Dynamics of Facial Expression: Recognition of Facial Actions and Their Temporal Segments From Face Profile Image Sequences

Maja Pantic, *Member, IEEE,* and Ioannis Patras, *Member, IEEE*

***Abstract*—Automatic analysis of human facial expression is a challenging problem with many applications. Most of the existing automated systems for facial expression analysis attempt to recognize a few prototypic emotional expressions, such as anger and happiness. Instead of representing another approach to machine analysis of prototypic facial expressions of emotion, the method presented in this paper attempts to handle a large range of human facial behavior by recognizing facial muscle actions that produce expressions. Virtually all of the existing vision systems for facial muscle action detection deal only with frontal-view face images and cannot handle temporal dynamics of facial actions. In this paper, we present a system for automatic recognition of facial action units (AUs) and their temporal models from long, profile-view face image sequences. We exploit particle filtering to track 15 facial points in an input face-profile sequence, and we introduce facial-action-dynamics recognition from continuous video input using temporal rules. The algorithm performs both automatic segmentation of an input video into facial expressions pictured and recognition of temporal segments (i.e., onset, apex, offset) of 27 AUs occurring alone or in a combination in the input face-profile video. A recognition rate of 87% is achieved.**

***Index Terms*—Computer vision, facial action units, facial expression analysis, facial expression dynamics analysis, particle filtering, rule-based reasoning, spatial reasoning, temporal reasoning.**

I. Introduction

THE human face is involved in a large variety of different activities. It houses the apparatus for speech production as well as the majority of our sensors (eyes, nose, mouth). Besides these biological functions, the human face provides a number of social signals essential for our public life. The face mediates person identification, attractiveness, and facial communicative cues, that is, facial expressions. Our utterances are accompanied by the appropriate facial expressions, which clarify what is said and whether it is supposed to be important, funny or serious. Facial expressions reveal our current focus of attention, synchronize the dialogue, signal comprehension or disagreement; in brief, they regulate our interactions with the environment and other persons in our vicinity [1]. As indicated by Mehrabian [2], whether the listener feels liked or disliked depends only for 7% on the spoken word, for 38% on vocal utterances, while facial expressions determine 55% of this feeling. Finally, facial expressions are our direct and naturally preeminent means of communicating emotions [1], [3]. Hence, facial expressions play a very important role in human face-to-face interpersonal interaction. Automatic analysis of facial expressions would, therefore, be highly beneficial for fields as diverse as behavioral science, psychology, medicine, security, education, and computer science (facilitating lip reading, face and visual speech synthesis, videoconferencing, affective computing, and anticipatory human-machine interfaces). It is this wide range of applications that has produced a surge of interest in machine analysis of facial expressions.

Most of the facial expression analyzers developed so far attempt to recognize a small set of prototypic emotional facial expressions, i.e., fear, sadness, disgust, anger, surprise, and happiness (e.g., [4]–[8]; for an exhaustive survey, see [9]). This practice may follow from the large body of psychological research (from Darwin [10] to Ekman [3], [11]) which argues that these "basic" emotions have corresponding prototypic facial displays. However, there is also a growing body of psychological research that argues that it is not prototypic expressions but some components of those expressions (e.g., "squared" mouth, raised eyebrows, etc.) which are commonly displayed and universally linked with the emotion labels listed above [1], [12]. To detect such subtle facial expressions and to make the facial expression information available for usage in the various applications mentioned above, automatic recognition of facial muscle actions (i.e., atomic facial signals) is needed.

A. *Facial Action Coding System (FACS)*

There are several methods for measuring and describing facial muscular activity [13]. From these, the FACS is the most widely used method in psychological research [13]. Ekman and Friesen developed the original FACS in the 1970s by determining how the contraction of each facial muscle (singly and in combination with other muscles) changes the appearance of the face. They examined videotapes of facial behavior to identify specific changes that occur with muscular contractions and how to differentiate one from another. They associated the facial appearance changes with the action of muscles that produce them. Namely, the changes in the facial expression are described with FACS in terms of 44 different action units (AUs), each of which is anatomically related to the contraction of either a specific facial muscle or of a set of facial muscles. Along with the definition of various AUs, FACS also provides the rules for visual detection of AUs and their temporal segments (onset, apex, offset)

Manuscript received December 15, 2004; revised June 10, 2004. This work was supported by the Netherlands Organization for Scientific Research (NWO) under Grant EW-639.021.202. This paper was recommended by Associate Editor S. Sarkar.

M. Pantic is with Delft University of Technology, Electrical Engineering, Mathematics and Computer Science, The Netherlands (e-mail: mpantic@ieee.org).

I. Patras is with the Department of Computer Science, The University of York, York Y010 5DD, U.K. (e-mail: I.Patras@cs.york.ac.uk).

Digital Object Identifier 10.1109/TSMCB.2005.859075

in a video of the observed face. Using these rules, a FACS coder (i.e., a human observer having a formal training in using FACS) "dissects" a shown facial expression, decomposing it into the specific AUs and their temporal segments that produced the expression. The FACS Manual was first published in 1978 [14]. The latest version was published in 2002 [15].

B. Automated FACS: Frontal Face

Although FACS provides a good foundation for AU coding of face images by human observers, automatic recognition of AUs by computers remains difficult. One problem is that AUs can occur in more than 7000 different combinations [13], causing bulges (e.g., by the tongue pushed under one of the lips) and various in-and out-of-plane movements of facial components (e.g., jetted jaw) that are difficult to detect in 2D face images. Few methods have been reported for automatic AU detection in face image sequences [16]. Some researchers described patterns of facial motion that correspond to a few specific AUs but did not report on actual recognition of these AUs (e.g., [4]–[6], [8], [17], [18]). Only recently there has been an emergence of efforts toward automatic analysis of facial expressions into elementary AUs [19]. For instance, the Machine Perception group at UCSD has proposed several methods for automatic AU coding of facial expressions. To detect 6 individual AUs in face image sequences free of head motions, Bartlett *et al.* [20] used a $61 \times 10 \times 6$ feed-forward neural network. They achieved 91% accuracy by feeding the pertinent network with the results of a hybrid system combining holistic spatial analysis and optical flow with local feature analysis. To recognize eight individual AUs and four combinations of AUs in face image sequences free of head motions, Donato *et al.* [21] used Gabor wavelet representation and independent component analysis. They reported a 95.5% average recognition rate achieved by their method. The most recent work by Bartlett *et al.* [22] reports on accurate automatic recognition of 18 AUs (95% average recognition rate) from near frontal-view face image sequences using Gabor filters and Support Vector Machines. Another group that has focused on automatic FACS coding of face image sequences is that led by Cohn and Kanade. To recognize eight individual AUs and seven combinations of AUs in face image sequences free of head motions, Cohn *et al.* [23] used facial feature point tracking and discriminant function analysis and achieved an 85% average recognition rate. Tian *et al.* [24] used lip tracking, template matching and neural networks to recognize 16 AUs occurring alone or in combination in near frontal-view face image sequences. They reported an 87.9% average recognition rate. The authors' group also reported on multiple efforts toward automatic analysis of facial expressions into atomic facial actions. The majority of this previous work concerns automatic AU recognition in static face images [7], [25]. Only recently, the authors' group has focused on automatic FACS coding of face video. To recognize 15 AUs occurring alone or in combination in near frontal-view face image sequences, Valstar *et al.* [26] used temporal templates (i.e., motion history images) and a combined k-Nearest-Neighbor and rule-based classifier. An average recognition rate of 65% was reported.

C. Automated FACS: Profile Face

In contrast to these previous approaches to automatic AU detection, which deal only with frontal-view face images and cannot code temporal segments (i.e., onset, apex, offset) of AUs [19], the research reported here addresses the problem of automatic detection of AUs and their temporal segments from profile-view face image sequences. It was undertaken with the following motivations.

1) In a frontal-view face image, facial actions such as tongue pushed under the upper lip (AU36t) or pushing the jaw forward (AU29) represent out-of-plane nonrigid movements that are difficult to detect. Such facial actions are clearly observable in a profile view. Hence, the use of face-profile view promises a qualitative enhancement of AU detection performed (by enabling detection of AUs that are difficult to encode in a frontal view).
2) Existing AU detectors achieve good recognition rates, but virtually all of them perform well only when the user faces the camera and does not change his/her three-dimensional (3-D) head pose. Robust AU detection, independent of rigid head movements that can cause changes in the viewing angle and the visibility of the tracked face and its features, is yet to be attained. Perhaps the most promising method for achieving this aim is through the use of multiple cameras yielding multiple views of the face [27]. For example, the system could be trained using triplets of images per facial expression to be recognized shown at three orientations that differ by a rotation of 90° (portrait, left and right profile). Novel rotations at 30° and 45° from the nearest trained orientation can be interpolated between the trained orientations. Test images of facial displays shown at any orientation between the left and the right profile view of the face could be finally classified by generalizing from independent facial expression representations at each training/interpolated facial view. A basic understanding of how to achieve automatic AU detection from the profile view of the face is necessary if such a technological framework for automatic AU detection from multiple views of the face is to be established.
3) There is now a growing body of psychological research that argues that temporal dynamics of facial behavior (i.e., the timing and the duration of facial activity) is a critical factor for the interpretation of the observed behavior [1]. For example, Schmidt and Cohn [28] have shown that spontaneous smiles, in contrast to posed smiles, are fast in onset, can have multiple AU12 apexes (i.e., multiple rises of the mouth corners), and are accompanied by other AUs that appear either simultaneously with AU12 or follow AU12 within 1 s. Since it takes more than one hour to manually score 100 still images or a minute of videotape in terms of AUs and their temporal segments [14], it is obvious that automated tools for the detection of AUs and their temporal dynamics would be highly beneficial. Nevertheless no effort in automating the detection of the temporal segments of AUs in face image sequences has been reported so far.
4) Areas where machine tools for the analysis of human facial expressions from face profile could expand and enhance research include numerous specialized areas in scientific and professional sectors. Automatic analysis of expressions from face-profile view would facilitate research on human emotion, which is in turn important for areas such as behavioral science, psychology, neurology

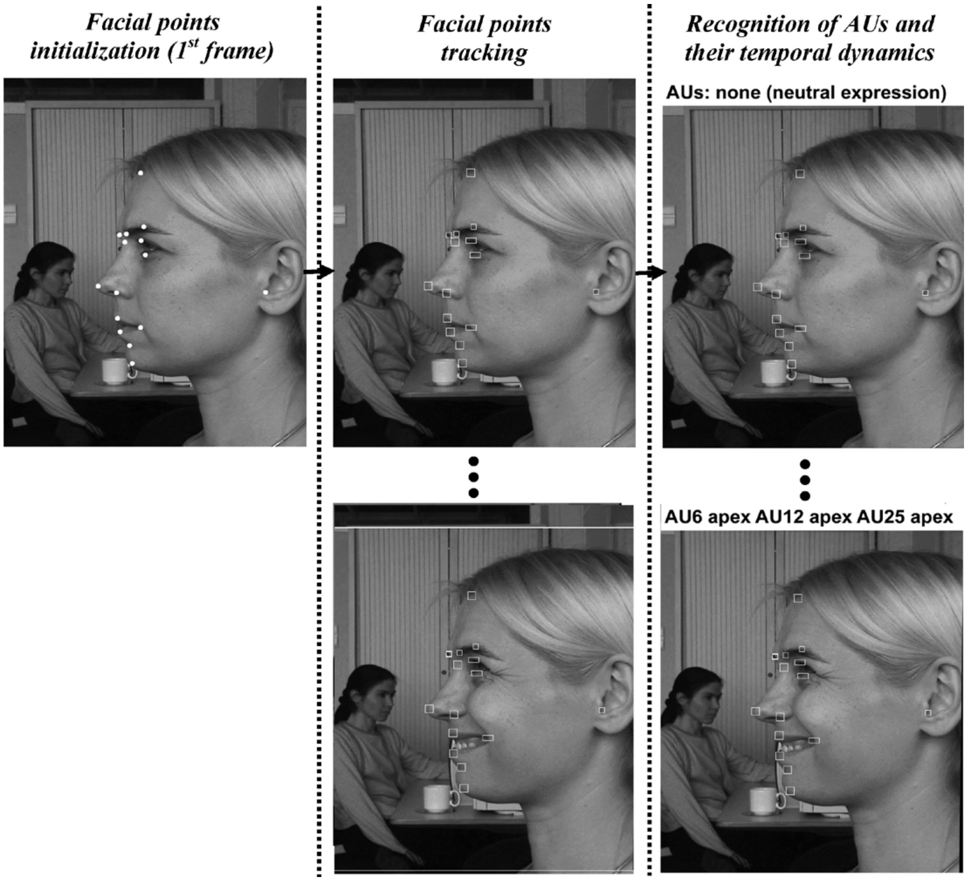
Facial points initialization (1st frame)
Facial points tracking
Recognition of AUs and their temporal dynamics
AUs: none (neutral expression)
AU6 apex AU12 apex AU25 apex

Fig. 1. Outline of the profile-face-based method for detection of AUs and their temporal dynamics.

(in studies on dependence between emotional abilities impairments and brain lesions), and psychiatry (in studies on autism and schizophrenia) [29]. It seems that negative emotions (where facial displays of AU2, AU4, AU9, etc., are often involved) are more easily perceivable from the left hemiface and the full face than from the right hemiface and that, in general, the left hemiface is perceived to display more emotion than the right hemiface [30]. Also, it seems that facial actions involved in spontaneous emotional expressions are more symmetrical, involving both the left and the right side of the face, than deliberate actions displayed on request [31]. Based upon these observations, Mitra and Liu [32] have shown that facial asymmetry has sufficient discriminating power to improve the performance of an automated emotion classifier significantly. Martinez [33] has shown that, by taking into account facial asymmetry caused by certain emotion, expression-invariant face recognition can be achieved. Finally, machine analysis of facial behavior from profile expressions could be of considerable value in any situation where issues concerning emotion, attention, deception, and attitude are of importance and frontal-face observations are not always feasible. Such situations occur often in security sectors, where the observed persons should not be aware of the video surveillance.

The authors have already built a first prototype of an automated profile-face-based AU detector [34], the novel version of which is presented in this paper. This prototype system was aimed at automatic recognition of 20 AUs from subtle changes in the contour of the face profile tracked in an input face-profile image sequence. This previous version of the profile-face-based AU detector had several limitations: 1) it was applicable only to images depicting the left profile of the face; 2) it did not apply temporal reasoning; 3) it could not recognize temporal dynamics of AUs; and 4) AU coding was based only upon changes in the contour of the face profile region (i.e., changes within the face profile region were disregarded).

The current version of the method, proposed in this paper, addresses these limitations. Fig. 1 outlines this novel method, the prelim of which was reported in [35]. It operates under two assumptions: 1) the input video sequence is a nonoccluded (left or right) near-profile view of the face with possible in-image-plane head rotations and 2) the first frame of it shows a neutral expression. After the facial points are initialized in the first frame of the input image sequence, we exploit particle filtering to track the 15 points automatically in the rest of the sequence. Based on the changes in the position of the points, we measure changes in facial expression. Changes in the position of the facial points are first transformed into a set of mid-level parameters for AU recognition. Based upon the temporal consistency of these parameters, a rule-based method encodes temporal segments (onset, apex, offset) of 27 AUs occurring alone or in combination. The usage of temporal information allows us not only to code a video segment to the corresponding AUs, but also to automatically segment an arbitrarily long video sequence into the segments that correspond to different expressions. Facial point tracking, parametric representation of the extracted information, recognition of AUs and their dynamics, and automatic

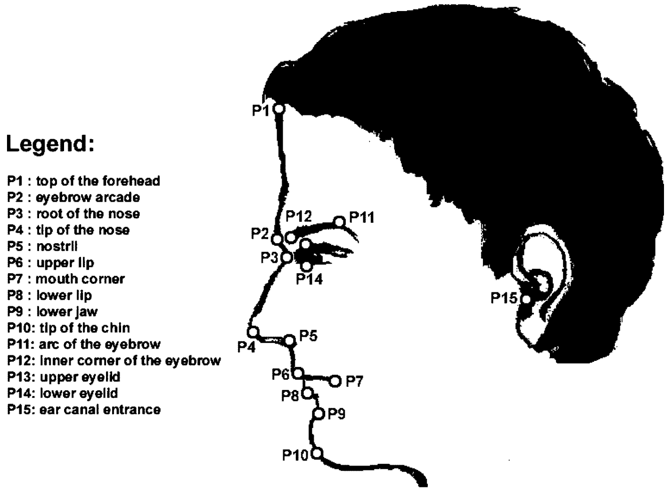

Fig. 2. Facial points (fiducial points of the face components).

segmentation of the video sequence are explained in Sections II, III, IV, and V. Evaluation studies and experimental results are discussed in Section VI.

II. Facial Point Tracking

Contractions of facial muscles induce movements of the facial skin and changes in the appearance of facial features (facial components) such as the eyebrows, nose, and mouth. Their shape and location, as visible in a face profile, can alter immensely with facial expressions (e.g., pursed lips versus jaw dropped). To be able to reason about the shown expression and the facial muscle actions that produced it, one must first detect the current appearance of the facial features. To do so, we track a set of facial points illustrated in Fig. 2, the locations of which alter as the current appearance of the facial features changes with the facial expression. In this paper, we do not address the problem of initially locating the facial points. We assume that they are initialized either manually or automatically in the first frame of the input face image sequence (e.g., using the method proposed in [25] and/or in [36]) and they are automatically tracked for the rest of the sequence by applying a particle filtering method.

In recent years, particle filtering has been the dominant paradigm for tracking the state α of a temporal event given a set of noisy observations $Y = \{y^1, \ldots, y^-, y\}$ up to the current time instant [37]–[43]. In our case, the state α is the location of a facial fiducial point while set $Y = \{y^1, \ldots, y^-, y\}$ is the set of image frames up to the current time instant. The main idea behind particle filtering is to maintain a set of solutions that are an efficient representation of the conditional probability $p(\alpha|Y)$. By maintaining a set of solutions instead of a single estimate (as is done by Kalman filtering, for example), particle filtering is able to track multimodal conditional probabilities $p(\alpha|Y)$, and it is therefore robust to missing and inaccurate data and particularly attractive for estimation and prediction in nonlinear, non-Gaussian systems. In this paper, we adapt the auxiliary particle filtering method that was introduced by Pitt and Shepard [39] to independently track the location of the 15 facial features depicted in Fig. 2. In order to make the tracking robust to in-plane head rotations and translations as well as to small translations along the z-axis, we estimate a global affine transformation ϑ for each frame and based on it we register the current frame to the first frame of the sequence. In order to estimate the global affine transformation, we track three referential points. These are: the top of the forehead (P1), the tip of the nose (P4), and the ear canal entrance (P15). We use these points as the referential points because of their stability with respect to nonrigid facial movements: contractions of facial muscles do not cause physical displacement of these points [44]. We estimate the global affine transformation ϑ as the one that minimizes the distance (in the least-squares sense) between the ϑ-based projection of the tracked locations of the referential points and these locations in the first frame of the sequence. The rest of the facial features are tracked in image frames that have been compensated for the transformation ϑ. In what follows, without loss of generality, we will describe the proposed color-based tracking scheme for tracking a single facial feature.

A. Auxiliary Particle Filtering

The tracking is initialized in the first frame of the input image sequence when a window is centered around the facial feature to be tracked. Let c denote the template that contains the color information in such a window. We will use α to denote the unknown location of the facial feature at the current time instant and $Y = \{y^1, \ldots, y^-, y\}$ will denote the observations (i.e., the images) up to the current time instant. In order to fully specify a particle filter, we need to model two probability densities. One is the observation likelihood $p(y|\alpha; c)$, which expresses in our case how similar the color information in a window in image y around the position α is to the color template c. The second density is the transition density $p(\alpha|\alpha^-)$ which, in our case, models the temporal dynamics of the facial feature. That is, $p(\alpha|\alpha^-)$ models the probability that the facial feature is at position α in the current frame, given that it was at position α^- in the previous frame.

The main idea of particle filtering is to maintain a particle-based representation of the *a posteriori* probability $p(\alpha|Y)$ of the state α given all the observations Y up to the current time instance. This means that the distribution $p(\alpha|Y)$ is represented by a set of pairs $\{(s_k, \pi_k)\}$ such that if s_k is chosen with probability equal to π_k, then it is as if s_k is drawn from $p(\alpha|Y)$. That is [40], the probability $p(\alpha|Y)$ is approximated by the discrete distribution $\sum_k \pi_k \delta(\alpha - s_k)$, where $\delta(.)$ is the Dirac function and $\sum_k \pi_k = 1$. Let the particles s_k be sampled from a sampling distribution $G(\alpha)$ which has a positive probability density function that (up to a normalization constant) is equal to a function $g(\alpha)$. Then calculate the weights π_k as $\pi_k = w_k / \sum_j w_j$, where $w_k \propto (p(\alpha|Y)/g(\alpha))$. It can be shown that, if the pairs $\{(s_k, \pi_k)\}$ are chosen in this way then, as the number of particles approaches infinity, an estimation $\sum_k \pi_k f(s_k)$ converges to the expected (under the distribution $p(\alpha|Y)$) value of the function $f(\alpha)$. Therefore, once a particle-based representation of the *a posteriori* probability $p(\alpha|Y)$ is available, we can estimate statistics such as the mean $(f(\alpha) = \alpha)$ and the variance $(f(\alpha) = \alpha^2)$ of the state. In our case, an estimation $\hat{\alpha}$ of the the position of the facial feature is obtained as the mean of the state α, that is

$$\hat{\alpha} = E_{p(a|Y)}\{a\} = \int_a ap(a|Y) \cong \sum_k s_k \pi_k. \qquad (1)$$

In the particle filtering framework, our representation of the *a posteriori* probability $p(\alpha|Y)$ of the state α is updated in a recursive way. More specifically, let us assume that at the current time instant we have a particle-based representation of the *a posteriori* probability $p(\alpha^{-}|Y^{-})$ of the state α^{-} at the previous time instant. That is, let us assume that we have a collection of K particles and their corresponding weights (i.e., $\{(s_k^{-}, \pi_k^{-})\}$) that represent the *a posteriori* probability at the previous time instant. Then, we can summarize a step of the Auxiliary Particle Filtering that will result in a collection of K particles and their corresponding weights (i.e., $\{(s_{k'}, \pi_{k'})\}$) that represent the *a posteriori* probability at the current time instant as follows.

1) Propagate all particles s_k^{-} via the transition probability $p(\alpha|\alpha^{-})$ in order to arrive at a collection of K particles μ_k.
2) Evaluate the likelihood associated with each particle μ_k, that is, let $\lambda_k = p(y|\mu_k; c)$.
3) Draw K particles from the probability density that is represented by the collection $\{(s_k^{-}, \lambda_k \pi_k^{-})\}$(see Fig. 4, lower left plot). Let k be the index of the particle that was drawn at the k' draw ($1 \leq k' \leq K$), that is, let the particle s_k be selected at the k' draw (in general $k \neq k'$). This is the essence of the auxiliary particle filtering; it favors particles with high λ_k (i.e., particles that end up in areas with high likelihood when propagated with the transition density).
4) Propagate each of the particles that were draw at step 3 with the transition probability $p(\alpha|\alpha^{-})$ in order to arrive at a collection of K particles $s_{k'}$.
5) Assign a weight $\pi_{k'}$ to each particle according to (2)

$$w_{k'} = \frac{p(y|s_{k'}; c)}{\lambda_k}, \quad \pi_{k'} = \frac{w_{k'}}{\sum_j w_j}. \tag{2}$$

This results in a collection of K particles and their corresponding weights (i.e., $\{(s_{k'}, \pi_{k'})\}$). This representation is an approximation of the density $p(\alpha|Y)$.[1]

An outline of the auxiliary particle filtering algorithm, in which the steps of the algorithm are visually depicted, is given in Fig. 3. At each subfigure, a set of circles depicts a set of particles, where larger circles depict particles with higher weights. At each subfigure, the continuous line depicts the probability density function that is represented by the corresponding set of particles. In addition, in the top-right subfigure the observation likelihood $p(y|\alpha; c)$ is depicted with a dashed line. Note that the horizontal axes of the two plots at the left represent the state α^{-} (i.e. the state at the previous time instant) while the horizontal axes of the two plots at the right depict the state α at the current time instant.

We proceed by modeling the observation likelihood $p(y|\alpha; c)$ and the transition density $p(\alpha|\alpha^{-})$. The observation likelihood is used at steps 2 and 5 and its role is to assign higher weights π_k to particles s_k according to how similar the color information around the position s_k is to the color template c. Note that we need an observation model that given an image y, a position α and a color template c can evaluate the scalar value $p(y|\alpha; c)$. The transition density $p(\alpha|\alpha^{-})$ is used at steps 1 and 4. Its role is to propagate a particle s_k^{-} from the previous frame to a position in which s_k is likely to be in the current frame. Note that we need a transition model from which we can sample, that is, a model that, given a particle s_k^{-}, can produce a particle s_k with a probability equal to $p(s_k|s_k^{-})$. In what follows we will formally define the two density models that we use.

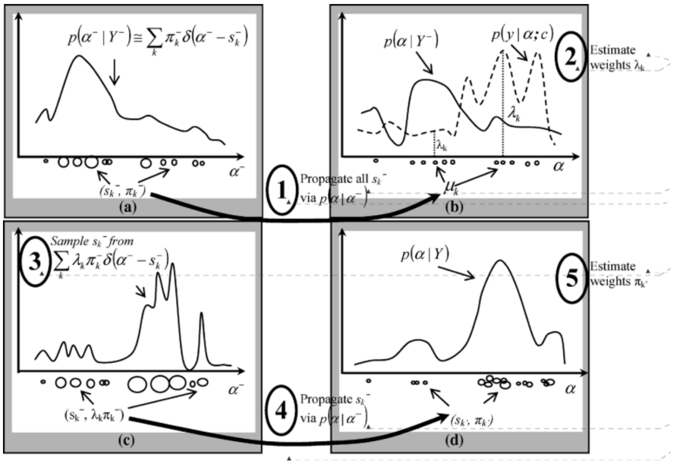
$p(\alpha^{-}|Y^{-}) \cong \sum_k \pi_k^{-} \delta(\alpha^{-} - s_k^{-})$
(s_k^{-}, π_k^{-})
α^{-}
(a)
1
Propagate all s_k^{-} via $p(\alpha|\alpha^{-})$
$p(\alpha|Y^{-})$
$p(y|\alpha; c)$
2
Estimate weights λ_k
λ_k
λ_k
μ_k
α
(b)
3
Sample s_k^{-} from $\sum_k \lambda_k \pi_k^{-} \delta(\alpha^{-} - s_k^{-})$
$(s_k^{-}, \lambda_k \pi_k^{-})$
α^{-}
(c)
4
Propagate s_k^{-} via $p(\alpha|\alpha^{-})$
$p(\alpha|Y)$
5
Estimate weights $\pi_{k'}$
$(s_{k'}, \pi_{k'})$
α
(d)

Fig. 3. Outline of the auxiliary particle filtering method [39].

B. Robust Color-Based Observation Model

Various observation models have been proposed for template-based tracking, where special attention is given to both the robustness in the presence of clutter and occlusions and the adaptation of the observation model [45], [46]. Recently, attention has been drawn to color-based tracking [42], [47]. In what follows, we propose a color-based observation model that is invariant to global illumination changes.

Our observation model is initialized in the first frame of the input image sequence when the user centers a window around the facial point to be tracked. Let c denote the template feature vector that contains the RGB color information in such a window in the first frame and let $c(i)$ denote the color at a pixel i. Clearly, c has a dimensionality equal to three times the number of pixels in the window. We need to define the probability density $p(y|\alpha; c)$. Let $y(\alpha)$ denote the data vector that contains the RGB color information at the image window around position α and let $y(\alpha, i)$ denote the color at a pixel i. We propose a color-based distance between the vectors c and $y(\alpha)$ that is invariant to global changes in the intensity. For each pixel i, the color distance is defined as

$$d_c\left(c, y(a), i\right) = \frac{c(i)}{E\{c\}} - \frac{y(a,i)}{E\left\{y(a)\right\}} \tag{3}$$

where $E\{c\}$ is the mean of vector c (i.e., the average intensity of the color template c) and $E\{y(a)\}$ is the mean of vector $y(a)$ (i.e., the average intensity of the color template $y(a)$). By dividing the color at each pixel with the average intensity of the color template to which it belongs, the color difference vector $d_c(c, y(a), i)$ becomes invariant to changes in the illumination

[1] To be more specific, with the above scheme, at step 4 we arrive at a collection of pairs $(s_{k'}, k)$ (i.e. a particle $s_{k'}$ and the index k of the particle that was drawn at the k' draw at step 3), each one of which is sampled from a probability $G(\alpha, k)$ with density proportional to $g(\alpha, k) = p(y|\mu_k)p(\alpha|s_k^{-})\pi_k^{-}$. With the weighting of step 5, the set $\{((s_{k'}, k), w_{k'})\}$ is a particle-based representation of $p(\alpha, k|Y)$ which up to proportionality is $p(y|\alpha)p(\alpha|s_k^{-})\pi_k^{-}$. By dropping the index k from the above set $\{((s_{k'}, k), w_{k'})\}$, we arrive at a particle-based representation of $p(\alpha|Y)$. See [39] for a complete proof.

intensity. Finally, we define the scalar color distance using a robust function ρ [48]. More specifically

$$d_c\left(c, y(a)\right) = E_i\left\{\rho\left(\left\|\frac{c(i)}{E\{c\}} - \frac{y(a,i)}{E\left\{y(a)\right\}}\right\|_1\right)\right\} \quad (4)$$

where $\|.\|_1$ is the L_1 norm and ρ is the absolute value in our experiments. Then, the observation likelihood is

$$p(y|\alpha;c) = \frac{1}{Z}e^{-\frac{d_c(c,y(a))}{\sigma_c}} \quad (5)$$

where Z is a normalization term that can be ignored since in the context of particle filtering the weights of the particles are renormalized at each iteration [see (2)]. The term σ_c is a scaling parameter which was set to 0.01 in all of our experiments.

C. Transition Model

Once the observation model has been defined, we need to model the transition probability that is used to generate a new set of particles given the current one. The transition probability models our knowledge of the dynamics of the feature, that is, it models our knowledge of the feature's position α in the current frame given its position α^- in the previous frame. We model the transition probability of each feature as a mixture of L Gaussians. The first few components model the feature's dynamics as a mixture of Gaussians around the previous position α^-. The last few components of the Gaussian mixture ignore the information about the position in the previous frame and model the static prior $p(\alpha)$. These last components are essentially used to recover the tracking by creating particles at positions with high priors, such as the position of the facial points in the expressionless face. More specifically

$$p(\alpha|\alpha^-) = w_1\sum_{l=1}^{L_1} m_l N(\alpha^- + \mu_l, \Sigma_l) + (1-w_1)\sum_{l=L_1+1}^{L} m_l N(\mu_l, \Sigma_l) \quad (6)$$

where the coefficient w_1 is set to 95%. This means that 95% of the samples are generated based on the feature's dynamics. The number of the Gaussians to be used for each feature is a design choice, which depends on the degree of freedom of each facial feature. In our implementation, we used a very simple model with 1 or 2 Gaussians for the dynamic components (i.e., $L_1 = 1-2$) and 1 to 3 Gaussians (i.e., $L = 2-5$) for the static components. To obtain valid static components, we first need to compensate for head motion using the global transformation ϑ. Then, we need to compensate for physiognomic variability. Namely, different people have different faces, and the facial features are not located at exactly the same position in each face. We handle this by translating the mean of each Gaussian component of the second term of (6) by a vector estimated based on the location of the facial feature in the first frame (neutral expression frame) of the input image sequence. The means and the variances of the components of the Gaussians are estimated using the EM algorithm on a semi-automatically annotated training dataset containing the coordinates of the facial features under consideration. This dataset contains images of two persons (other than the 19 persons whose images are used to test the performance of the system as a whole) showing various facial expressions. The parameters of the transition model of each facial feature are estimated independently of the transition models of the other facial features.

D. Tracking Multiple Facial Points

The application of auxiliary particle filtering for tracking the position α of each facial point results in a set of particles and their corresponding weights $\{(s_k, \pi_k)\}$ in each frame of the sequence. This set is a representation of the posterior $p(\alpha|Y)$. An estimate of the position of the facial point is then given by (1). Typical results of this algorithm are illustrated in Figs. 4 and 5. Finally, let us note that the computational complexity of the above algorithm is linear with respect to the number of particles and to the number of facial points. The main computational burdens of the algorithm are the evaluation of the likelihood $p(y|s_k;c)$ of the particles and the calculation of the color distance in (4). In our experiments, we used 100 particles for each of the 15 points which, for our Matlab code, resulted in the processing of 1 frame per 15 s on a 2.5-GHz Pentium. We expect that a careful C/C++ implementation can achieve a near-real time performance.

III. Mid-Level Parametric Representation

Contractions of facial muscles alter the shape and location of the facial components (eyebrows, eyes, mouth, chin). Some of these changes in facial expression are observable from the changes in the position of the tracked points. To classify the tracked changes in terms of AUs, these changes are transformed first into a set of mid-level parameters. We have defined two mid-level parameters in total: $\mathbf{up/down}(\boldsymbol{P})$ and $\mathbf{inc/dec}(\boldsymbol{PP'})$.

1) Parameter $\mathbf{up/down}(\boldsymbol{P}) = y(P_{t1}) - y(P_t)$, where $t1$ stands for the 1st frame and t for the current frame, describes upward and downward movements of point P. If $y(P_{t1}) - y(P_t) > \varepsilon$, point P moves upwards. If $y(P_{t1}) - y(P_t) < \varepsilon$, point P moves downwards. The value of $y(P)$ is the y-coordinate of point P and the value of ε is 1 pixel.
2) Parameter $\boldsymbol{inc/dec}(\boldsymbol{PP'}) = PP'_{t1} - PP'_t$, where $t1$ stands for the 1st frame and t for the current frame, describes the increase or decrease of the distance between points P and P'. If $PP'_{t1} - PP'_t < \varepsilon$, distance PP' increases. If $PP'_{t1} - PP'_t > \varepsilon$, distance PP' decreases. Distance PP' is calculated as the Euclidian distance between P and P'.

Originally, in the first prototype of our profile-face-based AU detector [34], we used another parameter as well. The parameter in question, $\mathbf{in/out}(\boldsymbol{P'}) = x(P'_{t1}) - x(P'_t)$, describes inward and outward movements of point P'. For a left-profile view of the face, $in/out(P') < \varepsilon$ describes an inward movement of point P'. On the other hand, for a right-profile view of the face, $in/out(P') < \varepsilon$ describes an outward movement of point P'. Thus, this parameter depends upon the facial view depicted in the input image, which is the main reason why we chose not to use this parameter in the current version of our profile-face-based AU detector. In the current version of the system we use $\boldsymbol{inc/dec}(\boldsymbol{P'P}\mathbf{15})$ instead. We represent an outward movement of point P' as $inc/dec(P'P15) < \varepsilon$ (i.e., as an increase of the distance between points P15 and P'). Similarly, an inward movement of point P' is represented as $inc/dec(P'P15) > \varepsilon$

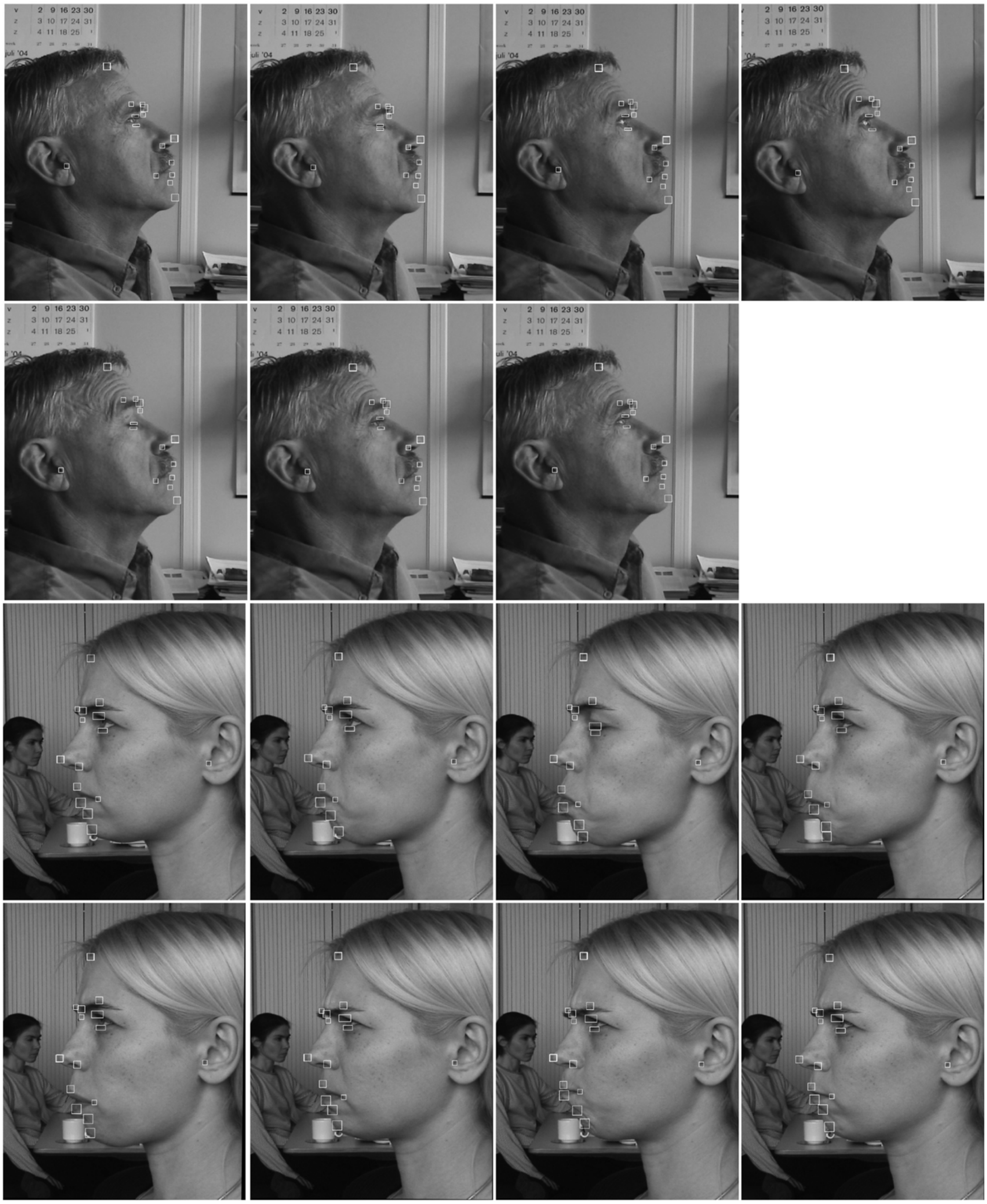

Fig. 4. Results of the facial point tracking. First and second rows: frames 1 (neutral), 14 (blink, i.e., apex AU45), 75 (onset AU1 + 2 + 5), 89 (apex AU1 + 2 + 5), 131 (apex AU1 + 2 + 45), 137 (offset AU1 + 2 + 45), 148 (neutral). 3rd and 4th rows: frames 1 (neutral), 19 (onset AU36T + 26), 23 (apex AU45, onset AU36T + 26), 38 (apex AU36T + 26), 76 (offset AU36T + 26), 159 (apex AU4, onset AU17 + 24), 194 (apex AU4 + 17 + 24), 237 (offset AU4 + 17 + 24).

(i.e., as an decrease of the distance between points P15 and P'). As explained in Section II, we use point P15 (Fig. 2) as the referential point because it is a stable facial point.

As can be seen from these definitions, the mid-level parameters are calculated for various points for each input frame by comparing the position of the points in the current frame with that of the relevant points in the first (neutral expression) frame. Before these calculations can be carried out, all rigid head motions in the input image sequence must be eliminated. Otherwise we would not be certain whether the value of a given parameter had changed due to the movement of the relevant points or due to a rigid head movement. As explained in Section II, to handle in-image-plane rotations and variations in scale of the observed face profile, we register each frame of the input image sequence using a global affine transformation ϑ that we estimate for each frame based on the tracked location of P1, P4, and P15. The feature parameters are then calculated for various points tracked in each frame of the registered sequence. As can be seen in Fig. 6, the values of the mid-level parameters change as a function of time and, thus, can be used to measure temporal dynamics of AUs.

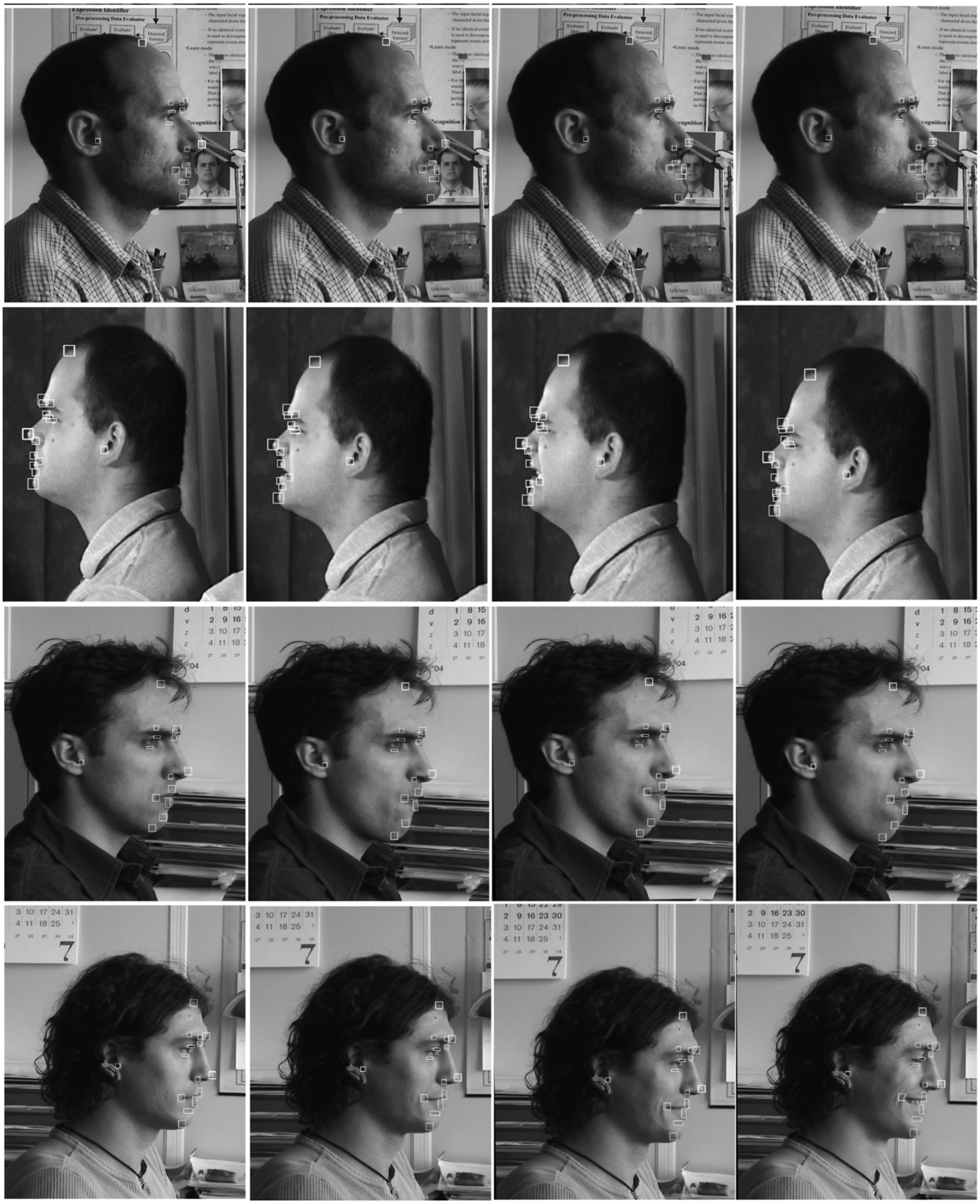

Fig. 5. Results of the facial point tracking. First row: frames 1 (neutral), 48 (onset AU29), 59 (apex AU29), 72 (offset AU29). Second row: frames 1 (neutral), 17 (onset AU44 + 9 + 10 + 20 + 25), 25 (apex AU44 + 9 + 10 + 20 + 25), 66 (apex AU45, offset AU44 + 9 + 10 + 20 + 25). Third row: frames 1 (neutral), 12 (onset AU36B + 26), 43 (apex AU36B + 26), 62 (offset AU36B + 26). Fourth row: frames 1 (neutral), 25 (onset AU12), 30 (onset AU6 + 12), 55 (apex AU6 + 12 + 25 + 45).

IV. Recognition of AUs and Their Temporal Dynamics

We transform the calculated mid-level feature parameters into a set of AUs describing the facial expression(s) captured in the input video. We use a set of temporal rules and a fast direct chaining inference procedure to encode 27 AUs occurring alone or in combination in an input face-profile image sequence. To minimize the effects of noise and inaccuracies in facial point tracking and to enable the recognition of the temporal dynamics of shown AUs, we consider the temporal consistency of the mid-level parameters.

We divide each facial action into three temporal segments: the onset (beginning), apex (peak), and offset (ending). We define each temporal rule for AU recognition in a unique way according to the relevant FACS rule and using the mid-level parameters explained in Section III. Tables I–III provide the list of the utilized rules. Fig. 6 illustrates the meaning of these rules for the case of AU1, AU2, and AU12. In Fig. 6, the horizontal

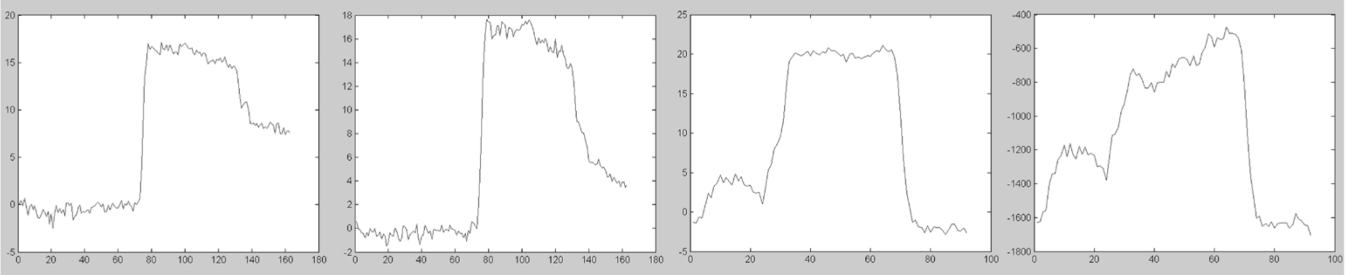

Fig. 6. Values of four mid-level feature parameters (in left-to-right order): $up/down(\mathrm{P12})$ and $up/down(\mathrm{P11})$ computed for 163 frames of AU1 + 2 + 5 face-profile video depicted in the first two rows of Fig. 4, and $up/down(\mathrm{P7})$ and $inc/dec(\mathrm{P5P7})$ computed for 92 frames of AU6 + 12 + 25 face-profile video depicted in the fourth row of Fig. 5.

axis represents the time dimension (i.e., the frame number) and the vertical axis represents values that the mid-level feature parameters take. As implicitly suggested by the two left-hand-side graphs of Fig. 6, P12 (respectively P11) should move upward and it should be above its neutral-expression location to label a frame as an "AU1 (respectively AU2)[2] onset". The upward motion should terminate, resulting in a (relatively) stable temporal location of P12 (P11), for a frame to be labeled as "AU1 (AU2) apex". Eventually, P12 (P11) should move downward toward its neutral-expression location to label a frame as an "AU1 (AU2) offset". Similarly, as implicitly suggested by the two right-hand-side graphs of Fig. 6, P7 should move upward, above its neutral-expression location, and the distance between points P5 and P7 should increase, exceeding its neutral-expression length, in order to label a frame as an "AU12[3] onset". In order to label a frame as "AU12 apex", the increase of the values of the relevant mid-level parameters should terminate. Once the values of these mid-level parameters begin to decrease, a frame can be labeled as "AU12 offset". Note that the two right-hand-side graphs of Fig. 6 show two distinct peaks in the increase of the pertinent mid-level parameters. As shown by Schmidt and Cohn [28], this is typical for spontaneous smiles and in contrast to posed smiles.

Generally, for each and every AU, it must be possible to detect a temporal segment (an onset, apex, or offset) continuously over at least five consecutive frames for the facial action in question to be scored. We determined this temporal duration empirically based on a video frame rate of 24 frames/s (i.e., five frames have a duration of less than 1/4 of a second) and based on research findings that suggest that temporal changes in neuromuscular facial activity last from 1/4 of a second (e.g., a blink) to several minutes (e.g., a jaw clench) [15].

The employed fast direct chaining inference procedure takes advantage of both a relational representation of the knowledge and a depth-first search to find as many conclusions as possible within a single "pass" through the knowledge base [49]. The use of R-list achieves a relational representation of the knowledge. The R-list is a 4-tuple list, where the first two columns identify the conclusion of a certain rule that forms the premise of another rule identified in the next two columns of the R-list. For example, the relation between rules 13 and 21 (Table II) is represented as (21, 1, 13, 2), which means that the 1st conclusion of rule 21 forms the 2nd premise of rule 13. The term *direct* indicates that as the inference process is executing, it creates the proper chain of reasoning.

A recursive process starts with the first rule of the knowledge base. Then, it searches the R-list for a link between the fired rule and the rule that the process will try to fire in the next loop. If such a relation does not exist, the procedure tries to fire the rule that in the knowledge base comes after the rule last fired.

Inaccuracies in facial point tracking and occurrences of non-prototypic facial activity may result in frames and temporal segments that are unlabeled (i.e., neither the onset, nor the apex, nor the offset) and in frames and temporal segments that are labeled incorrectly. The latter may arise, for example, when an apex frame or an apex temporal segment of an AU is detected either between two onset segments or between two offset segments. To handle such situations, we employ a memory-based process that takes into account the dynamics of facial expressions. More specifically, we examine the labels of both the previous and the next frame/temporal segment and re-label the current frame / temporal segment according to the ruled-based system summarized in Table IV.

For instance, any unlabeled temporal segment and/or any apex segment that have been detected between two onset segments are re-labeled as "onset". Finally, an AU should be recognized, in general, only when the full temporal pattern of that AU is observed (e.g., see Fig. 6 for cases of AU1, AU2, and AU12). However, in order to deal with fast transitions between onsets and offsets, we score AUs even if the relevant apexes are missing.

V. Automatic Segmentation of an Input Video Sequence

Virtually all the existing AU detectors only perform well on isolated or pre-segmented facial expression image sequences showing a single temporal activation pattern of one or more AUs. In reality, such segmentation is not available and, hence, there is a need to find an automatic way of segmenting face image sequences into the different facial expressions pictured. A way to achieve this has been proposed by Otsuka and Ohya [50] and Cohen *et al.* [51]. To cope with cases where two facial expressions of emotion are displayed contiguously, Otsuka and

[2] Since the upward motion of the inner corner of the eyebrow is the principle cue for the activation of AU1, the upward movement of the fiducial point P12 is used as the criterion for detecting the onset of the AU1 activation. Reversal of this motion is used to detect the offset of this facial expression. Similarly, the upward movement of the outer corner of the eyebrow (i.e., point P11) is used as the criterion for detecting the onset of the AU2 activation.

[3] The upward, oblique motion of the mouth corner is the principle cue for the activation of AU12. Hence, the upward movement of the fiducial point P7 and the increase of the distance between points P5 and P7, typical for oblique (AU12) rather than sharp (AU13) upward movement of the mouth corner, are used as the criteria for detecting the onset of the AU12 activation. Reversal of these motions is used to detect the offset of this facial expression.

TABLE I

Rules for Recognizing AU1, AU2, AU4–AU7, AU9, AU10, AU12, AU13, AU15, and AU16 From a Face-Profile Image Sequence. **Legend**: for Notational Simplicity, $(x_t < x_{t-1})$ Stands for $(x_t < x_{t-1} - \varepsilon)$, $(x_t = x_{t-1})$ for $(|x_t - x_{t-1}| \le \varepsilon)$, $(x_t > x_{t-1})$ for $(x_t > x_{t-1} + \varepsilon)$, $t1$ for the First Frame, t for the Current Frame, $t - 1$ for the Previous Frame. The Value Assigned to ε is 1 pixel. Threshold $T1 = 1/2\ P13_{t1}P14_{t1}$ Distinguishes Between the Activation of AU6, AU7, AU41 and That of AU44, AU43 and AU45. The Value of $T1$ Has Been Decided Based Upon the Threshold Description Provided by the Relevant FACS Rules

AU1
rule 1
	Pulls the eyebrows' inner corners upward, causes the skin of the center forehead to wrinkle horizontally.
	IF ***up/down***(P12) > ε THEN AU1-p
	IF [***up/down***(P12)]$_t$ > [***up/down***(P12)]$_{t-1}$ AND AU1-p THEN AU1-onset
	IF [***up/down***(P12)]$_t$ = [***up/down***(P12)]$_{t-1}$ AND AU1-p THEN AU1-apex
	IF [***up/down***(P12)]$_t$ < [***up/down***(P12)]$_{t-1}$ AND AU1-p THEN AU1-offset

AU2 rule 2	Pulls the eyebrows' outer corners upward, causes the skin of the lateral forehead to wrinkle horizontally. IF ***up/down***(P11) > ε THEN AU2-p The rules for recognition of temporal segments of AU2 are similar to those defined for AU1 (see rule 1)

AU4
rule 3
	Pulls the eyebrows closer together, produces a bulge between the eyebrows, lowers the eyebrows slightly.
	IF [***inc/dec***(P2P12)]$_t$ > [***inc/dec***(P2P12)]$_{t-1}$ AND ***inc/dec***(P2P12) > ε THEN AU4-onset
	IF [***inc/dec***(P2P12)]$_t$ = [***inc/dec***(P2P12)]$_{t-1}$ AND ***inc/dec***(P2P12) > ε THEN AU4-apex
	IF [***inc/dec***(P2P12)]$_t$ < [***inc/dec***(P2P12)]$_{t-1}$ AND ***inc/dec***(P2P12) > ε THEN AU4-offset

AU5
rule 4
	Raises the upper eyelid, widens the eye opening.
	IF [***inc/dec***(P13P14)]$_t$ < [***inc/dec***(P13P14)]$_{t-1}$ AND ***inc/dec***(P13P14) < ε THEN AU5-onset
	IF [***inc/dec***(P13P14)]$_t$ = [***inc/dec***(P13P14)]$_{t-1}$ AND ***inc/dec***(P13P14) < ε THEN AU5-apex
	IF [***inc/dec***(P13P14)]$_t$ > [***inc/dec***(P13P14)]$_{t-1}$ AND ***inc/dec***(P13P14) < ε THEN AU5-offset

AU6
rule 5
	Raises the cheeks, narrows the eye opening.
	IF ***inc/dec***(P13P14) > ε AND ***inc/dec***(P13P14) ≤ *T1* AND ***up/down***(P7) > ε THEN AU6-p
	IF [***inc/dec***(P13P14)]$_t$ > [***inc/dec***(P13P14)]$_{t-1}$ AND AU6-p AND ***up/down***(P14) > ε THEN AU6-onset
	IF [***inc/dec***(P13P14)]$_t$ = [***inc/dec***(P13P14)]$_{t-1}$ AND AU6-p AND ***up/down***(P14) > ε THEN AU6-apex
	IF [***inc/dec***(P13P14)]$_t$ < [***inc/dec***(P13P14)]$_{t-1}$ AND AU6-p AND ***up/down***(P14) > ε THEN AU6-offset

AU7 rule 6	Raises the lower eyelid, narrows the eye opening. IF ***inc/dec***(P13P14) > ε AND ***inc/dec***(P13P14) ≤ *T1* AND ***up/down***(P7) ≤ ε THEN AU7-p The rules for recognition of temporal segments of AU7 are similar to those defined for AU6 (see rule 5)

AU9
rule 7
	Wrinkles the nose, lowers the brows, produces a bulge between the brows and the root of the nose.
	IF ***inc/dec***(P2P15) < ε AND ***inc/dec***(P2P3) > ε AND ***inc/dec***(P2P12) ≤ ε THEN AU9-p
	IF [***inc/dec***(P2P15)]$_t$ < [***inc/dec***(P2P15)]$_{t-1}$ AND AU9-p THEN AU9-onset
	IF [***inc/dec***(P2P15)]$_t$ = [***inc/dec***(P2P15)]$_{t-1}$ AND AU9-p THEN AU9-apex
	IF [***inc/dec***(P2P15)]$_t$ > [***inc/dec***(P2P15)]$_{t-1}$ AND AU9-p THEN AU9-offset

AU10
rule 8
	Raises the upper lip, deepens the nasolabial furrow, does not wrinkle the nose.
	IF ***inc/dec***(P2P15) ≥ ε AND ***inc/dec***(P2P3) ≤ ε AND ***inc/dec***(P5P6) > ε THEN AU10-p
	IF [***inc/dec***(P5P6)]$_t$ > [***inc/dec***(P5P6)]$_{t-1}$ AND AU10-p THEN AU10-onset
	IF [***inc/dec***(P5P6)]$_t$ = [***inc/dec***(P5P6)]$_{t-1}$ AND AU10-p THEN AU10-apex
	IF [***inc/dec***(P5P6)]$_t$ < [***inc/dec***(P5P6)]$_{t-1}$ AND AU10-p THEN AU10-offset

AU12 rule 9	Pulls the lip corners upward obliquely. IF ***up/down***(P7) > ε AND ***inc/dec***(P5P7) < ε THEN AU12-p The rules for recognition of temporal segments of AU12 are similar to those defined for AU1 (see rule 1)
AU13 Rule 10	Pulls the lip corners sharply upward. IF ***up/down***(P7) > ε AND ***inc/dec***(P5P7) ≥ ε THEN AU13-p The rules for recognition of temporal segments of AU13 are similar to those defined for AU1 (see rule 1)

AU15
rule 11
	Pulls the corners of the lips downward, stretches the lips slightly.
	IF ***up/down***(P7) < ε THEN AU15-p
	IF [***up/down***(P7)]$_t$ < [***up/down***(P7)]$_{t-1}$ AND AU15-p THEN AU15-onset
	IF [***up/down***(P7)]$_t$ = [***up/down***(P7)]$_{t-1}$ AND AU15-p THEN AU15-apex
	IF [***up/down***(P7)]$_t$ > [***up/down***(P7)]$_{t-1}$ AND AU15-p THEN AU15-offset

AU16 rule 12	Pulls the lower lip downward laterally, causes the lower lip to protrude. IF ***inc/dec***(P8P10) > ε AND ***up/down***(P8) < ε THEN AU16-p The rules for recognition of temporal segments of AU16 are similar to those defined for AU15 (see rule 11)

Ohya applied a heuristic approach and modified the employed Hidden Markov model (HMM) computation such that when the peak of a facial motion is detected, the current emotional expression is assumed to start from the previous frame with minimal facial motion. Similarly, Cohen *et al.* proposed a HMM-based method for recognition of six basic emotions. This assumes that the transitions between emotions pass through the neutral facial expression. Loosely speaking, we adopted a similar approach.

To automatically segment an arbitrarily long video sequence into the segments that correspond to expressive and expressionless facial displays, we use a sequential facial expression model. A display of expressive facial behavior in video corresponds to a temporal sequence of facial movement that we represent as a sequence of temporal patterns (onset-apex-offset) of one or more AUs. Since the presence of facial activity determines the shown facial expression, its absence can be used to delimit the transition between different expressions. The term "neutral facial expression" ("expressionless face") is usually used to designate the absence of facial activity. Thus, to solve the segmentation problem, we use a neutral-expressive-neutral sequential

TABLE II

Rules for Recognizing AU17, AU18, AU20, and AU23–AU29 From a Face-Profile Image Sequence. **Legend**: For Notational Simplicity, $(\mathrm{x}_t < \mathrm{x}_{t-1})$ Stands for $(\mathrm{x}_t < \mathrm{x}_{t-1} - \varepsilon)$, $(\mathrm{x}_t = \mathrm{x}_{t-1})$ for $(|\mathrm{x}_t - \mathrm{x}_{t-1}| \leq \varepsilon)$, $(\mathrm{x}_t > \mathrm{x}_{t-1})$ for $(\mathrm{x}_t > \mathrm{x}_{t-1} + \varepsilon)$, $t1$ for the First Frame, t for the Current Frame, and $t-1$ for the Previous Frame. The Value Assigned to ε is 1 pixel. Threshold $T2 = 1/2\mathrm{P}\ 8_{t1}\mathrm{P}10_{t1}$ Distinguishes Between the Activation of AU26 and That of AU27. The Value of $T2$ Has Been Decided Based Upon Earlier Studies on Automatic Analysis of Facial Expressions From Static-Face Images [25]

AU17
rule 13
	Pushes the chin boss and the lower lip upward and stretches the skin on the chin boss.
	IF NOT (AU28 OR AU28t OR AU28b) AND ***inc/dec***(P10P15) > ε THEN AU17-p
	IF [***inc/dec***(P10P15)]$_t$ > [***inc/dec***(P10P15)]$_{t-1}$ AND AU17-p THEN AU17-onset
	IF [***inc/dec***(P10P15))]$_t$ = [***inc/dec***(P10P15)]$_{t-1}$ AND AU17-p THEN AU17-apex
	IF [***inc/dec***(P10P15)]$_t$ < [***inc/dec***(P10P15)]$_{t-1}$ AND AU17-p THEN AU17-offset

AU18
rule 14
	Pushes the mouth forward medially, causes the lips to protrude forwards (as by saying "fool").
	IF ***inc/dec***(P6P15) < ε AND ***inc/dec***(P8P15) < ε AND ***inc/dec***(P7P15) < ε THEN AU18-p
	IF [***inc/dec***(P7P15)]$_t$ < [***inc/dec***(P7P15)]$_{t-1}$ AND AU18-p THEN AU18-onset
	IF [***inc/dec***(P7P15)]$_t$ = [***inc/dec***(P7P15)]$_{t-1}$ AND AU18-p THEN AU18-apex
	IF [***inc/dec***(P7P15)]$_t$ > [***inc/dec***(P7P15)]$_{t-1}$ AND AU18-p THEN AU18-offset

AU20 rule 15	Pulls the lips backward laterally, flattens the skin of the lips and the chin boss. IF ***inc/dec***(P7P15) > ε THEN AU20-p The rules for recognition of temporal segments of AU20 are similar to those defined for AU17 (see rule 13)

AU23
rule 16
	Tightens the lips slightly making the lips appear more narrow.
	IF NOT (AU28 OR AU28t OR AU28b) AND ***inc/dec***(P6P8) > ε AND ***inc/dec***(P7P15) ≥ ε THEN AU23-p
	IF [***inc/dec***(P6P8)]$_t$ > [***inc/dec***(P6P8)]$_{t-1}$ AND AU23-p THEN AU23-onset
	IF [***inc/dec***(P6P8)]$_t$ = [***inc/dec***(P6P8)]$_{t-1}$ AND AU23-p THEN AU23-apex
	IF [***inc/dec***(P6P8)]$_t$ < [***inc/dec***(P6P8)]$_{t-1}$ AND AU23-p THEN AU23-offset

AU24 rule 17	Presses the lips together, tightens and narrows the lips to a small extent. IF NOT (AU28 OR AU28t OR AU28b) AND ***inc/dec***(P6P8) > ε AND ***inc/dec***(P7P15) < ε THEN AU24-p The rules for recognition of temporal segments of AU24 are similar to those defined for AU23 (see rule 16)
AU25 rule 18	Parts the lips, does not parts the jaws. IF ***inc/dec***(P6P8) < ε AND ***inc/dec***(P4P10) ≥ ε THEN AU25-p The rules for recognition of temporal segments of AU24 are similar to those defined for AU23 (see rule 16)

AU26
rule 19
	Parts the lips, parts the jaws, does not stretches the mouth.
	IF ***inc/dec***(P4P10) < ε AND |***inc/dec***(P4P10)| ≤ *T2* THEN AU26-p
	IF [***inc/dec***(P4P10)]$_t$ < [***inc/dec***(P4P10)]$_{t-1}$ AND AU26-p THEN AU26-onset
	IF [***inc/dec***(P4P10)]$_t$ = [***inc/dec***(P4P10)]$_{t-1}$ AND AU26-p THEN AU26-apex
	IF [***inc/dec***(P4P10)]$_t$ > [***inc/dec***(P4P10)]$_{t-1}$ AND AU26-p THEN AU26-offset

AU27 rule 20	Stretches the mouth as lower jaw is pulled down. IF ***inc/dec***(P4P10) < ε AND |***inc/dec***(P4P10)| > *T2* THEN AU27-p The rules for recognition of temporal segments of AU27 are similar to those defined for AU26 (see rule 19)

AU28t
rule 21
	Upper lip sucked into the mouth.
	IF ***up/down***(P8) > ε AND ***inc/dec***(P6P8) > ε AND P6$_t$P8$_t$ > ε THEN AU28t-p
	IF [***up/down***(P8)]$_t$ > [***up/down***(P8)]$_{t-1}$ AND AU28t-p THEN AU28t-onset
	IF [***up/down***(P8)]$_t$ = [***up/down***(P8)]$_{t-1}$ AND AU28t-p THEN AU28t-apex
	IF [***up/down***(P8)]$_t$ < [***up/down***(P8)]$_{t-1}$ AND AU28t-p THEN AU28t-offset

AU28b
rule 22
	Bottom lip sucked into the mouth.
	IF ***up/down***(P6) < ε AND ***inc/dec***(P6P8) > ε AND P6$_t$P8$_t$ > ε THEN AU28b-p
	IF[***up/down***(P6)]$_t$ < [***up/down***(P6)]$_{t-1}$ AND AU28b-p THEN AU28b-onset
	IF[***up/down***(P6)]$_t$ = [***up/down***(P6)]$_{t-1}$ AND AU28b-p THEN AU28b-apex
	IF[***up/down***(P6)]$_t$ > [***up/down***(P6)]$_{t-1}$ AND AU28b-p THEN AU28b-offset

AU28 rule 23	Lips sucked into the mouth. IF ***inc/dec***(P6P15) > ε AND ***inc/dec***(P8P15) > ε AND ***inc/dec***(P6P8) > ε THEN AU28-p The rules for recognition of temporal segments of AU28 are similar to those defined for AU23 (see rule 16)

AU29
rule 24
	Pushes the jaw forward making the chin to stick out and the lower teeth to extend in front of upper teeth.
	IF [***inc/dec***(P10P15)]$_t$ < [***inc/dec***(P10P15)]$_{t-1}$ AND ***inc/dec***(P10P15) < ε THEN AU29-onset
	IF [***inc/dec***(P10P15)]$_t$ = [***inc/dec***(P10P15)]$_{t-1}$ AND ***inc/dec***(P10P15) < ε THEN AU29-apex
	IF [***inc/dec***(P10P15)]$_t$ > [***inc/dec***(P10P15)]$_{t-1}$ AND ***inc/dec***(P10P15) < ε THEN AU29-offset

model, where an "expressive" segment contains temporal patterns (onset-apex-offset) of one or more AUs encoded by our AU recognizer.

Since we assume that the input to our system consists of facial expression sequences that always start with a neutral facial expression, the neutral-expressive-neutral sequential facial expression model suffices. The model will also be applicable to the data contained in the Cohn–Kanade Face Database [54], which is one of the most commonly used data-sets in the research on automatic facial expression analysis. This is because all facial expression sequences in the Cohn–Kanade Face Database start with a neutral expression. However, in cases where no constraints are posed on input facial expression sequences, the proposed model will not suffice. Also, if one wants to segment input face video in terms of specific facial displays such as the emotional facial expressions, the proposed model will suffice only if each of these specific states begins and ends with a neutral facial expression. This is because the model has been developed to differentiate facial activity (presence of AUs) from inactivity; it delimits the transition between different facial displays

TABLE III

Rules for Recognizing AU36, AU41, AU43, AU44, and AU45 From a Face-Profile Image Sequence. **Legend**: For Notational Simplicity, $(\mathrm{x}_t < \mathrm{x}_{t-1})$ Stands for $(\mathrm{x}_t < \mathrm{x}_{t-1} - \varepsilon)$, $(\mathrm{x}_t = \mathrm{x}_{t-1})$ for $(|\mathrm{x}_t - \mathrm{x}_{t-1}| \leq \varepsilon)$, $(\mathrm{x}_t > \mathrm{x}_{t-1})$ for $(\mathrm{x}_t > \mathrm{x}_{t-1} + \varepsilon)$, $t1$ for the 1st Frame, t for the Current Frame, and $t - 1$ for the Previous Frame. The Value Assigned to ε is 1 pixel. For $T1$ See Table I

AU36t
rule 25
	Pushes the tongue under the upper lip, causes a bulge above the upper lip.
	IF ***up/down***(P6) < ε AND ***up/down***(P8) < ε AND ***inc/dec***(P6P8) ≤ ε THEN AU36t-p
	IF[***up/down***(P6)]$_t$ < [***up/down***(P6)]$_{t-1}$ AND AU36t-p THEN AU36t-onset
	IF[***up/down***(P6)]$_t$ = [***up/down***(P6)]$_{t-1}$ AND AU36t-p THEN AU36t-apex
	IF[***up/down***(P6)]$_t$ > [***up/down***(P6)]$_{t-1}$ AND AU36t-p THEN AU36t-offset

AU36b
rule 26
	Pushes the tongue under the lower lip, causes a bulge below the lower lip.
	IF ***up/down***(P8) > ε AND ***inc/dec***(P9P15) < ε THEN AU36b-p
	IF [***inc/dec***(P9P15)]$_t$ < [***inc/dec***(P9P15)]$_{t-1}$ AND AU36b-p THEN AU36b-onset
	IF [***inc/dec***(P9P15)]$_t$ = [***inc/dec***(P9P15)]$_{t-1}$ AND AU36b-p THEN AU36b-apex
	IF [***inc/dec***(P9P15)]$_t$ > [***inc/dec***(P9P15)]$_{t-1}$ AND AU36b-p THEN AU36b-offset

AU41
rule 27
	Causes the upper eyelid to drop down, narrows the eye opening.
	IF ***inc/dec***(P13P14) > ε AND ***inc/dec***(P13P14) ≤ *T1* AND ***up/down***(P14) ≤ ε THEN AU41-p
	IF [***inc/dec***(P13P14)]$_t$ > [***inc/dec***(P13P14)]$_{t-1}$ AND AU41-p AND ***up/down***(P13) < ε THEN AU41-onset
	IF [***inc/dec***(P13P14)]$_t$ = [***inc/dec***(P13P14)]$_{t-1}$ AND AU41-p AND ***up/down***(P13) < ε THEN AU41-apex
	IF [***inc/dec***(P13P14)]$_t$ < [***inc/dec***(P13P14)]$_{t-1}$ AND AU41-p AND ***up/down***(P13) < ε THEN AU41-offset

AU43 rule 28	Causes the upper eyelid to drop down completely, closes the eye. The duration of the apex is greater than ¼ of a second (i.e., grater than the duration of 5 frames). IF ***inc/dec***(P13P14) > ε AND ***inc/dec***(P13P14) > *T1* AND ***up/down***(P14) ≤ ε THEN AU43-45-p The rules for recognition of temporal segments of AU43 are similar to those defined for AU41 (see rule 27)

AU44
rule 29
	Raises the lower eyelid, narrows the eye opening, squints the eye.
	IF ***inc/dec***(P13P14) > ε AND ***inc/dec***(P13P14) > *T1* THEN AU44-p
	IF [***inc/dec***(P13P14)]$_t$ > [***inc/dec***(P13P14)]$_{t-1}$ AND AU44-p AND ***up/down***(P14) > ε THEN AU44-onset
	IF [***inc/dec***(P13P14)]$_t$ = [***inc/dec***(P13P14)]$_{t-1}$ AND AU44-p AND ***up/down***(P14) > ε THEN AU44-apex
	IF [***inc/dec***(P13P14)]$_t$ < [***inc/dec***(P13P14)]$_{t-1}$ AND AU44-p AND ***up/down***(P14) > ε THEN AU44-offset

AU45 rule 30	Blink. Causes the upper eyelid to drop down completely, closes the eye. The duration of the apex is lesser than ¼ of a second (i.e., lesser than the duration of 5 frames). The rule for recognition of AU45 is for the rest exactly the same as that one defined for AU43 (see rule 28)

TABLE IV

Rules for Resolving Temporal Conflicts and Uncertainties. Except for Rule 3, the Rules are Utilized in Both Cases: If Single Frames are Unlabeled or Labeled Incorrectly and if Temporal Segments (a Sequence of at Least Five Consecutive Frames) are Unlabeled or Labeled Incorrectly. Rule 3 is Utilized for Temporal Segments Only. Rule 4 Has a More Complex Form for the Case of Temporal Segments. Namely, Only if a Sequence Onset-Apex-Unlabeled-Apex-Offset is Encountered, the Unlabeled Temporal Segment Will be Re-Labeled as "Apex"

	Previous labeling	Current labeling (old label)	Subsequent labeling	Current labeling (new label)
Rule 1	Onset	Unlabeled or Apex	Onset	Onset
Rule 2	Onset	Unlabeled	Apex	Apex
Rule 3	Onset	Unlabeled	Offset	Apex
Rule 4	Apex	Unlabeled	Apex	Apex
Rule 5	Apex	Unlabeled	Offset	Apex
Rule 6	Offset	Unlabeled or Apex	Offset	Offset

based on the absence of facial activity rather than the difference in facial activity. Thus, in the case of neutral → sad → smile → neutral facial display, the proposed model will handle the "sad" and "smile" segments as a single expressive segment rather than two distinct facial expressions. To achieve segmentation in specific facial displays and to handle cases where no constraints are posed on input facial expression sequences, an extended variable-neutral-expressive-variable sequential model, where a "variable" segment contains either an expressive or neutral facial appearance, should be used. However, appropriate handling of these "variable" segments and the associated problems including the registration of the input video sequence and cancellation of noise is not an easy task.

VI. Experimental Evaluation

In spite of repeated calls for the need of a comprehensive, readily accessible reference set of face images that could provide a basis for benchmarks for all different efforts in research on machine analysis of facial expressions, no such database has been yet created that is shared by all diverse facial-expression-research communities [9], [16], [29]. In general, only isolated pieces of such a facial database exist. An example is the unpublished database of Ekman–Hager Facial Action Exemplars [52]. It has been used by several research groups (e.g., [20], [21], [24]) to train and test their methods for AU detection from frontal-view facial expression sequences. The facial expression image databases that have been made publicly available but are still not used by all diverse facial-expression-research groups are the JAFFE database [53] and the Cohn–Kanade AU-coded face image database [54].

None of these existing databases contains images of faces in profile view and none contains images of all possible single-AU activations. Also, the metadata (labels) associated with each database object usually do not identify the temporal segments (onset, apex, offset) of AUs and emotional facial displays shown

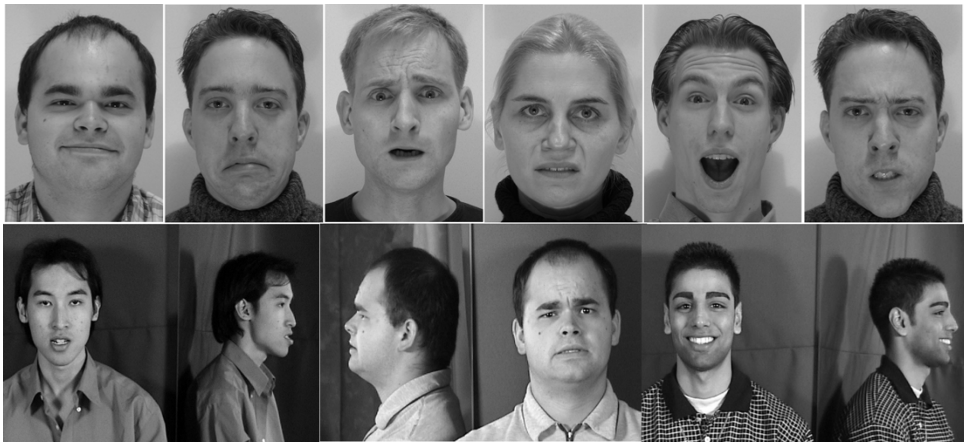

Fig. 7. Examples of MMI-Face-Database images. First row: static frontal-view images. Second row: apex frames of dual-view video sequences.

in the face video in question. Finally, these databases are not easily accessible and searchable. Once permission to use one of these databases has been issued, large, unstructured files of material are sent. As an attempt to address these issues, we have created a novel facial-expression-image database, which we call the MMI Face Database [55].

The MMI Face Database has been developed to address all the issues mentioned above. It contains more than 1500 samples of both static images and image sequences of faces in frontal and in profile view displaying various facial expressions of emotion, single AU activation, and multiple AU activation. It is publicly available and it has been developed as a web-based direct-manipulation application, allowing easy access and easy search of the available images. All data samples stored in the database have been acquired in the following way.

- *Sensing*: The static facial-expression images are all true color (24-bit) images which, when digitized, measure 720 × 576 pixels. There are approximately 600 frontal-view images and 140 dual-view images (i.e., combining frontal and profile view of the face, recorded using a mirror) of facial expressions. All video sequences have been recorded at a rate of 24 frames/s using a standard PAL camera. There are approximately 30 profile-view and 750 dual-view facial-expression sequences. The sequences are of variable length, lasting between 40 and 520 frames. Examples of recordings stored in the MMI Face Database are illustrated in Figs. 4, 5, and 7.
- *Subjects*: Our database includes 52 different faces of students and research staff members of both sexes (44% female), ranging in age from 19 to 62, having either a European, Asian, or South American ethnic background.
- *Samples*: The subjects were asked to display expressions that included either a single AU or a prototypic combination of AUs (such as in expressions of emotion). They were instructed by an expert (a FACS coder) on how to display the required facial expressions, and they were asked to include a short neutral state at the beginning and at the end of each expression. The subjects were asked to display the required expressions while minimizing out-of-plane head motions.
- *Metadata*: Two experts (FACS coders) were asked to depict the AUs displayed in the images constituting the MMI Face Database. In the case of facial-expression video sequences, they were also asked to depict the temporal segments of displayed AUs. When in doubt, decisions were made by consensus.

In order to test the AU recognition method described in the previous sections, we used 26 profile-view and 70 dual-view video sequences of the MMI Face Database (19 different subjects in total). In the case of dual-view video sequences, we used only the profile view of the face as the actual data. The metadata associated with these 96 image sequences represent the ground-truth with which we compared the judgments generated by our method. According to the neutral-expressive-neutral sequential facial expression model described in Section V, the sequences were first segmented into the different facial expressions pictured. Then, we initialized nine profile-contour facial points (P1–P6 and P8–P10, Fig. 2) as the extremities of the profile contour as proposed in [25]. Other six facial points (P7 and P11–P15, Fig. 2) were manually initialized in the first frame of each of 96 test sequences. (Note, however, that the method proposed in [36] can be easily trained to localize points P7 and P11–P15 automatically.) The accuracy of the method was measured with respect to the misclassification rate of each "expressive" segment of the input sequence, not with respect to each frame.

The results are summarized in Table V. The first column of Table V lists all different AUs occurring in 96 test image sequences according to the ground-truth. The second column identifies the total *number of occurrences* of each AU in the test data set according to the ground-truth. *Correct* means that the AUs detected by our method were identical to AUs indicated by the ground-truth. *Partially correct* denotes either that some, but not all, of the AUs indicated by the ground-truth were not recognized by our method (*Missing AUs*), or that some AUs that were not indicated by the ground-truth were recognized in addition to those that were (*Extra AUs*). *Incorrect* means that none of the AUs indicated by the ground-truth were recognized by the method. The overall recognition rate of the system has been calculated with respect to both the number of input AUs indicated by the ground-truth and the number of input samples (i.e., the number of "expressive" segments in the input video sequence). The average recognition rate of the system with respect to the number of AUs has been calculated as the ratio between the number of correctly recognized AUs and the number of input AUs. The average recognition rate of the system with respect to

TABLE V
Method's AU Recognition Performance for 96 Test Face-Profile Image Sequences. **Legend**: The Average Recognition Rate of the System With Respect to the Number of Input AUs: *Correct/Nr. of Occurrences*. The Average Recognition Rate of the System With Respect to the Number of Input Samples (i.e., to the Number of "Expressive" Segments of Input Video Sequences): *Nr. of Correctly Recognized Input Samples/the Total of 119 Input Samples*

Actual AUs	Nr. of occurrences	Recognized AUs			
		Correct	*Partially correct*		*Incorrect*
			Missing AUs	*Extra AUs*	
AU1	11	11	-	-	-
AU2	10	10	-	-	-
AU4	7	6	1(AU4)	-	-
AU5	9	9	-	-	-
AU6	7	7	-	-	-
AU7	3	2	-	-	-
AU9	5	4	1(AU9)	-	-
AU10	8	8	-	-	-
AU12	10	9	-	1 (AU6)	-
AU13	8	8	-	-	-
AU15	10	9	-	1 (AU26)	-
AU16	5	5	-	-	-
AU17	11	10	1(AU17)	-	-
AU18	6	5	1(AU18)	-	-
AU20	5	4	-	1 (AU26)	-
AU23	9	7	-	-	2(AU24)
AU24	6	6	-	-	-
AU25	14	14	-	-	-
AU26	21	21	-	-	-
AU27	13	13	-	-	-
AU28t	3	3	-	-	-
AU28b	5	5	-	-	-
AU28	7	6	-	1(AU17)	-
AU29	4	4	-	-	-
AU36t	2	2	-	-	-
AU36b	3	3	-	-	-
AU41	6	4	-	-	2(AU43)
AU43	4	4	-	-	-
AU44	6	5	-	-	1(AU43)
AU45	79	73	6(AU45)	-	-
With respect to the number of input AUs	299	280 (93.6%)			
With respect to the number of input samples	119	103 (**86.6%**)			

the number of input samples has been calculated as the ratio between the number of correctly recognized input samples and the total of 119 input samples. We achieved an average recognition rate of 93.6% input AUs-wise and an average recognition rate of 86.6% input samples-wise (Table V).

As far as misidentifications produced by our method are concerned, most of them arose from confusion between similar AUs (AU41 and AU43, AU23 and AU24) and from omission of very fast blinks (AU45 having a duration of less than five frames in either onset or offset). Both AU41 and AU43 cause the upper eyelid to drop down and narrow the eye opening. Only the height of the eye opening distinguishes AU41 from AU43, causing misidentification of AU41 in the case where the observed subject has long eyelashes or an eye opening that is naturally narrow. Since both AU23 and AU24 tighten the lips and reduce the height of the lips (vertical direction), only the length of the lips (horizontal direction) distinguishes AU24 from AU23, causing misidentification of any AU23 that is accompanied by an unintentional, small, out-of-plane head motion that makes the mouth appear shorter. Note that AU23 and AU24 are also often confused by human FACS coders [15] and by other automated AU analyzers (e.g., [22]). In addition, note that the temporal pattern of feature motion in AU23 activation is very similar to the one occurring in AU24 activation. Hence, the distinction between these two AUs may be more amenable to appearance-based analysis than to feature motion analysis.

In addition to the misidentifications listed above, the mistaken identifications of AU26 merit an explanation as well. In two cases, AU26 was present but the slightly parted teeth in a closed mouth remained undetected by human observers. In these cases, our method coded the input samples correctly, unlike the human observers.

As can be seen from Fig. 8, the temporal segments of the AUs indicated by the ground-truth varied slightly from those detected by our method. In Fig. 8, the full line represents the values calculated by the method for the relevant mid-level parameters over the number of frames defined at the horizontal axis. The dotted line represents the temporal segments of the

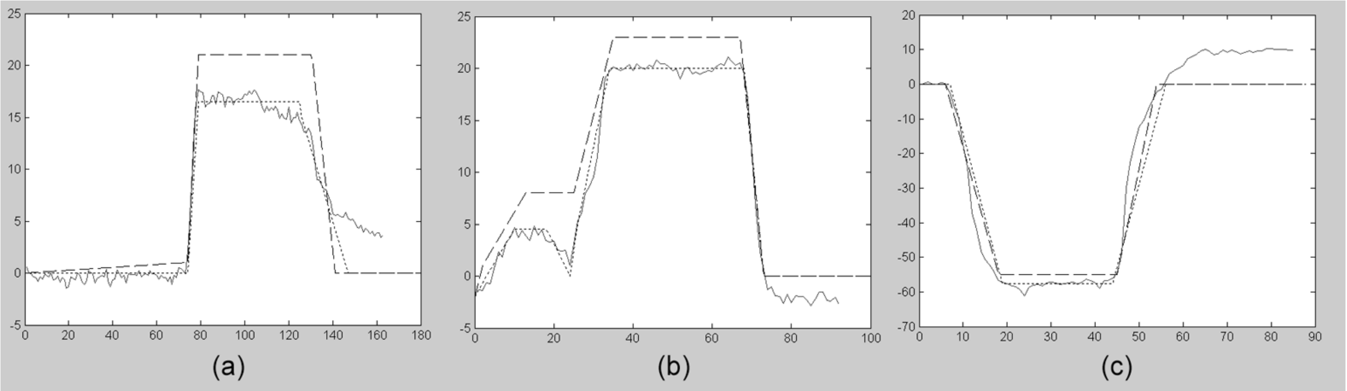

Fig. 8. Temporal segments of the AUs indicated by the ground truth (dashed line) and those detected by the method (dotted line). (a) AU2 activation in 163 frames of an "expressive" segment of AU1 + 2 + 5 video sequence (first two rows of Fig. 4). (b) AU12 activation in 92 frames of an "expressive" segment of AU6 + 12 + 25 video sequence (the fourth row of Fig. 5). (c) AU27 activation in 85 frames of an "expressive" segment of AU27 video sequence (not illustrated).

AUs calculated by the method. The dashed line represents an abstraction of the temporal segments of the AUs indicated by the ground-truth. For most AUs the boundaries of temporal segments were detected either at the same moment or a little bit later than prescribed by the ground-truth [Fig. 8(a)]. The measured delays take up to three frames on average, that is, up to 1/8 of a second. However, in the case of AUs whose activation becomes apparent from the movement of the mouth corner (i.e., AU12, AU13, AU15, and AU20), the temporal segments were almost always detected later than indicated by the ground-truth. The measured delays have an average duration of three to six frames, that is, up to 1/4 of a second [Fig. 8(b)]. The reason for these delays is the temporal rules used for recognition of AU activation. It seems that human observers detect activation of the AUs in question not only based on the presence of a certain movement (e.g., an upward movement of the mouth corner in the case of AU12) but also based on the appearance of the facial region around the mouth corner. Since appearance-based analysis is not performed by the system, only the movement of the mouth corner, which is detected usually later than the actual occurrence of the movement (due to thresholding), indicates the presence of the AUs in question, causing a delayed detection of these AUs. In addition, it is interesting to note that in cases of spontaneous smiles, the human observers indicated the presence of multiple apexes of AU12 but, in contrast to the analysis performed by our system, did not indicate the presence of multiple full temporal patterns (onset-apex-offset) of AU12 [Fig. 8(b)]. However, whether this difference is just a matter of a human coder blindly applying an accepted coding scheme according to which such a "dampened" smile is represented as a multiple-apex-AU12 [28] or a matter of a genuine insensitivity of the human eye to subtle offsets of AU12 occurring in between the apexes of AU12 remains an interesting research question.

Finally, upon a close inspection of the temporal rule used to recognize AU27 activation (rule 20, Table II), one may conclude that the onset of AU27 will always be detected later than indicated by the ground-truth. Namely, since both AU26 and AU27 pull down the lower jaw, only the extent of that pull distinguishes AU27 from AU26, causing misidentifications in the onset of AU27, that is, it causes a delayed detection of the onset of AU27. This is consistent with experimental data that show correlation between the extent of facial motion involved in a facial expression and the delay in the recognition of that expression [8], [56]: the larger the motion (and, in turn, the deformation in facial expression), the longer the response time. To handle this, any "onset AU26" segment that has been detected before the "onset AU27" segment is re-labeled as "onset AU27". In turn, the onset of AU27 is detected without delays [Fig. 8(c)].

VII. Conclusions

Automating the analysis of facial signals, especially rapid facial signals (i.e., AUs), is important to advance studies on human emotion and nonverbal communication, to design multimodal human-machine interfaces, and to boost numerous applications in fields as diverse as security, medicine, and education. In this paper, we presented a novel method for AU detection based upon changes in the position of the facial points tracked in a video of a near profile view of the face. The significance of this contribution are the following.

- The presented approach to automatic AU recognition extends the state of the art in automatic AU detection from face image sequences in several ways, including the facial view (profile), the temporal segments of AUs (onset, apex, offset), the number (27 in total), and the difference in AUs (e.g., AU29, AU36) handled. To wit, the automated systems for AU detection from face video that have been reported so far do not deal with the profile view of the face, cannot handle temporal dynamics of AUs, cannot detect out-of-plane movements such as thrusting the jaw forward (AU29), and, at best, can detect 16 to 18 AUs (from in total 44 AUs).
- This paper provides a basic understanding of how to achieve automatic detection of AUs and their temporal segments in a face-profile image sequence. Further research on facial expression symmetry, spontaneous vs. posed facial expressions, and facial expression recognition from multiple facial views can be based upon it.

Based upon the validation study presented in Section VI, it can be concluded that the proposed method exhibits an acceptable level of expertise. The achieved results are similar to those reported for other automated FACS coders of face video. Compared to the AFA system [24], our method achieves an average recognition rate of 86.6% for encoding of 27 AU codes and their combinations in 119 test samples, while the AFA system achieves an average recognition rate of 87.9% for encoding of

16 AUs and their combinations in 113 test samples. In comparison to the system proposed recently by Bartlett *et al.* [22], our method achieves an average recognition rate of 93.6% AU-wise for encoding of 27 AUs and their combinations, while their system achieves an average recognition rate of 94.5% AU-wise for encoding of 18 AUs and their combinations.

Except the profile view, the number of AUs, the difference in AUs, and the temporal dynamics handled, our method has also improved other aspects of automated AU detection compared to previously reported systems. In contrast to earlier approaches to automated AU detection, our system facilitates automatic segmentation of input image sequences into expressive and expressionless facial behavior pictured. Also, the performance of the proposed method is invariant to occlusions like glasses and facial hair as long as these do not entirely occlude facial fiducial points (e.g., P10 in the case of a long beard). Finally, due to the usage of the color-based observation model (Section II-B), the method performs well independently of changes in the illumination intensity.

However, the method cannot recognize the full range of facial behavior (i.e., all 44 AUs defined in FACS); it detects 27 AUs occurring alone or in combination in a near profile-view face image sequence. Although it has been reported that feature-based methods are usually outperformed by holistic template-based methods using Gabor wavelets, Independent Component Analysis, and Eigenfaces [20], [53], the comparison given above indicates that our feature-based method performs just as well as the best template-based method proposed up to date (i.e. [22]). We believe, however, that further research efforts toward combining both approaches are necessary if the full range of human facial behavior is to be coded in an automatic way.

If we consider the state of the art in face detection and facial point localization and tracking, noisy and partial data should be expected. As remarked by Pantic *et al.* [9], [19], a facial expression analyzer should be able to deal with these imperfect data and to generate its conclusion so that the certainty associated with it varies with the certainty of face and facial point localization and tracking data. To deal with inaccuracies in facial point tracking, our method employs a memory-based process that takes into account the dynamics of facial expressions (Table IV). However, our method does not calculate the output data certainty by propagating the input data certainty (i.e., the certainty of facial point tracking). Future work on this issue aims at investigating on the use of measures that can express the confidence to facial point tracking and that can facilitate both more robust AU recognition and the assessment of the certainty of the performed AU recognition.

Finally, our method assumes that the input data are near profile-view face image sequences showing facial displays that always begin with a neutral state. In reality, such an assumption cannot be made; variations in the viewing angle should be expected. Also, human facial behavior is more complex and transitions from a facial display to another do not have to involve intermediate neutral states. As a consequence, the proposed facial expression analyzer cannot deal with spontaneously occurring facial behavior. Yet, answering the question of how to achieve parsing the stream of facial and head movements not under volitional control is essential for the realization of multimodal human-machine interfaces and for advancing studies on human emotion and nonverbal communication [57]. This forms the main focus of our current and future research efforts.

Acknowledgment

The authors would like to thank the referees and M. J. Nieman for their helpful comments and suggestions.

References

[1] J. Russell and J. Fernandez-Dols, *The Psychology of Facial Expression*. New York: Cambridge Univ. Press, 1997.

[2] A. Mehrabian, "Communication without words," *Psych. Today*, vol. 2, no. 4, pp. 53–56, 1968.

[3] D. Keltner and P. Ekman, "Facial expression of emotion," in *Handbook of Emotions*, M. Lewis and J. M. Haviland-Jones, Eds. New York: Guilford, 2000, pp. 236–249.

[4] K. Mase, "Recognition of facial expression from optical flow," *IEICE Trans.*, vol. E74, no. 10, pp. 3474–3483, 1991.

[5] M. Black and Y. Yacoob, "Recognizing facial expressions in image sequences using local parameterized models of image motion," *Comput. Vis.*, vol. 25, no. 1, pp. 23–48, 1997.

[6] I. Essa and A. Pentland, "Coding, analysis, interpretation and recognition of facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 757–763, Jul. 1997.

[7] M. Pantic and L. J. M. Rothkrantz, "Expert system for automatic analysis of facial expression," *Image Vis. Comput. J.*, vol. 18, no. 11, pp. 881–905, 2000.

[8] A. M. Martinez, "Matching expression variant faces," *Vis. Res.*, vol. 43, no. 9, pp. 1047–1060, 2003.

[9] M. Pantic and L. J. M. Rothkrantz, "Toward an affect-sensitive multimodal human-computer interaction," *Proc. IEEE*, vol. 91, no. 9, pp. 1370–1390, Sep. 2003.

[10] C. Darwin, *The Expression of the Emotions in Man and Animals*. Chicago, IL: Univ. of Chicago Press, 1965.

[11] P. Ekman, *Emotions Revealed*. New York: Times Books, 2003.

[12] A. Ortony and T. J. Turner, "What is basic about basic emotions?," *Psych. Rev.*, vol. 74, pp. 315–341, 1990.

[13] K. R. Scherer and P. Ekman, *Handbook of Methods in Non-Verbal Behavior Research*. Cambridge, U.K.: Cambridge Univ. Press, 1982.

[14] P. Ekman and W. V. Friesen, *Facial Action Coding System*. Palo Alto, CA: Consulting Psychologist Press, 1978.

[15] P. Ekman, W. V. Friesen, and J. C. Hager, *Facial Action Coding System*. Salt Lake City, UT: A Human Face, 2002.

[16] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1424–1445, Dec. 2000.

[17] H. Tao and T. S. Huang, "Connected vibrations: a modal analysis approach to nonrigid motion tracking," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, 1998, pp. 735–740.

[18] S. B. Gokturk, J. Y. Bouguet, C. Tomasi, and B. Girod, "Model-based face tracking for view-independent facial expression recognition," in *Proc. IEEE Int. Conf. Face and Gesture Recognition*, 2002, pp. 272–278.

[19] M. Pantic, "Face for interface," in *The Encyclopedia of Multimedia Technology and Networking*, M. Pagani, Ed. Hershey, PA: Idea Group Reference, 2005, vol. 1, pp. 308–314.

[20] M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, "Measuring facial expressions by computer image analysis," *Psychophysiology*, vol. 36, pp. 253–263, 1999.

[21] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, "Classifying facial actions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 10, pp. 974–989, Oct. 1999.

[22] M. S. Bartlett, G. Littlewort, C. Lainscsek, I. Fasel, and J. R. Movellan, "Machine learning methods for fully automatic recognition of facial expressions and facial actions," in *Proc. IEEE Int. Conf. Systems, Man, and Cybernetics*, 2004, pp. 592–597.

[23] J. F. Cohn, A. J. Zlochower, J. Lien, and T. Kanade, "Automated face analysis by feature point tracking has high concurrent validity with manual faces coding," *Psychophysiology*, vol. 36, pp. 35–43, 1999.

[24] Y. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 97–115, Feb. 2001.

[25] M. Pantic and L. Rothkrantz, "Facial action recognition for facial expression analysis from static face images," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 34, no. Jun., pp. 1449–1461, 2004.

[26] M. F. Valstar, M. Pantic, and I. Patras, "Motion history for facial action detection from face video," in *Proc. IEEE Int. Conf. Systems, Man, and Cybernetics*, 2004, pp. 635–640.

[27] Y. Yacoob, L. Davis, M. Black, D. Gavrila, T. Horprasert, and C. Morimoto, "Looking at people in action," in *Computer Vision for Human–Machine Interaction*, R. Cipolla and A. Pentland, Eds. Cambridge, U.K.: Cambridge Univ. Press, 1998, pp. 171–187.

[28] K. L. Schmidt and J. F. Cohn, "Dynamics of facial expression: Normative characteristics and individual differences," in *Proc. IEEE Int. Conf. Multimedia and Expo*, 2001, pp. 547–550.

[29] Human Interaction Laboratory, "Final Report to NSF of the Planning Workshop on Facial Expression Understanding," Univ. of California, San Francisco, CA, P. Ekman, T. S. Huang, T. J. Sejnowski, and J. C. Hager, Eds., 1993.

[30] M. Mendolia and R. E. Kleck, "Watching people talk about their emotions—Inferences in respons to full-face vs. profile expressions," *Motiv. Emotion*, vol. 15, no. 4, pp. 229–242, 1991.

[31] J. C. Hager, "Asymmetry in facial muscular actions," in *What the Face Reveals*, P. Ekman and E. L. Rosenberg, Eds. New York: Oxford Univ. Press, 1997, pp. 58–62.

[32] S. Mitra and Y. Liu, "Local facial asymmetry for expression classification," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, 2004, pp. 889–894.

[33] A. M. Martinez, "Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 6, pp. 748–763, Jun. 2002.

[34] M. Pantic, I. Patras, and L. J. M. Rothkrantz, "Facial action recognition in face profile image sequences," in *Proc. IEEE Int. Conf. Multimedia and Expo*, 2002, pp. 37–40.

[35] M. Pantic and I. Patras, "Temporal modeling of facial actions from face profile image sequences," in *Proc. IEEE Int. Conf. Multimedia and Expo*, vol. 1, 2004, pp. 49–52.

[36] D. Vukadinovic and M. Pantic, "Fully automatic facial feature point detection using Gabor feature based boosted classifiers," in *Proc. IEEE Int. Conf. Systems, Man, and Cybernetics*, Vancouver, BC, Canada, Oct. 2005, pp. 1692–1698.

[37] M. Isard and A. Blake, "Condensation—conditional density propagation for visual tracking," *Int. J. Comput. Vis.*, vol. 29, no. 1, pp. 5–28, 1998.

[38] ——, "Icondensation: Unifying low-level and high-level tracking in a stochastic framework," in *Proc. Eur. Conf. Computer Vision*, 1998, pp. 893–908.

[39] M. K. Pitt and N. Shephard, "Filtering via simulation: auxiliary particle filtering," *J. Amer. Stat. Assoc.*, vol. 94, pp. 590–599, 1999.

[40] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. 173–188, Feb. 2002.

[41] J. MacCormick and A. Blake, "Probabilistic exclusion and partitioned sampling for multiple object tracking," *Int. J. Comput. Vis.*, vol. 39, no. 1, pp. 57–71, 2000.

[42] P. Perez, C. Hue, J. Vermaak, and M. Gangnet, "Color-based probabilistic tracking," in *Proc. Eur. Conf. Computer Vision*, 2002, pp. 661–675.

[43] "Special issue on sequential state estimation: from Kalman filters to particle filters," *Proc. IEEE*, vol. 92, no. 3, pp. 399–574, Mar. 2004.

[44] L. D. Harmon, M. K. Khan, R. Lash, and P. F. Raming, "Machine identification of human faces," *Pattern Recognit.*, vol. 13, pp. 97–110, 1981.

[45] H. T. Nguyen, M. Worring, and R. vd. Boomgaard, "Occlusion robust adaptive template tracking," in *Proc. IEEE Int. Conf. Computer Vision*, vol. 1, 2001, pp. 678–683.

[46] J. Vermaak, P. Perez, M. Gangnet, and A. Blake, "Toward improved observation models for visual tracking: Selective adaptation," in *Proc. Eur. Conf. Computer Vision*, 2002, pp. 645–660.

[47] Y. Wu and T. Huang, "A co-inference approach to robust tracking," in *Proc. IEEE Int. Conf. Computer Vision*, vol. 2, 2001, pp. 26–33.

[48] P. J. Huber, *Robust Statistics*. New York: Wiley, 1981.

[49] M. Schneider, A. Kandel, G. Langholz, and G. Chew, *Fuzzy expert system tools*. West Sussex, U.K.: Wiley, 1997.

[50] T. Otsuka and J. Ohya, "Recognizing multiple persons' facial expressions using HMM based on automatic extraction of significant frames from image sequences," in *Proc. IEEE Int. Conf. Image Processing*, vol. 2, 1997, pp. 546–549.

[51] I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang, "Facial expression recognition from video sequences: temporal and static modeling," *Comput. Vis. Image Understand.*, vol. 91, pp. 160–187, 2003.

[52] P. Ekman, J. Hager, C. H. Methvin, and W. Irwin, "Ekman–Hager Facial Action Exemplars," Human Interaction Lab., Univ. of California, San Francisco.

[53] M. J. Lyons, J. Budynek, and S. Akamatsu, "Automatic classification of single facial images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 12, pp. 1357–1362, Dec. 1999.

[54] T. Kanade, J. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition*, 2000, pp. 46–53.

[55] M. Pantic, M. F. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *Proc. IEEE Conf. Multimedia and Expo*, 2005, [Online] Available at: http://www.mmifacedb.com/, pp. 317–321.

[56] A. W. Young, D. Rowland, A. J. Calder, N. L. Etcoff, A. Seth, and D. I. Perrett, "Facial expression megamix: Test of dimensional and category accounts of emotion recognition," *Cognition*, vol. 63, pp. 271–313, 1997.

[57] M. Pantic, N. Sebe, J. F. Cohn, and T. Huang, "Affective multimodal human-computer interaction," in *Proc. ACM Int. Conf. Multimedia*, 2005.

Maja Pantic (S'98–M'02) received the M.S. and Ph.D. degrees in computer science from Delft University of Technology, Delft, The Netherlands, in 1997 and 2001, respectively.

From 2001 to 2005, she was an Assistant Professor in the Department of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology. She is currently an Associate Professor in the same department, where she is doing research in the area of machine analysis of human interactive cues for achieving a natural, multimodal human–machine interaction. She has published over 40 technical papers in the areas of machine analysis of facial expressions and emotions, artificial intelligence, and human-computer interaction and has served as an invited speaker and a program committee member at several conferences in these areas.

Dr. Pantic received the Innovational Research Award of the Dutch Scientific Organization for her research on Facial Information For Advanced Interface in 2002, as one of the seven best young scientists in exact sciences in the Netherlands.

Ioannis Patras (S'97–M'02) received the B.Sc. and M.Sc. degrees in computer science from the Computer Science Department, University of Crete, Heraklion, Greece, in 1994 and in 1997, respectively, and the Ph.D. degree from the Department of Electrical Engineering, Delft University of Technology, Delft, The Netherlands, in 2001.

From 2001 to 2003, he was a Postdoctorate Researcher in the area of multimedia analysis at the University of Amsterdam, Amsterdam, The Netherlands. From 2003 to 2005, he was a Postdoctorate Researcher in the area of vision-based human machine interaction (focusing on facial and body gesture analysis) at Delft University of Technology. Since 2005, he has been a Lecturer at the Computer Vision and Pattern Recognition Group, The University of York, York, U.K. His research interests lie mainly in the areas of computer vision and pattern recognition and their applications in multimedia data management, multimodal human machine interaction, and visual communications.