Automatic Detecting Neutral Face for Face Authentication and Facial Expression Analysis

Ying-li Tian and Ruud M. Bolle
Exploratory Computer Vision Group
IBM Thomas J. Watson Research Center
PO Box 704, Yorktown Heights, NY 10598
{yltian, bolle}@us.ibm.com

Abstract

Automatic detecting neutral faces will improve the accuracy of face recognition/authentication, speaker ID analysis, and make it is possible to do facial expression analysis automatically. This paper describes an automatic system to find neutral faces in images by using location and shape features. Using these features, a window is placed in the detected and normalized face region. In that fashion, zones in this window correspond to facial regions that are loosely invariant from subject to subject. Within these zones, shape features in the form of histograms and ellipses are extracted. These features, in addition to more global distance measures, are input to a classifier to arrive at a neutral/non-neutral decision. The system has achieved an average detection rate of 97.2% based on 536 test images of 24 subjects. Tests on an independent image database of 64 images of 13 subjects achieved success rate of 95.3%, giving some indication of the robustness of the system.

1 Introduction

A neutral face is a relaxed face without contraction of facial muscles and without facial movements. It is the state of a person's face most of the time, i.e., it is the facial appearance without any dramatic expression. In contrast, for a face with an expression, the facial muscles are somehow contracted. Hence, facial expressions are deformations of the neutral face due to a person's psychological state. When speaking, a person's expression is deformed from the neutral face because of the movement of the mouth and other muscle motions induced by the dramatic content of the speech.

Automatic detecting neutral face has various benefits:

- Face recognition systems can achieve high recognition rate for good quality, frontal view, constant lighting and only subtle expression or expressionless face images. The performance of face recognition system significantly decreases when there is a dramatic expression on the face [15] (and, of course, for side views and bad-lighting images). Therefore, it is important to automatically find the best face of a subject from the images. Using the neutral face during enrollment and when authenticating, speak ID analysis, video indexing will increase the accuracy than choosing arbitrary face images of the subject.

  Yacoob *et. al*[15] evaluate the sensitivity of face recognition with respect of facial expression and minor head motions using two approaches: the eigenface approach[13] and the feature-graph approach[7]. They collected the images based on six universal expressions. These universal expressions are happiness, sadness, disgust, anger, surprise and fear (see Figure 1) and these expressions do *not* change too much from culture to culture[4]. They separate the statistical results into segments with expressions and segments with neutral expressions. In both approaches it is observed that the performance on expression segments is worse than the neutral segments.

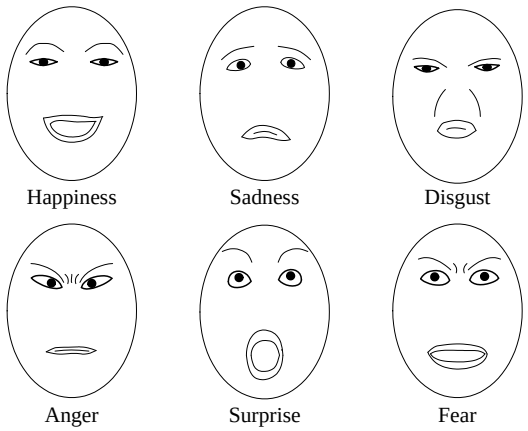


Figure 1: The six universal expressions.

- In the last decade, much progress has been made toward

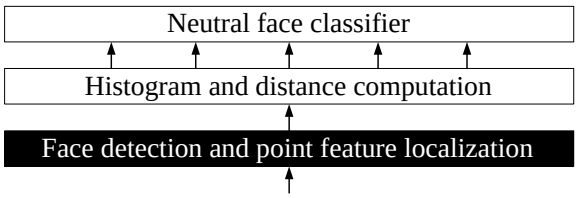


Figure 2: The stages of neutral face detection.

building computer systems to understand and use communication through facial expressions. However, that *most* of the existing expression analysis systems focus on individual expression recognition by comparing an unknown expression with a hand-labeled neutral face (e.g., [3, 8, 12]). The generic neutral or expressionless face appearance for each subject is provided by a human operator and expressions are classified based on how they differ from the neutral expression. The ability to automatically detect the neutral expression, enables facial expression analysis systems to operate without human intervention.

Both Tian *et. al*[12] and Donato *et. al*[3] are recognized facial expression based on Action Units. Ekman and Friesen [5] developed the Facial Action Coding System (FACS) for describing facial expressions by action units (AUs). Of 44 FACS AUs that they defined, 30 AUs are anatomically related to the contractions of specific facial muscles: 12 are for upper face, and 18 are for lower face. AUs can occur either singly or in combination. When AUs occur in combination they may be *additive*, in which the combination does not change the appearance of the constituent AUs, or *non-additive*, in which the appearance of the constituents does change. Although the number of atomic action units is relatively small, more than 7,000 different AU combinations have been observed [10]. Compare to the six universal expressions, FACS provides descriptive power necessary to describe the details of facial expression. For example, When AU 1 occurs, it raises the inner brow. When AU 4 occurs, the brows are drawn together and lowered.

- The error rate of speech recognition systems can be lowered by visually detecting if a person is speaking or not, especially in environments with high ambient noise. In addition to detecting that a person's lips are moving or not, detecting that the face is neutral can give valuable cues that the person is *not* speaking. If it is determined that the face is not speaking, the speech recognition engine is not employed on the acquired sound, the ambient noise.

Automatically detecting if a face is neutral can be achieved by determining that none, or as few as possible, of the facial features associated with the six universal expressions of Figure 1 are present in the face image. These features are primarily the shape and location of the mouth, the eyes, and the eyebrows. Examining these features is precisely the approach of our neutral face detection system shown in Figure 2. This system has an image as input. The image may contain one or more faces.

First, a face is detected in the image and facial features are extracted (the black box in Figure 2). Using the detected face features, the system normalizes a carefully selected window within the face to an image of predetermined resolution and size. This image is subsequently divided up into 3×3 zones from which expression features are computed. The original window of face image data is chosen in such a fashion that the upper-left zone and the upper-right zone contain the left and right eye and brow, respectively, and the bottom row of zones contain the mouth. New face expression features are then determined by computing shape histograms of the edges in the three bottom zones and by fitting ellipses to each eye in the top-left and top-right zones. Additionally, normalized distances between face features are used as relative location features. These are the distances between the line connecting the pupils and the corners of the mouth, the distances between this line and the inner eyebrows, and the distance between the two corners of the mouth. Both relative location and shape features are input to a classifier, which is trained with face images labeled as *neutral* face and *non-neutral* face. In this paper, the classifier is a neural network.

This paper is organized as follows. Section 2 describes the face and facial feature detection system. Section 3 introduces the location features (normalized distances) and new facial shape features for neutral face detection. Section 4 describes the neutral face classifier. Section 5 contains experiment results. Finally, in Section 6, we give some conclusions.

2 Face and feature detection

As a front-end to the neutral face detection system, the method proposed and implemented by Senior [11] is used to find face and facial features.

2.1 Face detection

Senior [11] finds the face using a series of detectors. A first detector is a skin-tone filter that determines if the pixels of the candidate region have coloring that is consistent with skin color. Each pixel is independently classified as skin tone or not, and the candidate region is rejected if it contains an insufficient proportion of skin-tone pixels (typically 60-80%). Subsequently a linear discriminant trained on face and non-face exemplars is applied to the gray-level pixels. The linear discriminant removes a significant proportion of

Figure 3: Face detection and facial feature extraction.

the non-face images in a computationally efficient fashion. Next a Distance From Face Space [13] measure is used to further filter the exemplars and finally a combination of the two scores is used to rank overlapping candidates that have been retained, with only the highest scoring candidate being selected. For gray-scale images, the skin tone detector is omitted.

2.2 Facial point feature detection

After a face is found in the image, both the Fisher discriminant and Distance From Feature Space measurements are used to locate point features, i.e., specific landmarks, in the face. Feature templates are generated by resampling all training images at a predetermined scale, location and orientation. Here the resampling parameters are determined by the scale and orientation of the detected face in the training images. In the test image, all possible feature locations are searched and points with the highest likelihood are retained. Feature collocation statistics are subsequently used to remove spurious features. A total of 19 facial features are extracted [11].

In our neutral face detection system, the six facial point features that can be most reliably extracted are used. These point features are: pupil centers (2), eyebrow inner endpoints (2), and corners of the mouth (2). The appearances of the surrounding images of these points are least affected by image noise and distortions caused by hair obscuring the face. An example of detected face and facial features is shown in Figure 3

3 Location and shape features extraction

First the facial image is normalized to a canonical face size based on two of the facial point features, i.e., the eye-separation. In our system, all faces are normalized to 128 $\times$ 128 pixels by re-sampling. To detect if a face is a *neutral* face, both the relative location of the above six points and shape features in facial zones determined by these points are extracted. As far as possible, location features are transformed into distances between the detected facial point features. These distances are relative measures, indicating relative locations, with respect to the canonical face size.

The openness of the eyes distinctly contributes to the differences in appearance of the universal expressions. However, the openness, the distance between upper- and lower-eyelids of the eyes cannot be obtained directly from the facial point features. We use an ellipse fitting method in eye regions based on a binary image of the eye. Another distinguishing feature is the shape of the mouth. Global shape features are not adequate to describe the shape of the mouth. Therefore, in order to extract the mouth shape features, an edge detector is applied to the normalized face to get an edge map. This edge map is divided into 3×3 zones. The mouth shape features are computed from zonal shape histograms of the edges in the mouth region.

3.1 Relative location representation

Previously, we [12] used the normalized geometric features, such as distances between points in normalized face images, for facial expression recognition. Good recognition results were obtained. Similarly, in order to detect neutral faces, normalized distances between face features are used as discriminating features. These distances, $L_1, L_2, ..., L_7$, are shown in Figure 4. Four of these distances are determined with respect to the line connecting the pupils, that is, the distances between the line and the corners of the mouth and the distances between the line and the inner eyebrows. The other distances are the width of the mouth (the distance between two corners of the mouth) and the distances between upper- and lower-eyelids.

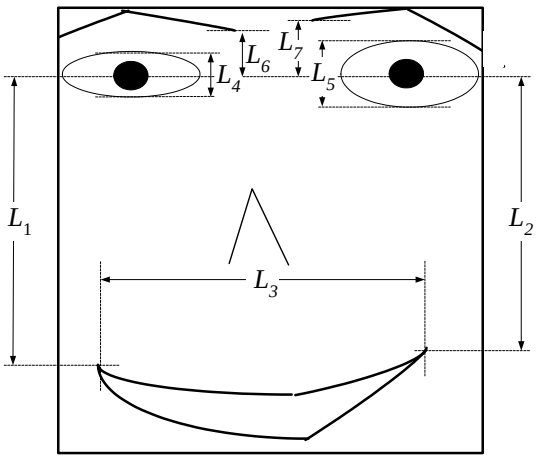


Figure 4: Distance features in the neutral face detection system.

Location features of mouth and eyebrows

After facial feature detection, the distances between the line connecting the pupils and the corners of the mouth (L_1 and L_2), the distances between this line and the inner eyebrows (L_6 and L_7), the distance between two corners of the mouth (L_3) can be easily calculated based on the facial point features. The measurements of the opening of the eyes, i.e., the distance between the upper- and lower-eyelids cannot be calculated directly. Previous work shows that it is difficult to detect or track eyelid positions accurately and robustly [2, 14, 16]. In this paper, we propose a new method to extract the measurements of the opening of the eyes by ellipse fitting based on the binary image by adjusting the threshold of the face image.

Location features of eyes

First, a binary face image of the face is obtained by using a threshold to the brightness. This threshold was obtained by experiment training. As shown in Figure5b, the binary image of both the closed eye and the open eye are pretty clear. The eye regions in the binary image are determined by the facial point features. An ellipse-specific direct least-square fitting method[9] is applied to the eye regions in the binary image. The ellipse fitting method estimates the height of the eye very well, on the other hand, it may overestimate the width of the eye. The fitting results for an open and closed eye are shown in Figure5c. In our system, the distances between the upper- and lower-eyelids are measured as the minor axis of the ellipse if the ratio of the minor and major axes is larger than a threshold. Otherwise, the eye is marked as closed and the distance is set to zero:

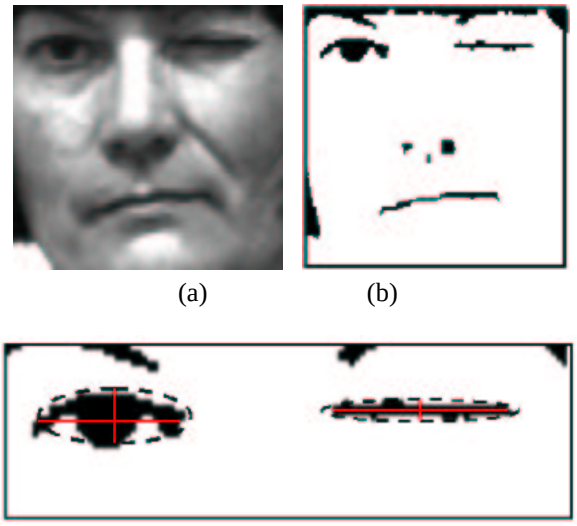
$$Eyelid\ distance = \begin{cases} minor\ axis\ of\ ellipse & \text{if } \frac{minor\ axis}{major\ axis} \geq 0.15 \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

3.2 Shape features of mouth

To calculate shape features, an edge map of the normalized face is divided into 3×3 rectangular zones of $a \times a$ pixels each, as shown in Figure 6b. The size of the zones a is selected to be $a = d/2$, with d the distance between the eyes. To estimate d, the detected pupil positions are chosen as the location of the eyes. (We selected the zone size as $d/2$ because in most faces the distance between the eyes is approximately two times the distance between the inner corners of the eyes.)

To place the 3×3 zones onto the face image, the centers of the top-left and top-right zone are placed at the pupil locations. In the vertical direction, then, the face is divided into three equal parts which correspond to the eyes and eyebrows region, the nose region and the mouth region, respectively. The top-left and right zones contain the eye regions.

The coarsely quantized edge directions are represented as local shape features and more global shape features are presented as histograms of local shape (edge directions) along the shape contour. The edge directions are quantized into 4 angular segments (Figure 6c). Representing the whole mouth as one histogram does not capture the local shape properties that are needed to distinguish neutral from non-neutral expressions and we use the zones to compute three histograms of the edge directions. A histogram is then a description of the shape within a zone. In the current system, only the three zones in the mouth region are used to obtain shape histograms. Hence, the mouth is represented as a feature vector of 12 components (3 histograms of 4 components). An example of the histogram of edge directions corresponding to the middle zone of the mouth is shown in Figure 6d.



(a) (b)

(c)

Figure 5: Eye features extracted by fitting ellipses. (a) A face image where the right eye is open and the left eye is closed. (b) The binary face image, which is used to extract eye features by ellipse fitting. (c) The ellipse fitting results for the open eye and the closed eye.

4 Neutral face detection

We use a neural network of the structure shown in Figure 7 as the classifier. Hence standard back-propagation in the form of a three-layer neural network with one hidden layer is used to detect if a face is a neutral face *or not*. The inputs to the network are the 7 location features (Figure 4) and the 12

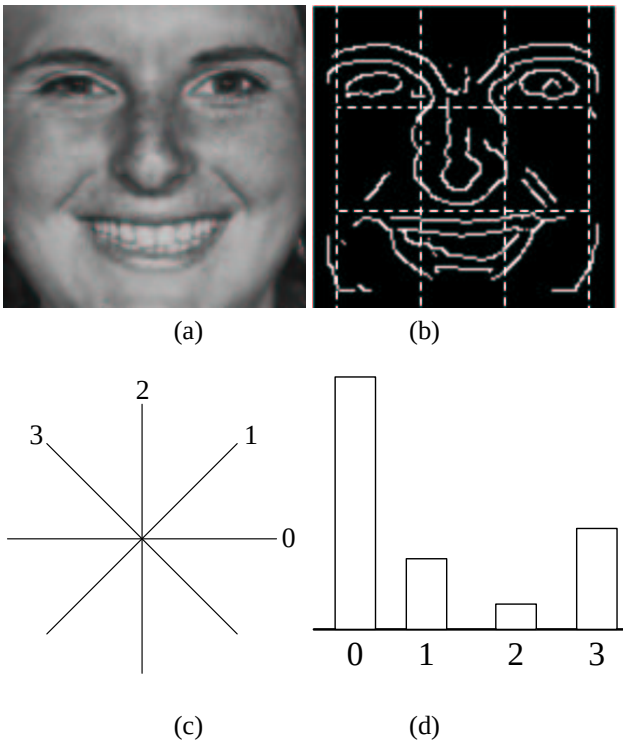


Figure 6: Zonal-histogram features. (a) Normalized face. (b) Zones of the edge map of the normalized face. (c) Four quantization levels for calculating histogram features. (d) Histogram corresponding to the middle zone of the mouth.

zone components of shape features of the mouth (Figure 6). Hence, a total of 19 features is used to represent the amount of expression in a face image.

We tested various numbers of hidden units and found that 4 hidden units is enough to reach the best performance. The outputs are two classes corresponding to the *neutral* and *non-neutral* face, respectively.

5 Experiment Results

To test the algorithms, we specify the problem in terms of a classification system with the hypotheses

$$H_0: \quad \text{the face is neutral,}$$
$$H_1: \quad \text{the face is } not \text{ neutral.}$$

Such a system can make two kind of mistakes:

1. *A false negative*: a non-neutral face is classified as a neutral face, and

2. *A false positive*: a neutral face is classified as a non-neutral face.

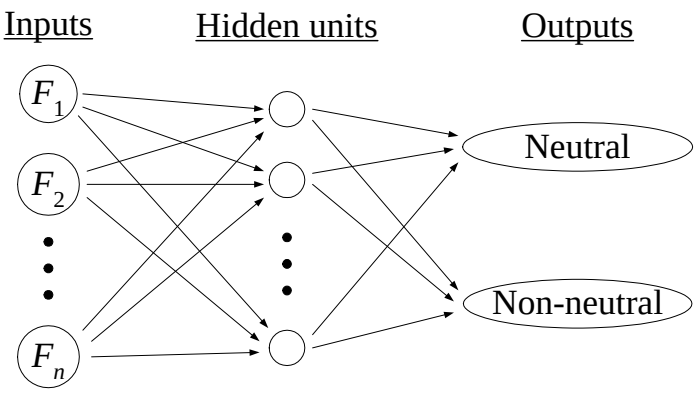


Figure 7: Neural network-based neutral face detector.

We express the system performance in terms of estimates of the false negative and false positive rates. Another measure of system performance we use, is the classification rate, which is the percentage of face images that is correctly classified.

Two databases are used to evaluate our system:

1. The first portion of the Cohn-Kanade AU-Coded Face Expression Image Database [6], hereafter denoted as database A.

2. The Yale face Database [1], denoted as database B.

Database A is divided into a training set A_1 and a test set A_2 of images. None of the subjects in the test set appears in the training set.

Classification results on A_2 and on B are reported. Note that databases A and B are an independent data sets, in the sense that lighting, imaging geometry and the subjects of database A and database B, are not purposely duplicated.

Results on database A

The subjects in the first portion of the Cohn-Kanade database A are 100 university students enrolled in introductory psychology classes. The demographics of this population are 65% female and 35% male, and 82% Caucasian, 15% African and 3% Asian or Latino ranging ranging in age from 18 to 30 years. The subjects were instructed by an experimenter to perform single action unit and action unit combinations. Subjects' facial behavior was recorded in an observation room using a frontal camera. The image sequences begin with a neutral face and are digitized into 640 pixel arrays with 8-bit gray-scale.

A total of 988 images from database A are used for neutral face detection. Of these, the training set A_1 contains 452 images and the test set A_2 contains 536 images, with:

1. Set A_1: 27 subjects, 195 neutral face images and 257 non-neutral images,

2. Set A_2: 24 subjects, 204 neutral face images and 332 non-neutral images.

Table 1: Neutral and non-neutral face classification results on Cohn-Kanade data set A_2.

	Total	Correct	Wrong
Non-neutral	332	325	7
Neutral	204	196	8
Total	536	521	15
Classification rate	97.2%		

Table 1 shows the neutral face classification results for database A_2. An average detection rate of 97.2% is achieved. The false negative rate is 7 in 536, i.e., in the order of 1.3 percent, while the false positive is of the same order, 8 in 536.

Results on database B

To get some idea about the robustness of the system, we tested it on the Yale face database B. This database contains 165 images of 15 subjects. There are 11 images per subject, one for each of the following facial expressions or configurations: center-light, glasses, happy, left-light, normal, right-light, sad, sleepy, surprised and wink.

From the Yale face database, a total of 64 images acquired from 13 subjects are selected for testing. These are 25 face images with a neutral expression and 39 images with non-neutral expression. The center-light face images and the normal face images are selected as neutral expressions, while the happy, surprised and wink face images are selected as non-neutral expressions. We omitted the glasses, left-light, right-light, sad and sleepy face images, because

1. There are not enough images of faces with glasses in the training set A_1.
2. In the left/right-light face images, half of the face is shadowed. They are too dark to extract facial features.
3. The sad and the sleepy face images show little expression change compared to the neutral faces.

The details of the result are shown in Table 2. The average detection rate is 95.3%. The false negative rate is 1 in 64 (1.5%) and the false positive rate is 2 in 64 (3%). Other than there is no significant breakdown of the algorithm when it is tested on an independent data set, there are not many conclusive things one can say based on these rates.

Discussion

Some examples of misclassified faces are shown in Figure 8. Figure 8a shows the examples of the neutral faces that are classified as non-neutral faces. This happens probably because the beard of some subjects causes the mouth shape features to be computed incorrectly. It is a weakness of the approach and needs to be addressed by somehow detecting facial hair. Figure 8b shows the examples of the non-neutral faces that are classified as neutral faces. Most of these have only subtle expressions.

Table 2: Face classification results on database B.

	Total	Correct	Wrong
Non-neutral	39	38	1
Neutral	25	23	2
Total	64	61	3
Classification rate	95.3%		

(a) Neutral faces are classified as non-neutral

(b) Non-neutral faces are classified as neutral

Figure 8: Examples of wrong classified faces. For display purpose, images have been cropped to reduce space.

6 Conclusions

Automatically detecting neutral faces is important for face recognition, facial expression analysis, speaker ID analysis, driver awareness systems, and related applications such as multi-modal user interfaces. We developed a new system to classify a face as neutral or non-neutral by combining both location and shape features in the regions of the mouth and the eyes. After localizing face and facial point features for each face, the eye features are extracted by ellipse fitting based on the binary image of the face. The shape features of the mouth are extracted by a zonal-histogram in 4 directions

based on edge maps of the face. All the location and shape features are input to a neural-network-based detector to arrive at a neutral/non-neutral decision. An average detection rate of 97.2% is achieved for 536 images from 24 subjects. The robustness of the system has been tested by using an independent image database where an average classification rate of 95.3% is achieved.

We examine the classification of the expression on a single face appearance as neutral or *not* neutral. This problem is not very well-defined and even for humans the use of one static snapshot for neutral face detection is hard. If a video sequence of the face is available, integrating the classification results over the sequence will enable the use of more information and classification results will improve.

Acknowledgements

We would like to thank Andrew W. Senior of IBM Watson Research Center for providing the face and facial feature finding code. We also acknowledge the use of the "Cohn-Kanade AU-Coded Face Expression Image Database" and the "Yale face Database". We appreciate the helpful comments and suggestions of anonymous reviewers.

References

[1] Yale university face database. In *http://cvc.yale.edu/projects/yalefaces/yalefaces.html*.

[2] G. Chow and X. Li. Towards a system for automatic facial feature detection. *Pattern Recognition*, 26(12):1739–1755, 1993.

[3] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski. Classifying facial actions. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 21(10):974–989, October 1999.

[4] P. Ekman. *The argument and evidence about universals in facial expressions of emotion*, pages 143–164. Wiley & Sons, New York, 1989.

[5] P. Ekman and W.V. Friesen. *The Facial Action Coding System: A Technique For The Measurement of Facial Movement*. Consulting Psychologists Press, Inc., San Francisco, CA, 1978.

[6] T. Kanade, J.F. Cohn, and Y.L. Tian. Comprehensive database for facial expression analysis. In *Proceedings of International Conference on Face and Gesture Recognition*, pages 46–53, March, 2000.

[7] M. S. Manjunath, R. Chellappa, and C. Malsburg. A feature based approach to face recognition. In *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pages 373–378.

[8] M. Pantic and Leon J. M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, December 2000.

[9] M. Pilu, A. W. Fitzgibbon, and R. B. Fisher. Ellipse-specific direct least-square fitting. In *IEEE International Conference on Image Processing*, 1996.

[10] K.R. Scherer and P. Ekman. *Handbook of methods in nonverbal behavior research*. Cambridge University Press, Cambridge, UK, 1982.

[11] A.W. Senior. Face and feature finding for a face recognition system. In *Proc. Conf. Audio- and Video-based Biometric Person Authentication (AVBPA)*, pages 154–159, 1999.

[12] Y.L. Tian, T. Kanade, and J.F. Cohn. Recognizing action units for facial expression analysis. *IEEE Trans. on Pattern Analysis and Machine Intell.*, 23(2):1–19, February 2001.

[13] M.A. Turk and A.P. Pentland. Face recognition using eigenfaces. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 586–591, 1991.

[14] X. Xie, R. Sudhakar, and H. Zhuang. On improving eye feature extraction using deformable templates. *Pattern Recognition*, 27(6):791–799, 1994.

[15] Y. Yacoob, H-M Lam, and L.S. Davis. Recognizing faces showing expressions. In *Proc. Int. Workshop on Automatic Face- and Gesture-Recognition*, pages 278–283, Zurich, Switserland, 1995.

[16] A.L. Yuille, P.W. Haallinan, and D. S. Cohen. Feature extraction from faces using deformable templates. *International Journal of Computer Vision*, 8(2):99–111, 1992.