

Recognizing Action Units for Facial Expression Analysis

Ying-li Tian, *Member, IEEE*, Takeo Kanade, *Fellow, IEEE*, and Jeffrey F. Cohn, *Member, IEEE*

Abstract—Most automatic expression analysis systems attempt to recognize a small set of prototypic expressions, such as happiness, anger, surprise, and fear. Such prototypic expressions, however, occur rather infrequently. Human emotions and intentions are more often communicated by changes in one or a few discrete facial features. In this paper, we develop an Automatic Face Analysis (AFA) system to analyze facial expressions based on both permanent facial features (brows, eyes, mouth) and transient facial features (deepening of facial furrows) in a nearly frontal-view face image sequence. The AFA system recognizes fine-grained changes in facial expression into action units (AUs) of the Facial Action Coding System (FACS), instead of a few prototypic expressions. Multistate face and facial component models are proposed for tracking and modeling the various facial features, including lips, eyes, brows, cheeks, and furrows. During tracking, detailed parametric descriptions of the facial features are extracted. With these parameters as the inputs, a group of action units (neutral expression, six upper face AUs and 10 lower face AUs) are recognized whether they occur alone or in combinations. The system has achieved average recognition rates of 96.4 percent (95.4 percent if neutral expressions are excluded) for upper face AUs and 96.7 percent (95.6 percent with neutral expressions excluded) for lower face AUs. The generalizability of the system has been tested by using independent image databases collected and FACS-coded for ground-truth by different research teams.

Index Terms—Computer vision, multistate face and facial component models, facial expression analysis, facial action coding system, action units, AU combinations, neural network.

1 INTRODUCTION

FACIAL expression is one of the most powerful, natural, and immediate means for human beings to communicate their emotions and intentions. The face can express emotion sooner than people verbalize or even realize their feelings. In the past decade, much progress has been made to build computer systems to understand and use this natural form of human communication [4], [3], [8], [10], [16], [18], [24], [26], [28], [32], [37], [38], [36], [40]. Most such systems attempt to recognize a small set of prototypic emotional expressions, i.e., joy, surprise, anger, sadness, fear, and disgust. This practice may follow from the work of Darwin [9] and more recently Ekman and Friesen [13], Friesen [12], and Izard et al. [19] who proposed that basic emotions have corresponding prototypic facial expressions. In everyday life, however, such prototypic expressions occur relatively infrequently. Instead, emotion more often is communicated by subtle changes in one or a few discrete facial features, such as a tightening of the lips in anger or obliquely lowering the lip corners in sadness [7]. Change in isolated features, especially in the area of the eyebrows or eyelids, is typical of paralinguistic displays; for instance, raising the brows signals greeting [11]. To capture such subtlety of human emotion and paralinguistic communication, automated recognition of fine-grained changes in facial expression is needed.

1.1 Facial Action Coding System

Ekman and Friesen [14] 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 *nonadditive*, 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 [30]. FACS provides the descriptive power necessary to describe the details of facial expression.

Commonly occurring AUs and some of the additive and nonadditive AU combinations are shown in Tables 1 and 2. As an example of a nonadditive effect, AU 4 appears differently depending on whether it occurs alone or in combination with AU 1 (as in AU 1 + 4). When AU 4 occurs alone, the brows are drawn together and lowered. In AU 1 + 4, the brows are drawn together but are raised due to the action of AU 1. AU 1 + 2 is another example of nonadditive combinations. When AU 2 occurs alone, it not only raises the outer brow, but also often pulls up the inner brow which results in a very similar appearance to AU 1 + 2. These effects of the nonadditive AU combinations increase the difficulties of AU recognition.

1.2 Automated Facial Expression Analysis

Most approaches to automated facial expression analysis so far attempt to recognize a small set of prototypic emotional expressions. Suwa et al. [31] presented an early attempt to analyze facial expressions by tracking the motion of 20 identified spots on an image sequence. Essa and

• Y.-l. Tian and T. Kanade are with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213. E-mail: {yltian, tk}@cs.cmu.edu.






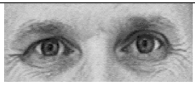









• J.F. Cohn is with the Robotics Institute, Carnegie Mellon University, Pittsburgh, and the Department of Psychology, University of Pittsburgh, Pittsburgh, PA 15260. E-mail: jeffcohn@pitt.edu.

Manuscript received 26 Apr. 2000; revised 5 Oct. 2000; accepted 14 Oct. 2000.

Recommended for acceptance by K.W. Bowyer.

For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 112006.

TABLE 1
Upper Face Action Units and Some Combinations

<i>NEUTRAL</i>	AU 1	AU 2	AU 4	AU 5
				
Eyes, brow, and cheek are relaxed.	Inner portion of the brows is raised.	Outer portion of the brows is raised.	Brows lowered and drawn together	Upper eyelids are raised.
AU 6	AU 7	AU 1+2	AU 1+4	AU 4+5
				
Cheeks are raised.	Lower eyelids are raised.	Inner and outer portions of the brows are raised.	Medial portion of the brows is raised and pulled together.	Brows lowered and drawn together and upper eyelids are raised.
AU 1+2+4	AU 1+2+5	AU 1+6	AU 6+7	AU 1+2+5+6+7
				
Brows are pulled together and upward.	Brows and upper eyelids are raised.	Inner portion of brows and cheeks are raised.	Lower eyelids cheeks are raised.	Brows, eyelids, and cheeks are raised.

Pentland [16] developed a dynamic parametric model based on a 3D geometric mesh face model to recognize five prototypic expressions. Mase [26] manually selected facial regions that corresponded to facial muscles and computed motion within these regions using optical flow. The work by Yacoob and Davis [37] used optical flow like Mase's work, but tracked the motion of the surface regions of facial features (brows, eyes, nose, and mouth) instead of that of the underlying muscle groups. Zhang [40] investigated the use of two types of facial features: the geometric positions of 34 fiducial points on a face and a set of multiscale, multiorientation Gabor wavelet coefficients at these points for facial expression recognition.

Automatic recognition of FACS action units (AU) is a difficult problem, and relatively little work has been reported. AUs have no quantitative definitions and, as noted, can appear in complex combinations. Mase [26] and Essa [16] described patterns of optical flow that corresponded to several AUs, but did not attempt to recognize them. Bartlett et al. [2] and Donato et al. [10] reported some of the most extensive experimental results of upper and lower face AU recognition. They both used image sequences that were free of head motion, manually aligned faces using three coordinates, rotated the images so that the eyes were in horizontal, scaled the images and, finally, cropped a window of 60×90 pixels. Their system was trained and tested using the leave-one-out cross-validation procedure, and the mean classification accuracy was calculated across all of the test cases. Bartlett et al. [2] recognized six single upper face AUs (AU 1, AU 2, AU 4, AU 5, AU 6, and AU 7) but no AUs occurring in combinations. They achieved 90.9 percent accuracy by combining holistic spatial analysis and optical flow with local feature analysis in a hybrid system. Donato et al. [10] compared several techniques for recognizing action























units. These techniques included optical flow, principal component analysis, independent component analysis, local feature analysis, and Gabor wavelet representation. The best performances were obtained by using Gabor wavelet representation and independent component analysis with which a 95.5 percent average recognition rate was reported for six single upper face AUs (AU 1, AU 2, AU 4, AU 5, AU 6, and AU 7) and two lower face AUs and four AU combinations (AU 17, AU 18, AU 9 + 25, AU 10 + 25, AU 16 + 25, AU 20 + 25). For analysis purpose, they treated each combination as if it were a separate new AU.

The authors' group has developed a few versions of the facial expression analysis system. Cohn et al. [8] and Lien et al. [24] used dense-flow, feature-point tracking, and edge extraction to recognize four upper face AUs and two combinations (AU 4, AU 5, AU 6, AU 7, AU 1 + 2, and AU 1 + 4) and four lower face AUs and five combinations (AU 12, AU 25, AU 26, AU 27, AU 12 + 25, AU 20 + 25 \pm 16, AU 15 + 17, AU 17 + 23 + 24, and AU 9 + 17 \pm 25). Again, each AU combination was regarded as a separate new AU. The average recognition rate ranged from 80 percent to 92 percent depending on the method used and AUs recognized.

These previous versions have several limitations:

1. They require manual marking of 38 to 52 feature points around face landmarks in the initial input frame. A more automated system is desirable.
2. The initial input image is aligned with a standard face image by affine transformation, which assumes that any rigid head motion is in-plane.
3. The extraction of dense flow is relatively slow, which limits its usefulness for large databases and real-time applications.

TABLE 2
Lower Face Action Units and Some Combinations

NEUTRAL	AU 9	AU 10	AU 12	AU 20
				
Lips relaxed and closed.	The infraorbital triangle and center of the upper lip are pulled upwards. Nasal root wrinkling is present.	The infraorbital triangle is pushed upwards. Upper lip is raised. Causes angular bend in shape of upper lip. Nasal root wrinkle is absent.	Lip corners are pulled obliquely.	The lips and the lower portion of the nasolabial furrow are pulled pulled back laterally. The mouth is elongated.
AU15	AU 17	AU 25	AU 26	AU 27
				
The corners of the lips are pulled down.	The chin boss is pushed upwards.	Lips are relaxed and parted.	Lips are relaxed and parted; mandible is lowered.	Mouth stretched open and the mandible pulled downwards.
AU 23+24	AU 9+17	AU9+25	AU9+17+23+24	AU10+17
				
Lips tightened, narrowed, and pressed together.				
AU 10+25	AU 10+15+17	AU 12+25	AU12+26	AU 15+17
				
AU 17+23+24	AU 20+25			
				

Single AU 23 and AU 24 are not included in this table because our database happens to contain only occurrences of their combination, but not individual ones.

4. Lip and eye feature tracking is not reliable because of the aperture problem and when features undergo a large amount of change in appearance, such as open to tightly closed mouth or eyes.
5. While they used three separate feature extraction modules, they were not integrated for the purpose of AU recognition. By integrating their outputs, it is likely that even higher accuracy could be achieved.
6. A separate hidden Markov model is necessary for each single AU and each AU combination. Because FACS consists of 44 AUs and potential combinations numbering in the thousands, a more efficient approach will be needed.

The current AFA system addresses many of these limitations:

1. Degree of manual preprocessing is reduced by using automatic face detection [29]. Templates of face components are quickly adjusted in the first frame and then tracked automatically.
2. No image alignment is necessary, and in-plane and limited out-of-plane head motion can be handled.
3. To decrease processing time, the system uses a more efficient facial feature tracker instead of a computationally intensive dense-flow extractor. Processing now requires less than 1 second per frame pair.
4. To increase the robustness and accuracy of the feature extraction, multistate face-component models are devised. Facial feature tracking can cope with a large change of appearance and limited out-of-plane head motion.

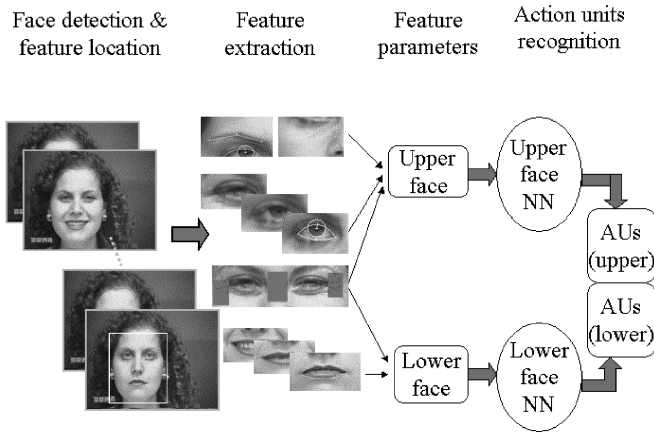


Fig. 1. Feature-based automatic facial action analysis (AFA) system.

5. Extracted features are represented and normalized based on an explicit face model that is invariant to image scale and in-plane head motion.
6. More AUs are recognized and they are recognized whether they occur alone or in combinations. Instead of one HMM for each AU or AU combination, the current system employs two Artificial Neural Networks (one for the upper face and one for the lower face) for AU recognition. It recognizes 16 of the 30 AUs that have a specific anatomic basis and occur frequently in emotion and paralinguistic communication.

2 MULTISTATE FEATURE-BASED AU RECOGNITION

An automated facial expression analysis system must solve two problems: facial feature extraction and facial expression classification. In this paper, we describe our multistate feature-based AU recognition system, which explicitly analyzes appearance changes in localized facial features in a nearly frontal image sequence. Since each AU is associated with a specific set of facial muscles, we believe that accurate geometrical modeling and tracking of facial features will lead to better recognition results. Furthermore, the knowledge of exact facial feature positions could be useful for the area-based [37], holistic analysis [2], and optical-flow-based [24] classifiers.

Fig. 1 depicts the overall structure of the AFA system. Given an image sequence, the region of the face and approximate location of individual face features are detected automatically in the initial frame [29]. The contours of the face features and components then are adjusted manually in the initial frame. Both permanent (e.g., brows, eyes, lips) and transient (lines and furrows) face feature changes are automatically detected and tracked in the image sequence. Informed by FACS AUs, we group the facial features into separate collections of feature parameters because the facial actions in the upper and lower face are relatively independent for AU recognition [14]. In the upper face, 15 parameters describe shape, motion, eye state, motion of brow and cheek, and furrows. In the lower face, nine parameters describe shape, lip state, and furrows. These parameters are geometrically

normalized to compensate for image scale and in-plane head motion.

The facial feature parameters are fed to two neural-network-based classifiers. One recognizes six upper face AUs (AU 1, AU 2, AU 4, AU 5, AU 6, AU 7) and *NEUTRAL*, and the other recognizes 10 lower face AUs (AU 9, AU 10, AU 12, AU 15, AU 17, AU 20, AU 25, AU 26, AU 27, AU 23 + 24) and *NEUTRAL*. These classifiers are trained to respond to the designated AUs whether they occur singly or in combination. When AUs occur in combination, multiple output nodes could be excited. For the upper face, we have achieved an average recognition rate of 96.4 percent for 50 sample sequences of 14 subjects performing seven AUs (including *NEUTRAL*) singly or in combination. For the lower face, our system has achieved an average recognition rate of 96.7 percent for 63 sample sequences of 32 subjects performing 11 AUs (including *NEUTRAL*) singly or in combination. The generalizability of AFA has been tested further on an independent database recorded under different conditions and ground-truth coded by an independent laboratory. A 93.3 percent average recognition rate has been achieved for 122 sample sequences of 21 subjects for neutral expression and 16 AUs whether they occurred individually or in combinations.

3 FACIAL FEATURE EXTRACTION

Contraction of the facial muscles produces changes in the direction and magnitude of the motion on the skin surface and in the appearance of permanent and transient facial features. Examples of permanent features are the lips, eyes, and any furrows that have become permanent with age. Transient features include facial lines and furrows that are not present at rest but appear with facial expressions. Even in a frontal face, the appearance and location of the facial features can change dramatically. For example, the eyes look qualitatively different when open and closed. Different components require different extraction and detection methods. Multistate models of facial components have been introduced to detect and track both transient and permanent features in an image sequence.

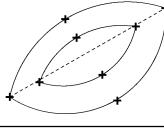
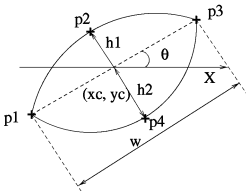

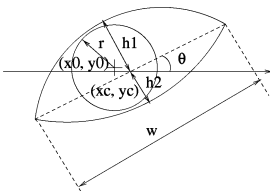

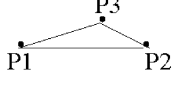
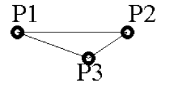
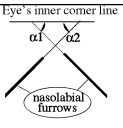
3.1 Multistate Face Component Models

To detect and track changes of facial components in near frontal images, we develop multistate facial component models. The models are illustrated in Table 3, which includes both permanent (i.e., lips, eyes, brows, and cheeks) and transient components (i.e., furrows). A three-state lip model describes lip state: open, closed, and tightly closed. A two-state model (open or closed) is used for each of the eyes. Each brow and cheek has a one-state model. Transient facial features, such as nasolabial furrows, have two states: present and absent.

3.2 Permanent Features

Lips: A three-state lip model represents open, closed, and tightly closed lips. A different lip contour template is prepared for each lip state. The open and closed lip contours are modeled by two parabolic arcs, which are described by six parameters: the lip center position (x_c , y_c), the lip shape (h_1 , h_2 , and w), and the lip orientation (θ). For

TABLE 3
Multistate Facial Component Models of a Frontal Face

Component	State	Description/Feature
Lip	Open	
	Closed	
	Tightly closed	
Eye	Open	
	Closed	
Brow	Present	
Cheek	Present	
Furrow	Present	
	Absent	

tightly closed lips, the dark mouth line connecting the lip corners represents the position, orientation, and shape.

Tracking of lip features uses color, shape, and motion. In the first frame, the approximate position of the lip template is detected automatically. Then, it is adjusted manually by moving four key points. A Gaussian mixture model represents the color distribution of the pixels inside of the lip template [27]. The details of our lip tracking algorithm have been presented in [33].

Eyes: Most eye trackers developed so far are for open eyes and simply track the eye locations [23], [39]. To recognize facial AUs, however, we need to detect whether the eyes are open or closed, the degree of eye opening, and the location and radius of the iris. For an open eye, the eye template (Table 3), is composed of a circle with three

parameters (x_0, y_0, r) to model the iris and two parabolic arcs with six parameters $(x_c, y_c, h_1, h_2, w, \theta)$ to model the boundaries of the eye. This template is the same as Yuille's [39] except for the two points located at the center of the whites of the eyes. For a closed eye, the template is reduced to four parameters: two for the position of each of the eye corners.

The open-eye template is adjusted manually in the first frame by moving six points for each eye. We found that the outer corners are more difficult to track than the inner corners; for this reason, the inner corners of the eyes are tracked first. The outer corners then are located on the line that connects the inner corners at a distance of the eye width as estimated in the first frame.

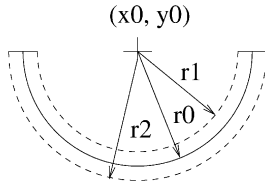


Fig. 2. Half circle iris mask. (x_0, y_0) is the iris center, r_0 is the iris radius r_1 is the minimum radius of the mask, and r_2 is the maximum radius of the mask.

The iris provides important information about the eye state. Part of the iris is normally visible if the eye is open. Intensity and edge information are used to detect the iris. We have observed that the eyelid edge is noisy even in a good quality image. However, the lower part of the iris is almost always visible and its edge is relatively clear if the eye is open. Thus, we use a half circle mask to filter the iris edge (Fig. 2). The radius of the iris circle template r_0 is determined in the first frame, since it is stable except for large out-of-plane head motion. The radius of the circle is increased or decreased slightly (δr) from r_0 so that it can vary between minimum radius ($r_0 - \delta r$) and maximum radius ($r_0 + \delta r$). The system determines that the iris is found when the following two conditions are satisfied: One is that the edges in the mask are at their maximum. The other is that the change in the average intensity is less than a threshold. Once the iris is located, the eye is determined to be open and the iris center is the iris mask center (x_0, y_0) . The eyelid contours then are tracked. For a closed eye, a line connecting the inner and outer corners of the eye is used as the eye boundary. The details of our eye-tracking algorithm have been presented in [34].

Brow and cheek: Features in the brow and cheek areas are also important for expression analysis. Each left or right brow has one model—a triangular template with six parameters (x_1, y_1) , (x_2, y_2) , and (x_3, y_3) . Also, each cheek has a similar six parameter downward triangular template model. Both brow and cheek templates are tracked using the Lucas-Kanade algorithm [25].

3.3 Transient Features

In addition to permanent features that move and change their shape and positions, facial motion also produces transient features that provide crucial information for recognition of certain AUs. Wrinkles and furrows appear perpendicular to the direction of the motion of the activated muscles. Contraction of the *corrugator* muscle, for instance, produces vertical furrows between the brows, which is coded in FACS as AU 4, while contraction of the medial portion of the *frontalis* muscle (AU 1) causes horizontal wrinkling in the center of the forehead.

Some of these transient features may become permanent with age. Permanent crow's-feet wrinkles around the outside corners of the eyes, which are characteristic of AU 6, are common in adults but not in children. When wrinkles and furrows become permanent, contraction of the corresponding muscles produces only changes in their appearance, such as deepening or lengthening. The presence or absence of the furrows in a face image can be determined by edge feature analysis [22], [24], or by eigen-image analysis [21], [35]. Terzopoulos and Waters [32]

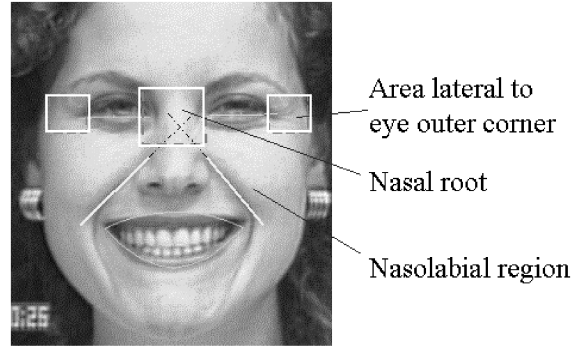


Fig. 3. The areas for nasolabial furrows, nasal root, and outer eye corners.

detected the nasolabial furrows for driving a face animator, but with artificial markers. Kwon and Lobo [22] detected furrows using snakes to classify pictures of people into different age groups. Our previous system [24] detected horizontal, vertical, and diagonal edges using a complex face template.

In our current system, we detect wrinkles in the nasolabial region, the nasal root, and the areas lateral to the outer corners of the eyes (Fig. 3). These areas are located using the tracked locations of the corresponding permanent features. We classify each of the wrinkles into one of two states: present and absent. Compared with the neutral frame, the wrinkle state is classified as present if wrinkles appear, deepen, or lengthen. Otherwise, it is absent.

We use a Canny edge detector to quantify the amount and orientation of furrows [6]. For nasal root wrinkles and crow's-feet wrinkles, we compare the number of edge pixels E in the wrinkle areas of the current frame with the number of edge pixels E_0 of the first frame. If the ratio E/E_0 is larger than a threshold, the furrows are determined to be present. Otherwise, the furrows are absent. For nasolabial furrows, the existence of vertical to diagonal connected edges is used for classification. If the connected edge pixels are larger than a threshold, the nasolabial furrow is determined to be present and is modeled as a line. The orientation of the furrow is represented as the angle between the furrow line and line connecting the eye inner corners. This angle changes according to different AUs. For example, the nasolabial furrow angle of AU 9 or AU 10 is larger than that of AU 12.

3.4 Examples of Feature Extraction

Permanent Features: Fig. 4 shows the results of tracking permanent features for the same subject with different expressions. In Figs. 4a, 4b, and 4d, the lips are tracked as they change in state from open to closed and tightly closed. The iris position and eye boundaries are tracked while the eye changes from widely opened to tightly closed and blink (Figs. 4b, 4c, and 4d). Notice that the semicircular iris model tracks the iris even when the iris is only partially visible. Figs. 5 and 6 show examples of tracking in subjects who vary in age, sex, skin color, and in amount of out-plane head motion. Difficulty occurs in eye tracking when the eye becomes extremely narrow. For example, in Fig. 5a, the left eye in the last image is mistakenly determined to be closed because the iris was too small to be detected. In these examples, face size varies between 80×90 and

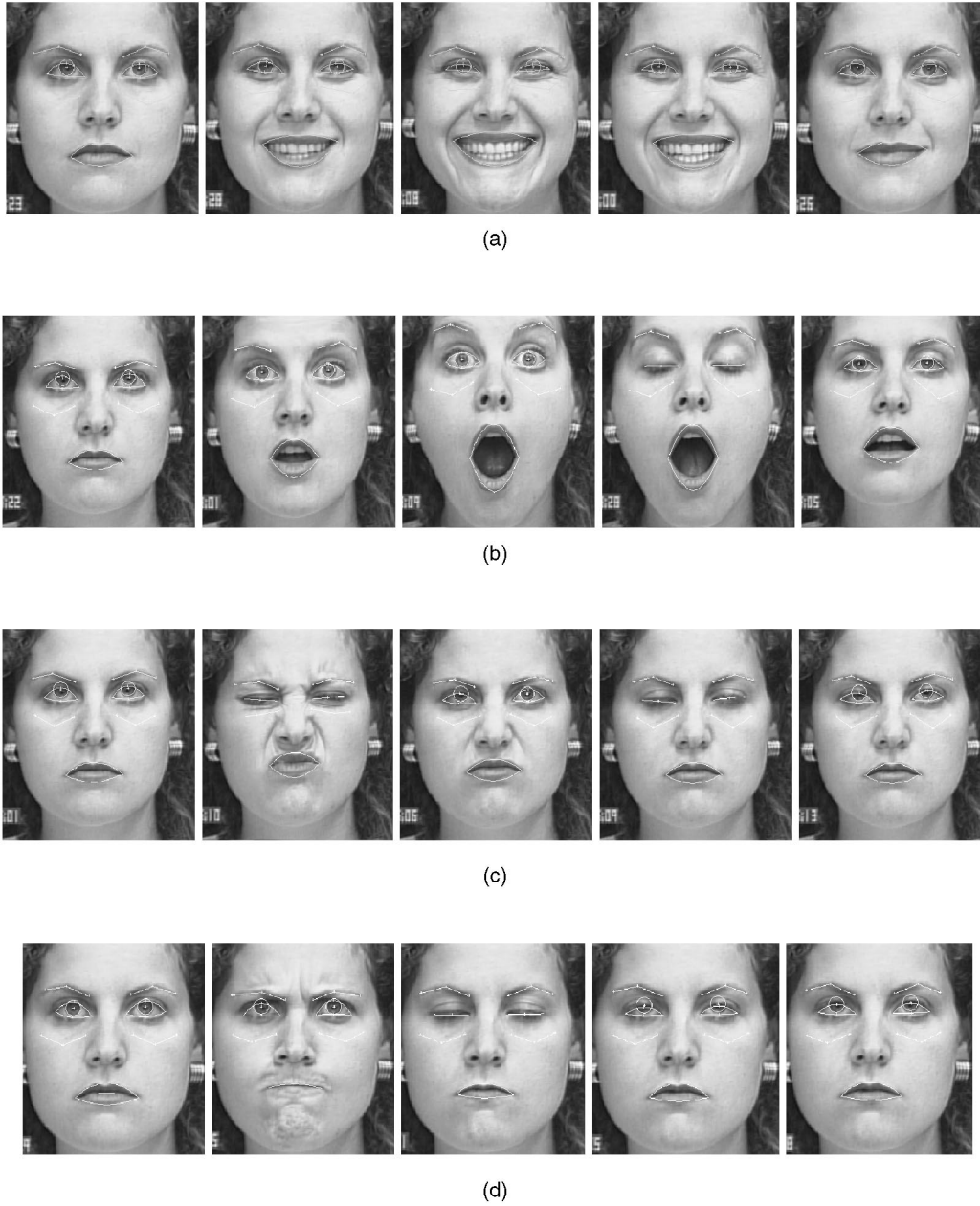


Fig. 4. Permanent feature tracking results for different expressions of same subject. Note appearance changes in eye and mouth states. In this and the following figures, images have been cropped for display purpose. Face size varies between 80×90 and 200×220 pixels.

200×220 pixels. For display purpose, images have been cropped to reduce space. Additional results can be found at <http://www.cs.cmu.edu/~face>.

Transient Features: Fig. 7 shows the results of nasolabial furrow detection for different subjects and AUs. The nasolabial furrow angles systematically vary between AU 9 and AU 12 (Fig. 7a and 7b). For some images, the nasolabial furrow is detected only on one side. In the first image of Fig. 7d, only the left nasolabial furrow exists, and it is correctly detected. In the middle image of Fig. 7b, the right nasolabial furrow is missed because the length of the detected edges is less than threshold. The results of nasal root and crow's-feet wrinkle detection are shown in Fig. 8. Generally, the crow's-feet wrinkles are present for AU 6, and the nasal root wrinkles appear for AU 9.

4 FACIAL FEATURE REPRESENTATION AND AU RECOGNITION by NEURAL NETWORKS

We transform the extracted features into a set of parameters for AU recognition. We first define a face coordinate system. Because the inner corners of the eyes are most reliably detected and their relative position is unaffected by muscle contraction, we define the x -axis as the line connecting two inner corners of eyes and the y -axis as perpendicular to it. We split the facial features into two groups (upper face and lower face) of parameters because facial actions in the upper face have little interaction with facial motion in lower face and vice versa [14].

Upper Face Features: We represent the upper face features by 15 parameters, which are defined in Table 4. Of these, 12 parameters describe the motion and shape of the eyes,



Fig. 5. Permanent feature tracking results for different subjects.

brows, and cheeks, two parameters describe the state of the crow's-feet wrinkles, and one parameter describes the distance between the brows. To remove the effects of variation in planar head motion and scale between image sequences in face size, all parameters are computed as ratios of their current values to that in the initial frame. Fig. 9 shows the coordinate system and the parameter definitions.

Lower Face Features: Nine parameters represent the lower face features (Table 5 and Fig. 10). Of these, six parameters describe lip shape, state and motion, and three describe the furrows in the nasolabial and nasal root regions. These parameters are normalized by using the ratios of the current feature values to that of the neutral frame.

AU Recognition by Neural Networks: We use three-layer neural networks with one hidden layer to recognize AUs by a standard back-propagation method [29]. Separate networks are used for the upper- and lower face. For AU recognition in the upper face, the inputs are the 15 parameters shown in Table 4. The outputs are the six single AUs (AU 1, AU 2, AU 4, AU 5, AU 6, and AU 7) and *NEUTRAL*. In the lower face, the inputs are the seven parameters shown in Table 5 and the outputs are 10 single AUs (AU 9, AU 10, AU 12, AU 15, AU 17, AU 20, AU 23 + 24, AU 25, AU 26, and AU 27) and *NEUTRAL*. These networks are trained to respond to the designated AUs whether they occur singly or in combination. When AUs occur in combination, multiple output nodes are excited.

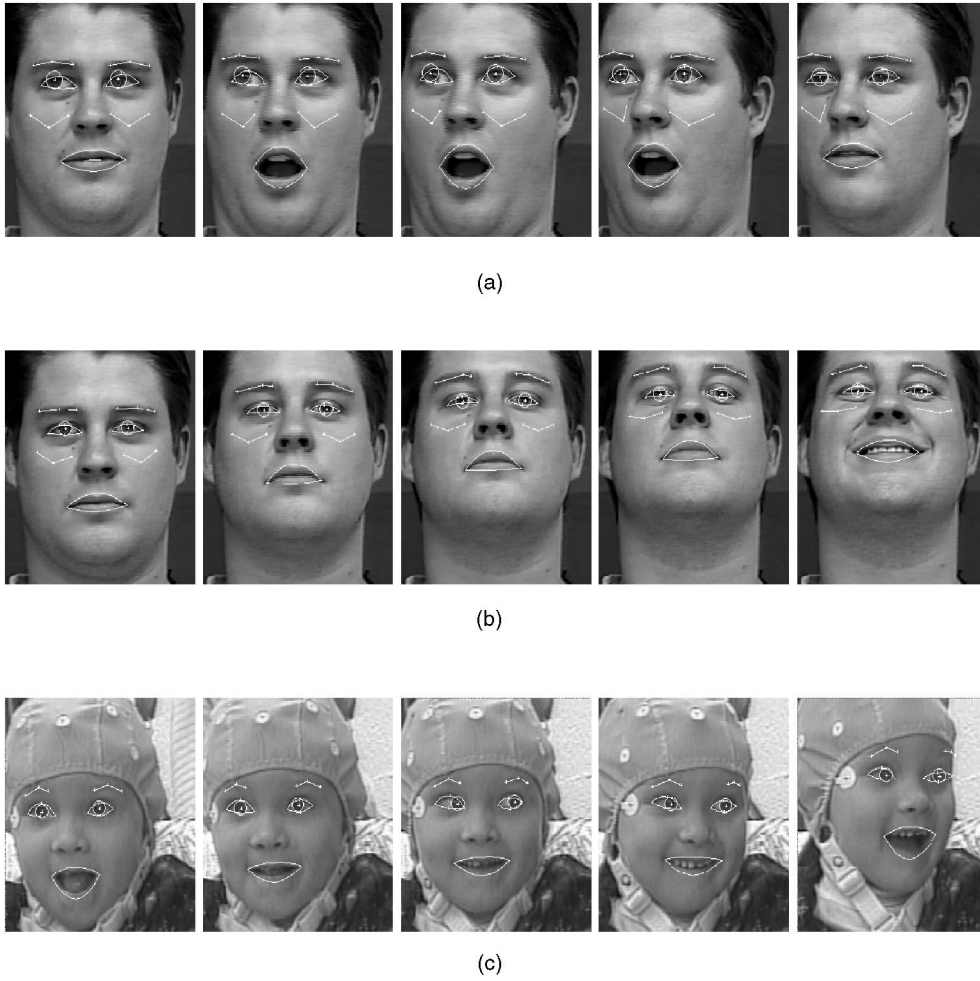


Fig. 6. Permanent feature tracking results with head motions. (a) Head yaw. (b) Head pitch. (c) Head up and left with background motion.

5 EXPERIMENTAL EVALUATIONS

We conducted three experiments to evaluate the performance of our system. The first is AU recognition in the upper face when image data contain only single AUs. The second is AU recognition in the upper and lower face when image data contain both single AUs and combinations. The third experiment evaluates the generalizability of our system by using completely disjointed databases for training and testing, while image data contain both single AUs and combinations. Finally, we compared the performance of our system with that of other AU recognition systems.

5.1 Facial Expression Image Databases

Two databases were used to evaluate our system: the Cohn-Kanade AU-Coded Face Expression Image Database [20] and Ekman-Hager Facial Action Exemplars [15].

Cohn-Kanade AU-Coded Face Expression Image Database: We have been developing a large-scale database for promoting quantitative study of facial expression analysis [20]. The database currently contains a recording of the facial behavior of 210 adults who are 18 to 50 years old, 69 percent female and 31 percent male, and 81 percent Caucasian, 13 percent African, and 6 percent other groups. Over 90 percent of the subjects had no prior experience in FACS. Subjects were instructed by an experimenter to

perform single AUs and AU combinations. Subjects' facial behavior was recorded in an observation room. Image sequences with in-plane and limited out-of-plane motion were included.

The image sequences began with a neutral face and were digitized into 640×480 pixel arrays with either 8-bit gray-scale or 24-bit color values. To date, 1,917 image sequences of 182 subjects have been FACS coded by certified FACS coders for either the entire sequence or target AUs. Approximately 15 percent of these sequences were coded by two independent certified FACS coders to validate the accuracy of the coding. Interobserver agreement was quantified with coefficient kappa, which is the proportion of agreement above what would be expected to occur by chance [17]. The mean kappas for interobserver agreement were 0.82 for target AUs and 0.75 for frame-by-frame coding.

Ekman-Hager Facial Action Exemplars: This database was provided by P. Ekman at the Human Interaction Laboratory, University of California, San Francisco, and contains images that were collected by Hager, Methvin, and Irwin. Bartlett et al. [2] and Donato et al. [10] used this database to train and test their AU recognition systems. The Ekman-Hager database includes 24 Caucasian subjects (12 males and 12 females). Each image sequence consists of six to eight frames that were sampled from a longer image sequence. Image sequences begin with a neutral

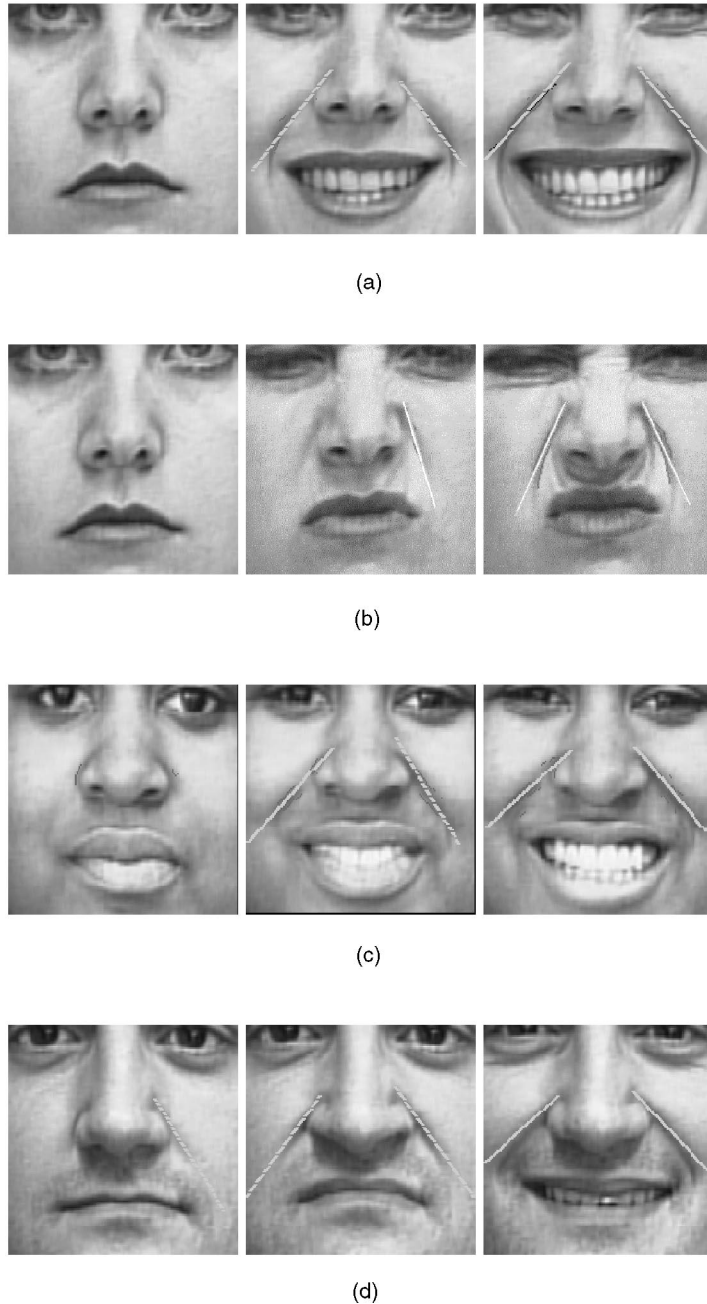


Fig. 7. Nasolabial furrow detection results.

expression (or weak facial actions) and end with stronger facial actions. AUs were coded for each frame. Sequences containing rigid head motion detectable by a human observer were excluded. Some of the image sequences contain large lighting changes between frames and we normalized intensity to keep the average intensity constant throughout the image sequence.

5.2 Upper Face AU Recognition for Image Data Containing Only Single AUs

In the first experiment, we used a neural network-based recognizer having the structure shown in Fig. 11. The inputs to the network were the upper face feature parameters shown in Table 4. The outputs were the same set of six single AUs (AU 1, AU 2, AU 4, AU 5, AU 6, AU 7); these are

the same set that were used by Bartlett and Donato. In addition, we included an output node for *NEUTRAL*. The output node that showed the highest value was interpreted as the recognized AU. We tested various numbers of hidden units and found that six hidden units gave the best performance.

From the Ekman-Hager database, we selected image sequences in which only a single AU occurred in the upper face. 99 image sequences from 23 subjects met this criterion. These 99 image sequences we used are the superset of the 80 image sequences used by Bartlett and Donato. The initial and final two frames in each image sequence were used. As shown in Table 6, the image sequences were assigned to training and testing sets in two ways. In *S1*, the sequences

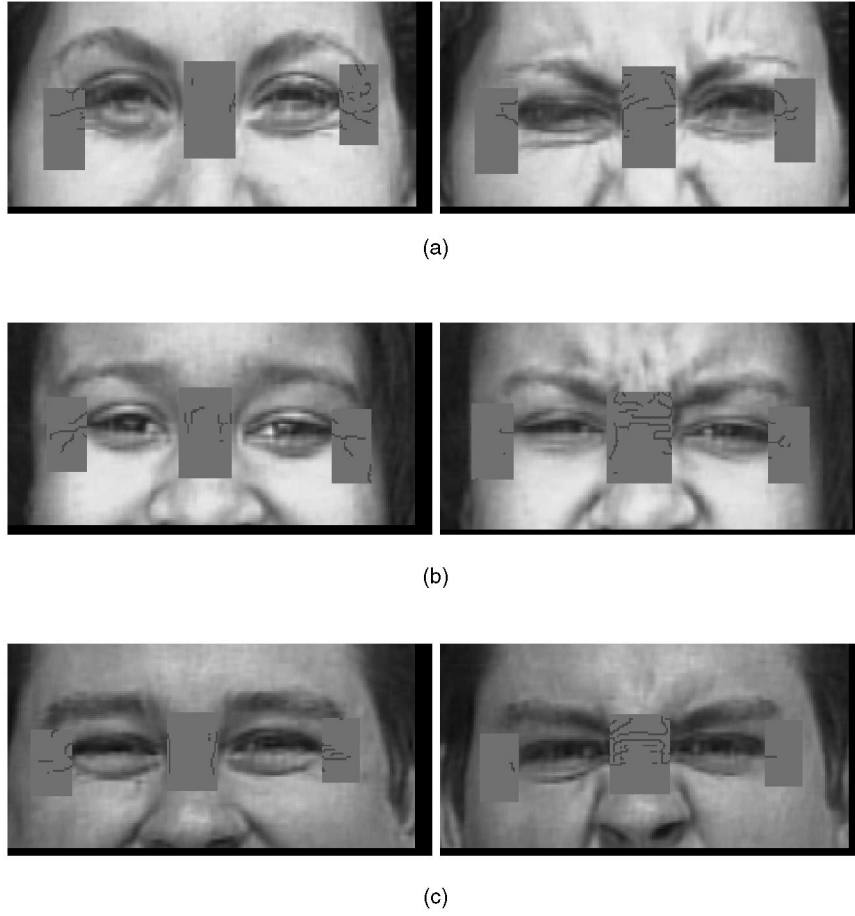


Fig. 8. Nasal root and crow's-feet wrinkle detection. For the left image of (a), (b), and (c), crow's-feet wrinkles are present. For the right image of (a), (b), and (c), the nasal root wrinkles appear.

were randomly selected, so the same subject was allowed to appear in both training and testing sets. In *S2*, no subject could appear in both training and testing sets; testing was performed done with novel faces.

Table 7 shows the recognition results with the *S1* testing set. The average recognition rate was 88.5 percent when

samples of *NEUTRAL* were excluded (Recognizing neutral faces is easier), and 92.3 percent when samples of *NEUTRAL* were included. For the *S2* test set (i.e., novel faces), the recognition rate remained virtually identical: 89.4 percent (*NEUTRAL* exclusive) and 92.9 percent (*NEUTRAL* inclusive), which is shown in Table 8.

TABLE 4
Upper Face Feature Representation for AU Recognition

Permanent features (Left and right)			Other features
Inner brow motion (r_{binner})	Outer brow motion (r_{bouter})	Eye height ($r_{eheight}$)	Distance of brows (D_{brow})
$r_{binner} = \frac{bi-bi_0}{bi_0}$ <p>If $r_{binner} > 0$, Inner brow move up.</p>	$r_{bouter} = \frac{bo-bo_0}{bo_0}$ <p>If $r_{bouter} > 0$, Outer brow move up.</p>	$r_{eheight} = \frac{(h1+h2)-(h1_0+h2_0)}{(h1_0+h2_0)}$ <p>If $r_{eheight} > 0$, Eye height increases.</p>	$D_{brow} = \frac{D-D_0}{D_0}$ <p>If $D_{brow} < 0$ Two brows drawn together.</p>
Eye top lid motion (r_{top})	Eye bottom lid motion (r_{btm})	Cheek motion (r_{cheek})	crows-feet wrinkles $W_{left/right}$
$r_{top} = \frac{h1-h1_0}{h1_0}$ <p>If $r_{top} > 0$, Eye top lid move up.</p>	$r_{btm} = -\frac{h2-h2_0}{h2_0}$ <p>If $r_{btm} > 0$, Eye bottom lid move up.</p>	$r_{cheek} = -\frac{c-c_0}{c_0}$ <p>If $r_{cheek} > 0$, Cheek move up.</p>	<p>If $W_{left/right} = 1$, Left/right crows feet wrinkle present.</p>

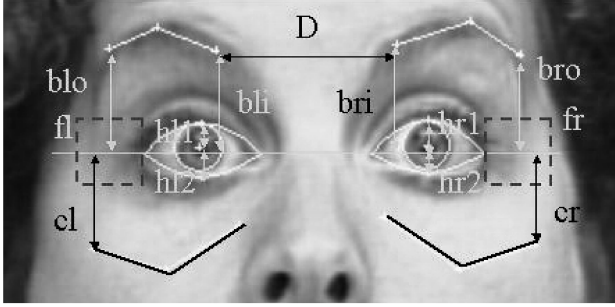


Fig. 9. Upper face features. $hl(hl1 + hl2)$ and $hr(hr1 + hr2)$ are the height of left eye and right eye; D is the distance between brows; cl and cr are the motion of left cheek and right cheek. bli and bri are the motion of the inner part of left brow and right brow. blo and bro are the motion of the outer part of left brow and right brow. fl and fr are the left and right crow's-feet wrinkle areas.

5.3 Upper and Lower Face AU Recognition for Image Sequences Containing Both Single AUs and Combinations

Because AUs can occur either singly or in combinations, an AU recognition system must have the ability to recognize them however they occur. All previous AU recognition systems [2], [10], [24] were trained and tested on single AUs only. In these systems, even when AU combinations were included, each combination was treated as if it were a separate AU. Because potential AU combinations number in the thousands, this method of separately treating AU combinations is impractical. In our second experiment, we trained a neural network to recognize AUs singly and in

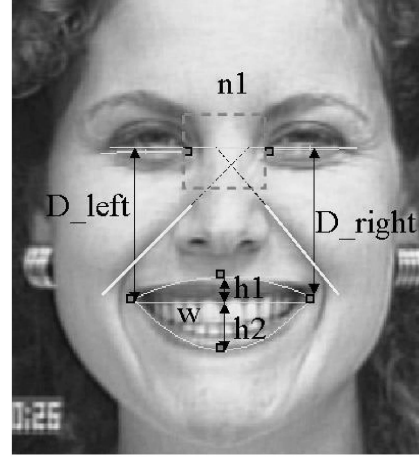


Fig. 10. Lower face features. $h1$ and $h2$ are the top and bottom lip heights; w is the lip width; D_{left} is the distance between the left lip corner and eye inner corners line; D_{right} is the distance between the right lip corner and eye inner corners line; $n1$ is the nasal root area.

combinations by allowing multiple output units of the networks to fire when the input consists of AU combinations.

Upper Face AUs: The neural network-based recognition system for AU combination is shown in Fig. 12. The network has a similar structure to that used in Experiment 1, where the output nodes correspond to six single AUs plus *NEUTRAL*. However, the network for recognizing AU combinations is trained so that when an AU combination is presented, multiple output nodes that correspond to the component AUs are excited. In training, all of the output nodes that correspond to the input AU components are set to have the same value. For example, when a training input is AU 1 + 2 + 4, the output values are trained to be 1.0 for AU 1, AU 2, and AU 4; 0.0 for the remaining AUs and *NEUTRAL*. At the runtime, AUs whose output nodes show values higher than the threshold are considered to be recognized.

A total of 236 image sequences of 23 subjects from the Ekman-Hager database (99 image sequences containing only single AUs and 137 image sequences containing AU combinations) were used for recognition of AUs in the upper face. We split them into training (186 sequences)

TABLE 5
Lower Face Feature Representation for AUs Recognition

Permanent features		
Lip height (r_{height})	Lip width (r_{width})	Left lip corner motion (r_{left})
$r_{height} = \frac{(h1+h2)-(h1_0+h2_0)}{(h1_0+h2_0)}$ If $r_{height} > 0$, lip height increases.	$r_{width} = \frac{w-w_0}{w_0}$ If $r_{width} > 0$, lip width increases.	$r_{left} = -\frac{D_{left}-D_{left0}}{D_{left0}}$ If $r_{left} > 0$, left lip corner moves up.
Right lip corner (r_{right})	Top lip motion (r_{top})	Bottom lip motion (r_{btm})
$r_{right} = -\frac{D_{right}-D_{right0}}{D_{right0}}$ If $r_{right} > 0$, right lip corner moves up.	$r_{top} = -\frac{D_{top}-D_{top0}}{D_{top0}}$ If $r_{top} > 0$, top lip moves up.	$r_{btm} = -\frac{D_{btm}-D_{btm0}}{D_{btm0}}$ If $r_{btm} > 0$, bottom lip moves up.
Transient features		
Left nasolabial furrow angle ($Angle_{left}$)	Right nasolabial furrow angle ($Angle_{right}$)	State of nasal root wrinkles (S_{nosew})
Left nasolabial furrow present with angle $Angle_{left}$.	Left nasolabial furrow present with angle $Angle_{right}$.	If $S_{nosew} = 1$, nasal root wrinkles present.

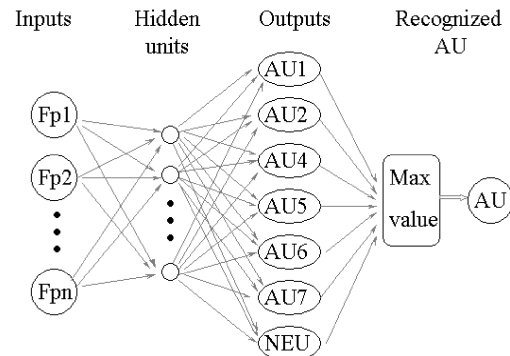


Fig. 11. Neural network-based recognizer for single AUs in the upper face. The inputs are the feature parameters, and the output is one label out of six single AUs and *NEUTRAL*.

TABLE 6

Details of Training and Testing Data from Ekman-Hager Database that Are Used for Single AU Recognition in the Upper Face

Data Set		number of Sequences	Single AUs							Total
			AU1	AU2	AU4	AU5	AU6	AU7	NEUTRAL	
S1	Train	47	14	12	16	22	12	18	47	141
	Test	52	14	12	20	24	14	20	52	156
S2	Train	50	18	14	14	18	22	16	50	152
	Test	49	10	10	22	28	4	22	49	145

In S1, some subjects appear in both training and testing sets. In S2, no subject appears in both training and testing sets.

TABLE 7

AU Recognition for Single AUs on S1 Training and Testing Sets in Experiment 1

		Recognition outputs						
		AU1	AU2	AU4	AU5	AU6	AU7	NEUTRAL
Human	AU1	12	2	0	0	0	0	0
	AU2	3	9	0	0	0	0	0
	AU4	0	0	20	0	0	0	0
	AU5	0	0	0	22	0	0	2
	AU6	0	0	0	0	12	2	0
	AU7	0	0	0	0	2	17	1
	NEUTRAL	0	0	0	0	0	0	52
Recognition Rate		88.5% (excluding NEUTRAL)						
		92.3% (including NEUTRAL)						

A same subject could appear in both training and testing sets. The numbers in bold are results excluding NEUTRAL.

TABLE 8

AU Recognition for Single AUs on S2 Train and Testing Sets in Experiment 1

		Recognition outputs						
		AU1	AU2	AU4	AU5	AU6	AU7	NEUTRAL
Human	AU1	10	0	0	0	0	0	0
	AU2	2	7	0	0	0	0	1
	AU4	0	0	20	0	0	0	2
	AU5	0	0	0	26	0	0	2
	AU6	0	0	0	0	4	0	2
	AU7	0	0	0	0	0	21	1
	NEUTRAL	0	0	0	0	0	0	49
Recognition Rate		89.4% (excluding NEUTRAL)						
		92.9% (including NEUTRAL)						

No subject appears in both training and testing sets. The numbers in bold are results excluding NEUTRAL.

and testing (50 sequences) sets by subjects (9 subjects for training and 14 subjects for testing) to ensure that the same

subjects did not appear in both training and testing. Testing, therefore, was done with “novel faces.” From experiments, we have found that it was necessary to increase the number of hidden units from six to 12 to obtain optimized performance.

Because input sequences could contain one or more AUs, several outcomes were possible. *Correct* denotes that the recognized results were completely identical to the input samples. *Partially correct* denotes that some, but not all of the AUs were recognized (*Missing AUs*) or that AUs that did not occur were misrecognized in addition to the one(s) that did (*Extra AUs*). If none of the AUs that occurred were recognized, the result was *Incorrect*.

Using (1) and (2), we calculated recognition- and false-alarm rates for input samples and input AU components, respectively. Human FACS coders typically use the latter to calculate percentage agreement. We believe, however, that

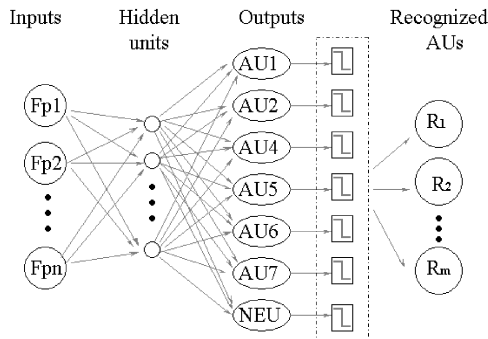


Fig. 12. Neural network-based recognizer for AU combinations in the upper face.

TABLE 9
Upper Face AU Recognition with AU Combinations in Experiment 2

Actual AUs		Samples	Recognized AUs			
			Correct	Partially correct		Incorrect
				Missing AUs	Extra AUs	
AU 1		8	4	-	4(AU 1 + AU 2)	-
AU 2		4	-	-	2(AU 1 + AU 2) 2(AU 1 + AU 2 + AU 4)	-
AU 4		8	8	-	-	-
AU 5		8	8	-	-	-
AU 6		8	8	-	-	-
AU 7		4	2	-	2(AU 6 + AU 7)	-
AU 1+2		16	16	-	-	-
AU 1+2+4		8	8	-	-	-
AU 1+2+5		4	2	2(AU 1 + AU 2 + AU 4)*		-
AU 1+4		4	4	-	-	-
AU 1+6		4	2	2(AU 1)	-	-
AU 4+5		8	6	2(AU 4)	-	-
AU 6+7		16	14	2(AU 6)	-	-
NEUTRAL		50	50	-	-	-
With respect to samples	Total	100	82	18		
		150	132			
	Recognition rate	82% (excluding NEUTRAL)				
		88% (including NEUTRAL)				
	False alarm	12% (excluding NEUTRAL)				
6.7% (including NEUTRAL)						
With respect to AU components	Total	172	164	8	14	-
		222	214			
	Recognition rate	95.4% (excluding NEUTRAL)				
		96.4% (including NEUTRAL)				
	False alarm	8.2% (excluding NEUTRAL)				
		6.3% (including NEUTRAL)				

The numbers in bold are results excluding *NEUTRAL*. The Missing AUs column shows the AUs that are missed. The Extra AUs column lists the extra AUs that are misrecognized. The recognized AU with "*" indicates that it includes both Missing AUs and Extra AUs.

the recognition rates based on input samples are the more conservative measures.

Recognition rate =

$$\left\{ \begin{array}{ll} \frac{\text{Total number of correctly recognized samples}}{\text{Total number of samples}} & \text{based on input samples} \\ \frac{\text{Total number of correctly recognized AUs}}{\text{Total number of AUs}} & \text{based on AU components} \end{array} \right. \quad (1)$$

False alarm rate =

$$\left\{ \begin{array}{ll} \frac{\text{Total number of recognized samples with extra AUs}}{\text{Total number of samples}} & \text{based on input samples} \\ \frac{\text{Total number of extra AUs}}{\text{Total number of AUs}} & \text{based on AU components.} \end{array} \right. \quad (2)$$

Table 9 shows a summary of the AU combination recognition results of 50 test image sequences of 14 subjects from the Ekman-Hager database. For input samples, we achieved average recognition and false alarm rates of 88 percent and 6.7 percent, respectively, when *NEUTRAL* was included, and 82 percent and 12 percent, respectively, when *NEUTRAL* was excluded. AU component-wise, an average recognition rate of 96.4 percent and a false alarm rate

of 6.3 percent were achieved when *NEUTRAL* was included and a recognition rate of 95.4 percent and a false alarm rate of 8.2 percent was obtained when *NEUTRAL* was excluded.

Recognition rates in Experiment 2 were slightly higher than those in Experiment 1. There are two possible reasons: One is that in the neural network used in Experiment 2, multiple output nodes could be excited to allow for recognition of AUs occurring in combinations. Another reason maybe that a larger training data set was used in Experiment 2.

Lower Face AUs: The same structure of the neural network-based recognition scheme, as shown in Fig. 12, was used, except that the input feature parameters and the output component AUs now are those for the lower face. The inputs were the lower face feature parameters shown in Table 5. The outputs of the neural network were the 11 single AUs (AU 9, AU 10, AU 12, AU 15, AU 17, AU 20, AU 25, AU 26, AU 27, AU 23 + 24, and *NEUTRAL*) (see Table 2). Note that AU 23 + 24 is modeled as a single unit, instead of as AU 23 and AU 24 separately, because they almost always occurred together in our data. Use of 12 hidden units achieved the best performance in this experiment.

TABLE 10
Lower Face AU Recognition Results in Experiment 2

Actual AUs		Samples	Recognized AUs			
			<i>Correct</i>	<i>Partially correct</i>		<i>Incorrect</i>
				<i>Missing AUs</i>	<i>Extra AUs</i>	
AU 9		2	2	-	-	-
AU 10		4	4	-	-	-
AU 12		4	4	-	-	-
AU 15		2	2	-	-	-
AU 17		6	6	-	-	-
AU 20		4	4	-	-	-
AU 25		30	30	-	-	-
AU 26		12	9	-	-	3(AU 25)
AU 27		8	8	-	-	-
AU 23+24		0	-	-	-	-
AU 9+17		12	12	-	-	-
AU 9+17+23+24		2	2	-	-	-
AU 9+25		2	2	-	-	-
AU 10+17		4	1	1(AU 17)	-	-
				2(AU 10 + AU 12)*		
AU 10+15+17		2	2	-	-	-
AU 10+25		2	2	-	-	-
AU 12+25		8	8	-	-	-
AU 12+26		2	-	2(AU 12 + AU 25)*		-
AU 15+17		8	8	-	-	-
AU 17+23+24		4	4	-	-	-
AU 20+25		8	8	-	-	-
NEUTRAL		63	63	-	-	-
With respect to samples	Total No. of input samples	126	118	8		
		189	181			
	Recognition rate of samples	93.7% (excluding NEUTRAL)				
		95.8% (including NEUTRAL)				
	False alarm of samples	6.4% (excluding NEUTRAL)				
4.2% (including NEUTRAL)						
With respect to AU components	Total No. of AUs	180	172	5	7	3
		243	235			
	Recognition rate of AUs	95.6% (excluding NEUTRAL)				
		96.7% (including NEUTRAL)				
	False alarm of AUs	3.9% (excluding NEUTRAL)				
2.9% (including NEUTRAL)						

A total of 463 image sequences from the Cohn-Kanade AU-Coded Face Expression Image Database were used for lower face AU recognition. Of these, 400 image sequences were used as the training data and 63 sequences were used as the testing data. The test data set included 10 single AUs, *NEUTRAL*, and 11 AU combinations (such as AU 12 + 25, AU 15 + 17 + 23, AU 9 + 17 + 23 + 24, and AU 17 + 20 + 26) from 32 subjects; none of these subjects appeared in training data set. Some of the image sequences contained limited planar and out-of-plane head motions.

Table 10 shows a summary of the AU recognition results for the lower face when image sequences contain both single AUs and AU combinations. As above, we report the recognition and false alarm rates based on both the number of input samples and the number of AU components (see

(1) and (2)). With respect to the input samples, an average recognition rate of 95.8 percent was achieved with a false alarm rate of 4.2 percent when *NEUTRAL* was included and a recognition rate of 93.7 percent and a false alarm rate of 6.4 percent when *NEUTRAL* was excluded. With respect to AU components, an average recognition rate of 96.7 percent was achieved with a false alarm rate of 2.9 percent when *NEUTRAL* was included, and a recognition rate of 95.6 percent with a false alarm rate of 3.9 percent was obtained when *NEUTRAL* was excluded.

Major Causes of the Misidentifications: Most of the misidentifications come from confusions between similar AUs: AU1 and AU2, AU6 and AU7, and AU25 and AU26. The confusions between AU 1 and AU 2 were caused by the strong correlation between them. The action of AU 2, which

TABLE 11
Generalizability to Independent Databases

		Test databases		Train databases
		Cohn-Kanade	Ekman-Hager	
Recognition Rate	upper face	93.2%	96.4% (Table 9)	Ekman-Hager
	lower face	96.7% (Table 10)	93.4%	Cohn-Kanade

The numbers in bold are results from independent databases.

raises the outer portion of the brow, tends to pull the inner brow up as well (see Table 1). Both AU 6 and AU 7 raise the lower eyelids and are often confused by human AU coders as well [8]. All the mistakes of AU 26 were due to confusion with AU 25. AU 25 and AU 26 contain parted lips but differ only with respect to motion of the jaw, but jaw motion was not detected or used in the current system.

5.4 Generalizability between Databases

To evaluate the generalizability of our system, we trained the system on one database and tested it on another independent image database that was collected and FACS coded for ground-truth by a different research team. One was Cohn-Kanade database and the other was the Ekman-Hager database. This procedure ensured a more rigorous test of generalizability than more usual methods which divide a single database into training and testing sets. Table 11 summarizes the generalizability of our system.

For upper face AU recognition, the network was trained on 186 image sequences of nine subjects from the Ekman-Hager database and tested on 72 image sequences of seven subjects from the Cohn-Kanade database. Of the 72 image sequences, 55 consisted of single AUs (AU 1, AU 2, AU 4, AU 5, AU 6, and AU 7) and the others contained AU combinations such as AU 1 + 2, AU 1 + 2 + 4, and AU 6 + 7. We achieved a recognition rate of 93.2 percent and a false alarm of 2 percent (when samples of *NEUTRAL* were included), which is only slightly (3-4 percent) lower than the case when the Ekman-Hager database was used for both training and testing.

For lower face AU recognition, the network was trained on 400 image sequences of 46 subjects from the Cohn-Kanade database and tested on 50 image sequences of 14 subjects from the Ekman-Hager database. Of the 50 image sequences, half contained AU combinations, such as AU 10 + 17, AU 10 + 25, AU 12 + 25, AU 15 + 17, and AU 20 + 25. No instances of AU 23 + 24 were available in the Ekman-Hager database. We achieved a recognition rate of 93.4 percent (when samples of *NEUTRAL* were included). These results were again only slightly lower than those of using the same database. The system showed high generalizability.

5.5 Comparison with Other AU Recognition Systems

We compare the current AFA system's performance with that of Cohn et al. [8], Lien et al. [24], Bartlett et al. [2], and

Donato et al. [10]. The comparisons are summarized in Table 12. When performing comparison of recognition results in general, it is important to keep in mind differences in experimental procedures between systems. For example, scoring methods may be either by dividing the data set into training and testing sets [8], [24] or by using a leave-one-out cross-validation procedure [2], [10]. Even when the same data set is used, the particular AUs that were recognized or the specific image sequence that were used for evaluation are not necessarily the same. Therefore, minor differences in recognition rates between systems are not meaningful.

In Table 12, the systems were compared along several characteristics: feature extraction methods, recognition rates, treatment of AU combinations, AUs recognized, and databases used. The terms "old faces" and "novel faces" in the third column requires some explanation. "Old faces" means that in obtaining the recognition rates, some subjects appear in both training and testing sets. "Novel faces" means no same subject appears in both training and testing sets; this is obviously a little more difficult case than "Old faces." In the fourth column, the terms "No," "Yes/Yes," and "Yes/No" are used to describe how the AU combinations are treated. "No" means that no AU combination was recognized. "Yes/Yes" means that AU combinations were recognized and AUs in combination were recognizable individually. "Yes/No" means that AU combinations were recognized but each AU combination was treated as if it were a separate new AU. Our current AFA system, while being able to recognize a larger number of AUs and AU combinations, shows the best or near the best recognition rates even for the tests with "novel faces" or in tests where independent different databases are used for training and testing.

6 CONCLUSION

Automatically recognizing facial expressions is important to understand human emotion and paralinguistic communication, to design multimodal user interfaces, and to relate applications, such as human identification. The facial action coding system (FACS) developed by Ekman and Friesen [14] is considered to be one of the best and accepted foundations for recognizing facial expressions. Our feature-based automatic face analysis (AFA) system has shown improvement in AU recognition over previous systems.

It has been reported [2], [5], [40] that holistic template-based methods (including image decomposition with image

TABLE 12
Comparison with Other AU Recognition Systems

Systems	Methods	Recognition rates	Treatment of AU combinations	AUs to be recognized	Databases
Current AFA system	Feature-based	88.5% (old faces)	No	AU 1, 2, 4, AU 5, 6, 7.	Ekman-Hager
		89.4% (novel faces)			
		95.4%	Yes/Yes		
		95.6%	Yes/Yes	AU 9,10,12,15, AU17,20,25,26, AU27,23+24.	Cohn-Kanade
Bartlett et al. [2]	Feature-based	85.3% (old faces)	No	AU 1, 2, 4, AU 5, 6, 7.	Ekman-Hager
		57% (novel faces)			
	Optic-flow	84.5%			
	Hybrid	90.9%			
Donato et al. [10]	ICA or Gabor wavelet	96.9%	No	AU 1, 2, 4, AU 5, 6, 7.	Ekman-Hager
	Others	70.3%-85.6%			
	ICA or Gabor	95.5%	Yes/No	AU17,18,9+25 AU10+25,16+25	
	Others	70.3%-85.6%			
Cohn et al.[8]	Feature-Tracking	89%	Yes/No	AU1+2,1+4, 4, AU 5, 6, 7.	Cohn-Kanade (Subset)
		82.3%	Yes/No	AU12,6+12+25, AU20+25,15+17, AU17+23+24,9+17. AU 25, 26, 27	
Lien et al.[24]	Dense-flow	91%	Yes/No	AU 1+2,1+4, AU 4.	Cohn-Kanade (Subset)
	Edge-detection	87.3%			
	Feature-Tracking	89%	Yes/No	AU1+2,1+4, 4, AU 5, 6, 7.	
	Dense-flow	92.3%	Yes/No	AU12,6+12+25, AU20+25,15+17, AU17+23+24,9+17.	
	Feature-Tracking	88%	Yes/No	AU12,6+12+25, AU20+25,15+17, AU17+23+24,9+17. AU 25, 26, 27	
	Edge-detection	80.5%	Yes/No	AU9+17,12+25.	

In the fourth column, “*No*” means that no AU combination was recognized. “*Yes/Yes*” means that AU combinations were recognized and AUs in combination were recognizable individually. “*Yes/No*” means that AU combinations were recognized but each AU combination was treated as if it were a separate new AU.

kernels such as Gabors, Eigenfaces, and Independent Component Images) outperform explicit parameterization of facial features. Our comparison indicates that a feature-based method performs just as well as the best holistic template-based method and in more complex data. It may be premature to conclude that one or the other approach is superior. Recovering FACS-AUs from video using automatic computer vision techniques is not an easy task and numerous challenges remain [20]. We feel that further efforts will be required for combining both approaches in order to achieve the optimal performance, and that tests with a substantially large database are called for [1].

ACKNOWLEDGMENTS

The authors would like to thank Paul Ekman, at the Human Interaction Laboratory, University of California, San Francisco for providing the Ekman-Hager database. The authors also thank Zara Ambadar, Bethany Peters, and Michelle Lemenager for processing the images. The authors appreciate the helpful comments and suggestions of Marian Bartlett, Simon Baker, Karen Schmidt, and anonymous reviewers. This work was supported by the National Institute of Mental Health grant R01 MH51435.

REFERENCES

- [1] *Facial Expression Coding Project*, cooperation and competition between Carnegie Mellon Univ. and Univ. of California, San Diego, unpublished, 2000.
- [2] M. Bartlett, J. Hager, P. Ekman, and T. Sejnowski, "Measuring Facial Expressions by Computer Image Analysis," *Psychophysiology*, vol. 36, pp. 253-264, 1999.
- [3] M.J. Black and Y. Yacoob, "Tracking and Recognizing Rigid and Nonrigid Facial Motions Using Local Parametric Models of Image Motion," *Proc. Int'l Conf. Computer Vision*, pp. 374-381, 1995.
- [4] M.J. Black and Y. Yacoob, "Recognizing Facial Expressions in Image Sequences Using Local Parameterized Models of Image Motion," *Int'l J. Computer Vision*, vol. 25, no. 1, pp. 23-48, Oct. 1997.
- [5] R. Brunelli and T. Poggio, "Face Recognition: Features versus Templates," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 15, no. 10, pp. 1042-1052, Oct. 1993.
- [6] J. Canny, "A Computational Approach to Edge Detection," *IEEE Trans. Pattern Analysis Machine Intelligence*, vol. 8, no. 6, June 1986.
- [7] J.M. Carroll and J. Russell, "Facial Expression in Hollywood's Portrayal of Emotion," *J. Personality and Social Psychology*, vol. 72, pp. 164-176, 1997.
- [8] J.F. Cohn, A.J. Zlochower, J. Lien, and T. Kanade, "Automated Face Analysis by Feature Point Tracking has High Concurrent Validity with Manual Faces Coding," *Psychophysiology*, vol. 36, pp. 35-43, 1999.
- [9] C. Darwin, *The Expression of Emotions in Man and Animals*, John Murray, reprinted by Univ. of Chicago Press, 1965, 1872.
- [10] G. Donato, M.S. Bartlett, J.C. Hager, P. Ekman, and T.J. Sejnowski, "Classifying Facial Actions," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 21, no. 10, pp. 974-989, Oct. 1999.
- [11] I. Eibl-Eibesfeldt, *Human Ethology*. New York: Aldine de Gruyter, 1989.
- [12] P. Ekman, "Facial Expression and Emotion," *Am. Psychologist*, vol. 48, pp. 384-392, 1993.
- [13] P. Ekman and W.V. Friesen, *Pictures of Facial Affect*. Palo Alto, Calif.: Consulting Psychologist, 1976.
- [14] P. Ekman and W.V. Friesen, *The Facial Action Coding System: A Technique for The Measurement of Facial Movement*. San Francisco: Consulting Psychologists Press, 1978.
- [15] P. Ekman, J. Hager, C.H. Methvin, and W. Irwin, "Ekman-Hager Facial Action Exemplars," unpublished data, Human Interaction Laboratory, Univ. of California, San Francisco.
- [16] I.A. Essa and A.P. Pentland, "Coding, Analysis, Interpretation, and Recognition of Facial Expressions," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 757-763, July 1997.
- [17] J. Fleiss, *Statistical Methods for Rates and Proportions*. New York: Wiley, 1981.
- [18] K. Fukui and O. Yamaguchi, "Facial Feature Point Extraction Method Based on Combination of Shape Extraction and Pattern Matching," *Systems and Computers in Japan*, vol. 29, no. 6, pp. 49-58, 1998.
- [19] C. Izard, L. Dougherty, and E.A. Hembree, "A System for Identifying Affect Expressions by Holistic Judgments," unpublished manuscript, Univ. of Delaware, 1983.
- [20] T. Kanade, J. Cohn, and Y. Tian, "Comprehensive Database for Facial Expression Analysis," *Proc. Int'l Conf. Face and Gesture Recognition*, pp. 46-53, Mar. 2000.
- [21] M. Kirby and L. Sirovich, "Application of the k-l Procedure for the Characterization of Human Faces," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 12, no. 1, pp. 103-108, Jan. 1990.
- [22] Y. Kwon and N. Lobo, "Age Classification from Facial Images," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 762-767, 1994.
- [23] K. Lam and H. Yan, "Locating and Extracting the Eye in Human Face Images," *Pattern Recognition*, vol. 29, no. 5, pp. 771-779, 1996.
- [24] J.-J. Lien, T. Kanade, J.F. Cohn, and C.C. Li, "Detection, Tracking, and Classification of Action Units in Facial Expression," *J. Robotics and Autonomous System*, vol. 31, pp. 131-146, 2000.
- [25] B. Lucas and T. Kanade, "An Interactive Image Registration Technique with an Application in Stereo Vision," *Proc. Seventh Int'l Joint Conf. Artificial Intelligence*, pp. 674-679, 1981.
- [26] K. Mase, "Recognition of Facial Expression from Optical Flow," *IEICE Trans.*, vol. E74, no. 10, pp. 3474-3483, Oct. 1991.
- [27] R.R. Rao, "Audio-Visual Interaction in Multimedia," PhD thesis, Electrical Eng., Georgia Inst. of Technology, 1998.
- [28] M. Rosenblum, Y. Yacoob, and L.S. Davis, "Human Expression Recognition from Motion Using a Radial Basis Function Network Architecture," *IEEE Trans. Neural Network*, vol. 7, no. 5, pp. 1121-1138, 1996.
- [29] H.A. Rowley, S. Baluja, and T. Kanade, "Neural Network-Based Face Detection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, pp. 23-38, Jan. 1998.
- [30] K. Scherer and P. Ekman, *Handbook of Methods in Nonverbal Behavior Research*. Cambridge, UK: Cambridge Univ. Press, 1982.
- [31] M. Suwa, N. Sugie, and K. Fujimora, "A Preliminary Note on Pattern Recognition of Human Emotional Expression," *Proc. Int'l Joint Conf. Pattern Recognition*, pp. 408-410, 1978.
- [32] D. Terzopoulos and K. Waters, "Analysis of Facial Images Using Physical and Anatomical Models," *Proc. IEEE Int'l Conf. Computer Vision*, pp. 727-732, 1990.
- [33] Y. Tian, T. Kanade, and J. Cohn, "Robust Lip Tracking by Combining Shape, Color, and Motion," *Proc. Asian Conf. Computer Vision*, pp. 1040-1045, 2000.
- [34] Y. Tian, T. Kanade, and J. Cohn, "Dual-State Parametric Eye Tracking," *Proc. Int'l Conf. Face and Gesture Recognition*, pp. 110-115, Mar. 2000.
- [35] M. Turk and A. Pentland, "Face Recognition Using Eigenfaces," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 586-591, 1991.
- [36] Y. Yacoob and M.J. Black, "Parameterized Modeling and Recognition of Activities," *Proc. of the Sixth Int'l Conf. Computer Vision*, pp. 120-127, 1998.
- [37] Y. Yacoob and L.S. Davis, "Recognizing Human Facial Expression from Long Image Sequences Using Optical Flow," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 18, no. 6, pp. 636-642, June 1996.
- [38] Y. Yacoob, H. Lam, and L. Davis, "Recognizing Face Showing Expressions," *Proc. Int'l Workshop Automatic Face and Gesture Recognition*, 1995.
- [39] A. Yuille, P. Haallinan, and D.S. Cohen, "Feature Extraction from Faces Using Deformable Templates," *Int'l J. Computer Vision*, vol. 8, no. 2, pp. 99-111, 1992.
- [40] Z. Zhang, "Feature-Based Facial Expression Recognition: Sensitivity Analysis and Experiments with a Multilayer Perceptron," *Int'l J. Pattern Recognition and Artificial Intelligence*, vol. 13, no. 6, pp. 893-911, 1999.



Ying-li Tian received the BS and MS degrees in optical engineering from TianJin University, China, in 1987 and 1990, and the PhD degree in electrical engineering from the Chinese University of Hong Kong, in 1996. After holding a faculty position at National Laboratory of Pattern Recognition, Chinese Academy of Sciences, Beijing, she joined Carnegie Mellon University in 1998, where she is currently a postdoctoral fellow of the Robotics Institute. Her

current research focuses on a wide range of computer vision problems from photometric modeling and shape from shading, to human identification, 3D reconstruction, motion analysis, and facial expression analysis. She is a member of the IEEE.



Takeo Kanade received the Doctoral degree in electrical engineering from Kyoto University, Japan, in 1974. After holding a faculty position at Department of Information Science, Kyoto University, he joined Carnegie Mellon University in 1980. He is currently U.A. Helen Whitaker University Professor of computer science and director of the Robotics Institute. Dr. Kanade has worked in multiple areas of robotics: computer vision, manipulators, autonomous mobile robots, and sensors. He has written more than 200 technical papers and reports in these areas, as well as more than 10 patents. He has been the principal investigator of a dozen major vision and robotics projects at Carnegie Mellon. Dr. Kanade has been elected to the National Academy of Engineering. He is a fellow of the IEEE, the ACM, a founding fellow of the American Association of Artificial Intelligence, and the founding editor of the *International Journal of Computer Vision*. He has received several awards including the C&C Award, the Joseph Engelberger Award, JARA Award, Otto Franc Award, Yokogawa Prize, and Marr Prize Award. Dr. Kanade has served the government, industry, and as a university advisory or on consultant committees, including Aeronautics and Space Engineering Board (ASEB) of the National Research Council, NASA's Advanced Technology Advisory Committee, and the Advisory Board of Canadian Institute for Advanced Research.



Jeffrey F. Cohn received the PhD degree in clinical psychology from the University of Massachusetts at Amherst in 1983. He is an associate professor of psychology and psychiatry at the University of Pittsburgh and an adjunct faculty member at the Robotics Institute, Carnegie Mellon University. Dr. Cohn's primary research interests are emotion and paralinguistic communication, developmental psychopathology, face image and prosodic analysis, and human-computer interaction. He has published more than 60 peer reviewed articles and conference proceedings on these topics. He is member of the IEEE.