
Chapter 11. Facial Expression Analysis

Ying-Li Tian,¹ Takeo Kanade², and Jeffrey F. Cohn^{2,3}

¹ IBM T. J. Watson Research Center, Hawthorne, NY 10532, USA. yltian@us.ibm.com

² Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA. tk@cs.cmu.edu

³ Department of Psychology, University of Pittsburgh, Pittsburgh, PA 15260, USA.
jeffcohn@pitt.edu

1 Principles of Facial Expression Analysis

1.1 What Is Facial Expression Analysis?

Facial expressions are the facial changes in response to a person's internal emotional states, intentions, or social communications. Facial expression analysis has been an active research topic for behavioral scientists since the work of Darwin in 1872 [18, 22, 25, 71]. Suwa et al. [76] presented an early attempt to automatically analyze facial expressions by tracking the motion of 20 identified spots on an image sequence in 1978. After that, much progress has been made to build computer systems to help us understand and use this natural form of human communication [6, 7, 17, 20, 28, 39, 51, 55, 65, 78, 81, 92, 93, 94, 96].

In this chapter, facial expression analysis refers to computer systems that attempt to automatically analyze and recognize facial motions and facial feature changes from visual information. Sometimes the facial expression analysis has been confused with emotion analysis in the computer vision domain. For emotion analysis, higher level knowledge is required. For example, although facial expressions can convey emotion, they can also express intention, cognitive processes, physical effort, or other intra- or interpersonal meanings. Interpretation is aided by context, body gesture, voice, individual differences, and cultural factors as well as by facial configuration and timing [10, 67, 68]. Computer facial expression analysis systems need to analyze the facial actions regardless of context, culture, gender, and so on.

The accomplishments in the related areas such as psychological studies, human movement analysis, face detection, face tracking, and recognition make the automatic facial expression analysis possible. Automatic facial expression analysis can be applied in many areas such as emotion and paralinguistic communication, clinical psychology, psychiatry, neurology, pain assessment, lie detection, intelligent environments, and multimodal human computer interface (HCI).

1.2 Basic Structure of Facial Expression Analysis Systems

Facial expression analysis includes both measurement of facial motion and recognition of expression. The general approach to automatic facial expression analysis (AFEA) consists of

three steps (Fig. 11.1): face acquisition, facial data extraction and representation, and facial expression recognition.

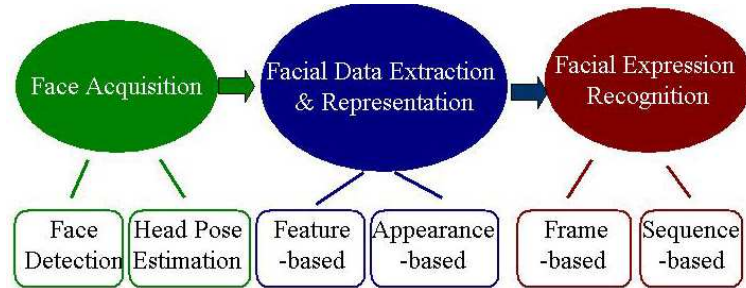


Fig. 11.1. Basic structure of facial expression analysis systems.

Face acquisition is a processing stage to automatically find the face region for the input images or sequences. It can be a detector to detect face for each frame or just detect face in the first frame and then track the face in the remainder of the video sequence. To handle large head motion, the the head finder, head tracking, and pose estimation can be applied to a facial expression analysis system.

After the face is located, the next step is to extract and represent the facial changes caused by facial expressions. In facial feature extraction for expression analysis, there are mainly two types of approaches: geometric feature-based methods and appearance-based methods. The geometric facial features present the shape and locations of facial components (including mouth, eyes, brows, and nose). The facial components or facial feature points are extracted to form a feature vector that represents the face geometry. With appearance-based methods, image filters, such as Gabor wavelets, are applied to either the whole-face or specific regions in a face image to extract a feature vector. Depending on the different facial feature extraction methods, the effects of in-plane head rotation and different scales of the faces can be eliminated by face normalization before the feature extraction or by feature representation before the step of expression recognition.

Facial expression recognition is the last stage of AFEA systems. The facial changes can be identified as facial action units or prototypic emotional expressions (see Section 2.1 for definitions). Depending on if the temporal information is used, in this chapter we classified the recognition approaches as frame-based or sequence-based.

1.3 Organization of the Chapter

This chapter introduces recent advances in facial expression analysis. The first part discusses general structure of AFEA systems. The second part describes the problem space for facial expression analysis. This space includes multiple dimensions: level of description, individual differences in subjects, transitions among expressions, intensity of facial expression, deliberate versus spontaneous expression, head orientation and scene complexity, image acquisition and resolution, reliability of ground truth, databases, and the relation to other facial behaviors or

nonfacial behaviors. We note that most work to date has been confined to a relatively restricted region of this space. The last part of this chapter is devoted to a description of more specific approaches and the techniques used in recent advances. They include the techniques for face acquisition, facial data extraction and representation, and facial expression recognition. The chapter concludes with a discussion assessing the current status, future possibilities, and open questions about automatic facial expression analysis.



Fig. 11.2. Emotion-specified facial expression (posed images from database [43]). 1, disgust; 2, fear; 3, joy; 4, surprise; 5, sadness; 6, anger. From Schmidt and Cohn [72], with permission.

2 Problem Space for Facial Expression Analysis































2.1 Level of Description

With few exceptions [17, 20, 30, 81], most AFEA systems attempt to recognize a small set of prototypic emotional expressions as shown in Fig. 11.2, (i.e., disgust, fear, joy, surprise, sadness, anger). This practice may follow from the work of Darwin [18] and more recently Ekman and Friesen [23, 24] and Izard et al. [42] who proposed that emotion-specified expressions 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 tightening of the lips in anger or obliquely lowering the lip corners in sadness [11]. 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 [21]. To capture such subtlety of human emotion and paralinguistic communication, automated recognition of fine-grained changes in facial expression is needed. The facial action coding system (FACS: [25]) is a human-observer-based system designed to detect subtle changes in facial features. Viewing videotaped facial behavior in slow motion, trained observers can manually FACS code all possible facial displays, which are referred to as action units and may occur individually or in combinations.

FACS consists of 44 action units. Thirty are anatomically related to contraction of a specific set of facial muscles (Table 11.1) [22]. The anatomic basis of the remaining 14 is unspecified (Table 11.2). These 14 are referred to in FACS as miscellaneous actions. Many action units may be coded as symmetrical or asymmetrical. For action units that vary in intensity, a 5-point ordinal scale is used to measure the degree of muscle contraction. Table 11.3 shows some examples of combinations of FACS action units.

Although Ekman and Friesen proposed that specific combinations of FACS action units represent prototypic expressions of emotion, emotion-specified expressions are not part of FACS; they are coded in separate systems, such as the emotional facial action system (EMFACS) [37].

Table 11.1. FACS action units (AU). AUs with "*" indicate that the criteria have changed for this AU, that is, AU 25, 26, and 27 are now coded according to criteria of intensity (25A-E), and AU 41, 42, and 43 are now coded according to criteria of intensity.

Upper Face Action Units					
AU 1	AU 2	AU 4	AU 5	AU 6	AU 7
					
Inner Brow Raiser	Outer Brow Raiser	Brow Lowerer	Upper Lid Raiser	Cheek Raiser	Lid Tightener
*AU 41	*AU 42	*AU 43	AU 44	AU 45	AU 46
					
Lid Droop	Slit	Eyes Closed	Squint	Blink	Wink
Lower Face Action Units					
AU 9	AU 10	AU 11	AU 12	AU 13	AU 14
					
Nose Wrinkler	Upper Lip Raiser	Nasolabial Deepener	Lip Corner Puller	Cheek Puffer	Dimpler
AU 15	AU 16	AU 17	AU 18	AU 20	AU 22
					
Lip Corner Depressor	Lower Lip Depressor	Chin Raiser	Lip Puckerer	Lip Stretcher	Lip Funneler
AU 23	AU 24	*AU 25	*AU 26	*AU 27	AU 28
					
Lip Tightener	Lip Pressor	Lips Part	Jaw Drop	Mouth Stretch	Lip Suck

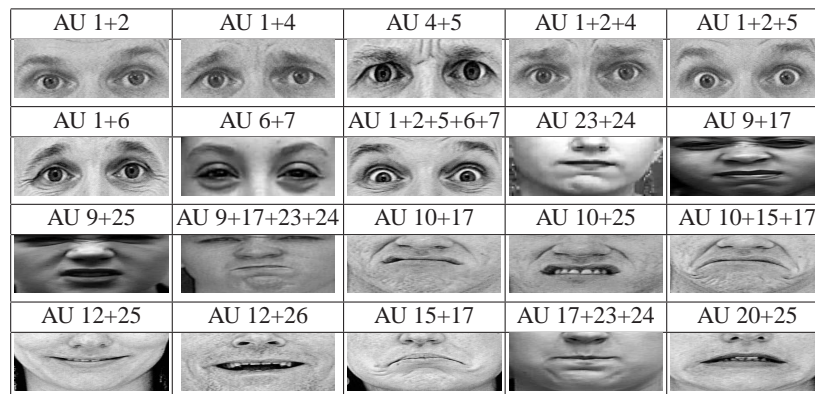
FACS itself is purely descriptive and includes no inferential labels. By converting FACS codes to EMFACS or similar systems, face images may be coded for emotion-specified expressions (e.g., joy or anger) as well as for more molar categories of positive or negative emotion [56].

2.2 Individual Differences in Subjects

Face shape, texture, color, and facial and scalp hair vary with sex, ethnic background, and age [29, 99]. Infants, for instance, have smoother, less textured skin and often lack facial hair in the brows or scalp. The eye opening and contrast between iris and sclera differ markedly between Asians and Northern Europeans, which may affect the robustness of eye tracking and facial feature analysis more generally. Beards, eyeglasses, or jewelry may obscure facial features. Such individual differences in appearance may have important consequences for face analysis. Few attempts to study their influence exist. An exception was a study by Zlochower et al. [99], who found that algorithms for optical flow and high-gradient component detection that had been optimized for young adults performed less well when used in infants. The reduced texture

Table 11.2. Miscellaneous Actions.

AU	Description
8	Lips toward
19	Tongue show
21	Neck tighten
29	Jaw thrust
30	Jaw sideways
31	Jaw clench
32	Bite lip
33	Blow
34	Puff
35	Cheek suck
36	Tongue bulge
37	Lip wipe
38	Nostril dilate
39	Nostril compress

Table 11.3. Some examples of combination of FACS action units.

of infants' skin, their increased fatty tissue, juvenile facial conformation, and lack of transient furrows may all have contributed to the differences observed in face analysis between infants and adults.

In addition to individual differences in appearance, there are individual differences in expressiveness, which refers to the degree of facial plasticity, morphology, frequency of intense expression, and overall rate of expression. Individual differences in these characteristics are well established and are an important aspect of individual identity [53] (these individual differences in expressiveness and in biases for particular facial actions are sufficiently strong that they may be used as a biometric to augment the accuracy of face recognition algorithms [16]). An extreme example of variability in expressiveness occurs in individuals who have incurred damage either to the facial nerve or central nervous system [63, 85]. To develop algorithms that are robust to individual differences in facial features and behavior, it is essential to include

a large sample of varying ethnic background, age, and sex, which includes people who have facial hair and wear jewelry or eyeglasses and both normal and clinically impaired individuals.

2.3 Transitions Among Expressions

A simplifying assumption in facial expression analysis is that expressions are singular and begin and end with a neutral position. In reality, facial expression is more complex, especially at the level of action units. Action units may occur in combinations or show serial dependence. Transitions from action units or combination of actions to another may involve no intervening neutral state. Parsing the stream of behavior is an essential requirement of a robust facial analysis system, and training data are needed that include dynamic combinations of action units, which may be either additive or nonadditive.

As shown in Table 11.3, an example of an additive combination is smiling (AU 12) with mouth open, which would be coded as AU 12+25, AU 12+26, or AU 12+27 depending on the degree of lip parting and whether and how far the mandible was lowered. In the case of AU 12+27, for instance, the facial analysis system would need to detect transitions among all three levels of mouth opening while continuing to recognize AU 12, which may be simultaneously changing in intensity.

Nonadditive combinations represent further complexity. Following usage in speech science, we refer to these interactions as co-articulation effects. An example is the combination AU 12+15, which often occurs during embarrassment. Although AU 12 raises the cheeks and lip corners, its action on the lip corners is modified by the downward action of AU 15. The resulting appearance change is highly dependent on timing. The downward action of the lip corners may occur simultaneously or sequentially. The latter appears to be more common [73]. To be comprehensive, a database should include individual action units and both additive and nonadditive combinations, especially those that involve co-articulation effects. A classifier trained only on single action units may perform poorly for combinations in which co-articulation effects occur.

2.4 Intensity of Facial Expression

Facial actions can vary in intensity. Manual FACS coding, for instance, uses a 3- or more recently a 5-point intensity scale to describe intensity variation of action units (for psychometric data, see Sayette et al. [70]). Some related action units, moreover, function as sets to represent intensity variation. In the eye region, action units 41, 42, and 43 or 45 can represent intensity variation from slightly drooped to closed eyes. Several computer vision researchers proposed methods to represent intensity variation automatically. Essa and Pentland [28] represented intensity variation in smiling using optical flow. Kimura and Yachida [44] and Lien et al. [50] quantified intensity variation in emotion-specified expression and in action units, respectively. These authors did not, however, attempt the more challenging step of discriminating intensity variation within types of facial actions. Instead, they used intensity measures for the more limited purpose of discriminating between different types of facial actions. Bartlett and colleagues [4] tested their algorithms on facial expressions that systematically varied in intensity as measured by manual FACS coding. Although they failed to report results separately for each level of intensity variation, their overall findings suggest some success. Tian et al. [83] may be the only group to compare manual and automatic coding of intensity variation. Using Gabor features

and an artificial neural network, they discriminated intensity variation in eye closure as reliably as did human coders. These findings suggest that it is feasible to automatically recognize intensity variation within types of facial actions. Regardless of whether investigators attempt to discriminate intensity variation within facial actions, it is important that the range of variation be described adequately. Methods that work for intense expressions may generalize poorly to ones of low intensity.

2.5 Deliberate Versus Spontaneous Expression

Most face expression data have been collected by asking subjects to perform a series of expressions. These directed facial action tasks may differ in appearance and timing from spontaneously occurring behavior [27]. Deliberate and spontaneous facial behavior are mediated by separate motor pathways, the pyramidal and extrapyramidal motor tracks, respectively [63]. As a consequence, fine-motor control of deliberate facial actions is often inferior and less symmetrical than what occurs spontaneously. Many people, for instance, are able to raise their outer brows spontaneously while leaving their inner brows at rest; few can perform this action voluntarily. Spontaneous depression of the lip corners (AU 15) and raising and narrowing the inner corners of the brow (AU 1+4) are common signs of sadness. Without training, few people can perform these actions deliberately, which incidentally is an aid to lie detection [27]. Differences in the temporal organization of spontaneous and deliberate facial actions are particularly important in that many pattern recognition approaches, such as hidden Markov modelling, are highly dependent on the timing of the appearance change. Unless a database includes both deliberate and spontaneous facial actions, it will likely prove inadequate for developing face expression methods that are robust to these differences.

2.6 Head Orientation and Scene Complexity

Face orientation relative to the camera, the presence and actions of other people, and background conditions may influence face analysis. In the face recognition literature, face orientation has received deliberate attention. The FERET database [64], for instance, includes both frontal and oblique views, and several specialized databases have been collected to try to develop methods of face recognition that are invariant to moderate change in face orientation [86]. In the face expression literature, use of multiple perspectives is rare; and relatively less attention has been focused on the problem of pose invariance. Most researchers assume that face orientation is limited to in-plane variation [4] or that out-of-plane rotation is small [51, 58, 65, 81]. In reality, large out-of-plane rotation in head position is common and often accompanies change in expression. Kraut and Johnson [48] found that smiling typically occurs while turning toward another person. Camras et al. [9] showed that infant surprise expressions often occur as the infant pitches her head back. To develop pose invariant methods of face expression analysis, image data are needed in which facial expression changes in combination with significant non-planar change in pose. Some efforts have been made to handle large out-of-plane rotation in head position [80, 90, 91].

Scene complexity, such as background and the presence of other people, potentially influences accuracy of face detection, feature tracking, and expression recognition. Most databases use image data in which the background is neutral or has a consistent pattern and only a single

person is present in the scene. In natural environments, multiple people interacting with each other are likely, and their effects need to be understood. Unless this variation is represented in training data, it will be difficult to develop and test algorithms that are robust to such variation.

2.7 Image Acquisition and Resolution

The image acquisition procedure includes several issues, such as the properties and number of video cameras and digitizer, the size of the face image relative to total image dimensions, and the ambient lighting. All of these factors may influence facial expression analysis. Images acquired in low light or at coarse resolution can provide less information about facial features. Similarly, when the face image size is small relative to the total image size, less information is available. NTSC cameras record images at 30 frames per second, The implications of down-sampling from this rate are unknown. Many algorithms for optical flow assume that pixel displacement between adjacent frames is small. Unless they are tested at a range of sampling rates, the robustness to sampling rate and resolution cannot be assessed.

Within an image sequence, changes in head position relative to the light source and variation in ambient lighting have potentially significant effects on face expression analysis. A light source above the subject's head causes shadows to fall below the brows, which can obscure the eyes, especially for subjects with pronounced bone structure or hair. Methods that work well in studio lighting may perform poorly in more natural lighting (e.g., through an exterior window) when the angle of lighting changes across an image sequence. Most investigators use single-camera setups, which is problematic when a frontal orientation is not required. With image data from a single camera, out-of-plane rotation may be difficult to standardize. For large out-of-plane rotation, multiple cameras may be required. Multiple camera setups can support three dimensional (3D) modelling and in some cases ground truth with which to assess the accuracy of image alignment. Pantic and Rothkrantz [60] were the first to use two cameras mounted on a headphone-like device; one camera is placed in front of the face and the other on the right side of the face. The cameras are moving together with the head to eliminate the scale and orientation variance of the acquired face images.

Table 11.4. A face at different resolutions. All images are enlarged to the same size. At 48 x 64 pixels the facial features such as the corners of the eyes and the mouth become hard to detect. Facial expressions are not recognized at 24 x 32 pixels [80].





Face				
Process	96 x 128	69 x 93	48 x 64	24 x 32
Detect?	Yes	Yes	Yes	Yes
Pose?	Yes	Yes	Yes	Yes
Recognize?	Yes	Yes	Yes	Maybe
Features?	Yes	Yes	Maybe	No
Expressions?	Yes	Yes	Maybe	No

Image resolution is another concern. Professional grade PAL cameras, for instance, provide very high resolution images. By contrast, security cameras provide images that are seriously degraded. Although postprocessing may improve image resolution, the degree of potential improvement is likely limited. Also the effects of post processing for expression recognition are not known. Table 11.4 shows a face at different resolutions. Most automated face processing tasks should be possible for a 69 x 93 pixel image. At 48 x 64 pixels the facial features such as the corners of the eyes and the mouth become hard to detect. The facial expressions may be recognized at 48 x 64 and are not recognized at 24 x 32 pixels. Algorithms that work well at optimal resolutions of full face frontal images and studio lighting can be expected to perform poorly when recording conditions are degraded or images are compressed. Without knowing the boundary conditions of face expression algorithms, comparative performance is difficult to assess. Algorithms that appear superior within one set of boundary conditions may perform more poorly across the range of potential applications. Appropriate data with which these factors can be tested are needed.

2.8 Reliability of Ground Truth

When training a system to recognize facial expression, the investigator assumes that training and test data are accurately labeled. This assumption may or may not be accurate. Asking subjects to perform a given action is no guarantee that they will. To ensure internal validity, expression data must be manually coded, and the reliability of the coding verified. Interobserver reliability can be improved by providing rigorous training to observers and monitoring their performance. FACS coders must pass a standardized test, which ensures (initially) uniform coding among international laboratories. Monitoring is best achieved by having observers independently code a portion of the same data. As a general rule, 15% to 20% of data should be comparison-coded. To guard against drift in coding criteria [54], re-standardization is important. When assessing reliability, coefficient kappa [32] is preferable to raw percentage of agreement, which may be inflated by the marginal frequencies of codes. Kappa quantifies interobserver agreement after correcting for the level of agreement expected by chance.

2.9 Databases

Because most investigators have used relatively limited data sets, the generalizability of different approaches to facial expression analysis remains unknown. In most data sets, only relatively global facial expressions (e.g., joy or anger) have been considered, subjects have been few in number and homogeneous with respect to age and ethnic background, and recording conditions have been optimized. Approaches to facial expression analysis that have been developed in this way may transfer poorly to applications in which expressions, subjects, contexts, or image properties are more variable. In the absence of comparative tests on common data, the relative strengths and weaknesses of different approaches are difficult to determine. In the areas of face and speech recognition, comparative tests have proven valuable [64], and similar benefits would likely accrue in the study of facial expression analysis. A large, representative test-bed is needed with which to evaluate different approaches. We have built a common database (Cohn-Kanade AU-Coded Face Expression Image Database) with which multiple laboratories may conduct comparative tests of their methods [43]. The details of the Cohn-Kanade AU-Coded Face Expression Image Database can be found in Chapter 12.

2.10 Relation to Other Facial Behavior or Nonfacial Behavior

Facial expression is one of several channels of nonverbal communication. Contraction of the muscle *zygomaticus major* (AU 12), for instance, often is associated with positive or happy vocalizations, and smiling tends to increase vocal fundamental frequency [15]. Also facial expressions often occur during conversations. Both expressions and conversations can cause facial changes. Few research groups, however, have attempted to integrate gesture recognition broadly defined across multiple channels of communication. An important question is whether there are advantages to early rather than late integration [35]. Databases containing multimodal expressive behavior afford the opportunity for integrated approaches to analysis of facial expression, prosody, gesture, and kinetic expression.

2.11 Summary and Ideal Facial Expression Analysis Systems

The problem space for facial expression includes multiple dimensions. An ideal facial expression analysis system has to address all these dimensions, and it outputs accurate recognition results. In addition, the ideal facial expression analysis system must perform automatically and in real-time for all stages (Fig. 11.1). So far, several systems can recognize expressions in real time [47, 58, 80]. We summarize the properties of an ideal facial expression analysis system in Table 11.5.

Table 11.5. Properties of an ideal facial expression analysis system

Robustness	
Rb1	Deal with subjects of different age, gender, ethnicity
Rb2	Handle lighting changes
Rb3	Handle large head motion
Rb4	Handle occlusion
Rb5	Handle different image resolution
Rb6	Recognize all possible expressions
Rb7	Recognize expressions with different intensity
Rb8	Recognize asymmetrical expressions
Rb9	Recognize spontaneous expressions
Automatic process	
Am1	Automatic face acquisition
Am2	Automatic facial feature extraction
Am3	Automatic expression recognition
Real-time process	
Rt1	Real-time face acquisition
Rt2	Real-time facial feature extraction
Rt3	Real-time expression recognition
Autonomic Process	
An1	Output recognition with confidence
An2	Adaptive to different level outputs based on input images

3 Recent Advances in Automatic Facial Expression Analysis

For automatic facial expression analysis, Suwa et al. [76] presented an early attempt in 1978 to analyze facial expressions by tracking the motion of 20 identified spots on an image sequence. Considerable progress had been made since 1990 in related technologies such as image analysis and pattern recognition that make AFEA possible. Samal and Iyengar [69] surveyed the early work (before 1990) about automatic recognition and analysis of human face and facial expression. Recently, two survey papers summarized the work (before year 1999) of facial expression analysis [31, 59]. In this chapter, instead of giving a comprehensive survey of facial expression analysis literature, we explore the recent advances in facial expression analysis based on three problems: (1) face acquisition, (2) facial feature extraction and representation, and (3) facial expression recognition.

The most extensive work about facial expression analysis includes the systems of CMU (Carnegie Mellon University) [81, 82, 14, 57, 90], UCSD (University of California, San Diego) [3, 20, 33], UIUC (University of Illinois at Urbana-Champaign) [13, 88], MIT (Massachusetts Institute of Technology) [28], UMD (University of Maryland) [7, 93], TUDELFT (Delft University of Technology) [60, 59], IDIAP (Dalle Molle Institute for Perceptual Artificial Intelligence) [30, 31], and others [52, 97]. Because most of the work of MIT, UMD, TUDELFT, IDIAP, and others are summarized in the survey papers [31, 59], here we summarize the systems of CMU, UCSD, and UIUC. We focus on the new developments after year 2000. The recent research in automatic facial expression analysis tends to follow these directions:

- Build more robust systems for face acquisition, facial data extraction and representation, and facial expression recognition to handle head motion (in-plane and out-of-plane), occlusion, lighting changes, and lower intensity of expressions
- Employ more facial features to recognize more expressions and to achieve a higher recognition rate
- Recognize facial action units and their combinations rather than emotion-specified expressions
- Recognize action units as they occur spontaneously
- Develop fully automatic and real-time AFEA systems

Table 11.6 summarizes the basic properties of the systems from the above three universities. Altogether, six systems are compared. They are CMU S1 [81, 82, 83] and CMU S2 [14, 57, 90], UCSD S1 [33] and UCSD S2 [3], and UIUC S1 [13] and UIUC S2 [88]. For each system, we list the methods used for face acquisition, feature extraction, and facial expression recognition.

3.1 Face Acquisition

With few exceptions, most AFEA research attempts to recognize facial expressions only from frontal-view or near frontal-view faces [45, 60]. Kleck and Mendolia [45] first studied the decoding of profile versus full-face expressions of affect by using three perspectives (a frontal face, a 90° right profile, and a 90° left profile). Forty-eight decoders viewed the expressions from 64 subjects in one of the three facial perspectives. They found that the frontal faces elicited higher intensity ratings than profile views for negative expressions. The opposite was found for positive expressions. Pantic and Rothkrantz [60] used dual-view facial images (a full-face and

Table 11.6. Recent Advances of Facial Expression Analysis

	Method	Property		
		Real time	Fully automatic	Others
CMU S1 [81, 82, 83]	Detect face in 1st frame	Yes	Yes	Handle limited head motion; faces with glasses and hair; recognize more than 30 AUs and combinations; 1st frame is frontal and expressionless
	Initialize features in 1st frame	No	No	
	Track geometric features	Yes	Yes	
	Extract appearance features	Yes	Yes	
	Neural network classifier	Yes	Yes	
CMU S2 [14, 57, 90]	Detect face in 1st frame	Yes	Yes	Handle large head motion; handle lighting changes; handle occlusions; spontaneous expressions; 1st frame is frontal and expressionless
	Map face model in 1st frame	Yes	Yes	
	Track 3D head	Yes	Yes	
	Stabilize facial region	Yes	Yes	
	Extract geometric features	Yes	Yes	
	Rule-based classifier	Yes	Yes	
UCSD S1 [33]	Detect face in each frame	Yes	Yes	Frontal view; recognize basic expressions; neutral face needed
	Resize and crop face	Yes	Yes	
	Extract appearance features	Yes	Yes	
	SVM classifier	Yes	Yes	
UCSD S2 [3]	Estimate 3D head pose	No	No	Handle large head motion; neutral face needed; recognize single AUs; spontaneous expressions
	Stabilize facial region	Yes	Yes	
	Extract appearance features	Yes	Yes	
	HMM-SVM classifier	Yes	Yes	
UIUC S1 [13]	Initialize face model	No	No	Handle limited head motion; recognize basic expressions; 1st frame is frontal and expressionless
	Track face in 3D	Yes	Yes	
	Extract geometric features	Yes	Yes	
	HMM classifier	Yes	Yes	
UIUC S2 [88]	Initialize face model	No	No	Handle limited head motion; handle lighting changes; recognize basic expressions; 1st frame is frontal and expressionless
	Track face in 3D	Yes	Yes	
	Extract geometric features	Yes	Yes	
	Extract appearance features	No	Yes	
	NN-HMM classifier	Yes	Yes	

a 90° right profile) which are acquired by two cameras mounted on the user's head. They did not compare the recognition results by using only the frontal view and the profile. So far, it is unclear how many expressions can be recognized by side-view or profile faces. Because the frontal-view face is not always available in real environments, the face acquisition methods should detect both frontal and nonfrontal view faces in an arbitrary scene.

To handle out-of-plane head motion, face can be obtained by face detection, 2D or 3D face tracking, or head pose detection. Nonfrontal view faces are warped or normalized to frontal view for expression analysis.

Face Detection

Many face detection methods has been developed to detect faces in an arbitrary scene [41, 49, 61, 66, 74, 75, 87]. Most of them can detect only frontal and near-frontal views of faces. Heisele

et al. [41] developed a component-based, trainable system for detecting frontal and near-frontal views of faces in still gray images. Rowley et al. [66] developed a neural network based system to detect frontal-view face. Viola and Jones [87] developed a robust real-time face detector based on a set of rectangle features.

To handle out-of-plane head motion, some researchers developed face detectors to detect face from different views [61, 74, 49]. Pentland et al. [61] detected faces by using the view-based and modular eigenspace method. Their method runs real-time and can handle varying head positions. Schneiderman and Kanade [74] proposed a statistical method for 3D object detection that can reliably detect human faces with out-of-plane rotation. They represent the statistics of both *object* appearance and *nonobject* appearance using a product of histograms. Each histogram represents the joint statistics of a subset of wavelet coefficients and their position on the object. Li et al. [49] developed an AdaBoost-like approach to detect faces with multiple views. A detail survey about face detection can be found in paper [95]. Chapter 2 is about face detection, so we here do not describe the details.

The UCSD S1 uses the face detector developed by Viola and Jones [87] to detect faces for each frame. The CMU systems (CMU S1 and CMU S2) assume that the first frame of the sequence is frontal and expressionless. They detect faces only in the first frame by a neural network-based face detector [66]. Then they perform feature tracking or head tracking for the remaining frames of the sequence.

Head Pose Estimation

In a real environment, out-of-plane head motion is common for facial expression analysis. To handle the out-of-plane head motion, head pose estimation can be employed. The methods for estimating head pose can be classified as 3D model-based methods [1, 77, 84, 90] and 2D image-based methods [8, 80, 89, 98].

3D Model-Based Method. The CMU S2, UCSD S2, UIUC S1, and UIUC S2 systems employ a 3D model based method to estimate head pose. UCSD S2 uses a canonical wire-mesh face model to estimate face geometry and 3D pose from hand-labeled feature points [3]. The UIUC systems use an explicit 3D wireframe face model to track geometric facial features defined on the model [77]. The 3D model is fitted to the first frame of the sequence by manually selecting landmark facial features such as corners of the eyes and mouth. The generic face model, which consists of 16 surface patches, is warped to fit the selected facial features. To estimate the head motion and deformations of facial features, a two-step process is used. The 2D image motion is tracked using template matching between frames at different resolutions. From the 2D motions of many points on the face model, the 3D head motion then is estimated by solving an overdetermined system of equations of the projective motions in the least-squares sense [13].

In CMU S2 [90], a cylindrical head model is used to automatically estimate the 6 degrees of freedom (dof) of head motion in realtime. An active appearance model (AAM) method is used to automatically map the cylindrical head model to the face region, which is detected by face detection [66], as the initial appearance template. For any given frame, the template is the head image in the previous frame that is projected onto the cylindrical model. Then the template is registered with the head appearance in the given frame to recover the full motion of the head. They first use the iteratively reweighted least squares technique [5] to deal with nonrigid motion and occlusion. Second, they update the template dynamically in order to deal

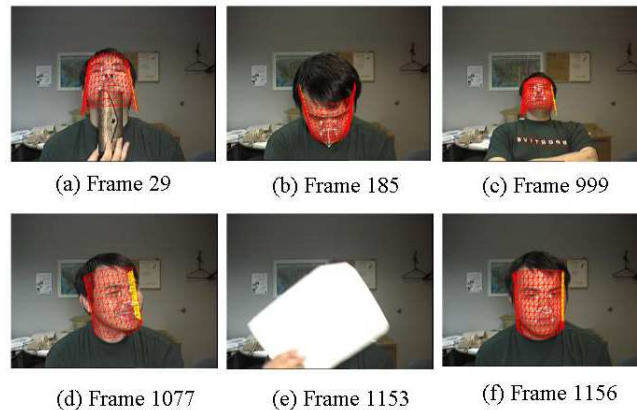



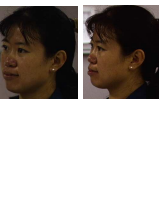

Fig. 11.3. Example of the CMU S2 3D head tracking, including re-registration after losing the head [90].

with gradual changes in lighting and self-occlusion. This enables the system to work well even when most of the face is occluded. Because head poses are recovered using templates that are constantly updated and the pose estimated for the current frame is used in estimating the pose in the next frame, errors would accumulate unless otherwise prevented. To solve this problem, the system automatically selects and stores one or more frames and associated head poses from the tracked images in the sequence (usually including the initial frame and pose) as references. Whenever the difference between the estimated head pose and that of a reference frame is less than a preset threshold, the system rectifies the current pose estimate by re-registering this frame with the reference. The re-registration prevents errors from accumulating and enables the system to recover head pose when the head reappears after occlusion, such as when the head moves momentarily out of the camera's view. On-line tests suggest that the system could work robustly for an indefinite period of time. It was also quantitatively evaluated in image sequences that include maximum pitch and yaw as large as 40 and 75 degrees, respectively. The precision of recovered motion was evaluated with respect to the ground truth obtained by a precise position and orientation measurement device with markers attached to the head and found to be highly consistent (e.g., for maximum yaw of 75 degrees, absolute error averaged 3.86 degrees). An example of the CMU S2 3D head tracking is shown in Figure 11.3 including re-registration after losing the head. For details, see Xiao et al. [90].

2D Image-Based Method. To handle the full range of head motion for expression analysis, Tian et al. [80] detected the head instead of the face. The head detection uses the smoothed silhouette of the foreground object as segmented using background subtraction and computing the *negative curvature minima* (NCM) points of the silhouette. Other head detection techniques that use silhouettes can be found elsewhere [38, 40].

After the head is located, the head image is converted to gray-scale, histogram-equalized, and resized to the estimated resolution. Then a three-layer neural network (NN) is employed to estimate the head pose. The inputs to the network are the processed head image. The outputs are the three head poses: (1) frontal or near frontal view, (2) side view or profile, (3) others,

Table 11.7. Definitions and examples of the three head pose classes: frontal or near frontal view, side view or profile, and others, such as back of the head or occluded faces. The expression analysis process is applied to only the frontal and near-frontal view faces [80].

Poses	Frontal or near frontal	Side view or profile	Others
Definitions	Both eyes and lip corners are visible	One eye or one lip corner is occluded	Not enough facial features
Examples			

such as back of the head or occluded face (Table 11.7). In the frontal or near frontal view, both eyes and lip corners are visible. In the side view or profile, at least one eye or one corner of the mouth becomes self-occluded because of the head. The expression analysis process is applied only to the frontal and near-frontal view faces. Their system performs well even with very low resolution of face images.

3.2 Facial Feature Extraction and Representation

After the face is obtained, the next step is to extract facial features. Two types of features can be extracted: geometric features and appearance features. Geometric features present the shape and locations of facial components (including mouth, eyes, brows, nose). The facial components or facial feature points are extracted to form a feature vector that represents the face geometry. The appearance features present the appearance (skin texture) changes of the face, such as wrinkles and furrows. The appearance features can be extracted on either the whole-face or specific regions in a face image.

To recognize facial expressions, an AEFA system can use geometric features only [13, 60], appearance features only [3, 33, 52], or hybrid features (both geometric and appearance features) [20, 81, 82, 88, 97]. The research shows that using hybrid features can achieve better results for some expressions.

To remove the effects of variation in face scale, motion, lighting, and other factors, one can first align and normalize the face to a standard face (2D or 3D) manually or automatically [3, 20, 33, 51, 88, 97], and then obtain normalized feature measurements by using a reference image (neutral face) [81].

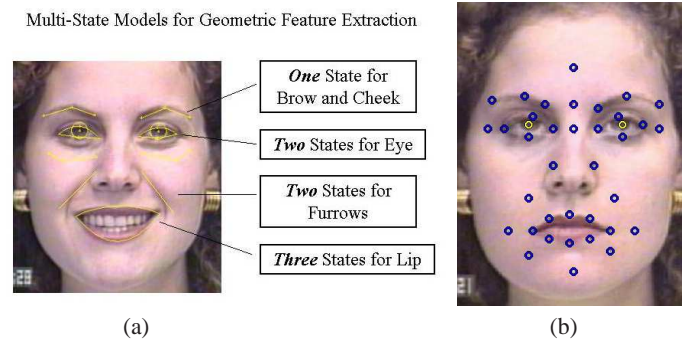


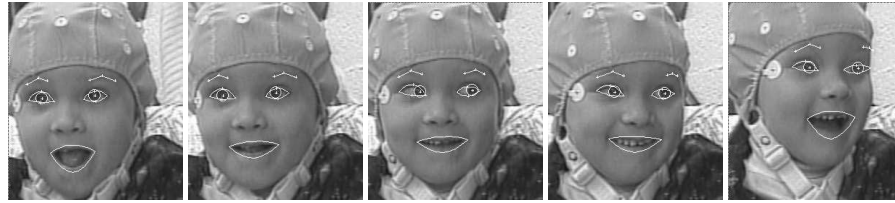
Fig. 11.4. Feature extraction of CMU S1. a. Multistate models for geometric feature extraction. b. Locations for calculating appearance features.

Geometric Feature Extraction

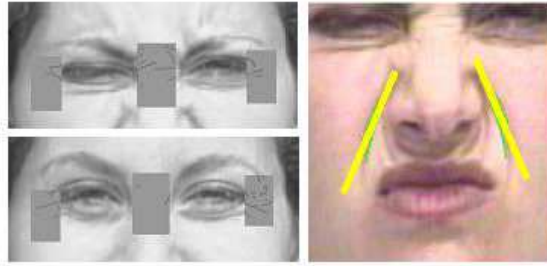
In CMU S1, to detect and track changes of facial components in near frontal face images, multi-state models are developed to extract the geometric facial features (Fig. 11.4). A three-state lip model describes the lip state: open, closed, tightly closed. A two-state model (open or closed) is used for each of the eyes. Each brow and cheek has a one-state model. Some appearance features, such as *nasolabial furrows* and *crows-feet wrinkles* (Fig. 11.5b), are represented explicitly by using two states: present and absent. Given an image sequence, the region of the face and approximate location of individual face features are detected automatically in the initial frame [66]. The contours of the face features and components then are adjusted manually in the initial frame. After the initialization, all face feature changes are automatically detected and tracked in the image sequence. CMU S1 system groups 15 parameters for the upper face and 9 parameters for the lower face, which describe shape, motion, and state of face components and furrows. 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 reference frame. Details of geometric feature extraction and representation can be found in paper [81].

In CMU S2, the 3D head tracking is applied to handle large out-of plane head motion (Section 3.1) and track nonrigid features. The manual preprocessing of the geometric feature initialization in CMU S1 is reduced by using automatic active appearance model (AAM) mapping. Once the head pose is recovered, the face region is stabilized by transforming the image to a common orientation for expression recognition. Currently, only the geometric features of the eyebrow are extracted [14, 57].

The UIUC systems use an explicit 3D wireframe face model to track geometric facial features defined on the model [77]. The 3D model is fitted to the first frame of the sequence by manually selecting landmark facial features such as corners of the eyes and mouth. The generic face model, which consists of 16 surface patches, is warped to fit the selected facial features. Figure 11.6b shows an example of the geometric feature extraction of UIUC systems [88].



(a) Permanent features



(b) Transient features

Fig. 11.5. Results of feature extraction in CMU S1 [81]. a. Permanent feature extraction (eyes, brows, and mouth). b. Transient feature extraction (crows-foot wrinkles, wrinkles at nasal root, and nasolabial furrows).

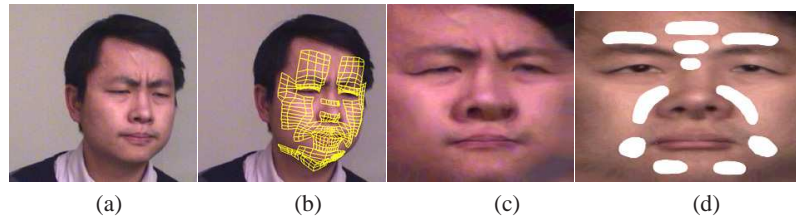


Fig. 11.6. Feature extraction of UIUC S1. a. Input video frame. b. Snapshot of the geometric tracking system. c. Extracted texture map. d. Selected facial regions for appearance feature extraction [88].

Appearance Feature Extraction

Gabor wavelets [19] are widely used to extract the facial appearance changes as a set of multiscale and multiorientation coefficients. The Gabor filter may be applied to specific locations on a face [52, 82, 83, 97] or to the whole face image [3, 20, 33]. Zhang et al. [97] was the first to compare two type of features to recognize expressions, the geometric positions of 34 fiducial points on a face and 612 Gabor wavelet coefficients extracted from the face image at these 34 fiducial points. The recognition rates for six emotion-specified expressions (e.g., joy and anger) were significantly higher for Gabor wavelet coefficients. Donato et al. [20] compared several techniques for recognizing six single upper face AUs and six lower face AUs.

These techniques include optical flow, principal component analysis, independent component analysis, local feature analysis, and Gabor wavelet representation. The best performances were obtained using a Gabor wavelet representation and independent component analysis. All of these systems [20, 97] used a manual step to align each input image with a standard face image using the center of the eyes and mouth.

In CMU S1, Tian et al. [82] studied geometric features and Gabor coefficients to recognize single AU and AU combinations. Following Zhang et al. [97], they used 480 Gabor coefficients in the upper face for 20 locations and 432 Gabor coefficients in the lower face for 18 locations (Figure 11.4). They found that Gabor wavelets work well for single AU recognition for homogeneous subjects without head motion. However, for recognition of AU combinations when image sequences include nonhomogeneous subjects with small head motions, the recognition results are relatively poor if we use only Gabor appearance features. Several factors may account for the difference. First, the previous studies used homogeneous subjects. For instance, Zhang et al. [97] included only Japanese and Donato et al. [20] included only Euro-Americans. The CMU S1 uses diverse subjects of European, African, and Asian ancestry. Second, the previous studies recognized emotion-specified expressions or only single AUs. The CMU S1 tested the Gabor-wavelet-based method on both single AUs and AU combinations, including nonadditive combinations in which the occurrence of one AU modifies another. Third, the previous studies manually aligned and cropped face images. The CMU S1 omitted this preprocessing step. In summary, using Gabor wavelets alone, recognition is adequate only for AU6, AU43, and AU0. Using geometric features alone, recognition is consistently good and shows high AU recognition rates with the exception of AU7. Combining both Gabor wavelet coefficients and geometric features, the recognition performance increased for all AUs.

In UCSD S1, the 2D frontal face image is detected by a face detector for each frame [87]. In UCSD S2, 3D pose and face geometry is estimated from hand-labeled feature points by using a canonical wire-mesh face model [62]. Once the 3D pose is estimated, faces are rotated to the frontal view and warped to a canonical face geometry. Then, the face images are automatically scaled and cropped to a standard face with a fixed distance between the two eyes in both UCSD S1 and S2. Difference images are obtained by subtracting a neutral expression face. UCSD S1 and S2 employed a family of Gabor wavelets at five spatial frequencies and eight orientations to a different image. Instead of specific locations on a face, they apply the Gabor filter to the whole face image. To provide robustness to lighting conditions and to image shifts they employed a representation in which the outputs of two Gabor filters in quadrature are squared and then summed. This representation is known as Gabor energy filters and it models complex cells of the primary visual cortex. Figure 11.7 shows an example of the UCSD systems Gabor appearance feature extraction [3].

In UIUC S2, Wen and Huang [88] use the ratio-image based method to extract appearance features, which is independent of a person's face albedo. To limit the effects of the noise in tracking and individual variation, they extracted the appearance features in facial regions instead of points, and then used the weighted average as the final feature for each region. Eleven regions were defined on the geometric-motion-free texture map of the face (Fig. 11.6d). Gabor wavelets with two spatial frequency and six orientations are used to calculate Gabor coefficients. A 12-dimension appearance feature vector is computed in each of the 11 selected regions by weighted averaging of the Gabor coefficients. To track the face appearance variations, an appearance model (texture image) is trained using a Gaussian mixture model based on exemplars.

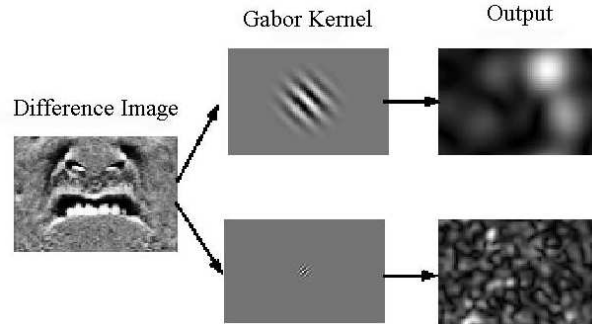


Fig. 11.7. Gabor appearance feature extraction in the UCSD systems. From Bartlett et al. [3], with permission.

Then an online adaptation algorithm is employed to progressively adapt the appearance model to new conditions such as lighting changes or differences in new individuals. See elsewhere for details [88].

3.3 Facial Expression Recognition

The last step of AFEA systems is to recognize facial expression based on the extracted features. Many classifiers have been applied to expression recognition such as neural network (NN), support vector machines (SVM), linear discriminant analysis (LDA), K-nearest neighbor, multinomial logistic ridge regression (MLR), hidden Markov models (HMM), tree augmented naive Bayes, and others. Some systems use only a rule-based classification based on the definition of the facial actions. Here, we summarize the expression recognition methods to frame-based and sequence-based expression recognition methods. The frame-based recognition method uses only the current frame with or without a reference image (mainly it is a neutral face image) to recognize the expressions of the frame. The sequence-based recognition method uses the temporal information of the sequences to recognize the expressions for one or more frames. Table 11.8 summarize of the recognition methods, recognition rates, recognition outputs, and the databases used in the most recent systems. For the systems that used more classifiers, the best performance for person-independent test has been selected.

Frame-Based Expression Recognition. Frame-based expression recognition does not use temporal information for the input images. It uses the information of current input image with/without a reference frame. The input image can be a static image or a frame of a sequence that is treated independently. Several methods can be found in the literature for facial expression recognition such as *neural networks* [81, 82, 97], *support vector machines* [3, 33], *linear discriminant analysis* [17], Bayesian network [13], and rule-based classifiers [60]. The CMU S1, UCSD S1, and UIUC S2 systems used frame-based classifiers.

CMU S1 employed a neural network-based recognizer to recognize FACS AUs. They used three-layer neural networks with one hidden layer to recognize AUs by a standard back-propagation method [66]. Separate networks are used for the upper and lower face. The inputs can be the normalized geometric features, the appearance feature, or both as shown in Figure

Table 11.8. FACS AU or expression recognition of recent advances. SVM, support vector machines; MLR, multinomial logistic ridge regression; HMM, hidden Markov models; BN, Bayesian network; GMM, Gaussian mixture model.

Systems	Recognition methods	Recognition rate	Recognized outputs	Databases
CMU S1 [81, 82, 83]	Neural network (frame)	95.5%	16 single AUs and their combinations	Ekman-Hager [26], Cohn-Kanade [43]
CMU S2 [14, 57]	Rule-based (sequence)	100%	Blink, nonblink	Frank-Ekman [36]
		57%	Brow up, down, and non-motion	
UCSD S1 [33]	SVM+MLR (frame)	91.5%	6 Basic expressions	Cohn-Kanade [43]
UCSD S2 [3]	SVM + HMM (sequence)	98%	blink, non-blink	Frank-Ekman [36]
		70%	brow up, down, and nonmotion	
UIUC S1 [13]	BN + HMM (frame & sequence)	73.22%	6 Basic expressions	Cohn-Kanade [43]
		66.53%	6 Basic expressions	UIUC-Chen [12]
UIUC S2 [88]	NN + GMM (frame)	71%	6 Basic expressions	Cohn-Kanade [43]

11.8. The outputs are the recognized AUs. The network is trained to respond to the designated AUs whether they occur alone or in combination. When AUs occur in combination, multiple output nodes are excited. To our knowledge, CMU S1 is the first system to handle AU combinations. Although several other systems tried to recognize AU combinations [17, 20, 51], they treated each combination as if it were a separate AU. More than 7,000 different AU combinations have been observed [71], and a system that can handle AU combinations is more efficient. A overall recognition rate of 95.5% had been achieved for neutral expression and 16 AUs whether they occurred individually or in combinations.

In UCSD S1 [33], they used a two-stage classifier to recognize neutral expression and six emotion-specified expressions. First, SVMs were used for the pairwise classifiers, i.e. each SVM is trained to distinguish two emotions. Then they tested several approaches, such as nearest neighbor, a simple voting scheme, and multinomial logistic ridge regression (MLR) to convert the representation produced by the first stage into a probability distribution over six emotion-specified expressions and neutral. The best performance at 91.5% was achieved by MLR.

UIUC S2 also employed a two-stage classifier to recognize neutral expression and six emotion-specified expressions. First, a neural network is used to classify *neutral* and *nonneutral*-like [79]. Then Gaussian mixture models (GMMs) were used for the remaining expressions. The overall average recognition rate was 71% for a people-independent test. See details elsewhere [88].

Sequence-Based Expression Recognition. The sequence-based recognition method uses the temporal information of the sequences to recognize the expressions of one or more frames. To use the temporal information, the techniques such as HMM [3, 13, 17, 51], recurrent neural networks [46, 65], and rule-based classifier [14] were used in facial expression analysis. The CMU S2, UCSD S2, and UIUC S1 systems employed a sequence-based classifier. Note that

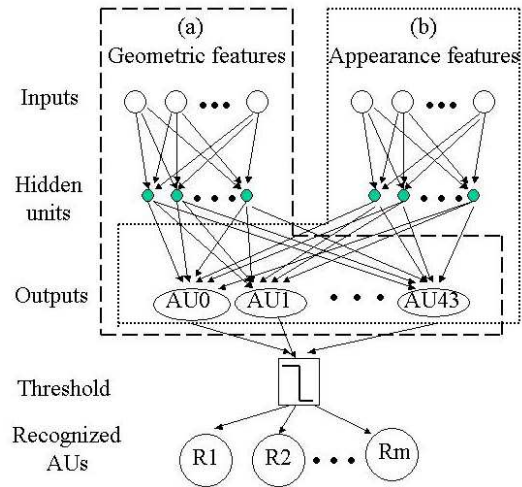


Fig. 11.8. Neural network-based recognizer for AU combinations in CMU S1. From Tian et al. [82], with permission.

the CMU S2 and UCSD S2 systems are comparative studies for FACS AU recognition in spontaneously occurring behavior by using the same database [36]. In that database, subjects were ethnically diverse, AUs occurred during speech, and out-of-plane motion and occlusion from head motion and glasses were common. So far, only several systems tried to recognize AUs or expression in spontaneously occurring behavior [3, 14, 80].

CMU S2 [14] employed a rule-based classifier to recognize AUs of eye and brow in spontaneously occurring behavior by using a number of frames in the sequence. The algorithm achieved an overall accuracy of 98% for three eye behaviors: blink (AU 45), flutter, and no blink (AU 0). *Flutter* is defined as two or more rapidly repeating blinks (AU 45) with only partial eye opening (AU 41 or AU 42) between them. 100% accuracy is achieved between blinks and non-blinks. Accuracy across the three categories in the brow region (brow-up, brow-down, nonbrow motion) was 57%. The number of brow-down actions was too small for reliable point estimates. Omitting brow-down from the analysis, recognition accuracy would increase to 80%. Human FACS coders had similar difficulty with brow-down, agreeing only about 50% in this database. The small number of occurrences was no doubt a factor for FACS coders as well. The combination of occlusion from eyeglasses and correlation of forward head pitch with brow-down complicated FACS coding. See Figure 11.9 for an example of the upper face in the database that shows the difficulties.

UCSD S2 [3] first employed SVMs to recognize AUs by using Gabor representations. Then they used hidden Markov models (HMMs) to deal with AU dynamics. HMMs were applied in two ways: (1) taking Gabor representations as input, and (2) taking the outputs of SVM as input. When they use Gabor representations as input to train HMMs, the Gabor coefficients were reduced to 100 dimensions per image using PCA. Two HMMs, one for blinks and one for nonblinks were trained and tested using leave-one-out cross-validation. A best performance of 95.7% recognition rate was obtained using five states and three Gaussians. They achieved a 98.1% recognition rate for blink and non-blink using SVM outputs as input to train HMMs

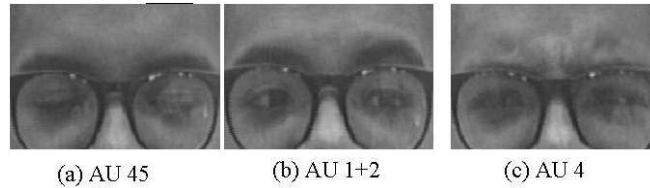


Fig. 11.9. An example of the Frank-Ekman database [36] was used by CMU S2 and UCSD S2 for spontaneous AU recognition. The cropped and stabilized upper face from the original video is displayed to show the difficulties. (From Bartlett et al. [3], with permission.)

for five states and three Gaussians. Accuracy across the three categories in the brow region (brow-up, brow-down, nonbrow motion) was 70.1% (HMMs trained on PCA-reduced Gabors) and 66.9% (HMMs trained on SVM outputs) respectively. Omitting brow-down, the accuracy increases to 90.9% and 89.5% respectively.

In UIUC S1, Cohen et al. [13] first evaluated Bayesian network (frame-based) classifiers such as Gaussian naive Bayes (NB-Gaussian), Cauchy naive Bayes (NB-Cauchy), and tree-augmented-naive Bayes (TAN), focusing on changes in distribution assumptions and feature dependency structures. They also proposed a new architecture of HMMs to segment and recognize neutral and six emotion-specified expressions from video sequences. For the person-independent test in the Cohn-Kanade database [43], the best performance at recognition rate of 73.2% was achieved by the TAN classifier. See details in Cohen et al. [13].

4 Discussion

Four recent trends in automatic facial expression analysis are (1) diversity of facial features in an effort to increase the number of expressions that may be recognized; (2) recognition of facial action units and their combinations rather than more global and easily identified emotion-specified expressions; (3) more robust systems for face acquisition, facial data extraction and representation, and facial expression recognition to handle head motion (both in-plane and out-of-plane), occlusion, lighting change, and low intensity expressions, all of which are common in spontaneous facial behavior in naturalistic environments; and (4) fully automatic and real-time AFEA systems. All of these developments move AFEA toward real-life applications. A public database (Cohn-Kanade AU-Coded Face Expression Image Database [43]) that addresses most problems for deliberate facial expression analysis has become widely used by many researchers to conduct comparative tests of their methods. Comparable image databases with ground-truth labels, preferably both action units and emotion-specified expressions, are needed for the next generation of systems, which are intended for naturally occurring behavior in real-life settings. For directed facial action tasks or other posed facial expressions, the CMU S1 system [81, 82, 83] can recognize the largest number of AUs whether they occur alone or in combinations. The system uses two neural networks (one for the upper face and one for the lower face). In recognizing whether AUs occur alone or in combinations, the system is performing a perceptual task analogous to that of human observers, who can recognize facial actions occurring in novel contexts. For spontaneous expression analysis, promising results have been

achieved for a few AUs by CMU [14, 57] and UCSD [3] systems. Work in spontaneous facial expression analysis is just now emerging and potentially will have significant impact across a range of theoretical and applied topics.

Although many recent advances and successes in automatic facial expression analysis have been achieved, as described in the previous sections, many questions remain open, for which answers must be found. Some major points are considered here.

1. How do humans correctly recognize facial expressions?

Research on human perception and cognition has been conducted for many years, but it is still unclear how humans recognize facial expressions. Which types of parameters are used by humans and how are they processed? By comparing human and automatic facial expression recognition we may be able advance our understanding of each and discover new ways of improving automatic facial expression recognition.

2. Is it always better to analyze finer levels of expression?

Although it is often assumed that more fine-grained recognition is preferable, the answer depends on both the quality of the face images and the type of application. Ideally, an AFEA system should recognize all action units and their combinations. In high quality images, this goal seems achievable; emotion-specified expressions then can be identified based on emotion prototypes identified in the psychology literature. For each emotion, prototypic action units have been identified. In lower quality image data, only a subset of action units and emotion-specified expression may be recognized. Recognition of emotion-specified expressions directly may be needed. We seek systems that become 'self aware' about the degree of recognition that is possible based on the information of given images and adjust processing and outputs accordingly. Recognition from coarse-to-fine, for example from emotion-specified expressions to subtle action units, depends on image quality and the type of application. Indeed, for some purposes, it may be sufficient that a system is able to distinguish between positive, neutral, and negative expression, or recognize only a limited number of target action units, such as brow lowering to signal confusion, cognitive effort, or negative affect.

3. Is there any better way to code facial expressions for computer systems?

Almost all the existing work has focused on recognition of facial expression, either emotion-specified expressions or FACS coded action units. The emotion-specified expressions describe expressions at a coarse level and are not sufficient for some applications. Although the FACS was designed to detect subtle changes in facial features, it is a human-observer-based system with only limited ability to distinguish intensity variation. Intensity variation is scored at an ordinal level; the interval level measurement is not defined and anchor points may be subjective. Challenges remain in designing a computer-based facial expression coding system with more quantitative definitions.

4. How do we obtain reliable ground truth?

Whereas some approaches have used FACS, which is a criterion measure widely used in the psychology community for facial expression analysis, most vision-based work uses emotion-specified expressions. A problem is that emotion-specified expressions are not well defined. The same label may apply to very different facial expressions, and different labels may refer to the same expressions, which confounds system comparisons. Another problem is that the reli-

ability of labels typically is unknown. With few exceptions, investigators have failed to report interobserver reliability and the validity of the facial expressions they have analyzed. Often there is no way to know whether subjects actually showed the target expression or whether two or more judges would agree that the subject showed the target expression. At a minimum, investigators should make explicit labeling criteria and report interobserver agreement for the labels. When the dynamics of facial expression are of interest, temporal resolution should be reported as well. Because intensity and duration measurements are critical, it is important to include descriptive data on these features as well. Unless adequate data about stimuli are reported, discrepancies across studies are difficult to interpret. Such discrepancies could be due to algorithms or to errors in ground truth determination.

5. How do we recognize facial expressions in real life?

Real-life facial expression analysis is much more difficult than the posed actions studied predominantly to date. Head motion, low resolution input images, absence of a neutral face for comparison, and low intensity expressions are among the factors that complicate facial expression analysis. Recent work in 3D modeling of spontaneous head motion and action unit recognition in spontaneous facial behavior are exciting developments. How elaborate a head model is required in such work is as yet a research question. A cylindrical model is relatively robust and has proven effective as a part of blink detection system [90], but highly parametric, generic, or even custom-fitted head models may prove necessary for more complete action unit recognition.

Most work to date has used a single, passive camera. Although there are clear advantages to approaches that require only a single passive camera or video source, multiple cameras are feasible in a number of settings and can be expected to provide improved accuracy. Active cameras can be used to acquire high resolution face images [40]. Also, the techniques of super-resolution can be used to obtain higher resolution images from multiple low resolution images [2]. At present, it is an open question how to recognize expressions in situations in which a neutral face is unavailable, expressions are of low intensity, or other facial or nonverbal behaviors, such as occlusion by the hands, are present.

6. How do we best use the temporal information?

Almost all work has emphasized recognition of discrete facial expressions, whether defined as emotion-specified expressions or action units. The timing of facial actions may be as important as their configuration. Recent work by our group has shown that intensity and duration of expression vary with context and that the timing of these parameters is highly consistent with automatic movement [73]. Related work suggests that spontaneous and deliberate facial expressions may be discriminated in terms of timing parameters [16], which is consistent with neuropsychological models [63] and may be important to lie detection efforts. Attention to timing also is important in guiding the behavior of computer avatars. Without veridical timing, believable avatars and ones that convey intended emotions and communicative intents may be difficult to achieve.

7. How may we integrate facial expression analysis with other modalities?

Facial expression is one of several modes of nonverbal communication. The message value of various modes may differ depending on context and may be congruent or discrepant with each other. An interesting research topic is the integration of facial expression analysis with

that of gesture, prosody, and speech. Combining facial features with acoustic features would help to separate the effects of facial actions due to facial expression and those due to speech-related movements. The combination of facial expression and speech can be used to improve speech recognition and multimodal person identification [34].

Acknowledgments

We sincerely thank Jing Xiao at Carnegie Mellon University, Dr. Marian Bartlett at MPLab of University of California San Diego, Zhen Wen at the University of Illinois at Urbana-Champaign for providing pictures and their permission to use them in this chapter. We also thank Dr. Zicheng Liu at Microsoft Research for reviewing this chapter.

References

1. J. Ahlberg and R. Forchheimer. Face tracking for model-based coding and face animation. *International Journal of Imaging Systems and Technology*, pages 8–22, Vol. 13, Issue 1, 2003.
2. S. Baker and T. Kanade. Limits on super-resolution and how to break them. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9):1167–1183, 2002.
3. M. Bartlett, B. Braathen, G. Littlewort-Ford, J. Hershey, I. Fasel, T. Marks, E. Smith, T. Sejnowski, and J. R. Movellan. Automatic analysis of spontaneous facial behavior: A final project report. Technical Report INC-MPLab-TR-2001.08, Machine Perception Lab, Institute for Neural Computation, University of California, San Diego, 2001.
4. M. Bartlett, J. Hager, P. Ekman, and T. Sejnowski. Measuring facial expressions by computer image analysis. *Psychophysiology*, 36:253–264, 1999.
5. M. Black. Robust incremental optical flow. In *PhD thesis, Yale University*, 1992.
6. M. Black and Y. Yacoob. Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. In *Proc. of International conference on Computer Vision*, pages 374–381, 1995.
7. M. Black and Y. Yacoob. Recognizing facial expressions in image sequences using local parameterized models of image motion. *International Journal of Computer Vision*, 25(1):23–48, October 1997.
8. L. Brown and Y.-L. Tian. Comparative study of coarse head pose estimation. In *IEEE Workshop on Motion and Video Computing, Orlando*, 2002.
9. L. Camras, L. Lambrecht, and G. Michel. Infant surprise expressions as coordinative motor structures. *Journal of Nonverbal Behavior*, 20:183–195, 1966.
10. J. Carroll and J. Russell. Do facial expression signal specific emotions? *Journal of Personality and Social Psychology*, 70:205–218, 1996.
11. J. Carroll and J. Russell. Facial expression in hollywood’s portrayal of emotion. *Journal of Personality and Social Psychology*, 72:164–176, 1997.
12. L. Chen. Joint processing of audio-visual information for the recognition of emotional expressions in human-computer interaction. In *PhD thesis, University of Illinois at Urbana-Champaign, Department of Electrical Engineering*, 2000.
13. I. Cohen, N. Sebe, F. Cozman, M. Cirelo, and T. Huang. Coding, analysis, interpretation, and recognition of facial expressions. *Journal of Computer Vision and Image Understanding Special Issue on Face Recognition*, 2003.
14. J. Cohn, T. Kanade, T. Moriyama, Z. Ambadar, J. Xiao, J. Gao, and H. Imamura. A comparative study of alternative face coding algorithms. Technical Report CMU-RI-TR-02-06, Robotics Institute, Carnegie Mellon University, Pittsburgh, November 2001.

15. J. Cohn and G. Katz. Bimodal expression of emotion by face and voice. In *ACM and ATR Workshop on Face/Gesture Recognition and Their Applications*, pages 41–44, 1998.
16. J. Cohn, K. Schmidt, R. Gross, and P. Ekman. Individual differences in facial expression: stability over time, relation to self-reported emotion, and ability to inform person identification. In *Proceedings of the International Conference on Multimodal User Interfaces (ICMI 2002)*, pages 491–496, 2002.
17. J. Cohn, A. Zlochower, J. Lien, and T. Kanade. Automated face analysis by feature point tracking has high concurrent validity with manual faces coding. *Psychophysiology*, 36:35–43, 1999.
18. C. Darwin. *The Expression of Emotions in Man and Animals*. John Murray, reprinted by University of Chicago Press, 1965, 1872.
19. J. Daugmen. Complete discrete 2d gabor transforms by neural networks for image analysis and compression. *IEEE Transaction on Acoustic, Speech and Signal Processing*, 36(7):1169–1179, 1988.
20. G. Donato, M. Bartlett, J. Hager, P. Ekman, and T. Sejnowski. Classifying facial actions. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 21(10):974–989, 1999.
21. Eihl-Eihesfeldt. *Human Ethology*. Aldine de Gruyter, New York, 1989.
22. P. Ekman. *The Argument and Evidence about Universals in Facial Expressions of Emotion*, pages 143–164. Wiley, New York, 1989.
23. P. Ekman. Facial expression and emotion. *American Psychologist*, 48:384–392, 1993.
24. P. Ekman and W. Friesen. *Pictures of Facial Affect*. Palo Alto, CA: Consulting Psychologist., 1976.
25. P. Ekman and W. Friesen. *The Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, San Francisco, 1978.
26. P. Ekman, J. Hager, C. Methvin, and W. Irwin. *Ekman-Hager Facial Action Exemplars*. Human Interaction Laboratory, University of California, San Francisco.
27. P. Ekman and E. e. Rosenberg. *What the Face Reveals*. Oxford University, New York, 1997.
28. I. Essa and A. Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Trans. on Pattern Analysis and Machine Intell.*, 19(7):757–763, 1997.
29. L. Farkas and I. Munro. *Anthropometric Facial Proportions in Medicine*. Charles C Thomas, Springfield, IL, 1987.
30. B. Fasel and J. Luttin. Recognition of asymmetric facial action unit activities and intensities. In *Proceedings of International Conference of Pattern Recognition*, 2000.
31. B. Fasel and J. Luttin. Automatic facial expression analysis: Survey. *Pattern Recognition*, 36(1):259–275, 2003.
32. J. Fleiss. *Statistical Methods for Rates and Proportions*. Wiley, New York, 1981.
33. G. Ford. Fully automatic coding of basic expressions from video. Technical Report INC-MPLab-TR-2002.03, Machine Perception Lab, Institute for Neural Computation, University of California, San Diego, 2002.
34. N. Fox, R. Gross, P. de Chazal, J. Cohn, and R. Reilly. Person identification using multi-modal features: speech, lip, and face. In *Proc. of ACM Multimedia Workshop in Biometrics Methods and Applications (WBMA 2003)*, CA, 2003.
35. N. Fox and R. Reilly. Audio-visual speaker identification. In *Proc. of the 4th International Conference on Audio-and Video-Based Biometric Person Authentication*, 2003.
36. M. Frank and P. Ekman. The ability to detect deceit generalizes across different types of high-stake lies. *Personality and Social Psychology*, 72:1429–1439, 1997.
37. W. Friesen and P. Ekman. Emfacs-7: emotional facial action coding system. *Unpublished manuscript, University of California at San Francisco*, 1983.
38. H. Fujiyoshi and A. Lipton. Real-time human motion analysis by image skeletonization. In *Proc. of the Workshop on Application of Computer Vision*, 1998.
39. K. Fukui and O. Yamaguchi. Facial feature point extraction method based on combination of shape extraction and pattern matching. *Systems and Computers in Japan*, 29(6):49–58, 1998.

40. A. Hampapur, S. Pankanti, A. Senior, Y. Tian, L. Brown, and R. Bolle. Face cataloger: multi-scale imaging for relating identity to location. In *Proceedings of IEEE Conference on Advanced Video and Signal Based Surveillance*, 2003.
41. B. Heisele, T. Serre, M. Pontil, and T. Poggio. Component-based face detection. In *Proc. IEEE Conf. on Computer Vision and Pattern Recogn. (CVPR)*, 2001.
42. C. Izard, L. Dougherty, and E. A. Hembree. A system for identifying affect expressions by holistic judgments. In *Unpublished Manuscript, University of Delaware*, 1983.
43. T. Kanade, J. Cohn, and Y.-L. Tian. Comprehensive database for facial expression analysis. In *Proceedings of International Conference on Face and Gesture Recognition*, pages 46–53, 2000.
44. S. Kimura and M. Yachida. Facial expression recognition and its degree estimation. In *Proc. Of the International Conference on Computer Vision and Pattern Recognition*, pages 295–300, 1997.
45. R. Kleck and M. Mendolia. Decoding of profile versus full-face expressions of affect. *Journal of Nonverbal Behavior*, 14(1):35–49, 1990.
46. H. Kobayashi, K. Tange, and F. Hara. Dynamic recognition of six basic facial expressions by discrete-time recurrent neural network. In *Proc. of the International Joint Conference on Neural Networks*, pages 155–158, 1993.
47. H. Kobayashi, K. Tange, and F. Hara. Real-time recognition of six basic facial expressions. In *Proc. IEEE Workshop on Robot and Human Communication*, pages 179–186, 1995.
48. R. Kraut and R. Johnson. Social and emotional messages of smiling: an ethological approach. *Journal of Personality and Social Psychology*, 37:1539–1523, 1979.
49. S. Li and L. Gu. Real-time multi-view face detection, tracking, pose estimation, alignment, and recognition. In *IEEE Conf. on Computer Visioin and Pattern Recognition Demo Summary*, 2001.
50. J.-J. Lien, T. Kanade, J. Cohn, and C. Li. Subtly different facial expression recognition and expression intensity estimation. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 853–859, 1998.
51. J.-J. Lien, T. Kanade, J. Cohn, and C. Li. Detection, tracking, and classification of action units in facial expression. *Journal of Robotics and Autonomous System*, 31:131–146, 2000.
52. M. Lyons, S. Akamasku, M. Kamachi, and J. Gyoba. Coding facial expressions with gabor wavelets. In *Proceedings of International Conference on Face and Gesture Recognition*, 1998.
53. A. Manstead. *Expressiveness as an individual difference*, pages 285–328. Cambridge University Press, Cambridge, UK, 1991.
54. P. Martin and P. Bateson. *Measuring behavior: An Introductory Guide*. Cambridge University Press, Cambridge, UK, 1986.
55. K. Mase. Recognition of facial expression from optical flow. *IEICE Transactions*, E. 74(10):3474–3483, 1991.
56. R. Matias, J. Cohn, and S. Ross. A comparison of two systems to code infants' affective expression. *Developmental Psychology*, 25:483–489, 1989.
57. T. Moriyama, T. Kanade, J. Cohn, J. Xiao, Z. Ambadar, J. Gao, and M. Imanura. Automatic recognition of eye blinking in spontaneously occurring behavior. In *Proceedings of the 16th International Conference on Pattern Recognition (ICPR '2002)*, volume 4, pages 78–81, 2002.
58. Y. Moses, D. Reynard, and A. Blake. Determining facial expressions in real time. In *Proc. of Int. Conf. On Automatic Face and Gesture Recognition*, pages 332–337, 1995.
59. M. Pantic and L. Rothkrantz. Automatic analysis of facial expressions: the state of the art. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, 2000.
60. M. Pantic and L. Rothkrantz. Expert system for automatic analysis of facial expression. *Image and Vision Computing*, 18(11):881–905, 2000.
61. A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eigenspaces for face recognition. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 84–91, 1994.
62. F. Pighin, H. Szeliski, and D. Salesin. Synthesizing realistic facial expressions from photographs. In *Proc of SIGGRAPH*, 1998.

63. W. Rinn. The neuropsychology of facial expression: a review of the neurological and psychological mechanisms for producing facial expressions. *Psychological Bulletin*, (95):52–77, 1984.
64. S. Rizvi, P. Phillips, and H. Moon. The feret verification testing protocol for face recognition algorithms. In *Proceedings of the Third International Conference on Automatic Face and Gesture Recognition*, pages 48–55, 1998.
65. M. Rosenblum, Y. Yacoob, and L. Davis. Human expression recognition from motion using a radial basis function network architecture. *IEEE Transactions On Neural Network*, 7(5):1121–1138, 1996.
66. H. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Transactions On Pattern Analysis and Machine intelligence*, 20(1):23–38, 1998.
67. J. Russell. Culture and the categorization. *Psychological Bulletin*, 110:426–450, 1991.
68. J. Russell. Is there universal recognition of emotion from facial expression? a review of the cross-cultural studies. *Psychological Bulletin*, 115:102–141, 1991.
69. A. Samal and P. Iyengar. Automatic recognition and analysis of human faces and facial expressions: A survey. *Pattern Recognition*, 25(1):65–77, 1992.
70. M. Sayette, J. Cohn, J. Wertz, M. Perrott, and D. Parrott. A psychometric evaluation of the facial action coding system for assessing spontaneous expression. *Journal of Nonverbal Behavior*, 25:167–186, 2001.
71. K. Scherer and P. Ekman. *Handbook of Methods in Nonverbal Behavior Research*. Cambridge University Press, Cambridge, UK, 1982.
72. K. Schmidt and J. Cohn. Human facial expressions as adaptations: Evolutionary questions in facial expression. *Yearbook of Physical Anthropology*, 44:3–24, 2002.
73. K. Schmidt, J. F. Cohn, and Y.-L. Tian. Signal characteristics of spontaneous facial expressions: Automatic movement in solitary and social smiles. *Biological Psychology*, 2003.
74. H. Schneiderman and T. Kanade. A statistical model for 3d object detection applied to faces and cars. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2000.
75. K. Sung and T. Poggio. Example-based learning for view-based human face detection. *IEEE Transactions on Pattern Analysis and Machine intelligence*, 20(1):39–51, 1998.
76. M. Suwa, N. Sugie, and K. Fujimora. A preliminary note on pattern recognition of human emotional expression. In *International Joint Conference on Pattern Recognition*, pages 408–410, 1978.
77. H. Tao and T. Huang. Explanation-based facial motion tracking using a piecewise bezier volume deformation model. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1999.
78. D. Terzopoulos and K. Waters. Analysis of facial images using physical and anatomical models. In *IEEE International Conference on Computer Vision*, pages 727–732, 1990.
79. Y.-L. Tian and R. Bolle. Automatic detecting neutral face for face authentication. In *Proceedings of AAAI-03 Spring Symposium on Intelligent Multimedia Knowledge Management, CA*, 2003.
80. Y.-L. Tian, L. Brown, A. Hampapur, S. Pankanti, A. Senior, and R. Bolle. Real world real-time automatic recognition of facial expressions. In *Proceedings of IEEE Workshop on Performance Evaluation of Tracking and Surveillance, Graz, Austria*, 2003.
81. Y.-L. Tian, T. Kanade, and J. Cohn. Recognizing action units for facial expression analysis. *IEEE Trans. on Pattern Analysis and Machine Intell.*, 23(2):1–19, 2001.
82. Y.-L. Tian, T. Kanade, and J. Cohn. Evaluation of gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity. In *Proceedings of the 5th IEEE International Conference on Automatic Face and Gesture Recognition (FG'02), Washington, DC*, 2002.
83. Y.-L. Tian, T. Kanade, and J. Cohn. Eye-state action unit detection by gabor wavelets. In *Proceedings of International Conference on Multi-modal Interfaces (ICMI 2000)*, pages 143–150, Sept, 2000.
84. K. Toyama. look, ma—no hands! hands-free cursor control with real-time 3d face tracking. In *Proc. Workshop on Perceptual User Interfaces(PUI'98)*, 1998.
85. J. VanSwearingen, J. Cohn, and A. Bajaj-Luthra. Specific impairment of smiling increases severity of depressive symptoms in patients with facial neuromuscular disorders. *Journal of Aesthetic Plastic Surgery*, 23:416–423, 1999.

86. T. Vetter. Learning novel views to a single face image. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pages 22–29, 1995.
87. P. Viola and M. Jones. Robust real-time object detection. In *International Workshop on Statistical and Computational Theories of Vision - Modeling, Learning, Computing, and Sampling*, 2001.
88. Z. Wen and T. Huang. Capturing subtle facial motions in 3d face tracking. In *Proc. of Int. Conf. On Computer Vision*, 2003.
89. Y. Wu and K. Toyama. Wide-range person and illumination-insensitive head orientation estimation. In *Proceedings of International Conference on Automatic Face and Gesture Recognition*, pages 183–188, 2000.
90. J. Xiao, T. Kanade, and J. Cohn. Robust full motion recovery of head by dynamic templates and re-registration techniques. In *Proceedings of International Conference on Automatic Face and Gesture Recognition*, pages 163–169, 2002.
91. J. Xiao, T. Moriyama, T. Kanade, and J. Cohn. Robust full-motion recovery of head by dynamic templates and re-registration techniques. *International Journal of Imaging Systems and Technology*, September, 2003.
92. Y. Yacoob and M. Black. Parameterized modeling and recognition of activities. In *Proc. 6th IEEE Int. Conf. on Computer Vision*, pages 120–127, Bombay, 1998.
93. Y. Yacoob and L. Davis. Recognizing human facial expression from long image sequences using optical flow. *IEEE Trans. on Pattern Analysis and Machine Intell.*, 18(6):636–642, June 1996.
94. Y. Yacoob, H.-M. Lam, and L. Davis. Recognizing faces showing expressions. In *Proc. Int. Workshop on Automatic Face- and Gesture-Recognition*, pages 278–283, Zurich, Switzerland, 1995.
95. M. Yang, D. Kriegman, and N. Ahuja. Detecting faces in images: a survey. *IEEE Trans. on Pattern Analysis and Machine Intell.*, 24(1), 2002.
96. Z. Zhang. Feature-based facial expression recognition: sensitivity analysis and experiments with a multi-layer perceptron. *International Journal of Pattern Recognition and Artificial Intelligence*, 13(6):893–911, 1999.
97. Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu. Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron. In *International Workshop on Automatic Face and Gesture Recognition*, pages 454–459, 1998.
98. L. Zhao, G. Pingali, and I. Carlbom. Real-time head orientation estimation using neural networks. In *Proc of the 6th International Conference on Image Processing*, 2002.
99. A. Zlochow, J. Cohn, J. Lien, and T. Kanade. A computer vision based method of facial expression analysis in parent-infant interaction. In *International Conference on Infant Studies, Atlanta*, 1998.

Index

Emotional-based expressions, 4

AFEA, 1

Appearance feature, 17

Database, 9

Deliberate expression, 7

Expression intensity, 6

Expression recognition, 19

Face Acquisition, 2

Face acquisition, 11

Face detection, 12

Facial Data Extraction and Representation, 2

Facial Expression Analysis, 1

Facial expression recognition, 2

Facial feature extraction, 15

Facial feature Representation, 15

FACS action units, 4

Geometric feature, 16

Ground truth, 9

Head orientation, 7

Head pose estimation, 13

Non-additive combinations, 6

Spontaneous expression, 7

Transitions, 5