Real-Time Facial Feature Point Extraction

Ce Zhan, Wanqing Li, Philip Ogunbona, and Farzad Safaei

University of Wollongong, Wollongong, NSW 2522, Australia
{cz847,wanqing,philipo,farzad}@uow.edu.au

Abstract. Localization of facial feature points is an important step for many subsequent facial image analysis tasks. In this paper, we proposed a new coarse-to-fine method for extracting 20 facial feature points from image sequences. In particular, the Viola-Jones face detection method is extended to detect small-scale facial components with wide shape variations, and linear Kalman filters are used to smoothly track the feature points by handling detection errors and head rotations. The proposed method achieved higher than 90% detection rate when tested on the BioID face database and the FG-NET facial expression database. Moreover, our method shows robust performance against the variation of face resolutions and facial expressions.

1 Introduction

Localization of facial feature points is often a critical step in many multimedia applications including face recognition, face tracking, facial expression recognition, gaze detection and face animation [1,2,3,4]. The fiducial facial feature points needed to be detected are usually the salient points on the face, such as eye corners, mouth corners, eyebrow corners and nostril corners. Various approaches have been proposed in the literature to extract these facial points from images or video sequences of faces. In general, these approaches can be categorized either as appearance-based or geometric-based. Appearance-based methods use feature vectors to model local texture around the facial feature points. To obtain the feature vectors, several methods such as Gabor wavelets [5], principal components analysis (PCA) [6] and Gaussian derivative filters [7] are often used to transform the local textures. The transformed features are then selected and processed by machine learning techniques such as multi-layer perceptrons (MLP) [8] and support vector machines (SVMs) [9]. In geometric-based methods, the prior knowledge of the face structure is used to constrain the facial feature point search, and the search is based on certain rules which can be learned from a set of labeled faces and often involve distance and angles [10,11].

It is interesting to note that most of the existing methods attempt to locate facial feature points from images/video captured in a highly controlled laboratory environment and with high spatial resolution. Furthermore, the face regions are always larger than 160 × 160 pixels. This resolution is equivalent to a person sitting only 30cm away from a webcam with focal length of 3cm and image capture resolution at 320 × 240 (see Table 1). It is obvious that many applications require

H.H.S. Ip et al. (Eds.): PCM 2007, LNCS 4810, pp. 88–97, 2007.

a much wider range of working distances than this, especially when wireless input devices and lower resolution cameras are used. In addition, most existing methods extract facial landmarks from expressionless face images which are unsuitable for facial expression recognition, in particular, when the recognition is based on local features. Finally, computational cost of the methods involving multiple classifications cannot be afforded by most real-time applications with limited computing resources.

To overcome the above mentioned problems, this paper proposes a new method for extracting 20 facial feature points. The method employs a number of modifications to the conventional Viola-Jones AdaBoost detection method and is relatively insensitive to the resolution of face images. It performs well within a practical range of working distances. With the goal of reducing the computational load, a coarse-to-fine strategy is adopted.

The rest of this paper is organized as follows: Section 2 describes the proposed method. Section 3 presents the experimental results obtained on a number of face databases. Conclusions are given in Section 4.

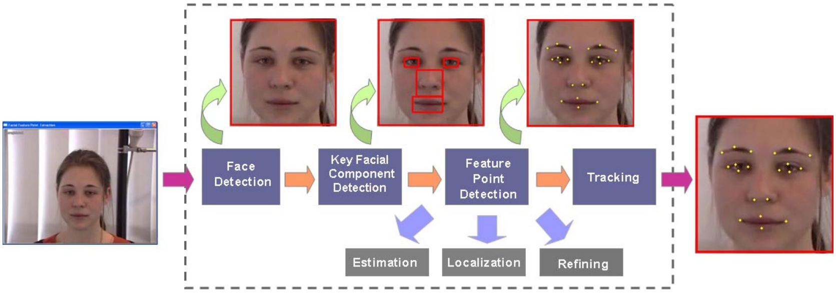


Fig. 1. The coarse-to-fine facial feature point extraction process

2 The Proposed Method

The proposed coarse-to-fine method consists of four stages, as shown in Figure 1: Face Detection, Key Facial Component Detection, Feature Point Detection and Feature Point Tracking.

2.1 Face Detection

The face region is detected and localized by the conventional Viola-Jones AdaBoost method. For details on the method, readers are referred to [12,13].

2.2 Key Facial Component Detection

The second stage of the coarse-to-fine process is to locate key facial components (nose, mouth and eyes) within the detected face area. To take advantage of the low computation overhead associated with Haar-like features and highly efficient cascade structure used in Viola-Jones AdaBoost face detection method, "AdaBoost" detection principle is adopted. However, low detection rate was observed when the conventional Viola-Jones method was trained with the facial components and employed in the detection process. This is probably due to the lack of significant structure information of the facial components (compared to the entire face). In general, the structure of the facial components become less detectable when the detected face is at low resolution. Table 1 shows approximate size of facial components at different distances for a webcam with focal length of 3cm and resolution of 320 × 240. Another cause of the low detection rate is probably the substantial variations in the shape of the components, especially mouth, among the different expressions conveyed by the same or different people. This is also true for high resolution face images. To solve these problems, we improve the "AdaBoost" detection method by employing: extended Haar-like features, modified training criteria, regional scanning and probabilistic selection of candidate sub-window.

Table 1. The approximate relationship between distance of user to camera and facial component resolution

	30cm	50cm	80cm	130cm
Face	165 × 165	100 × 100	65 × 65	45 × 45
Mouth	56 × 28	30 × 15	20 × 10	12 × 6
Eyes	36 × 18	22 × 11	14 × 7	8 × 4
Nose	42 × 42	26 × 26	16 × 16	9 × 9

Extended Haar-like Feature Set. An extended feature set with 14 Haar-like features (Figure 2) based on [14] is used in the facial component detection. Besides the basic upright rectangle features employed in face detection, 45° rotated rectangle features and center-surround features are added to the feature pool. The additional features are more representative for different shapes than the original Haar-feature set, and would therefore improve the detection performance.

High Hit Rate Cascade Training. In the conventional Viola-Jones method, the cascade classifier is trained based on the desirable hit rate and false positive rate. Additional stage is added to the cascade classifier if the false positive is higher. However, when the false positive rate decreases, the hit rate also decreases. In the case of facial components detection, hit rate will dramatically fall

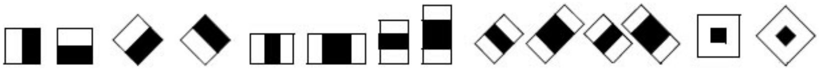


Fig. 2. The extended Haar-like feature set

for low resolution face images if the cascade classifier is training with low false positive rate.

To ensure that low resolution facial components could be detected, a minimum overall hit rate is set before training. For each stage in the training, the training goal is set to achieve a high hit rate and an acceptable false positive rate. The number of features used is then increased until the target hit rate and false positive rate are met for the stage. If the overall hit rate is still greater than the minimum value, another stage is added to the cascade to reduce the overall false positive rate. In this way, the trained detectors will detect the facial components at a guaranteed hit rate though some false positives will occur, which can be reduced or removed by the modifications introduced below.

Regional Scanning With a Fixed Classifier. Rather than rescaling the classifier as proposed by Viola and Jones, to achieve multiscale searching, input face images are resized to a range of predicted sizes and a fixed classifier is used for facial component detection. Due to the structure of face, we predict the face size according to the size of facial component used for training. In this way, the computation of the whole image pyramid is avoided. If the facial component size is bigger than the training size, fewer false positives would be produced due to down sampling; when the component is smaller than the training sample, the input image is scaled up to match the training size.

In addition, prior knowledge of the face structure is used to partition the region of scanning. The top region of the face image is used for eye detection; the central region of the face area is used for nose detection; and mouth is searched in the lower region of the face. The regional scanning not only reduces the false positives, but also lowers the computation.

Candidate sub-window selection. To select the true sub-window which contain the facial component, it is assumed that the central position of the facial components among different persons follows a normal distribution. Thus, the probability that a candidate component at $\mathbf{k} = \left[x\ y\right]^T$ is the true position can be calculated as:

$$P\left(\mathbf{k}\right) = \frac{1}{\left(2\pi\right)\left|s\mathbf{\Sigma}\right|^{1/2}} \exp\left(-\frac{1}{2}\left(\mathbf{k} - s\mathbf{m}\right)^T s\mathbf{\Sigma}^{-1}\left(\mathbf{k} - s\mathbf{m}\right)\right)$$

where the mean vector $\mathbf{m}$ and the covariance matrix $\mathbf{\Sigma}$ is estimated from normalized face image data set. The scale coefficient, s, can be computed as $s = w_d/w_n$, where w_d is the width of detected face and w_n is the width of normalized training faces. The candidate with maximum probability is selected as the true component.

Specialized classifiers. Two cascade classifiers are trained for mouth. One is for detecting closed mouths, and the other is for open mouths. During scanning, if the closed mouth detector failed to find a mouth, the open mouth detector is triggered. In addition, the left and right eye classifiers are trained separately.

2.3 Facial Feature Point Detection

The facial feature point detection process entails estimation, localization and refinement. First, positions of the 20 facial feature points are approximated based on the boundary box of the detected facial components, as shown in Figure 3. It is assumed that the actual landmark is localized within a $D \times D$ neighborhood of the approximated landmark, where D is determined by the size of facial components. For instance, 4 neighbourhoods of the mouth landmarks are indicated in Figure 3.

The localization is achieved by finding the position within the neighbourhood that maximizes eigenvalues of local structure matrix C,

$$C = w_G\left(r;\sigma\right) * \begin{bmatrix} f_x^2 & f_x f_y \\ f_x f_y & f_y^2 \end{bmatrix}$$

where $w_G\left(r;\sigma\right)$ is the Gaussian filter for smoothing the matrix entries and $f(x,y)$ is the intensity function. The classic Harris [15] corner finder is applied to refine the detected landmark positions so as to achieve sub-pixel accuracy.

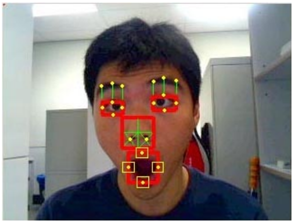


Fig. 3. Facial feature points estimation

2.4 Facial Feature Point Tracking

Occasionally, key facial components may not be reliably detected due to head rotations. There are also cases where the true facial feature points are not located in the $D \times D$ neighborhood of the estimated landmarks. With the goal of obtaining more accurate and smooth feature point positions, linear Kalman filters are employed to track the detected landmarks. In the linear Kalman filter the state vector consists of position, velocity and acceleration.

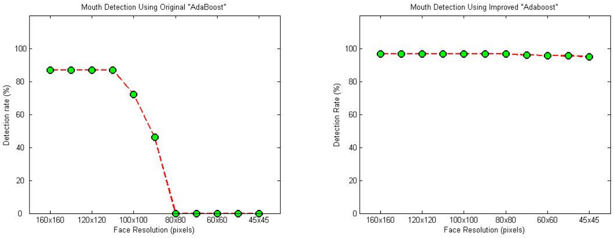
The Kalman filter predicts facial landmark positions in the next frame and corrects the localization results in the current frame. The prediction makes the feature points extraction process more stable when previous processing stages failed or some error occurred. At the same time, the correction enhances the accuracy.

3 Experimental Results

3.1 Facial Component Detection

As introduced in Section 2.2, five cascade classifiers were trained to detect the key facial components, one each, for left eye, right eye and nose, and two for mouth.

Positive training samples of eyes, mouths, noses and negative samples (non-facial components) were cropped from AR database [16] and AT&T database [17]. To accommodate low resolution facial components, the training samples were rescaled to small sizes: 10×6 for eyes, 16×8 for mouth and 15×15 for nose. For each detector about 1200 positive samples and 5000 negative samples were used for training. The trained detectors were first tested on BioID database [18]. The BioID database consists of 1521 images of frontal faces captured in uncontrolled conditions using a web camera in an office environment. For each image in the database, the ground truth of 20 facial feature points were obtained through manual annotation and supplied with the database. To evaluate the performance on low resolution input, the test images were downsized to different resolutions to simulate low resolution faces which are not included in the database. In this way, 300 images were tested at each face resolution. In the testing phase, a detection was regarded as SUCCESS if and only if the distance between the center of a detected and actual facial component was less than 30% of the width of the actual facial component as well as the width of the detected facial component was within ±50% of the actual width. To show the improvement obtained in comparison with the original detection method proposed by Viola and Jones, mouth detection results at different face resolutions are presented in Figure 4. The average detection rate for nose, left eye, right eye and mouth at different face resolutions is 91.3%, 95.7%, 97.2% and 95.6% respectively. A few detection examples are shown in Figure 5. The detectors were also tested on facial expression database FG-NET [19]. The database contains 399 video sequences of 6 basic emotions and a neutral expression from 18 individuals. The overall detection rate for all the detectors is 93.8%. Figure 6 shows typical detection examples of FG-NET. Figure 7 shows snap shots from a real-time example.

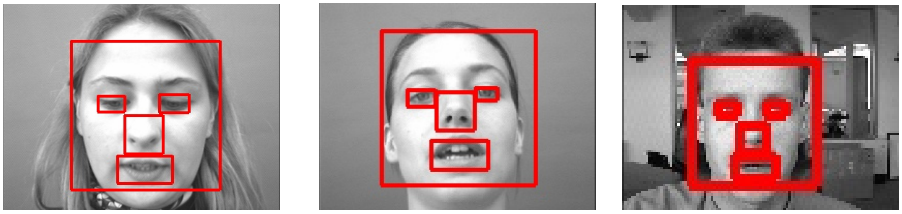


a. Original “AdaBoost” b. Improved “AdaBoost”

Fig. 4. Mouth detection result. Both detectors are trained using same dataset.

3.2 Facial Feature Point Extraction

The feature point extraction method were also tested on BioID database. In the testing phase, images from the same individual were reordered and treated as an image sequence. A detection is regarded as SUCCESS when the distance between the located facial point and the annotated true point was less than 10% of the inter-ocular distance (distance between left and right eye pupils). Unfortunately,



a. Face resolution: 160 × 160 b. Face resolution: 100 × 100 c. Face resolution: 50 × 50

Fig. 5. Facial component detection results from BioID database

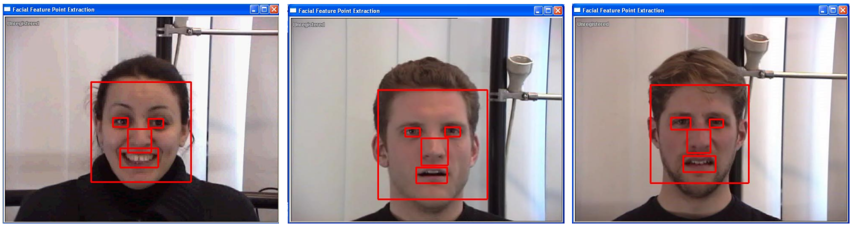

Facial Feature Point Extraction
Unregistered
Facial Feature Point Extraction
Unregistered
Facial Feature Point Extraction
Unregistered

Fig. 6. Facial component detection results from FG-NET database

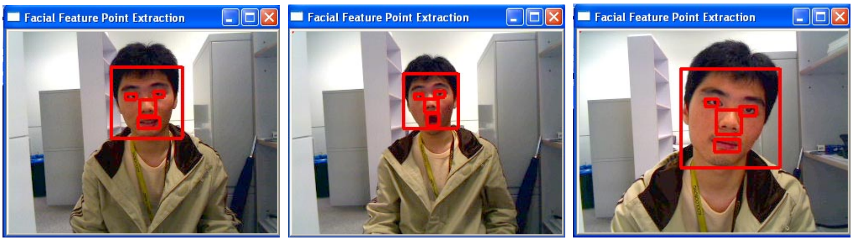

Facial Feature Point Extraction
Facial Feature Point Extraction
Facial Feature Point Extraction

Fig. 7. Real-time facial component detection results

only 14 of the facial points we detected are annotated in the BioID database. The testing result is presented in Table 2; the average detection rate for all of the 14 points is 93%. The same method was used as Section 3.1, to test the feature extraction approach on different resolution faces, the results is shown in Figure 8, and test examples are presented in Figure 9. When testing the proposed method on FG-NET database, each of the 20 automatically detected facial landmarks was compared to manually labeled facial point. The average detection rate for all of the points is 91%, and some examples are shown in Figure 10. During the real-time test, the proposed facial feature points extraction method exhibited robust performances against variations in face resolutions and facial expressions. The tracking module also enabled the proposed method to handle some degree of in-plane and out-of-plane rotations. Figure 11 are a few test examples.

Table 2. Facial feature point extraction results based on BioID database

Feature Point	Rate	Feature Point	Rate
1: Right mouth corner	96%	2: Left mouth corner	91%
3: Outer end of right eye brow	92%	4: Inner end of right eye brow	94%
5: Inner end of left eye brow	97%	6: Outer end of left eye brow	91%
7: Outer corner of right eye	90%	8: Inner corner of right eye	96%
9: Inner corner of left eye	97%	10: Outer corner of left eye	88%
11: Right nostril	95%	12: Left nostril	94%
13: Center point on outer edge of upper lip	87%	14: Center point on outer edge of lower lip	85%

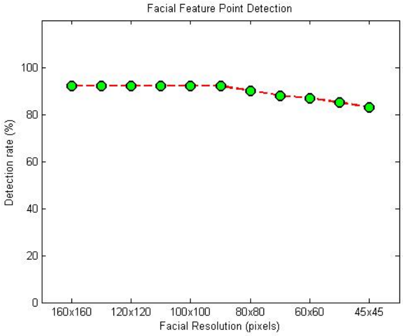
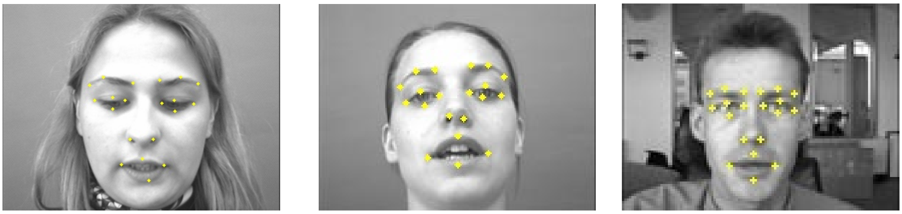


Fig. 8. Average feature point detection rates for different face resolutions



a. Face resolution: 160 × 160 b. Face resolution: 100 × 100 c. Face resolution: 50 × 50

Fig. 9. Facial feature point detection results from BioID database

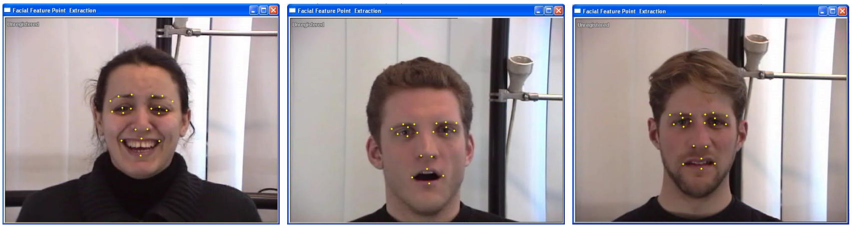

Facial Feature Point Extraction
Unregistered
Facial Feature Point Extraction
Unregistered
Facial Feature Point Extraction
Unregistered

Fig. 10. Feature point extraction results from FG-NET database

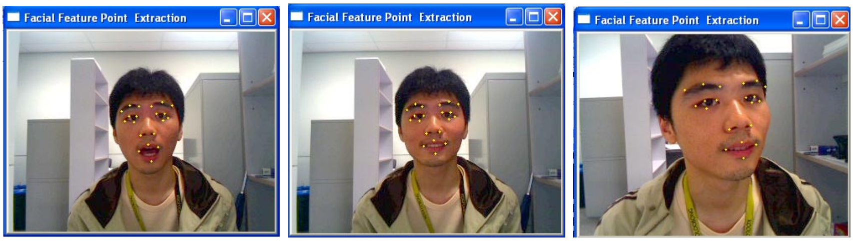

Facial Feature Point Extraction
Facial Feature Point Extraction
Facial Feature Point Extraction

Fig. 11. Real-time facial point extraction results

4 Conclusions

Various approaches have been proposed in the past to extract facial feature points from face images or video sequences. Most of the previous methods attempt to locate facial feature points from data collected in a highly controlled laboratory environment and characterized by high resolution and expressionless faces. This paper proposes a new method, based on previous research, for extracting 20 facial feature points from expressional face images at in varying resolutions. The proposed method can handle a certain degree of head rotations and achieved higher than 90% detection rate when tested on BioID face database and FG-NET facial expression databases.

References

1. Pantic, M., Rothkrantz, L.: Expert system for automatic analysis of facial expression. Image and Vision Computing Journal 18, 881–905 (2000)
2. Wiskott, L., Fellous, J.M., Krüger, N., von der Malsburg, C.: Face recognition by elastic bunch graph matching. In: Jain, L.C., Halici, U., Hayashi, I., Lee, S.B. (eds.) Intelligent Biometric Techniques in Fingerprint and Face Recognition, pp. 355–396. CRC Press, Boca Raton (1999)
3. Dailey, M.N., Cottrell, G.W.: PCA = gabor for expression recognition. Technical Report CS1999-0629 (1999)
4. Turk, M., Pentland, A.: Eigenfaces for recognition. Journal of Cognitive Neuroscience 3, 71–76 (1991)
5. Shih, F.Y., Chuang, C.F.: Automatic extraction of head and face boundaries and facial features. Information Sciences 158, 117–130 (2004)
6. Ryu, Y.S., Oh, S.Y.: Automatic extraction of eye and mouth fields from a face image using eigenfeatures and ensemble networks. Applied Intelligence 17, 171–185 (2002)
7. Arca, S., Campadelli, P., Lanzarotti, R.: A face recognition system based on automatically determined facial fiducial points. Pattern Recognition 39, 432–443 (2006)
8. Campadelli, P., Lanzarotti, R.: Localization of facial features and fiducial points. In: Proceedings of the International Conference on Visualisation, Imaging and image Processing, pp. 491–495 (2002)
9. Liao, C.T., Wu, Y.K., Lai, S.H.: Locating facial feature points using support vector machines. In: Proceedings of the 9th International Workshop on Cellular Neural Networks and Their Applications, pp. 296–299 (2005)

10. Zobel, M., Gebhard, A., Paulus, D., Denzler, J., Niemann, H.: Robust facial feature localization by coupled features. In: Proceedings of the 4th International Conference on Automatic Face and Gesture Recognition, pp. 2–7 (2000)
11. Yan, S., Hou, X., Li, S.Z., Zhang, H., Cheng, Q.: Face alignment using view-based direct appearance models. International Journal of Imaging Systems and Technology 13, 106–112 (2003)
12. Viola, P., Jones, M.: Robust real-time object detection. International Journal of Computer Vision (2002)
13. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of online learning and an application to boosting. In: Vitányi, P.M.B. (ed.) EuroCOLT 1995. LNCS, vol. 904, pp. 23–37. Springer, Heidelberg (1995)
14. Lienhart, R., Maydt, J.: An extended set of haar-like features for rapid object detection. In: Proceedings of the International Conference on Image Processing, vol. 1, pp. I–900–I–903 (2002)
15. Harris, C., Stephens, M.: A combined corner and edge detector. In: Alvey Vision Conference, pp. 147–151 (1998)
16. `http://cobweb.ecn.purdue.edu/aleix/aleix_face_DB.html`
17. `http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html`
18. `http://www.bioid.com/`
19. `http://www.mmk.ei.tum.de/waf/fgnet/feedtum.html`