Chapter 19
Facial Expression Analysis

Fernando De la Torre and Jeffrey F. Cohn

Abstract The face is one of the most powerful channels of nonverbal communication. Facial expression provides cues about emotion, intention, alertness, pain, personality, regulates interpersonal behavior, and communicates psychiatric and biomedical status among other functions. Within the past 15 years, there has been increasing interest in automated facial expression analysis within the computer vision and machine learning communities. This chapter reviews fundamental approaches to facial measurement by behavioral scientists and current efforts in automated facial expression recognition. We consider challenges, review databases available to the research community, approaches to feature detection, tracking, and representation, and both supervised and unsupervised learning.

19.1 Introduction

An automatic analysis of the facial expressions of people are highly important for automatic understanding of humans, their actions and their behavior in general. Facial expression has been a focus of research in human behavior for over a hundred years [30]. It is central to several leading theories of emotion [38, 116] and has been the focus of, at times, heated debate about issues in emotion science. Facial expression figures prominently in research on almost every aspect of emotion, including psychophysiology [66], neural correlates [39], development [84], perception [2], addiction [47], social processes [52], depression [27] and other emotion disorders [118]. Facial expression communicates physical pain [100], alertness, personality and interpersonal relations [46]. Applications of facial expression analysis include marketing [107], perceptual user interfaces, human–robot interaction [98, 126, 145], drowsy driver detection [128], telenursing [29], pain assessment [79],

F. De la Torre (✉)
Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA
e-mail: ftorre@cs.cmu.edu

J.F. Cohn
Department of Psychology, University of Pittsburgh, Pittsburgh, PA 15260, USA
e-mail: jeffcohn@pitt.edu

analyzing mother–infant interaction [45], autism [83], social robotics [6, 18], facial animation [72, 110] and expression mapping for video gaming [54] among others. A large number of examples are also provided in particular in Chaps. 22, 26 and 23.

In part because of its importance and potential uses as well as its inherent challenges, automated facial expression recognition has been of keen interest in computer vision and machine learning. Beginning with a seminal meeting sponsored by the US National Science Foundation [41], research on this topic has become increasingly broad, systematic, and productive. IEEE-sponsorship of international conferences (http://www.fg2011.org/), workshops, and a new journal in affective computing, among other outlets (e.g., IEEE journal System, Man, and Cybernetics and special issues of journals such as Image, Vision, and Computing Journal) speak to the vitality of research in this area. Automated facial expression analysis is critical as well to the emerging fields of Computational Behavior Science and Social Signal Processing.

Automated facial image analysis confronts a series of challenges. The face and facial features must be detected in video; shape or appearance information must be extracted and then normalized for variation in pose, illumination and individual differences; the resulting normalized features are used to segment and classify facial actions. Partial occlusion is a frequent challenge that may be intermittent or continuous (e.g., bringing an object in front of the face, self-occlusion from head turns, eyeglasses or facial jewelry). While human observers easily accommodate for changes in pose, scale, illumination, occlusion, and individual differences, these and other sources of variation represent considerable challenges for computer vision. Then there is the machine-learning challenge of automatically detecting actions that require significant training and expertise even for human coders. There is much good research to do.

We begin with a description of approaches to annotation and then review publicly available databases. Research in automated facial expression analysis depends on access to large, well-annotated, video data. We then review approaches to feature detection, representation, and registration, and both supervised and unsupervised learning of facial expression. We close with implications for future research in this area.

19.2 Annotation of Facial Expression

Two broad approaches to annotating facial expression are message–judgment and sign-based [25]. In the former, observers make inferences about the meaning of facial actions and assign corresponding labels. The most widely used approach of this sort makes inferences about felt emotion. Inspired by cross-cultural studies by Ekman [38] and related work by Izard [55], a number of expressions of what are referred to as basic emotions have been described. These include joy, surprise, anger, fear, disgust, sadness, embarrassment, and contempt. Examples of the first six are shown in Fig. 19.1. Message–judgment approaches tend to be holistic; that is, they typically combine information from multiple regions of the face, implicitly

Fig. 19.1 Basic facial expression phenotypes. 1, disgust; 2, fear; 3, joy; 4, surprise; 5, sadness; 6, anger. Figure reproduced with permission from [105]. © 2010 IEEE

acknowledge that the same emotion or cognitive state may be expressed in various ways, and they utilize the perceptual wisdom of human observers, which may include taking account of context. A limitation is that many of these emotions may occur infrequently in daily life and much human experience involves blends of two or more emotions. While a small set of specific expressions that vary in multiple regions of the face may be advantageous for training and testing, their generalizability to new image sources and applications is limited. Moreover, the use of emotion labels implies that posers are experiencing the actual emotion. This inference often is unwarranted, as when facial expression is posed or faked, and the same expression may map to different felt emotions. Smiles, for instance, occur in both joy and embarrassment [1].

In a sign-based approach, physical changes in face shape or texture are the descriptors. The most widely used approach is that of Ekman and colleagues. Their Facial Action Coding System (FACS) [40] segments the visible effects of facial muscle activation into "action units". Each action unit is related to one or more facial muscles. The Facial Action Coding System (FACS) is a comprehensive, anatomically based system for measuring nearly all visually discernible facial movement. FACS describes facial activity on the basis of 44 unique action units (AUs), as well as several categories of head and eye positions and movements. Facial movement is thus described in terms of constituent components, or AUs. Any facial event (for example, an emotion expression or paralinguistic signal) may be decomposed into one or more AUs. For example, what has been described as the felt or Duchenne smile typically includes movement of the zygomatic major (AU12) and orbicularis oculi, pars lateralis (AU6).

The FACS taxonomy was defined by manually observing graylevel variation between expressions in images and to a lesser extent by recording the electrical activity of underlying facial muscles [24]. Depending on which edition of FACS is used, there are 30 to 44 AUs and additional "action descriptors." Action descriptors are movements for which the anatomical basis is not established. More than 7000 AU combinations have been observed [104]. Figures 19.2 and 19.3 illustrate AUS from the upper and lower portions of the face, respectively. Figure 19.4 provides an example in which FACS action units have been used to label a prototypic expression of pain. Because of its descriptive power, FACS has become the standard for facial measurement in behavioral research and has supplanted use of message–judgment

Upper Face Action Units					
AU1	AU2	AU4	AU5	AU6	AU7
Inner Brow Raiser	Outer Brow Raiser	Brow Lowerer	Upper Lid Raiser	Cheek Raiser	Lid Tightener
*AU41	*AU42	*AU43	AU44	AU45	AU46
Lip Droop	Slit	Eyes Closed	Squint	Blink	Wink

Fig. 19.2 FACS action units (AU) for the upper face. Figure reproduced with permission from [24]

Lower Face Action Units					
AU9	AU10	AU11	AU12	AU13	AU14
Nose Wrinkler	Upper Lip Raiser	Nasolabial Deepener	Lip Corner Puller	Cheek Puffer	Dimpler
AU15	AU16	AU17	AU18	AU20	AU22
Lip Corner Depressor	Lower Lip Depressor	Chin Raiser	Lip Puckerer	Lip Stretcher	Lip Funneler
AU23	AU24	*AU25	*AU26	*AU27	AU28
Lip Tightener	Lip Pressor	Lips Parts	Jaw Drop	Mouth Stretch	Lip Suck

Fig. 19.3 Action units of the lower face. Figure reproduced with permission from [24]

AU 4: Brow lowering
AU 43: Closing eyes
AU 6: Cheek raising
AU 12: Pulling at corner lip
AU 7: Tightening of eyelids
AU 9: Wrinkling of nose
AU 10: Raising of upper lip
AU 25: Parting lips
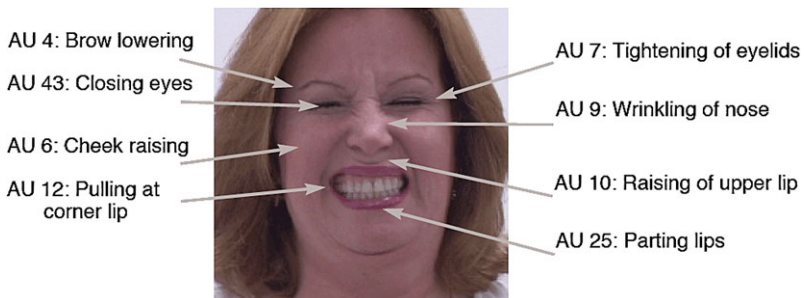
Fig. 19.4 An example of facial action units associated with a prototypic expression of pain. Figure reproduced with permission from [75]. © 2011 IEEE
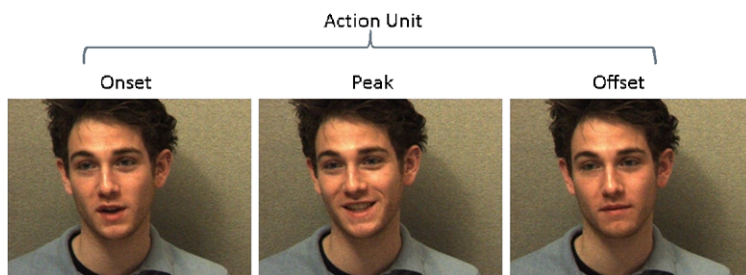
Fig. 19.5 FACS coding typically involves frame-by-frame inspection of the video, paying close attention to subtle cues such as wrinkles, bulges, and furrows to determine which facial action units have occurred and their intensity. Full labeling requires marking onset, peak and offset of the action unit and all changes in intensity. Full coding generally is too costly. *Left to right*, evolution of an AU 12 (involved in smiling), from onset, peak, to offset

approaches in automated facial image analysis. As well, FACS has become influential in the related area of computer facial animation. The MPEG-4 facial animation parameters [92] are derived from FACS.

Facial actions can vary in intensity, which FACS represents at an ordinal level of measurement. The original (1978) version of FACS included criteria for measuring intensity at three levels (X, Y, and Z). The more recent 2002 edition provides criteria for measuring intensity at five levels, ranging from A to E. FACS scoring produces a list of AU-based descriptions of each facial event in a video record. Figure 19.5 shows an example for FACS coding AU12 (Smile), where the onset, peak and offset are labeled.

For both message–judgment and sign-based approaches, the reliability of human coding has been a neglected topic in the automated facial expression recognition literature. With some exceptions, publicly available databases (Table 19.1) and research reports fail to provide information about inter-observer reliability or agreement. This is an important lack, in that inter-system agreement between manual and automated coding is inherently limited by intra-system agreement. If manual coding disagrees about the ground truth used to train classifiers, it is unlikely that classifiers will surpass them. Inter-system reliability can be considered in numerous ways [26]. These range from the precision of measurement of onsets, peaks, offsets, and changes in action unit intensity, to whether or not observers agree on action unit occurrence within some number of frames. More attention to reliability of coding would be useful in evaluating training data and test results. Sayette and Cohn [103] found inter-observer agreement varied among AU. Agreement for AU 7 (lower lid tightener) was relatively low, possibly due to confusion with AU 6 (cheek raiser). Some AU may occur too infrequently to measure reliably (e.g., AU 11). Investigators may want to consider pooling some AU to achieve more reliable units.

Agreement between human coders is better when temporal precision is relaxed. In behavioral research, it is common to expect coders to agree only within a $\frac{1}{2}$ second window. In automated facial image analysis, investigators typically assume exact agreement between classifiers and ground truth, which is a level of temporal precision beyond what may be feasible for many AU [24].

Table 19.1 Publicly available facial expression databases

Database	No. of Subjects	Elicitation	Imaging	Camera View	Labels	Requests
AR [85]	126	Posed	Static	Frontal Occlusion	Emotions	http://www2.ece.ohio-state.edu/~aleix/ARdatabase.html
Belfast	125	Interviews and TV	Video	Frontal	Emotion and dimensions	http://www.idiap.ch/mmm/corpora/emotion-corpus
Cohn–Kanade [58]	97	Posed	Video	Frontal	FACS AU	http://vasc.ri.cmu.edu/idb/html/face/facial_expression/
Cohn–Kanade+ [78]	123	Posed and Conversation	Video	Frontal and 15° to the side	FACS AU Emotion Landmarks	http://vasc.ri.cmu.edu/idb/html/face/facial_expression/
FABO	23	Posed	Video	Frontal	Emotion	http://research.it.uts.edu.au/cvrg/FABO.htm
GEMEP	10	Acted	Video	Frontal	Emotion	http://www.fg2011.org/fg.php?page=workshop
KDEF	70	Posed	Static	Five views	Emotion	http://www.facialstimuli.com/index_files/Page369.htm
JAFFE [82]	10	Posed	Static	Frontal	Emotion	http://kasrl.org/jaffe.html
MMI [101, 124]	101	Posed Spontaneous	Static Video (5 min)	Frontal 90° to the side	FACS AU	http://emotion-research.net/toolbox/ toolboxdatabase.2006-09-28.5469431043
Face Database MPI [99]		Posed	Video	11 views at 18° intervals	FACS AU	http://vdb.kyb.tuebingen.mpg.de/
Multi-PIE [48]	337	Posed	Static	15 views 19 illuminations	Emotion Landmarks	http://www.multipie.org/
Prkachin–Solomon Pain [80]	129	Pain induction	Video	Frontal	AU Landmarks	http://www.pitt.edu/~jeffcohn/?K
Multi-PIE [64]	72	Posed	Static	Five views	Emotion	http://facedb.blogspot.com/2008/07/ short-description-of-database-set.html
RU-FACS [7]	100	Interview	Video (2 min)	Mostly frontal	FACS AU	http://mplab.ucsd.edu/grants/project1/research/ rufacs1-dataset.html
University of Texas Video Database [91]	284	Viewing videoclip	Video (10 minutes)	Frontal	Emotion	http://portal.acm.org/citation.cfm?id=1053716
Bosphorous	105	Posed	Static	3D	FACS AU Emotion	http://bosphorus.ee.boun.edu.tr/
BU-3DFE [132]	100	Posed	Static	3D	Emotion	http://www.cs.binghamton.edu/~lijun/Research/3DFE/ 3DFE_Analysis.html
BU-4DFE	101	Posed	Dynamic	3D	Emotion	http://www.cs.binghamton.edu/~lijun/Research/3DFE/ 3DFE_Analysis.html

19.3 Databases

The development of robust facial recognition algorithms requires well labeled databases of sufficient size that include carefully controlled variations of pose, illumination and resolution. Publicity available databases are necessary to comparatively evaluate algorithms. Collecting a high quality database is a resource-intensive task. The availability of public facial expression databases is important for the advancement of the field. Table 19.1 illustrates the characteristics of publicly available databases.

Most face expression databases 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. Deliberate and spontaneous facial behavior are mediated by separate motor pathways, the pyramidal and extrapyramidal motor tracks, respectively. 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 [36]. Differences in the temporal organization of spontaneous and deliberate facial actions are particularly important in that many pattern recognition approaches, such as HMMs, 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.

19.4 Facial Feature Tracking, Registration and Feature Extraction

Prototypical expression and AU detection from video are challenging computer vision and pattern recognition problems. Some of the most important challenges are: (1) non-frontal pose and moderate to large head motion make facial image registration difficult, (2) classifiers can suffer from over-fitting when trained with relatively few examples for each AU; (3) many facial actions are inherently subtle making them difficult to be model; (4) individual differences among faces in shape and appearance make the classification task difficult to generalize across subjects; (5) temporal dynamics of AUs are highly variable. These differences can signal different communicative intentions [62], levels of distress [9], and presents a challenge for detection and classification; (6) and the number of possible combinations of 40+ individual action units numbers in the thousands (more than 7000 action unit combinations have been observed [42]). To address these issues over the last 20 years, a large number of facial expression and AU recognition/detection systems have been
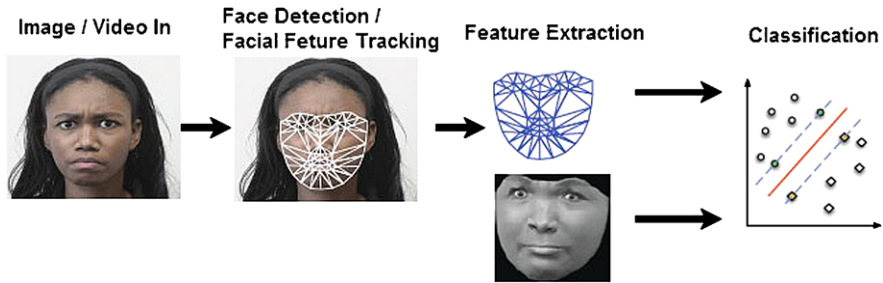
Fig. 19.6 Block diagram of our the CMU system. The face is tracked using an AAM; shape and appearance features are extracted, normalized, and output to a linear SVM for action unit or expression detection. Figure reproduced with permission from [78]. © 2010 IEEE

proposed. Some of the leading efforts include those at: Carnegie Mellon University [81, 108, 112, 142], University of California, San Diego [7, 68], University of Illinois at Urbana-Champaign [23, 129], Rensselaer Polytechnic Institute [117], Massachusetts Institute of Technology [43], University of Maryland [13, 131], Imperial College [59, 95, 123], IDIAP Dalle Molle Institute for Perceptual Artificial Intelligence [44], and others [82, 138].

Most facial expression recognition systems are composed of three main modules: (1) face detection, facial feature tracking and registration, (2) feature extraction and (3) supervised or unsupervised learning. Figure 19.6 illustrates an example of these three modules. In the following sections we will discuss each of these modules in more detail with emphasis in the current CMU system. For other systems see [44, 93, 113].

19.4.1 Facial Feature Detection and Tracking

Face detection is an initial step in most automatic facial expression recognition systems (see Chap. 5). For real-time, frontal face detection, the Viola and Jones [127] face detector is arguably the most commonly employed algorithm. See [137] for a survey of recent advances in face detection. Once the face is detected two approaches to registration are common. One performs coarse registration by detecting a sparse set of facial features (e.g., eyes) in each frame. The other detects detailed features (i.e. dense points around the eyes and other facial landmarks) in the video sequence. In this section we will describe a unified framework for the latter, which we refer to as Parameterized Appearance Models (PAMs). PAMs include the Lucas–Kanade method [74], Eigentracking [12], Active Appearance Models [28, 33, 34, 87] , and Morphable Models [14, 57], which have been popular approaches for facial feature detection, tracking and modeling faces in general.

PAMs are among the most popular methods for facial feature detection and face alignment in general. PAMs for faces build an appearance and/or shape representation from the principal components of labeled training data. Let $\mathbf{d}_i \in \Re^{m \times 1}$ (see an

Fig. 19.7 The figure shows the mean and first two modes of variation of 2D AAM shape **a** and appearance **b** variation and the mean and first two modes of 3D AAM shape. **c** Reconstructed face. Reproduced with permission from [88]

(a)

(b)

(c)
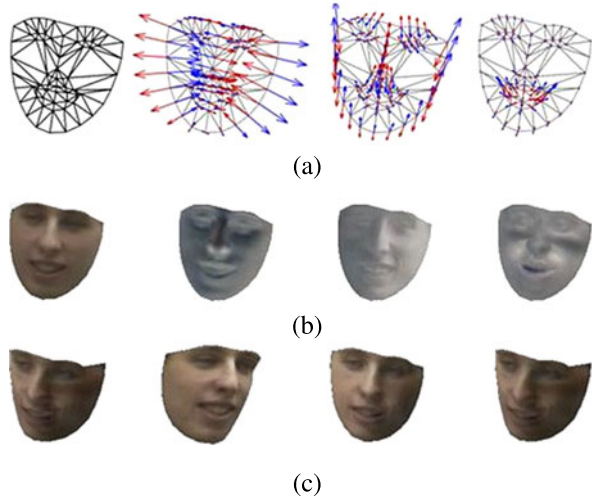
explanation of the notation[1]) be the ith sample of a training set $\mathbf{D} \in \Re^{m \times n}$ of n samples, where each vector $\mathbf{d}_i$ is a vectorized image of m pixels. In a training set, each face image is previously manually labeled with p landmarks. A $2p$-dimensional shape vector is constructed by stacking all (x, y) positions of the landmarks as $s = [x_1; y_1; x_2; y_2; \ldots; x_p; y_p]$. Figure 19.9a shows an example of several face images that have been labeled with 66 landmarks. Given the labeled training samples, Procrustes analysis [28] is applied to the shape vectors to remove two-dimensional rigid transformations. After removing rigid transformation with Procrustes, principal component analysis (PCA) is applied to the shape vectors to build a linear shape model. The shape model can reconstruct any shape on the training shape as the mean ($\mathbf{s}_0$) and linear combination of a shape basis ($\mathbf{U}^s$) (eigenvectors of the shape covariance matrix), that is, $\mathbf{s} \approx \mathbf{s}_0 + \mathbf{U}^s \mathbf{c}^s$, where $\mathbf{c}^s$ are the shape coefficients. $\mathbf{U}^s$ spans the shape space that accounts for identity, expression and pose variation in the training set. Figure 19.7a shows the shape mean and PCA basis. Similarly, after backwarping the texture to a canonical configuration, the appearance (normalized graylevel) is vectorized into an m-dimensional vector and stacked into the n columns of $\mathbf{D} \in \Re^{m \times n}$. The appearance model, $\mathbf{U} \in \Re^{m \times k}$ is computed by calculating the first k principal components [56] of $\mathbf{D}$. Figure 19.7b shows the mean appearance and the PCA basis. Figure 19.7c contains face images generated using the Active Appearance Model (AAM) by setting appropriate parameters of shape and texture.

Once the appearance and shape model have been learned from training samples (i.e., $\mathbf{U}$, $\mathbf{U}^s$ is known), alignment is achieved by finding the motion parameter $\mathbf{p}$ that

---

[1]Bold uppercase letters denote matrices (e.g., $\mathbf{D}$), bold lowercase letters denote column vectors (e.g., $\mathbf{d}$). $\mathbf{d}_j$ represents the jth column of the matrix $\mathbf{D}$. d_{ij} denotes the scalar in the row ith and column jth of the matrix $\mathbf{D}$. Non-bold letters represent scalar variables. $\text{tr}(\mathbf{D}) = \sum_i d_{ii}$ is the trace of square matrix $\mathbf{D}$. $\|\mathbf{d}\|_2 = \sqrt{\mathbf{d}^T \mathbf{d}}$ designates Euclidean norm of $\mathbf{d}$.

best aligns the image w.r.t. the subspace $\mathbf{U}$ by minimizing:

$$\min_{\mathbf{c}, \mathbf{p}} \left\| \mathbf{d}\big(\mathbf{f}(\mathbf{x}, \mathbf{p})\big) - \mathbf{U}\mathbf{c} \right\|_2^2, \tag{19.1}$$

where $\mathbf{c}$ is the vector for the appearance coefficients. $\mathbf{x} = [x_1, y_1, \ldots, x_l, y_l]^T$ is the coordinate vector with the pixels to track. $\mathbf{f}(\mathbf{x}, \mathbf{p})$ is the function for geometric transformation; the value of $\mathbf{f}(\mathbf{x}, \mathbf{p})$ is a vector denoted by $[u_1, v_1, \ldots, u_l, v_l]^T$. $\mathbf{d}$ is the image frame in consideration, and $\mathbf{d}(\mathbf{f}(\mathbf{x}, \mathbf{p}))$ is the appearance vector of which the ith entry is the intensity of image $\mathbf{d}$ at pixel (u_i, v_i). For affine and non-rigid transformations, (u_i, v_i) relates to (x_i, y_i) by:

$$\begin{bmatrix} u_i \\ v_i \end{bmatrix} = \begin{bmatrix} a_1 & a_2 \\ a_4 & a_5 \end{bmatrix} \begin{bmatrix} x_i^s \\ y_i^s \end{bmatrix} + \begin{bmatrix} a_3 \\ a_6 \end{bmatrix}. \tag{19.2}$$

Here $[x_1^s, y_1^s, \ldots, x_l^s, y_l^s]^T = \mathbf{x} + \mathbf{U}^s \mathbf{c}^s$. The affine and non-rigid motion parameters are $\mathbf{a}, \mathbf{c}^s$, respectively, and $\mathbf{p} = [\mathbf{a}; \mathbf{c}^s]$ a combination of both affine and non-rigid motion parameters. In the case of the Lukas–Kanade tracker [74], $\mathbf{c}$ is fixed to be one and $\mathbf{U}$ is the subspace that contains a single vector, the reference template which is the appearance of the tracked object in the initial/previous frame.

Given an unseen facial image $\mathbf{d}$, facial feature detection or tracking with PAM alignment algorithms optimize (19.1). Due to the high dimensionality of the motion space, a standard approach to efficiently search over the parameter space is to use gradient-based methods [5, 10, 12, 28, 31, 87]. To compute the gradient of the cost function given in (19.1), it is common to use Taylor series expansion to approximate:

$$\mathbf{d}\big(\mathbf{f}(\mathbf{x}, \mathbf{p} + \delta\mathbf{p})\big) \approx \mathbf{d}\big(\mathbf{f}(\mathbf{x}, \mathbf{p})\big) + \mathbf{J}_\mathbf{p}\mathbf{d}(\mathbf{p})\delta\mathbf{p}, \tag{19.3}$$

where $\mathbf{J}_\mathbf{p}\mathbf{d}(\mathbf{p}) = \frac{\partial \mathbf{d}(\mathbf{f}(\mathbf{x}, \mathbf{p}))}{\partial \mathbf{p}}$ is the Jacobian of the image $\mathbf{d}$ w.r.t. to the motion parameter $\mathbf{p}$ [74]. Once linearized, a standard approach is to use the Gauss–Newton method for optimization [10, 12]. Other approaches learn an approximation of the Jacobian matrix with linear [28] or non-linear [71, 102] regression. Figure 19.9a shows an example of tracking 66 facial features with an AAM in the RU-FACS database [7].

19.4.2 Registration and Feature Extraction

After the face has been detected and the facial feature points have been tracked, the next two steps registration and feature extraction follow.

Registration: The main goal of registration is to normalize the image to remove 3D rigid head motion, so features can be geometrically normalized. 3D transformations could be estimated from monocular (up to a scale factor) or multiple cameras using structure from motion algorithms [51, 130]. However, if there is not much out of plane rotation (i.e. less than about 15 to 20 degrees) and the face is relatively far
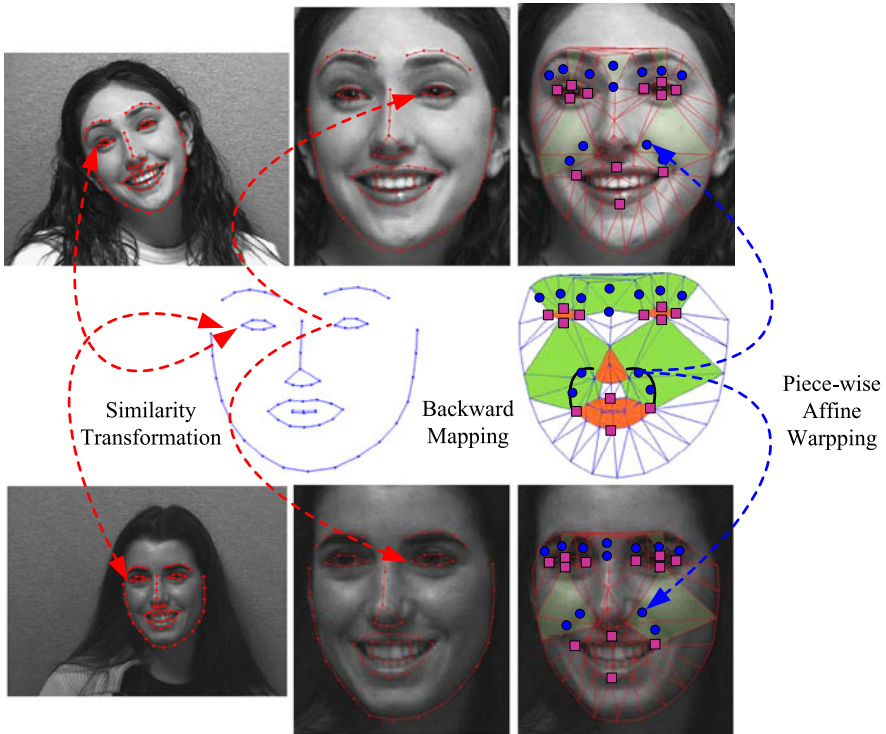
Fig. 19.8 Registration with two-step alignment. Figure reproduced with permission from [142]. © 2009 IEEE

away from the camera (assume orthographic projection), the 2D projected motion field of a 3D planar surface can be recovered with an affine model of six parameters. In this situation, simpler algorithms may be used to register the image to extract normalized facial features.

Following [108, 142] a similarity transform registers facial features with respect to an average face (see middle column in Fig. 19.8). To extract appearance representations in areas that have not been explicitly tracked (e.g. nasolabial furrow), we use a backward piece-wise affine warp with Delaunay triangulation. Fig. 19.8 shows the two step process for registering the face to a canonical pose for facial expression recognition. Purple squares represent tracked points and blue dots represent non-tracked meaningful points. The dashed blue line shows the mapping between the point in the mean shape and the corresponding points on the original image. Using an affine transformation plus backwarping, we can preserve the shape variation in appearance better than by geometric normalization alone. This two-step registration proves particularly important to detect low intensity AUs.

Geometric features: After the registration step, the shape and appearance features can be extracted from the normalized image. Geometric features contain information about shape and the locations of permanent facial features (e.g., eyes, brows,
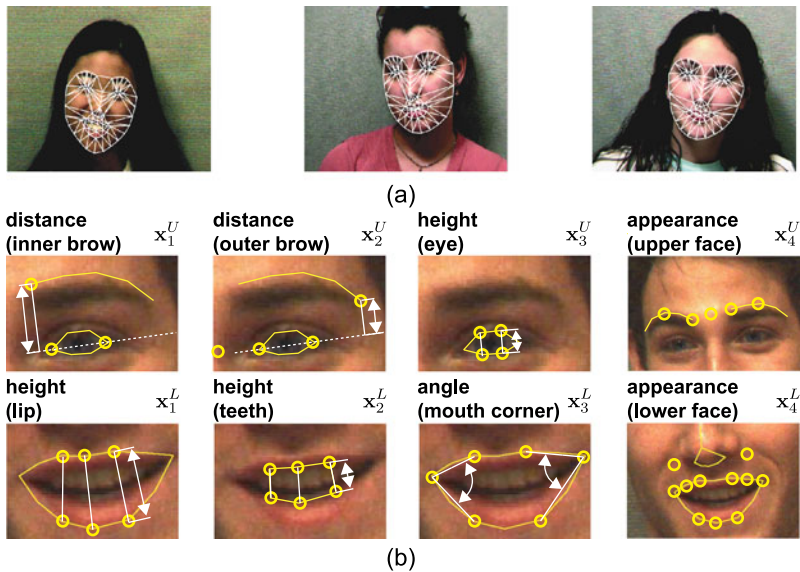
Fig. 19.9 **a** AAM fitting across different subjects. **b** Eight different features extracted from distance between tracked points, height of facial parts, angles for mouth corners, and appearance patches. Figure reproduced with permission from [144]. © 2010 IEEE

nose). Approaches that use only geometric features (or their derivatives) mostly rely on detecting sets of fiducial facial points [94, 96, 123], a connected face mesh or active shape model [20, 22, 61], or face component shape parametrization [113]. Some prototypical features include [108]: $\mathbf{x}_1^U$ the distance between inner brow and eye, $\mathbf{x}_2^U$ the distance between outer brow and eye, $\mathbf{x}_3^U$ the height of eye, $\mathbf{x}_1^L$ the height of lip, $\mathbf{x}_2^L$ the height of teeth, and $\mathbf{x}_3^L$ the angle of mouth corners, see Fig. 19.9b. However, shape features alone are unlikely to capture differences between subtle facial expressions or ones that are closely related. Many action units that are easily confusable by shape (e.g., AU 6 and AU 7 in FACS) can be discriminated by differences in appearance (e.g., furrows lateral to the eyes and cheek raising in AU 6 but not AU 7). Other AUs such as AU 11 (nasolabial furrow deepener), 14 (mouth corner dimpler), and 28 (inward sucking of the lips) cannot be detected from the movement of a sparse set of points alone but may be detected from changes in skin texture.

Appearance features: Represent the appearance (skin texture) changes and texture of the face, such as wrinkles and furrows. Appearance features for AU detection [3, 7, 50, 68, 111] outperformed shape only features for some action units, especially when registration is noisy see Lucey et al. [4, 77, 81] for a comparison.

Several approaches to appearance have been explored. Gabor wavelet coefficients are a popular approach. In several studies, Gabor wavelet coefficients outperformed optical flow, shape features, and Independent Component Analysis representations [3]. Tian [111, 113], however, reported that the combination of shape and appearance achieved better results than either shape or appearance alone. Recently,

Zhu et al. [142] have explored the use of SIFT [73] and DAISY [114] descriptors as appearance features. Given feature points tracked with AAMs, SIFT descriptors are first computed around the points of interest. SIFT descriptors are computed from the gradient vector for each pixel in the neighborhood to build a normalized histogram of gradient directions. For each pixel within a subregion, SIFT descriptors add the pixel's gradient vector to a histogram of gradient directions by quantizing each orientation to one of 8 directions and weighting the contribution of each vector by its magnitude. Similar in spirit to SIFT descriptors, DAISY descriptors are an efficient feature descriptor based on histograms. They are often used to match stereo images [114]. DAISY descriptors use circular grids instead of SIFT descriptors' regular grids; the former have been found to have better localization properties [89] and to outperform many state-of-the-art feature descriptors for sparse point matching [115]. At each pixel, DAISY builds a vector made of values from the convolved orientation maps located on concentric circles centered on the location. The amount of Gaussian smoothing is proportional to the radius of the circles. Donato [37] combined Gabor wavelet decomposition and independent component analysis. These representations use graylevel texture filters that share properties of spatial locality, independence, and have relationships to the response properties of visual cortical neurons. Zheng [138] investigated the use of two types of features extracted from face images for recognizing facial expressions. The first type is the geometric positions of a set of fiducial points on a face. The second type is a set of multi-scale and multi-orientation Gabor wavelet coefficients extracted from the face image at the fiducial points.

Other features: Other popular technique for feature extraction include more dynamic features such as optical flow [3], dynamic textures [21] and Motion History Images (MHIs) [16]. In an early exploration of facial expression recognition, Mase [86] used optical flow to estimate the activity in a subset of the facial muscles. Essa [43] extended this approach by using optic flow to estimate activity in a detailed anatomical and physical model of the face. Motion estimates from optic flow were refined by the physical model in a recursive estimation and control framework. The estimated forces were used to classify facial expressions. Yacoob and Davis [131] bypassed the physical model and constructed a mid-level representation of facial motion, such as a right mouth corner raise, directly from the optical flow. Ira et al. [22] implicitly recovered motion representations by building features such that each feature motion corresponded to a simple deformation of the face. MHIs were First proposed by Davis and Bobick [16]. MHIs compress the motion over a number of frames into a single image. This is done by layering the thresholded differences between consecutive frames one over the other. Valstar et al. [121] encoded face motion into Motion History Images. Zhao et al. [139] use volume Local Binary Patterns (LBPs), a temporal extension of local binary patterns often used in 2D texture analysis. The face is divided into overlapping blocks and the extracted LBP features in each block are concatenated into a single feature vector.

19.5 Supervised Learning

Supervised and more recently unsupervised approaches to action unit and expression detection have been pursued. In supervised learning event categories are defined in advance in labeled training data. In unsupervised learning no labeled training data are available and event categories must be discovered. In this section we discuss the supervised approach.

Early work in supervised learning sought to detect the six universal expressions of joy, surprise, anger, fear, disgust, and sadness; see Fig. 19.1. More recent work has attempted to detect expressions of pain [4, 69, 79], drowsiness, adult attachment [135], and indices of psychiatric disorder [27, 60]. Action unit detection remains a compelling challenge especially in unposed facial behavior. An open question is whether emotion and similar judgment-based categories are best detected by first detecting AU or by direct detection in which an AU detection step is bypassed. Work on this topic is just beginning [70, 79] and the question remains open.

Whether the focus is expression or AU, two main approaches have been pursued for supervised learning. These are (1) static modeling—typically posed as a discriminative classification problem in which each video frame is evaluated independently; and (2) temporal modeling—in which frames are segmented into sequences and typically modeled with a variant of DBNs (e.g. Hidden Markov Models, Conditional Random Fields).

19.5.1 Classifiers

In the case of static models, different feature representations and classifiers for frame-by-frame facial expression detection have been extensively studied. The pioneering work of Black and Yacoob [13] recognized facial expressions by fitting local parametric motion models to regions of the face and then feeding the resulting parameters to a nearest neighbor classifier for expression recognition. Tian et al. [111] made use of ANN classifiers for facial expression recognition. Barlett et al. [7, 8, 68] used Gabor filters in conjunction with AdaBoost feature selection followed by a Support Vector Machine (SVM) classifier. Lee and Elgammal [65] used multi-linear models to construct a non-linear manifold that factorizes identity from expression. Lucey et al. [76, 81] evaluated different shape and appearance representations derived from an AAM facial feature tracker, and an SVM for classification. Similarly, [139] made use of SVM.

More recent work has focused on incorporating the dynamics of facial expressions to improve performance (i.e. temporal modeling). De la Torre et al. [35] used condensation and appearance models to simultaneously track and recognize facial expression. Chang et al. [20] used a low-dimensional Lipschitz embedding to build a manifold of shape variation across several people and then used I-condensation to simultaneously track and recognize expressions. A popular strategy is to use HMMs to temporally segment expressions by establishing a correspondence between the action's onset, peak, and offset and an underlying latent state. Valstar and Pantic [123]

used a combination of SVM and HMM to temporally segment and recognize AUs. Valstar and Pantic [94, 122, 125] proposed a system that enables fully automated robust facial expression recognition and temporal segmentation of onset, peak and offset from video of mostly frontal faces. Koelstra and Pantic [59] used Gentle-Boost classifiers on motion from a non-rigid registration combined with an HMM. Similar approaches include a nonparametric discriminant HMM from Shang and Chan [106], and partially observed Hidden Conditional Random Fields by Chang et al. [19]. For other comprehensive surveys see [44, 95, 113, 136]. Tong et al. [117] used DBNs with appearance features to detect facial action units in posed facial behavior. The correlation among action units served as priors in action unit detection. Ira et al. [22] used a BN classifiers for classifying the six universal expressions from video. In particular they used a Naive-Bayes classifiers and change the distribution from Gaussian to Cauchy, and use Gaussian Tree-Augmented Naive Bayes (TAN) classifiers to learn the dependencies among different facial motion features.

19.5.2 Selection of Positive and Negative Samples During Training

Previous research in expression and AU detection has emphasized types of registration methods, features and classifiers (e.g., [67, 97, 113, 134, 140]). Little attention has been paid to make efficiently use of the training data for assignment of video frames to positive and negative classes. Typically, assignment has been done in one of two ways. One is to assign to the positive class those frames that occur at the peak of each AU or proximal to it. Peaks refer to the maximum intensity of an action unit between the frame at which begins ("onset") and ends ("offset"). Negative class then is chosen by randomly sampling other AUs, including AU 0 or neutral. This approach suffers at least three drawbacks: (1) the number of training examples will often be small, which results in a large imbalance between positive and negative frames; and (2) peak frames may provide too little variability to achieve good generalization. These problems may be circumvented by following an alternative approach; that is to include all frames from onset to offset in the positive class. This approach improves the ratio of positive to negative frames and increases representativeness of positive examples. The downside is confusability of positive and negative classes. Onset and offset frames and many of those proximal or even further from them may be indistinguishable from the negative class. As a consequence, the number of false positives may dramatically increase. Moreover, how to make use of all negative samples in an efficient manner? Is there a better approach to selecting positive and negative training samples?

In this section, we consider two approaches that have shown promise; one static and one dynamic. We illustrate the methods with particular classifiers and features, but the methods are not specific to the specific features or classifiers. As before, we distinguish between static and dynamic approaches. In the former, video frames are assumed to be independent. In the latter, first-order dependencies are assumed.

19.5.2.1 Static Approach

Recently, Zhu et al. [142] proposed an extension of cascade AdaBoost called Dynamic Cascade Bidirectional Bootstrapping (DCBB) to iteratively select positive samples and improve AU detection performance. In the first iteration, DCBB selected only the peaks and the two neighboring frames as positive frames, and randomly sample other AUs and non-AUs as negative samples. As in standard AdaBoost [127], DCBB defines the false positive target ratio, the maximum acceptable false positive ratio per cascade stage, and the minimum acceptable true positive ratio per each of the cascades. DCBB uses Classification and Regression Tree (CART) [17] as weak classifier. Once a cascade of peak frame detectors is learned in the first iteration, DCBB enlarges the positive set to increase the discriminative performance of the whole classifier. The new positive samples are selected after running the current classifier (learned in the previous iteration) in the original training data and selecting for the new positive training set the frames that were classified as positive. Recall that we have only trained with the peak frames in the first iteration. For more details see [142].

Figure 19.10 shows the improvement in the Receiver–Operator Characteristic (ROC) curve for testing data (subjects not in the training) using DCBB for three AUs (AU12, AU14, AU17). The ROC is obtained by plotting true positives ratios against false positives ratios for different decision threshold values of the classifier. In each subfigure there are five or six ROCs corresponding to alternative selection strategies: using only peak in the first step (same as standard Cascade AdaBoost), running three or four iterations in DCBB (spread x), and using all the frames between onset and offset (All+Boost). That is, there are three results shown using different positive training samples: 1) peak frames (first step); 2) all frames between onset and offset (All+Boost); and 3) iterations of DCBB (spread x). The first number between lines | denotes the area under the ROC, the second number is the size of positive samples in the testing dataset and separated by / is the size of negative samples in the testing dataset. The third number denotes the size of positive samples in training working sets and separated by / the total frames of target AU in training datasets. We can observe that the area under the ROC for frame-by-frame detection is improved gradually during each learning stage and the performance improves faster for some AU rather than others. Improvement rate appears to be influenced by the base rate of the AU. For AU14 and AU17, fewer potential training samples are available than for AU12.

Top of Fig. 19.11 shows the manual labeling for AU12 of the subject S015. We can see eight instances of AU12 with varying intensities ranging from A (weak) to E (strong). The strong AUs are represented by rectangles of height 4 and the weak ones with height 1. The remaining eight figures illustrates the sample selection process for each of the instances of the AU12. In the top right of each subfigure there is the corresponding AU instance number. The black curve in the bottom of the subfigures represents the similarity between the peak and the neighboring frames. The peak is the maximum of the curve. The positive samples in the first step are represented by green asterisks, in the second iteration by red crosses, in the third iteration by
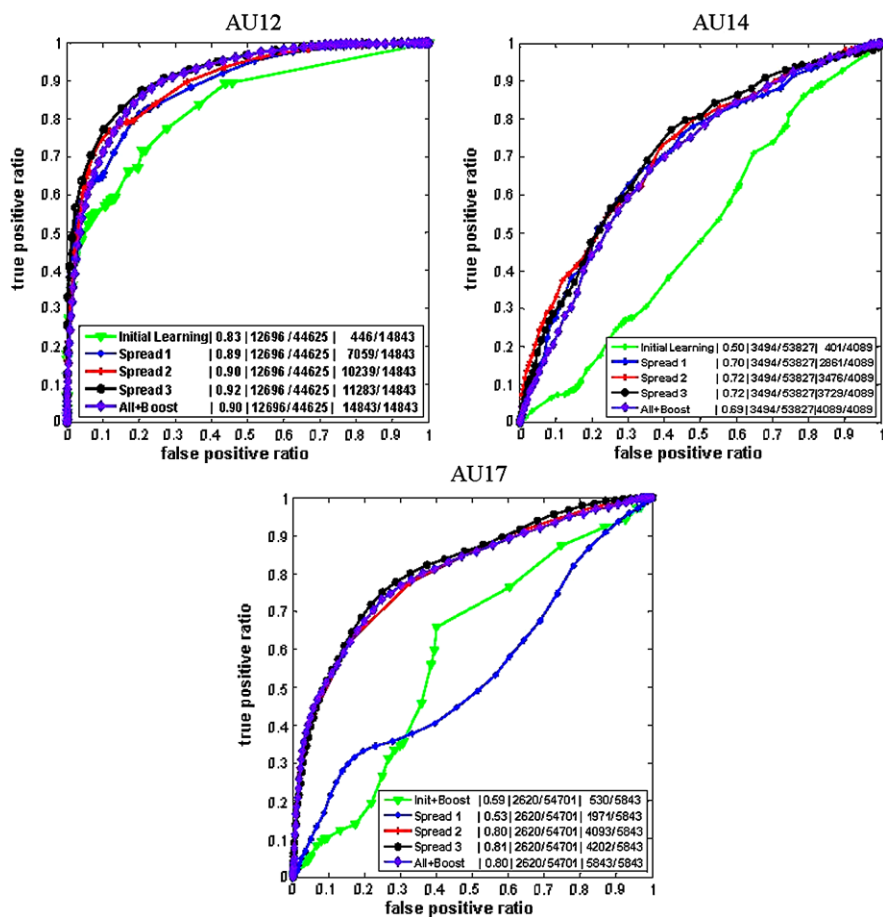
Fig. 19.10 ROCs for AU detection using DBCC: See text for the explanation of Init+Boost, spread x and All+Boost. Figure reproduced with permission from [142]. © 2009 IEEE

blue crosses, and in the final iteration by black circles. Observe that in the case of high peak intensity, subfigures 3 and 8 (top right number in the similarity plots), the final selected positive samples contain areas with low similarity values. However, when AU intensity is low, subfigure 7, the positive samples are only selected if they have a high similarity with the peak because otherwise we would select samples that will lead to many false positives. The ellipses and rectangles in the figures contain frames that are selected as positive samples, and correspond to strong and subtle AUs. The triangles correspond to frames between the onset and offset that are not selected as positive samples, and represent the ambiguous AUs.

Table 19.2 shows the area under the ROC for 14 AUs using DCBB and different set of features. The appearance features are based on SIFT descriptors. For all AUs the SIFT descriptor is built using a square of 48×48 pixels for twenty feature points for the lower face AUs or sixteen feature points for upper face. The shape
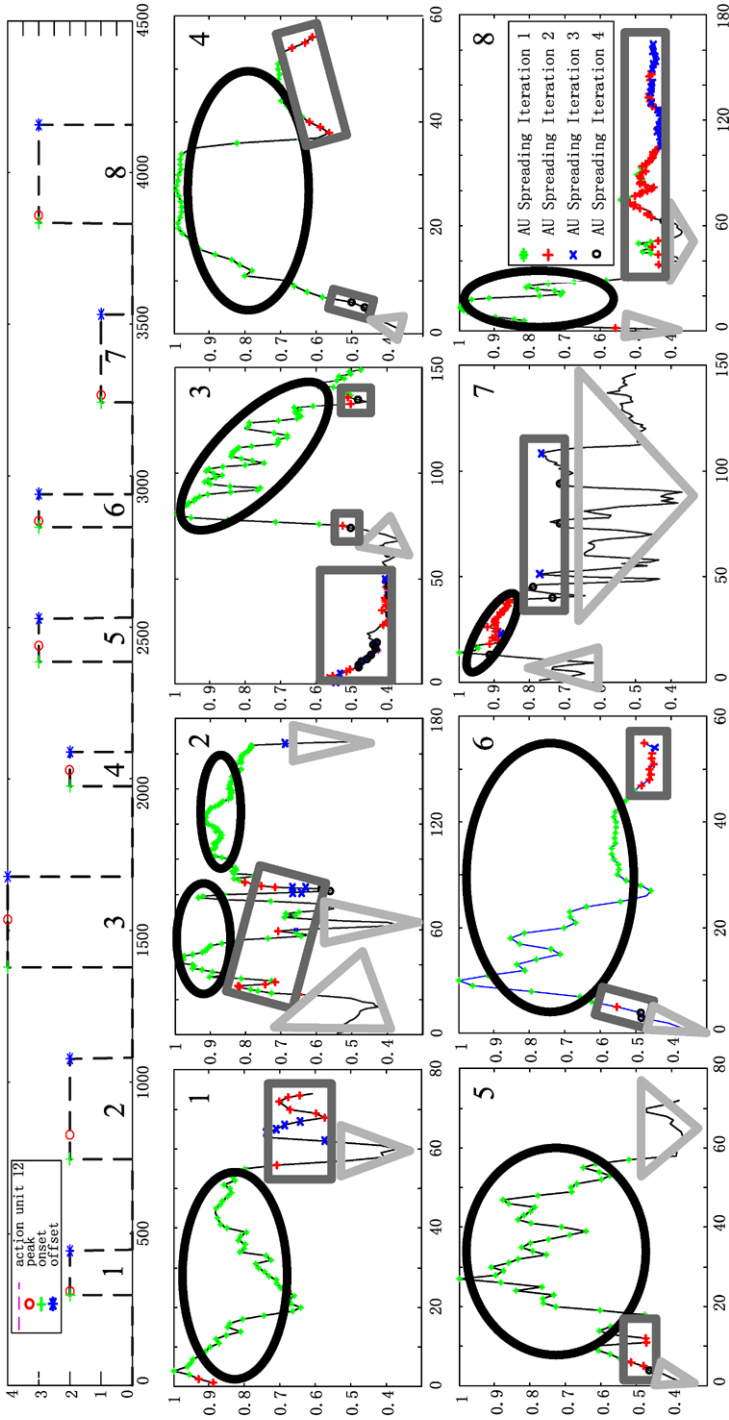
Fig. 19.11 The spreading of positive samples during each dynamic training step for AU12. See text for the explanation of the graphics. Figure reproduced with permission from [142]. © 2009 IEEE

Table 19.2 Area under the ROC for six different appearance and sampling strategies. AU peak frames with shape features and SVM (Peak+Shp+SVM), All frames between onset and offset with shape features and SVM (All+Shp+SVM), AU peak frames with appearance features and SVM (Peak+App+SVM), Sampling 1 frame in every 4 frames between onset and offset with PCA reduced appearance features and SVM (All+PCA+App+SVM), AU peak frames with appearance features and Cascade AdaBoost (Peak+App+Cascade Boost), DCBB with appearance features (DCBB)

|	AU1	AU2	AU4	AU5	AU6	AU7	AU10	AU12	AU14	AU15	AU17	AU18	AU23
Peak+Shp+SVM	0.71	0.62	0.76	0.58	0.93	0.64	0.61	0.89	0.57	0.73	0.66	0.86	0.74
All+Shp+SVM	0.68	0.65	0.78	0.55	0.88	0.64	0.67	0.77	0.72	0.61	0.72	0.88	0.67
Peak+App+SVM	0.43	0.45	0.61	0.86	0.77	0.67	0.63	0.90	0.50	0.69	0.53	**0.95**	0.62
All+PCA+App+SVM	0.74	0.74	**0.85**	**0.89**	0.96	0.63	0.54	0.91	**0.82**	**0.86**	0.78	0.94	0.54
Peak+App+Cascade Boost	0.75	0.71	0.53	0.49	0.93	0.56	0.52	0.83	0.50	0.52	0.59	0.57	0.34
DCBB	**0.76**	**0.75**	0.76	0.77	**0.97**	**0.69**	**0.72**	**0.92**	0.72	**0.86**	**0.81**	0.86	**0.75**

features are the landmarks of the AAM. For more details see [142]. It is important to notice that the results illustrated in this section are obtained using a particular set of features and classifiers, but the strategy of positive sample selection in principle can be used with any combination of classifiers and features.

19.5.2.2 Dynamic Approach

Extensions of DBNs have been a popular approach for expression analysis [22, 106, 117, 123]. A major challenge for DBNs based on generative models such as HMMs is how to effectively model the null class (none of the labeled classes) and how to train effectively on all possible segments of the video (rather than independent features). In this section, we review recent work on a temporal extensions of a bag-of-words (BoW) model called kSeg-SVM [108] that overcomes these drawbacks. kSeg-SVM is inspired by the success of the spatial BoW sliding-window model [15] that has been used in difficult object detection problems. We pose the AU detection problem as one of detecting temporal events (segments) in time series of visual features. Events correspond to AUs, including all frames between onset and offset (see Fig. 19.12). kSeg-SVM represents each segment as a BoW; however, the standard histogram of entries is augmented with a soft-clustering assignment of words to account for smoothly varying signals. Given several videos with AU labeled events, kSeg-SVM learns the SVM parameters that maximize the response on positive segments (AU to be detected) and minimize the response in the rest of the segments (all other positions and lengths). Figure 19.12 illustrates the main idea of kSeg-SVM.

kSeg-SVM can be efficiently trained on all available video using the Structure Output SVM (SO-SVM framework) [119]. Recent research [90] in the related area of sequence-labeling has shown that SO-SVMs can out-perform other algorithms including HMM, CRF[63] and Max-Margin Markov Networks [109]. SO-SVMs
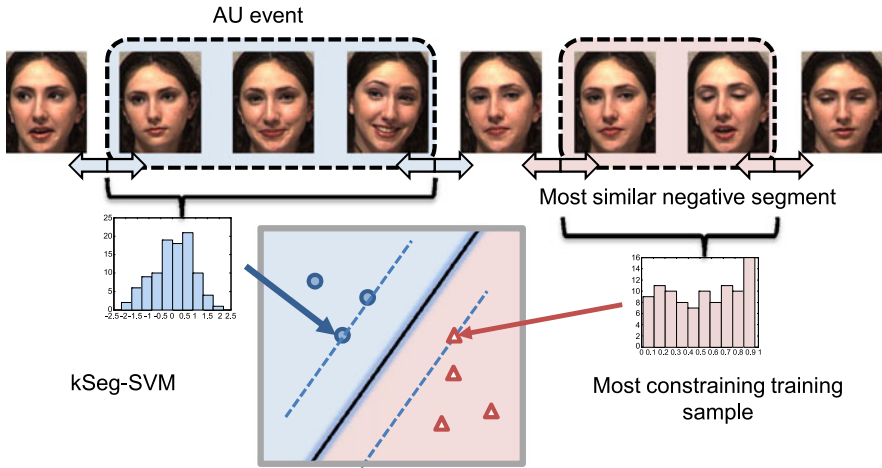
Fig. 19.12 During testing, the AU events are found by efficiently searching over the segments (position and length) that maximize the SVM score. During training, the algorithm searches over all possible negative segments to identify those hardest to classify, which improves classification of subtle AUs. Figure reproduced with permission from [108]. © 2009 IEEE

have several benefits in the context of AU detection: (1) they model the dependencies between visual features and the duration of AUs; (2) they can be trained effectively on all possible segments of the video (rather than on independent sequences); (3) they explicitly select negative examples that are most similar to the AU to be detected; and (4) they make no assumptions about the underlying structure of the AU events (e.g., i.i.d.). Finally, a novel parameterization of the output space is proposed to handle multiple AU event occurrences such that occur in long time series and search simultaneously for the k-or-fewer best matching segments in the time-series.

Given frame-level features, we will denote each processed video sequence i as $\mathbf{x}_i \in \mathbb{R}^{d \times m_i}$, where d is the number of features and m_i is the number of frames in the sequence. To simplify, we will assume that each sequence contains at most one occurrence of the AU event to be detected. For extensions to k-or-fewer occurrences see [108]. The AU event will be described by its corresponding onset to offset frame range and will be denoted by $\mathbf{y}_i \in \mathbb{Z}^2$. Let the full training set of video sequences be $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathcal{X}$, and their associated ground truth annotations for the occurrence of AUs $\mathbf{y}_1, \ldots, \mathbf{y}_n \in \mathcal{Y}$. We wish to learn a mapping $f : \mathcal{X} \to \mathcal{Y}$ for automatically detecting the AU events in unseen signals. This complex output space contains all contiguous time intervals; each label $\mathbf{y}_i$ consists of two numbers indicating the onset and the offset of an AU:

$$\mathcal{Y} = \left\{ \mathbf{y} \mid \mathbf{y} = \emptyset \text{ or } \mathbf{y} = [s, e] \in \mathbb{Z}^2, 1 \leq s \leq e \right\}. \tag{19.4}$$

Table 19.3 Performance on the RU-FACS-1 dataset, ROC metric and F1 metric. Higher numbers indicate better performance, and best results are printed in bold

AU	Area under ROC					Max F1 score				
	SVM	HMM2	HMM4	BoW-kSeg	kSeg-SVM	SVM	HMM2	HMM4	BoW-kSeg	kSeg-SVM
1	**0.86**	0.85	0.83	0.52	**0.86**	0.48	0.43	0.39	0.13	**0.59**
2	0.79	0.71	0.62	0.45	**0.81**	0.42	0.42	0.18	0.14	**0.56**
6	0.89	**0.92**	**0.92**	0.69	0.91	0.50	0.62	**0.63**	0.28	0.59
12	0.94	0.94	**0.95**	0.77	0.94	0.74	0.76	0.77	0.61	**0.78**
14	**0.70**	**0.70**	0.69	0.56	0.68	0.20	0.18	0.12	0.17	**0.27**
15	**0.90**	0.86	0.85	0.49	**0.90**	0.50	0.26	0.25	0.04	**0.59**
17	**0.90**	0.76	0.85	0.51	0.87	0.55	0.38	0.28	0.06	**0.56**
24	**0.85**	0.83	0.67	0.52	0.73	0.15	**0.18**	0.05	0.04	0.08
1+2	0.86	0.67	0.77	0.46	**0.89**	0.36	0.31	0.31	0.12	**0.56**
6+12	0.95	**0.98**	**0.98**	0.69	0.96	0.55	**0.64**	0.63	0.28	0.62

The empty label $\mathbf{y} = \emptyset$ indicates no occurrence of the AU. We will learn the mapping f as in the structured learning framework [15, 120] as

$$f(\mathbf{x}) = \underset{\mathbf{y} \in \mathcal{Y}}{\operatorname{argmax}}\, g(\mathbf{x}, \mathbf{y}), \tag{19.5}$$

where $g(\mathbf{x}, \mathbf{y})$ assigns a score to any particular labeling $\mathbf{y}$; the higher this value is, the closer $\mathbf{y}$ is to the ground truth annotation. For structured output learning, the choice of $g(\mathbf{x}, \mathbf{y})$ is often taken to be a weighted sum of features in the feature space:

$$g(\mathbf{x}, \mathbf{y}) = \mathbf{w}^T \varphi(\mathbf{x}, \mathbf{y}), \tag{19.6}$$

where $\varphi(\mathbf{x}, \mathbf{y})$ is a joint feature mapping for temporal signal $\mathbf{x}$ and candidate label $\mathbf{y}$, and $\mathbf{w}$ is the weight vector. Learning f can therefore be posed as an optimization problem:

$$\min_{\mathbf{w}, \boldsymbol{\xi}} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \xi_i,$$

$$\text{s.t. } \mathbf{w}^T \varphi(\mathbf{x}_i, \mathbf{y}_i) \geq \mathbf{w}^T \varphi(\mathbf{x}_i, \mathbf{y}) + \Delta(\mathbf{y}_i, \mathbf{y}) - \xi_i \; \forall \mathbf{y},$$

$$\xi_i \geq 0 \; \forall i. \tag{19.7}$$

Here, $\Delta(\mathbf{y}_i, \mathbf{y})$ is a loss function that decreases as a label $\mathbf{y}$ approaches the ground truth label $\mathbf{y}_i$. Intuitively, the constraints in (19.7) force the score of $g(\mathbf{x}, \mathbf{y})$ to be higher for the ground truth label $\mathbf{y}_i$ than for any other value of $\mathbf{y}$, and moreover, to exceed this value by a margin equal to the loss associated with labeling $\mathbf{y}$.

Table 19.3 shows the experimental results on the RU-FACS-1 dataset. As can be seen, kSeg-SVM consistently outperforms frame-based classification. It has the

highest ROC area for seven out of 10 AUs. Using the ROC metric, *k*Seg-SVM appears comparable to standard SVM. *k*Seg-SVM achieves highest F1 score on nine out of 10 test cases. As shown in Table 19.3, BoW-kSeg performs poorly. There are two possible reasons for this. First, clustering is done with *K*-means, an unsupervised, non-discriminative method that is not informed by the ground truth labels. Second, due to the hard dictionary assignment, each frame is forced to commit to a single cluster. While hard-clustering shows good performance in the task of object-detection, our time-series vary smoothly, resulting in large groups of consecutive frames being assigned to the same cluster.

At this point, it is worth pointing out that until now, a common measure of classifier performance for AU detection has been area under the curve (i.e. ROC). In object detection, the common measure represents the relation between recall and precision. The two approaches give very different views of classifier performance. This difference is not unanticipated in the object detection literature, but little attention has been paid to this issue in facial expression literature. In pattern recognition and machine learning, a common evaluation strategy is to consider correct classification rate (classification accuracy) or its complement error rate. However, this assumes that the natural distribution (prior probabilities) of each class are known and balanced. In an imbalanced setting, where the prior probability of the positive class is significantly less than the negative class (the ratio of these being defined as the skew), accuracy is inadequate as a performance measure since it becomes biased toward the majority class. That is, as the skew increases, accuracy tends toward majority class performance, effectively ignoring the recognition capability with respect to the minority class. This is a very common (if not the default) situation in facial expression recognition setting, where the prior probability of each target class (a certain facial expression) is significantly less than the negative class (all other facial expressions). Thus, when evaluating performance of automatic facial expression recognizer, other performance measures such as precision (this indicates the probability of correctly detecting a positive test sample and it is independent of class priors), recall (this indicates the fraction of the positives detected that are actually correct and, as it combines results from both positive and negative samples, it is class prior dependent), F1-measure (this is calculated as 2*recall*precision/(recall + precision)), and ROC (this is calculated as P(x|positive)/P(x|negative), where P(x|C) denotes the conditional probability that a data entry has the class label C, and where a ROC curve plots the classification results from the most positive to the most negative classification) are more appropriate.

19.6 Unsupervised Learning

With few exceptions, previous work on facial expression or action unit recognition has been supervised in nature. Little attention has been paid to the problem of unsupervised temporal segmentation or clustering facial events prior to recognition. Essa

and Pentland [43] proposed an unsupervised probabilistic flow-based method to describe facial expressions. Hoey [53] presented a multilevel BN to learn in a weakly supervised manner the dynamics of facial expression. Bettinger et al. [11] used AAMs to learn the dynamics of person-specific facial expression models. Zelnik-Manor and Irani [133] proposed a modification of structure-from-motion factorization to temporally segment rigid and non-rigid facial motion. De la Torre et al. [32] proposed a geometric-invariant clustering algorithm to decompose a stream of one person's facial behavior into facial gestures. Their approach suggested that unusual facial expressions might be detected through temporal outlier patterns. In recent work, Zhou et al. [143] proposed Aligned Cluster Analysis (ACA), an extension of spectral clustering for time series clustering and embedding. ACA was applied to discover in unsupervised manner facial actions across individuals that achieves moderate agreement with FACS. In this section, we briefly illustrate the applications of ACA for facial expression analysis, and refer the reader to [141, 143] for further details.

19.6.1 Facial Event Discovery for One Subject

Figure 19.13 shows the results of running unsupervised ACA on a video sequence of 1000 frames to summarize the facial expression of an infant into 10 temporal clusters. Appearance and shape features in the eyes and mouth, as described in Sect. 19.4.2, are used for temporal clustering. These 10 clusters provide a summarization of the infant's facial events. This visual summarization can be useful to automatically count the amount of time that the baby spends doing a particular facial expression (i.e. temporal cluster), such as crying, smiling or sleeping.

Extensions of ACA [143] can be used for facial expression indexing, given a sequence labeled by a user. Figure 19.14a on the left shows a frame of a sequence labeled by the user, and to the right there are six frames corresponding to six sequences returned by Supervised ACA (SACA). Next to the frames one can observe the matching score, which become higher the closer the retrieved sequence is to the user-specified sequence of facial expression.
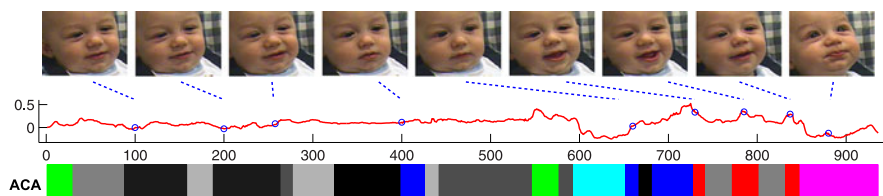

Fig. 19.13 Temporal clustering of infant facial behavior. Each color denotes a temporal unique cluster. Each facial gesture is coded with a different color. Observe how the frames of the same cluster correspond to similar facial expressions. Figure reproduced with permission from [108]. © 2010 IEEE
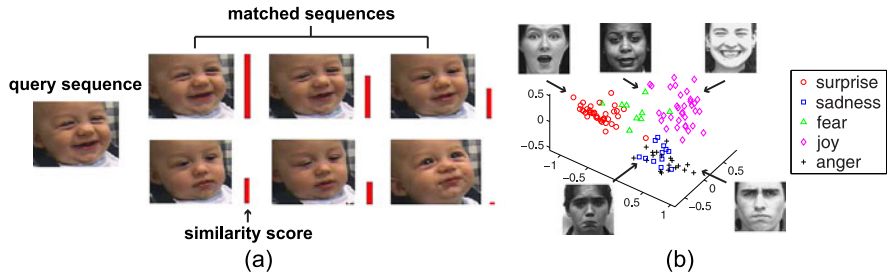
Fig. 19.14 a Facial expression indexing. The user specifies a query sequence and Supervised ACA returns six sequences with similar facial behavioral content as the video sequence selected by the user. **b** Three-dimensional embedding of 30 subjects with different facial expressions from the Cohn–Kanade database

ACA inherits the benefits of spectral clustering algorithms in that it provides a mechanism for finding a semantic low-dimensional embedding. In an evaluation, we tested the ability of unsupervised ACA to temporally cluster images and provide a visualization tool of several emotion-labeled sequences. Figure 19.14b shows the ACA embedding of 112 sequences from 30 randomly selected subjects from the Cohn–Kanade database [58]. The frames are labeled with five emotion labels: surprise, sadness, fear, joy and anger. The number of facial expressions varies across subjects. It is important to notice that, unlike traditional dimensionality reduction methods, each three-dimensional point in the embedding represents a video segment (of possibly different length) containing different facial expression. The ACA's embedding provides a natural mechanism for visualizing facial events and detecting outliers.

19.6.2 Facial Event Discovery for Sets of Subjects

This section illustrates the ability of ACA to discover dynamic facial events in the more challenging database RU-FACS [7] that contains naturally occurring facial behavior of multiple people. For this database the labels are AUs. We randomly selected 10 sets of 5 people and reported the mean clustering results and variance. The clustering accuracy is measured as the overlap between the temporal segmentation provided by ACA and the manually labeled FACS. ACA achieved an average accuracy of 52.2% in clustering the lower face and 68.8% in upper face using AUs labels. Figure 19.15a shows the results for temporal segmentation achieved by ACA on subjects S012, S028 and S049. Each color denotes a temporal cluster discovered by ACA. Figure 19.15 shows some of the dynamic vocabularies for facial expression analysis discovered by ACA. The algorithm correctly discovered smiling, with and without speech as different facial events. Visual inspection of all subjects' data suggests that the vocabulary of facial events is moderately consistent with human evaluation. More details are given in [143].
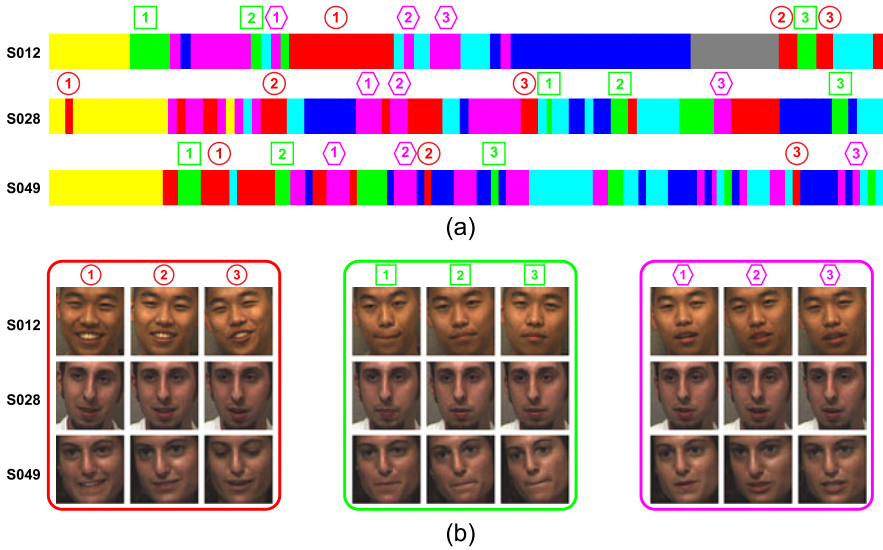
Fig. 19.15 **a** Results obtained by ACA for subjects S012, S028 and S049. **b** Corresponding video frames. Figure reproduced with permission from [108]. © 2010 IEEE

19.7 Conclusion and Future Challenges

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. Few challenges remain such as (1) how to detect subtle AUs: more robust 3D models that effectively decouple rigid and non-rigid motion and better models that normalize for subject variability are needed to be researched. (2) More robust real-time 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; new 3D sensors such as structure light cameras or time-of-flight cameras can are a promising direction for real-time segmentation (3) most work on facial expression analysis has been done in the area of recognition (temporal segmentation is provided), and more specialized machine learning algorithms are needed for the problem of detection in naturally occurring behavior.

Because most investigators have used relatively limited datasets (with typically unknown reliability), the generalizability of different approaches to facial expression analysis remains unknown. With few exceptions, investigators have failed to report inter-observer reliability and the validity of the facial expressions they have analyzed. 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 compara-

tive tests on common data, the relative strengths and weaknesses of different approaches are difficult to determine. In particular, there is need for fully FACS coded databases with natural occurring behavior. Because intensity and duration measurements are critical, it is important to include descriptive data on these features as well.

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 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.

At present, taxonomies of facial expression are based on FACS or other observer-based schemes. Consequently, approaches to automatic facial expression recognition are dependent on access to corpuses of FACS or similarly labeled video. This is a significant concern, in that recent work suggests that extremely large corpuses of labeled data may be needed to train robust classifiers. An open question in facial analysis is of whether facial actions can be learned directly from video in an unsupervised manner. That is, can the taxonomy be learned directly from video? And unlike FACS and similar systems that were initially developed to label static expressions, can we learn dynamic trajectories of facial actions? In our preliminary findings [143] on unsupervised learning using the using the RU-FACS database, agreement between facial actions identified by unsupervised analysis of face dynamics and FACS approached the level of agreement that has been found between independent FACS coders. These findings suggest that unsupervised learning of facial expression is a promising alternative to supervised learning of FACS-based actions. At least three benefits follow. One is the prospect that automatic facial expression analysis may be freed from its dependence on observer-based labeling. Second, because the current approach is fully empirical, it potentially can identify regularities in video that have not been anticipated by the top–down approaches such as FACS. New discoveries become possible. Three, similar benefits may accrue in other areas of image understanding of human behavior. Recent efforts by Guerra-Filho and Aloimonos [49] to develop vocabularies and grammars of human actions depend on advances in unsupervised learning. However, more robust and efficient algorithms that can learn from large databases are needed, as well as algorithms that can cluster more subtle facial behavior.

While research challenges in automated facial image and analysis remain, the time is near to apply these emerging tools to real-world problems in clinical science and practice, marketing, surveillance and human computer interaction.

References

1. Ambadar, Z., Cohn, J.F., Reed, L.I.: All smiles are not created equal: Morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous. J. Nonverbal Behav. **33**(1), 17–34 (2009) [379]

2. Ambadar, Z., Schooler, J.W., Cohn, J.F.: Deciphering the enigmatic face. Psychol. Sci. **16**(5), 403–410 (2005) [377]

3. Anderson, K., McOwan, P.W.: A real-time automated system for the recognition of human facial expressions. IEEE Trans. Syst. Man Cybern., Part B, Cybern. **36**(1), 96–105 (2006) [388,389]

4. Ashraf, A.B., Lucey, S., Cohn, J.F., Chen, T., Ambadar, Z., Prkachin, K.M., Solomon, P.E.: The painful face-pain expression recognition using active appearance models. Image Vis. Comput. **27**(12), 1788–1796 (2009) [388,390]

5. Baker, S., Matthews, I.: Lucas–Kanade 20 years on: A unifying framework. Int. J. Comput. Vis. **56**(3), 221–255 (2004) [386]

6. Bartlett, M., Littlewort, G., Fasel, I., Movellan, J.R.: Real time face detection and facial expression recognition: Development and applications to human computer interaction. In: CVPR Workshops for HCI (2003) [378]

7. Bartlett, M.S., Littlewort, G.C., Frank, M.G., Lainscsek, C., Fasel, I., Movellan, J.R.: Automatic recognition of facial actions in spontaneous expressions. J. Multimed. **1**(6), 22–35 (2006) [382,384,386,388,390,400]

8. Bartlett, M.S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., Movellan, J.: Fully automatic facial action recognition in spontaneous behavior. In: AFGR, pp. 223–230 (2006) [390]

9. Beebe, B., Badalamenti, A., Jaffe, J., Feldstein, S., Marquette, L., Helbraun, E.: Distressed mothers and their infants use a less efficient timing mechanism in creating expectancies of each other's looking patterns. J. Psycholinguist. Res. **37**(5), 293–307 (2008) [383]

10. Bergen, J.R., Anandan, P., Hanna, K.J., Hingorani, R.: Hierarchical model-based motion estimation. In: European Conference on Computer Vision, pp. 237–252 (1992) [386]

11. Bettinger, F., Cootes, T.F., Taylor, C.J.: Modelling facial behaviours. In: BMVC (2002) [399]

12. Black, M.J., Jepson, A.D.: Eigentracking: Robust matching and tracking of objects using view-based representation. Int. J. Comput. Vis. **26**(1), 63–84 (1998) [384,386]

13. Black, M.J., Yacoob, Y.: Recognizing facial expressions in image sequences using local parameterized models of image motion. Int. J. Comput. Vis. **25**(1), 23–48 (1997) [384]

14. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3D faces. In: SIGGRAPH (1999) [384]

15. Blaschko, M., Lampert, C.: Learning to localize objects with structured output regression. In: ECCV, pp. 2–15 (2008) [395,397]

16. Bobick, A., Davis, J.: The recognition of human movement using temporal templates. IEEE Trans. Pattern Anal. Mach. Intell. **23**(3), 257–267 (2001) [389]

17. Breiman, L.: Classification and Regression Trees. Chapman & Hall, London (1998) [392]

18. Bruce, V.: What the human face tells the human mind: Some challenges for the robot–human interface. In: IEEE Int. Workshop on Robot and Human Communication (1992) [378]

19. Chang, K.Y., Liu, T.L., Lai, S.H.: Learning partially-observed hidden conditional random fields for facial expression recognition. In: CVPR (2009) [391]

20. Chang, Y., Hu, C., Feris, R., Turk, M.: Manifold based analysis of facial expression. In: CVPR Workshops, p. 81 (2004) [388,390]

21. Chetverikov, D., Péteri, R.: A brief survey of dynamic texture description and recognition. In: Computer Recognition Systems, pp. 17–26 (2005) [389]

22. Cohen, I., Sebe, N., Cozman, F.G., Cirelo, M.C., Huang, T.S.: Learning Bayesian network classifiers for facial expression recognition using both labeled and unlabeled data. In: CVPR (2003) [388,389,391,395]

23. Cohen, I., Sebe, N., Garg, A., Chen, L.S., Huang, T.S.: Facial expression recognition from video sequences: Temporal and static modeling. Comput. Vis. Image Underst. **91**(1–2), 160–187 (2003) [384]

24. Cohn, J.F., Ambadar, Z., Ekman, P.: Observer-based measurement of facial expression with the facial action coding system. In: The Handbook of Emotion Elicitation and Assessment. Series in Affective Science. Oxford University Press, New York (2007) [379-381]

25. Cohn, J.F., Ekman, P.: Measuring facial action by manual coding, facial emg, and automatic facial image analysis. In: Handbook of Nonverbal Behavior Research Methods in the Affective Sciences, pp. 9–64 (2005) [378]

26. Cohn, J.F., Kanade, T.: Automated facial image analysis for measurement of emotion expression. In: The Handbook of Emotion Elicitation and Assessment, pp. 222–238 (2007) [381]

27. Cohn, J.F., Simon, T., Hoai, M., Zhou, F., Tejera, M., De la Torre, F.: Detecting depression from facial actions and vocal prosody. In: ACII (2009) [377,390]

28. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. IEEE Trans. Pattern Anal. Mach. Intell. **23**(6), 681–685 (2001) [384-386]

29. Dai, Y., Shibata, Y., Ishii, T., Hashimoto, K., Katamachi, K., Noguchi, K., Kakizaki, N., Ca, D.: An associate memory model of facial expressions and its application in facial expression recognition of patients on bed. In: ICME, pp. 591–594 (2001) [377]

30. Darwin, C.: The Expression of the Emotions in Man and Animals. Oxford University Press New York (1872/1998) [377]

31. De la Torre, F., Black, M.J.: Robust parameterized component analysis: theory and applications to 2d facial appearance models. Comput. Vis. Image Underst. **91**, 53–71 (2003) [386]

32. De la Torre, F., Campoy, J., Ambadar, Z., Cohn, J.: Temporal segmentation of facial behavior. In: International Conference on Computer Vision (2007) [399]

33. De la Torre, F., Collet, A., Cohn, J., Kanade, T.: Filtered component analysis to increase robustness to local minima in appearance models. In: CVPR (2007) [384]

34. De la Torre, F., Vitrià, J., Radeva, P., Melenchón, J.: Eigenfiltering for flexible eigentracking. In: ICPR (2000) [384]

35. De la Torre, F., Yacoob, Y., Davis, L.: A probabilistic framework for rigid and non-rigid appearance based tracking and recognition. In: AFGR, pp. 491–498 (2000) [390]

36. DePaulo, B., Lindsay, J., Malone, B., Muhlenbruck, L., Charlton, K., Cooper, H.: Cues to deception. Psychol. Bull. **129**(1), 74–118 (2003) [383]

37. Donato, G., Bartlett, M.S., Hager, J.C., Ekman, P., Sejnowski, T.J.: Classifying facial actions. IEEE Trans. Pattern Anal. Mach. Intell. **21**(10), 979–984 (1999) [389]

38. Ekman, P.: An argument for basic emotions. Cogn. Emot. **6**, 169–200 (1992) [377,378]

39. Ekman, P., Davidson, R.J., Friesen, W.V.: The Duchenne smile: Emotional expression and brain physiology II. J. Pers. Soc. Psychol. **58**(2), 342–353 (1990) [377]

40. Ekman, P., Friesen, W.: Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, Palo Alto (1978) [379]

41. Ekman, P., Huang, T.S., Sejnowski, T.J., Hager, J.C.: Final report to NSF of the planning workshop on facial expression understanding. Human Interaction Laboratory, University of California, San Francisco (1993) [378]

42. Ekman, P., Rosenberg, E.L.: What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS). Oxford University Press, London (2005) [383]

43. Essa, I.A., Pentland, A.P.: Coding, analysis, interpretation, and recognition of facial expressions. IEEE Trans. Pattern Anal. Mach. Intell. **19**(7), 757–763 (2002) [384,389,399]

44. Fasel, B., Luettin, J.: Automatic facial expression analysis: a survey. Pattern Recognit. **36**(1), 259–275 (2003) [384,391]

45. Forbes, E.E., Cohn, J.F., Allen, N.B., Lewinsohn, P.M.: Infant affect during parent–infant interaction at 3 and 6 months: Differences between mothers and fathers and influence of parent history of depression. Infancy **5**, 61–84 (2004) [378]

46. Gatica-Perez, D.: Automatic nonverbal analysis of social interaction in small groups: A review. Image Vis. Comput. **27**(12), 1775–1787 (2009) [377]

47. Griffin, K.M., Sayette, M.A.: Facial reactions to smoking cues relate to ambivalence about smoking. Psychol. Addict. Behav. **22**(4), 551 (2008) [377]
48. Gross, R., Matthews, I., Cohn, J.F., Kanade, T., Baker, S.: The cmu multi-pose, illumination, and expression (multi-pie) face database. Technical report, Carnegie Mellon University Robotics Institute, TR-07-08 (2007) [382]
49. Guerra-Filho, G., Aloimonos, Y.: A language for human action. Computer **40**, 42–51 (2007) [402]
50. Guo, G., Dyer, C.R.: Learning from examples in the small sample case: Face expression recognition. IEEE Trans. Syst. Man Cybern., Part B, Cybern. **35**(3), 477–488 (2005) [388]
51. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press, Cambridge (2000) [386]
52. Hatfield, E., Cacioppo, J.T., Rapson, R.L.: Primitive emotional contagion. Emotion and Social Behavior **13**, 151–177 (1992) [377]
53. Hoey, J.: Hierarchical unsupervised learning of facial expression categories. In: IEEE Workshop on Detection and Recognition of Events in Video, pp. 99–106 (2002) [399]
54. Huang, D., De la Torre, F.: Bilinear kernel reduced rank regression for facial expression synthesis. In: ECCV (2010) [378]
55. Izard, C.E., Huebner, R.R., Risser, D., Dougherty, L.: The young infant's ability to produce discrete emotion expressions. Dev. Psychol. **16**(2), 132–140 (1980) [378]
56. Jolliffe, I.T.: Principal Component Analysis. Springer, New York (1986) [385]
57. Jones, M.J., Poggio, T.: Multidimensional morphable models. In: ICCV (1998) [384]
58. Kanade, T., Cohn, J.F., Tian, Y.: Comprehensive database for facial expression analysis. In: AFGR (2000) [382,400]
59. Koelstra, S., Pantic, M.: Non-rigid registration using free-form deformations for recognition of facial actions and their temporal dynamics. In: AFGR (2008) [384,391]
60. Kohler, C.G., Martin, E.A., Stolar, N., Barrett, F.S., Verma, R., Brensinger, C., Bilker, W., Gur, R.E., Gur, R.C.: Static posed and evoked facial expressions of emotions in schizophrenia. Schizophr. Res. **105**, 49–60 (2008) [390]
61. Kotsia, I., Pitas, I.: Facial expression recognition in image sequences using geometric deformation features and support vector machines. IEEE Trans. Image Process. **16**, 172–187 (2007) [388]
62. Krumhuber, E., Manstead, A.S., Cosker, D., Marshall, D., Rosin, P.: Effects of dynamic attributes of smiles in human and synthetic faces: A simulated job interview setting. J. Nonverbal Behav. **33**(1), 1–15 (2009) [383]
63. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: ICML (2001) [395]
64. Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D.H.J., Hawk, S.T., van Knippenberg, A.: Presentation and validation of the Radboud Faces Database. Cogn. Emot. **24**(8), 1377–1388 (2010) [382]
65. Lee, C., Elgammal, A.: Facial expression analysis using nonlinear decomposable generative models. In: IEEE International Workshop on Analysis and Modeling of Faces and Gestures, pp. 17–31 (2005) [390]
66. Levenson, R.W., Ekman, P., Friesen, W.V.: Voluntary facial action generates emotion-specific autonomic nervous system activity. Psychophysiology **27**(4), 363–384 (1990) [377]
67. Li, S., Jain, A.: Handbook of Face Recognition. Springer, New York (2005) [391]
68. Littlewort, G., Bartlett, M.S., Fasel, I., Susskind, J., Movellan, J.: Dynamics of facial expression extracted automatically from video. Image Vis. Comput. **24**(6), 615–625 (2006) [384, 388,390]
69. Littlewort, G.C., Bartlett, M.S., Lee, K.: Automatic coding of facial expressions displayed during posed and genuine pain. Image Vis. Comput. **12**(27), 1797–1803 (2009) [390]
70. Littlewort, G., Bartlett, M.S., Whitehill, J., Wu, T.F., Butko, N., Ruvulo, P., et al.: The motion in emotion: A cert based approach to the fera emotion challenge. In: Paper presented at the 1st Facial Expression Recognition and Analysis challenge 2011, 9th IEEE International Conference on AFGR (2011) [390]

71. Liu, X.: Generic face alignment using boosted appearance model. In: CVPR (2007) [386]
72. Lo, H., Chung, R.: Facial expression recognition approach for performance animation. In: International Workshop on Digital and Computational Video (2001) [378]
73. Lowe, D.: Object recognition from local scale-invariant features. In: ICCV (1999) [389]
74. Lucas, B., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: Proceedings of Imaging Understanding Workshop (1981) [384,386]
75. Lucey, P., Cohn, J., Howlett, J., Lucey, S., Sridharan, S.: Recognizing emotion with head pose variation: Identifying pain segments in video. IEEE Trans. Syst. Man Cybern., Part B, Cybern. **41**(3), 664–674 (2011) [380]
76. Lucey, P., Cohn, J., Lucey, S., Sridharan, S., Prkachin, K.M.: Automatically detecting action units from faces of pain: Comparing shape and appearance features. In: CVPR Workshops (2009) [390]
77. Lucey, P., Cohn, J.F., Lucey, S., Sridharan, S., Prkachin, K.M.: Automatically detecting pain using facial actions. In: ACII (2009) [388]
78. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended Cohn–Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In: CVPR Workshops for Human Communicative Behavior Analysis (2010) [382,384]
79. Lucey, P., Cohn, J.F., Matthews, I., Lucey, S., Sridharan, S., Howlett, J., Prkachin, K.M.: Automatically detecting pain in video through facial action units. IEEE Trans. Syst. Man Cybern., Part B, Cybern. **PP**(99), 1–11 (2010) [377,390]
80. Lucey, P., Cohn, J.F., Prkachin, K.M., Solomon, P., Matthews, I.: Painful data: The UNBC-McMaster shoulder pain expression archive database. In: AFGR (2011) [382]
81. Lucey, S., Matthews, I., Hu, C., Ambadar, Z., De la Torre, F., Cohn, J.: AAM derived face representations for robust facial action recognition. In: AFGR (2006) [384,388,390]
82. Lyons, M., Akamatsu, S., Kamachi, M., Gyoba, J.: Coding facial expressions with gabor wavelets. In: AFGR (2002) [382,384]
83. Madsen, M., el Kaliouby, R., Eckhardt, M., Hoque, M., Goodwin, M., Picard, R.W.: Lessons from participatory design with adolescents on the autism spectrum. In: Proc. Computer Human Interaction (2009) [378]
84. Malatesta, C.Z., Culver, C., Tesman, J.R., Shepard, B., Fogel, A., Reimers, M., Zivin, G.: The Development of Emotion Expression During the First Two Years of Life. Monographs of the Society for Research in Child Development, pp. 97–136 (1989) [377]
85. Martinez, A.M., Benavente, R.: The AR face database. In: CVC Technical Report, number 24 (June 1998) [382]
86. Mase, K., Pentland, A.: Automatic lipreading by computer. Trans. Inst. Electron. Inf. Commun. Eng. **J73-D-II**(6), 796–803 (1990) [389]
87. Matthews, I., Baker, S.: Active appearance models revisited. Int. J. Comput. Vis. **60**(2), 135–164 (2004) [384,386]
88. Matthews, I., Xiao, J., Baker, S.: 2d vs 3d deformable face models: Representational power, construction, and real-time fitting. Int. J. Comput. Vis. **75**(1), 93–113 (2007) [385]
89. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Trans. Pattern Anal. Mach. Intell. **27**(10), 1615–1630 (2005) [389]
90. Nguyen, N., Guo, Y.: Comparisons of sequence labeling algorithms and extensions. In: ICML (2007) [395]
91. O'Toole, A.J., Harms, J., Snow, S.L., Hurst, D.R., Pappas, M.R., Ayyad, J.H., Abdi, H.: A video database of moving faces and people. IEEE Trans. Pattern Anal. Mach. Intell. **27**(5), 812–816 (2005) [382]
92. Pandzic, I.S., Forchheimer, R.R. (eds.): MPEG-4 Facial Animation: The Standard, Implementation and Applications. Wiley, New York (2002) [381]
93. Pantic, M., Bartlett, M.S.: Machine analysis of facial expressions. In: Face Recognition, pp. 377–416 (2007) [384]
94. Pantic, M., Patras, I.: Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences. IEEE Trans. Syst. Man Cybern., Part B, Cybern. **36**, 433–449 (2006) [388,391]

95. Pantic, M., Rothkrantz, L.J.M.: Automatic analysis of facial expressions: The state of the art. IEEE Trans. Pattern Anal. Mach. Intell. **22**(12), 1424–1445 (2002) [384,391]

96. Pantic, M., Rothkrantz, L.J.M.: Facial action recognition for facial expression analysis from static face images. IEEE Trans. Syst. Man Cybern., Part B, Cybern. **34**(3), 1449–1461 (2004) [388]

97. Pantic, M., Sebe, N., Cohn, J.F., Huang, T.: Affective multimodal human–computer interaction. In: ACM International Conference on Multimedia, pp. 669–676 (2005) [391]

98. Pentland, A.: Looking at people: Sensing for ubiquitous and wearable computing. IEEE Trans. Pattern Anal. Mach. Intell. **22**(1), 107–119 (2000) [377]

99. Pilz, S.K., Thornton, I.M., Bülthoff, H.H.: A search advantage for faces learned in motion. Exp. Brain Res. **171**(4) 436–447 (2006) [382]

100. Prkachin, K.M., Solomon, P.E.: The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain. Pain **139**(2), 267–274 (2008) [377]

101. Rademaker, R., Pantic, M., Valstar, M.F., Maat, L.: Web-based database for facial expression analysis. In: ICME (2005) [382]

102. Saragih, J., Goecke, R.: A nonlinear discriminative approach to AAM fitting. In: ICCV (2007) [386]

103. Sayette, M.A., Cohn, J.F., Wertz, J.M., Perrott, M.A., Parrott, D.J.: A psychometric evaluation of the facial action coding system for assessing spontaneous expression. J. Nonverbal Behav. **25**(3), 167–185 (2001) [381]

104. Scherer, K., Ekman, P.: Handbook of Methods in Nonverbal Behavior Research. Cambridge University Press, Cambridge (1982) [379]

105. Schmidt, K.L., Cohn, J.F.: Human facial expressions as adaptations: Evolutionary perspectives in facial expression research. Yearb. Phys. Antropol. **116**, 8–24 (2001) [379]

106. Shang, L.F., Chan, K.P.: Nonparametric discriminant HMM and application to facial expression recognition. In: CVPR (2009) [391,395]

107. Shergill, G.H., Sarrafzadeh, H., Diegel, O., Shekar, A.: Computerized sales assistants: The application of computer technology to measure consumer interest;a conceptual framework. J. Electron. Commer. Res. **9**(2), 176–191 (2008) [377]

108. Simon, T., Nguyen, M.H., De la Torre, F., Cohn, J.F.: Action unit detection with segment-based SVMs. In: Conference on Computer Vision and Pattern Recognition, pp. 2737–2744 (2010) [384,387,388,395,396,401]

109. Taskar, B., Guestrin, C., Koller, D.: Max-margin Markov networks. In: NIPS (2003) [395]

110. Theobald, B.J., Cohn, J.F.: Facial image synthesis. In: Oxford Companion to Emotion and the Affective Sciences, pp. 176–179. Oxford University Press, London (2009) [378]

111. Tian, Y., Kanade, T., Cohn, J.F.: Evaluation of Gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity. In: AFGR (2002) [388,390]

112. Tian, Y., Kanade, T., Cohn, J.F.: Recognizing action units for facial expression analysis. IEEE Trans. Pattern Anal. Mach. Intell. **23**(2), 97–115 (2002) [384]

113. Tian, Y., Kanade, T., Cohn, J.F.: Facial expression analysis. In: Handbook of Face Recognition, Springer, Berlin (2008) [384,388,391]

114. Tola, E., Lepetit, V., Fua, P.: A fast local descriptor for dense matching. In: CVPR (2008) [389]

115. Tola, E., Lepetit, V., Fua, P.: Daisy: An efficient dense descriptor applied to wide baseline stereo. IEEE Trans. Pattern Anal. Mach. Intell. **99**(1) (2009) [389]

116. Tomkins, S.S.: Affect, Imagery, Consciousness. Springer, New York (1962) [377]

117. Tong, Y., Liao, W., Ji, Q.: Facial action unit recognition by exploiting their dynamic and semantic relationships. IEEE Trans. Pattern Anal. Mach. Intell. **29** 1683–1699 (2007) [384, 391,395]

118. Tremeau, F., Malaspina, D., Duval, F., Correa, H., Hager-Budny, M., Coin-Bariou, L., Macher, J.P., Gorman, J.M.: Facial expressiveness in patients with schizophrenia compared to depressed patients and nonpatient comparison subjects. Am. J. Psychiatr. **162**(1), 92 (2005) [377]

119. Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y.: Large margin methods for structured and interdependent output variables. J. Mach. Learn. Res. **6**, 1453–1484 (2005) [395]

120. Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y.: Large margin methods for structured and interdependent output variables. J. Mach. Learn. Res. **6**, 1453–1484 (2005) [397]
121. Valstar, M., Pantic, M., Patras, I.: Motion history for facial action detection in video. In: IEEE Int'l Conf. on Systems, Man and Cybernetics, pp. 635–640 (2005) [389]
122. Valstar, M.F., Pantic, M.: Fully automatic facial action unit detection and temporal analysis. In: CVPR (2006) [391]
123. Valstar, M.F., Pantic, M.: Combined support vector machines and hidden Markov models for modeling facial action temporal dynamics. In: ICCV Workshop on HCI (2007) [384,388, 390,395]
124. Valstar, M.F., Pantic, M.: Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In: Proceedings of the EMOTION 2010 Workshop (2010) [382]
125. Valstar, M.F., Patras, I., Pantic, M.: Facial action unit detection using probabilistic actively learned support vector machines on tracked facial point data. In: CVPR Workshops (2005) [391]
126. van Dam, A.: Beyond wimp. IEEE Comput. Graph. Appl. **20**(1), 50–51 (2000) [377]
127. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: CVPR (2001) [384,392]
128. Vural, E., Bartlett, M., Littlewort, G., Cetin, M., Ercil, A., Movellan, J.: Discrimination of moderate and acute drowsiness based on spontaneous facial expressions. In: ICPR (2010) [377]
129. Wen, Z., Huang, T.S.: Capturing subtle facial motions in 3d face tracking. In: CVPR (2008) [384]
130. Xiao, J., Baker, S., Matthews, I., Kanade, T.: Real-time combined 2D+3D active appearance models. In: CVPR (2004) [386]
131. Yacoob, Y., Davis, L.S.: Recognizing human facial expressions from long image sequences using optical flow. IEEE Trans. Pattern Anal. Mach. Intell. **18**(6), 636–642 (2002) [384,389]
132. Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.J.: A 3d facial expression database for facial behavior research. In: AFGR (2006) [382]
133. Zelnik-Manor, L., Irani, M.: Temporal factorization vs. spatial factorization. In: ECCV (2004) [399]
134. Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Trans. Pattern Anal. Mach. Intell. **31**(1), 39–58 (2008) [391]
135. Zeng, Z., Hu, Y., Roisman, G.I., Wen, Z., Fu, Y., Huang, T.S.: Audio-visual emotion recognition in adult attachment interview. In: 8th International Conference on Multimodal Interfaces (2009) [390]
136. Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Trans. Pattern Anal. Mach. Intell. **31**(1), 31–58 (2009) [391]
137. Zhang, C., Zhango, Z.: A survey of recent advances in face detection. In: Technical Report, MSR-TR-2010-66 Microsoft Research (June 2010) [384]
138. Zhang, Z., Lyons, M., Schuster, M., Akamatsu, S.: Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron. In: AFGR (2002) [384,389]
139. Zhao, G., Pietikainen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans. Pattern Anal. Mach. Intell. **29**(6), 915–928 (2007) [389,390]
140. Zhao, W., Chellappa, R.: Face Processing: Advanced Modeling and Methods. Academic Press, San Diego (2006) [391]
141. Zhou, F., De la Torre, F., Hodgins, J.: Aligned cluster analysis for temporal segmentation of human motion. In: IEEE Automatic Face and Gesture Recognition (2008) [399]
142. Zhu, Y., De la Torre, F., Cohn, J.F.: Dynamic cascades with bidirectional bootstrapping for spontaneous facial action unit detection. In: ACII (2009) [384,387,389,392-395]

143. Zhou, F., De la Torre, F., Cohn, J.: Unsupervised discovery of facial events. In: CVPR (2010) [399,400,402]
144. Zhou, F., De la Torre, F., Cohn, J.F.: Unsupervised discovery of facial events. In: Conference on Computer Vision and Pattern Recognition, pp. 2574–2581 (2010) [388]
145. Zue, V.W., Glass, J.R.: Conversational interfaces: Advances and challenges. Proc. IEEE **88**(8), 1166–1180 (2002) [377]