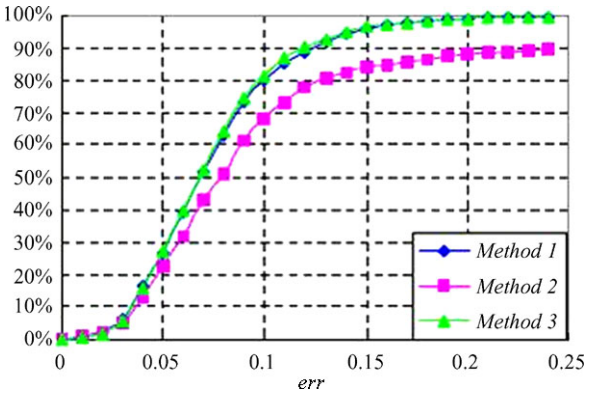

Fig. 12.10 Cumulative distribution of localization errors of three methods on test set

Three different eye localization methods are implemented and evaluated on the test set. Method 1: The proposed algorithm in this chapter. Method 2: Similar to the method proposed in [1]. After grossly locating the face area and determining the midline between left and right eye, connected components analysis and projection analysis are applied to the two areas separately to locate the eye center position. Method 3: Different from Method 1 only in that the step, subsampling eye candidates, is omitted.

The cumulative distribution function of localization error of three methods is shown in Fig. 12.10. From the figure, we can see that method 1 and method 3 achieve similar performance and about 99.1% of the test samples are with localization error below 0.20. Both are superior to method 2. But method 1 is 2–3 times faster than method 3. So the subsampling step does not degrade the location precision, but enhances the localization speed. The average processing time per face of method 1 on a PIV2.4 GHz PC system is 60 ms without special code optimization. In Fig. 12.11, we offer some examples out of the test sets for visual examination. The system appears to be robust to the presence of unbalanced illumination, eyeglasses, partial occlusion and even significant pose changes. This generalization ability is likely a consequence of the combination of local appearance and global appearance under probabilistic framework. Specially, in local appearance, the illumination influence can be effectively removed through local gray scale normalization; in global appearance, the influence caused by face rotation in image plane can be effectively removed through the aligning. We also compared method 1 with other newly published systems. In paper [35], the detection was considered to be correct if err < 0.25. Their detection performance on JAFFE database was 97.18%. We evaluate method 1 on JAFFE under the same test protocol. The detection rate of our method is 98.6% if err < 0.10, and the detection rate is 100% if err < 0.12.

12.5.2 Random Forest Embedded ASM

In order to verify our algorithm, experiments have been conducted on a large data set consisting of 3244 images from four databases for training. We collect and con-

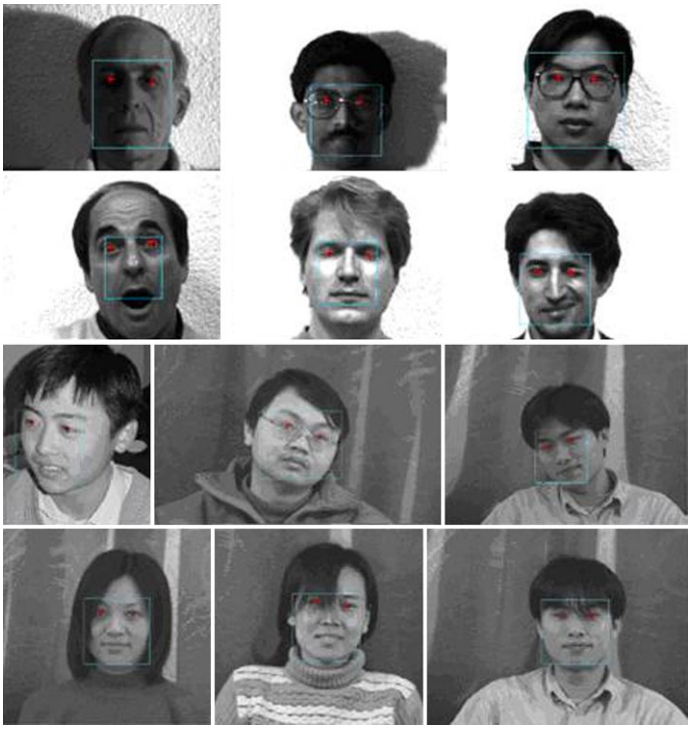

Fig. 12.11 Some eye localization results from test sets

struct the THFaceID database including 334 male and female aging from young to old with various facial expressions. The Yale database, FRGC database and JAFFE database are all publicly available. The Yale database includes illumination changes and facial expression changes; The FRGC database also includes facial expression and illumination changes under controlled and uncontrolled situations; The JAFFE database includes expression changes. All the 3244 images are manually labeled with 88 points as the ground truth landmarks. Test set 1 is constructed by the THFaceID database including 200 persons, totally 600 images. Test set 2 is IMM database. This method 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

$$e = \sum_{i=1}^{88} \| P_a - P_m \|_2 / (88 \cdot d_e). \tag{12.12}$$

We call it 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.

Table 12.1 Relative error of the algorithm

Number of random trees	Relative error on our database	Relative error on IMM database
1	6.79	7.28
5	3.84	4.41
10	3.76	4.28
30	3.76	4.23
50	3.77	4.27
80	3.82	4.30
100	3.82	4.27

Table 12.2 Speed of the algorithm

Number of random trees	Speed (ms)	Computer configuration
1	24	Intel Core2, 2.66 GHz, 3.25G RAM
5	30	
10	37	
30	69	
50	102	
80	150	
100	184	

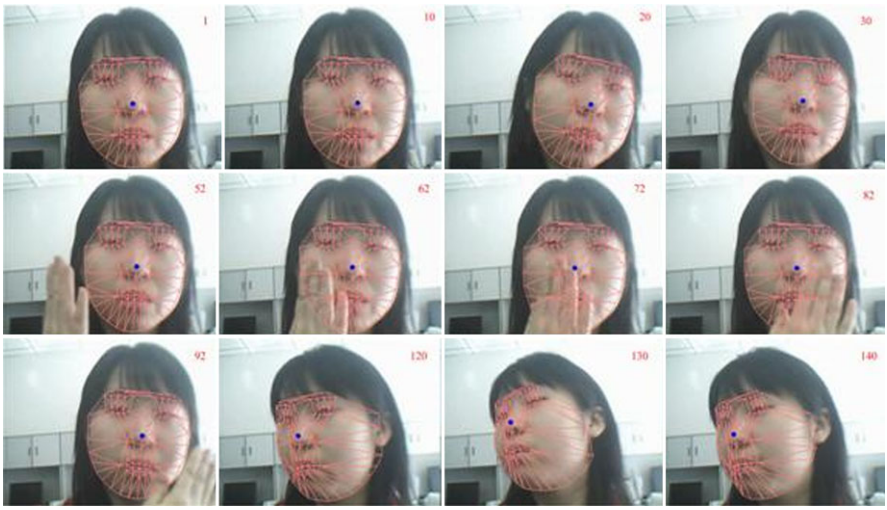

Fig. 12.12 Results of facial landmark localization towards video sequences

Table 12.1 shows the test results on Test set 1 and 2. Table 12.2 shows the speed of the algorithm. Figure 12.12 shows facial landmark localization algorithm towards video sequences.

Table 12.3 Face recognition results in different facial landmark localization results

Number of random trees	PCA dimension after reduction	Recognition rate (%)
10	500	63.7
20	500	64.9
30	600	65.1
40	500	64.3
50	600	63.7
60	500	63.8
80	500	64.0
100	600	63.6

In order to make sure what precision will make facial landmark localization meaningful in applications, the face recognition experiment on FRGC-V2 database is carried on. This experiment shows the relation of facial landmark localization's precision and face recognition rate. Gabor features are extracted at each facial landmark and put together as the whole feature vector. The training samples are 222 individuals from FRGC-V2 database, each individual has 10 images. The testing samples are 466 individuals from FRGC-V2 database, each individual has one image as face template. 466 individuals has totally 8014 images for testing. After feature extraction, PCA is used as the dimension reduction, LDA is used as the discriminant learning, and normalized correlation classifier is used. In addition, this experiment does not do illumination preprocessing. The face recognition results are listed in Table 12.3.

12.6 Conclusions

We have presented a facial landmark localization algorithm. Incorporating random forest classifier into ASM, this method works well when fitting to images which are unseen in the training set. Moreover, it runs in real time.

Face Recognition Vendor Test 2006 has shown that face recognition can achieve high accuracy under controlled conditions, for example, when the testing face samples are frontal. However, when face pose changes largely, the performance of existing methods drop drastically. The same difficulties are found in the literature of facial landmark localization. A reasonable way to improve multi-view facial landmark localization is to use 3D face morphable model. With the development of our further research, our studies will focus on fast and robust facial landmark localization algorithm by combining 2D deformable model with 3D morphable model.

Acknowledgements The author is indebted to the National Basic Research Program of China (973 program) under Grant No. 2007CB311004 for supporting this work, to Dr. Yong Ma for his works on face detection and eye localization, to Mr. Liu Ding who kindly helped do the face recognition experiment of this paper.

References

1. Baskan, S., Atalay, V.: Projection based method for segmentation of human face and its evaluation. Pattern Recognit. Lett. **23**, 1623–1629 (2002)
2. Brunet, N., Perez, F., de la Torre, F.: Learning good features for active shape models. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 206–211. IEEE Computer Society Press, Los Alamitos (2009)
3. Cootes, T.F., Taylor, C.J.: Active shape model search using local grey-level models: A quantitative evaluation. In: 4th British Machine Vision Conference, pp. 639–648. BMVA Press, Guildford (1993)
4. Cootes, T.F., Taylor, C.J.: Combining elastic and statistical models of appearance variation. In: Proceeding of 6th European Conference on Computer Vision, vol. 1, pp. 149–163. Springer, Berlin (2000)
5. Cootes, T.F., Taylor, C.J.: Constrained active appearance models. In: Proceeding of 8th International Conference on Computer Vision, vol. 1, pp. 748–754. IEEE Computer Society Press, Los Alamitos (2001)
6. Cootes, T.F., Taylor, C.J.: An algorithm for tuning an active appearance model to new data. In: Proceeding of British Machine Vision Conference, vol. 3, pp. 919–928. BMVA Press, Guildford (2006)
7. Cootes, T.F., Taylor, C.J., Cooper, D., Graham, J.: Training models of shape from sets of examples. In: 3rd British Machine Vision Conference, pp. 9–18. BMVA Press, Guildford (1992)
8. Cootes, T.F., Taylor, C.J., Lanitis, A.: Active shape models: Evaluation of a multi-resolution method for improving image search. In: 5th British Machine Vision Conference, pp. 327–336. BMVA Press, Guildford (1994)
9. Cootes, T.F., Taylor, C., Cooper, D., Graham, J.: Active shape models—their training and their applications. Comput. Vis. Image Underst. **61**(1), 38–59 (1995)
10. Cootes, T.F., Walker, K.N., Taylor, C.J.: View-based active appearance models. In: Proceeding of 4th International Conference on Automatic Face and Gesture Recognition, pp. 227–232. IEEE Computer Society Press, Los Alamitos (2000)
11. Cootes, T.F., Wheeler, G., Walker, K., Taylor, C.J.: Coupled-view active appearance models. In: 11th British Machine Vision Conference, pp. 52–61. BMVA Press, Guildford (2000)
12. Cootes, T.F., Edwards, G., Taylor, C.: Active appearance models. IEEE Trans. Pattern Anal. Mach. Intell. **23**(6), 681–685 (2001)
13. Cootes, T.F., Twining, C.J., Petrovic, V., Schestowitz, R., Taylor, C.J.: Groupwise construction of appearance models using piece-wise affine deformations. In: Proceeding of British Machine Vision Conference, vol. 2, pp. 879–888. BMVA Press, Guildford (2005)
14. Feng, G.C., Yuen, P.C.: Multi-cues eye detection on gray intensity image. Pattern Recognit. **34**, 1033–1046 (2001)
15. Gross, R., Matthews, I., Baker, S.: Generic vs. person specific active appearance models. Image Vis. Comput. **23**(1), 1080–1093 (2005)
16. Hamsici, O., Martinez, A.: Active appearance models with rotation invariant kernels. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1003–1009. IEEE Computer Society Press, Los Alamitos (2009)
17. Hess, M., Martinez, G.: Facial feature extraction based on the smallest univalue segment assimilating nucleus (susan) algorithm. In: Proceedings of Picture Coding Symposium, vol. 1, pp. 261–266. IEEE Computer Society Press, Los Alamitos (2004)
18. Hsu, R.L., Jain, A.K.: Generating discriminating cartoon faces using interacting snakes. IEEE Trans. Pattern Anal. Mach. Intell. **25**(11), 1388–1398 (2003)
19. Huang, C., Ai, H.Z., Li, Y., Lao, S.H.: High performance rotation invariant multiview face detection. IEEE Trans. Pattern Anal. Mach. Intell. **29**(4), 671–686 (2007)
20. Jiao, F., Li, S., Shum, H., Schuurmans, D.: Face alignment using statistical models and wavelet features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. I:321–327. IEEE Computer Society Press, Los Alamitos (2003)

21. Lee, H.-S., Kim, D.: Tensor-based aam with continuous variation estimation: Application to variation-robust face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **31**, 1102–1116 (2009)
22. Lepetit, V., Fua, P.: Keypoint recognition using randomized trees. IEEE Trans. Pattern Anal. Mach. Intell., 1465–1479 (2006)
23. Li, Y., Ito, W.: Shape parameter optimization for adaboosted active shape model. In: IEEE International Conference on Computer Vision, vol. 1, pp. 251–258 (2005)
24. Liebelt, J., Xiao, J., Yang, J.: Robust aam fitting by fusion of images and disparity data. In: Proceedings of IEEE Computer Vision and Pattern Recognition, vol. II, pp. 2483–2490. IEEE Computer Society Press, Los Alamitos (2006)
25. Liu, X.: Generic face alignment using boosted appearance model. In: Proceedings of IEEE Computer Vision and Pattern Recognition, p. 1079 (2007)
26. Liu, X.: Discriminative face alignment. IEEE Trans. Pattern Anal. Mach. Intell. **31**, 1941–1954 (2009)
27. Papandreou, G., Maragos, P.: Adaptive and constrained algorithms for inverse compositional active appearance model fitting. In: Proceedings of Computer Vision and Pattern Recognition, p. 1 (2008)
28. Sung, K.K., Poggio, T.: Example based learning for view-based human face detection. IEEE Trans. Pattern Anal. Mach. Intell. **20**, 39–51 (1995)
29. Tu, Z., Chen, X., Yuille, A.L., Zhu, S.-C.: Image parsing: Unifying segmentation, detection, and recognition. In: Proceeding of International Conference on Computer Vision, vol. 1, p. 18. IEEE Computer Society Press, Los Alamitos (2003)
30. Tu, J., Zhang, Z., Zeng, Z., Huang, T.: Face localization via hierarchical condensation with fisher boosting feature selection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. II, pp. 719–724. IEEE Computer Society Press, Los Alamitos (2004)
31. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple feature. In: Proceedings of Computer Vision and Pattern Recognition, vol. 1, pp. 511–518. IEEE Computer Society Press, Los Alamitos (2001)
32. Vogler, C., Li, Z., Kanaujia, A., Goldenstein, S., Metaxas, D.: The best of both worlds: Combining 3d deformable models with active shape models. In: IEEE International Conference on Computer Vision, pp. 1–7 (2007)
33. Xiao, J., Baker, S., Matthews, I., Kanade, T.: Real-time combined 2d+3d active appearance models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 535–542. IEEE Computer Society Press, Los Alamitos (2004)
34. Zhang, L., Ai, H.: Multi-view active shape model with robust parameter estimation. In: International Conference on Pattern Recognition, vol. 4, pp. 469–468 (2006)
35. Zhou, Z.H., Geng, X.: Projection functions for eye detection. Pattern Recognit. **37**(5), 1049–1056 (2004)