Facial Action Recognition for Facial Expression Analysis From Static Face Images

Maja Pantic, *Member, IEEE,* and Leon J. M. Rothkrantz

***Abstract*—Automatic recognition of facial gestures (i.e., facial muscle activity) is rapidly becoming an area of intense interest in the research field of machine vision. In this paper, we present an automated system that we developed to recognize facial gestures in static, frontal- and/or profile-view color face images. A multidetector approach to facial feature localization is utilized to spatially sample the profile contour and the contours of the facial components such as the eyes and the mouth. From the extracted contours of the facial features, we extract ten profile-contour fiducial points and 19 fiducial points of the contours of the facial components. Based on these, 32 individual facial muscle actions (AUs) occurring alone or in combination are recognized using rule-based reasoning. With each scored AU, the utilized algorithm associates a factor denoting the certainty with which the pertinent AU has been scored. A recognition rate of 86% is achieved.**

***Index Terms*—Facial action units, facial action unit combinations, facial expression analysis, image processing, rule-based reasoning, spatial reasoning, uncertainty.**

I. Introduction

FACIAL expressions play a significant role in our social and emotional lives. They are visually observable, conversational, and interactive signals that clarify our current focus of attention and regulate our interactions with the environment and other persons in our vicinity [22]. They are our direct and naturally preeminent means of communicating emotions [12], [22]. Therefore, automated analyzers of facial expressions seem to have a natural place in various vision systems, including automated tools for behavioral research, lip reading, bimodal speech processing, videoconferencing, face/visual speech synthesis, affective computing, and perceptual man-machine interfaces. It is this wide range of principle driving applications that has lent a special impetus to the research problem of automatic facial expression analysis and produced a surge of interest in this research topic.

Most approaches to automatic facial expression analysis attempt to recognize a small set of prototypic emotional facial expressions, i.e., fear, sadness, disgust, anger, surprise, and happiness (e.g., [2], [9], [13], [15]; for an exhaustive survey, see [17]). This practice may follow from the work of Darwin [4], and more recently Ekman [12], who suggested that basic emotions have corresponding prototypic expressions. In everyday life, however, such prototypic expressions occur relatively rarely; emotions are displayed more often by subtle changes in one or few discrete facial features, such as the raising of the eyebrows in surprise [14]. To detect such subtlety of human emotions and, in general, to make the information conveyed by facial expressions available for usage in the various applications mentioned above, automatic recognition of facial gestures (atomic facial signals) is needed.

From several methods for recognition of facial gestures, the facial action coding system (FACS) [6] is the best known and most commonly used in psychological research [23]. It is a system designed for human observers to describe changes in the facial expression in terms of visually observable activations of facial muscles. 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 a set of facial muscles. Along with the definition of various AUs, FACS also provides the rules for AU detection in a face image. Using these rules, a FACS coder (i.e., a human expert having a formal training in using FACS) encodes a shown facial expression in terms of the AUs that produce the expression.

Although FACS provides a good foundation for AU-coding of face images by human observers, achieving AU recognition by a computer remains difficult. A problematic issue is that AUs can occur in more than 7000 different combinations [23], causing bulges (e.g., by the tongue pushed under one of the lips) and various in- and out-plane movements of facial components (e.g., jetted jaw) that are difficult to detect in two-dimensional (2-D) face images.

Few approaches have been reported for automatic recognition of AUs in images of faces [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., [2], [9], [11], [13]). To detect six individual AUs in face image sequences free of head motions, Bartlett *et al.* [1] used a neural network (NN) approach. They achieved 91% accuracy by feeding the utilized NN 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.* [5] used Gabor wavelet representation and independent component analysis. They reported a 95.5% average recognition rate accomplished by their method. To recognize eight individual AUs and seven combinations of AUs in face image sequences free of head motions, Cohn *et al.* [3] used facial feature point tracking and discriminant function analysis. They reported an 85% average recognition rate. Tian *et al.* [27] used

Manuscript received July 29, 2003; revised October 17, 2003. This work was supported by the Netherlands Organization for Scientific Research (NWO) under Grant EW-639.021.202. This paper was recommend by Associate Editor P. Bhattacharya.

The authors are with the Department of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, 2628 CD Delft, The Netherlands (e-mail: M.Pantic@ewi.tudelft.nl).

Digital Object Identifier 10.1109/TSMCB.2004.825931

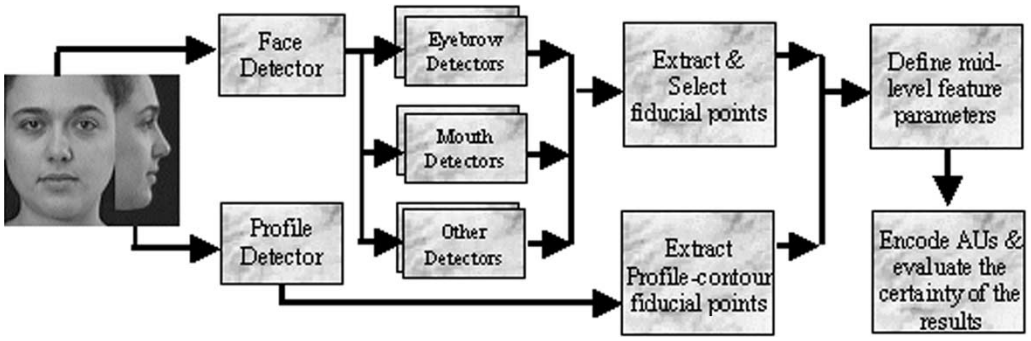


Fig. 1. Outline of the method for AU recognition from dual-view static face images.

lip tracking, template matching and NNs to recognize 16 AUs occurring alone or in combination in nearly frontal-view face image sequences. They reported an 87.9% average recognition rate attained by their method. Pantic *et al.* [19] used face-profile-contour tracking and rule-based reasoning to recognize 20 AUs occurring alone or in a combination in nearly left-profile-view face image sequences. They achieved an 84.9% average recognition rate by their method.

In contrast to these previous approaches to automatic AU detection, which deal neither with static face images nor with different facial views at the same time, the research reported here addresses the problem of automatic AU coding from frontal- and profile-view static face images. It was undertaken with three motivations.

1) While motion records are necessary for studying temporal dynamics of facial behavior, static images are important for obtaining configurational information about facial expressions, which is essential, in turn, for inferring the related meaning (e.g., in terms of emotions) [7], [23]. Since 100 still images or a minute of a video tape take approximately one hour to manually score in terms of AUs [6], it is obvious that automating facial expression measurement would be highly beneficial. While some efforts in automating FACS coding from face image sequences have been made, no such effort has been made for the case of static face images.
2) In a frontal-view face image, AUs such as showing the tongue or pushing the jaw forward represent out-of-plane nonrigid movements which are difficult to detect. Such AUs are clearly observable in a profile view of the face. On the other hand, changes in the appearance of the eyes and eyebrows cannot be detected from the nonrigid changes in the profile contour, but are clearly observable from a frontal facial view. The usage of both frontal and profile facial views promises, therefore, a quantitative increase in AUs that can be handled.
3) A basic understanding of how to achieve automatic facial gesture analysis from multiple views of the human face is necessary if facial expression analyzers capable of handling partial and inaccurate data are to be developed [20]. Based on such knowledge, procedures of greater flexibility and improved quality can evolve.

The authors' group has already built a first prototype of an automated facial action detector, the novel version of which is presented in this paper. This prototype system was aimed at automatic recognition of six basic emotions in static face images. Pantic and Rothkrantz [15] used different image processing techniques like edge detection, active contours and NNs in a combination with rule-based forward reasoning to recognize 27 AUs from a portrait and 20 AUs from a face profile image and then classify them in six basic emotion categories. The average recognition rate ranged from 62% to 100% for different AUs. This prototype system had several limitations.

1) It required manual detection of the features describing different shapes of the mouth.
2) Detection of the features related to the image intensity and brightness distribution in certain facial areas was not robust since it required highly constrained illumination conditions.
3) The algorithm used for localizing the face profile contour performed only well for face-profile images having a uniform dark background.
4) The system was not capable of dealing with minor inaccuracies of the utilized detectors.
5) The employed forward chaining inference procedure is relatively slow since it finds *one* solution in each "pass" through the knowledge base.

The current version of the automated facial action detector addresses many of these limitations. Fig. 1 outlines our novel method proposed in this paper. First, static frontal and/or profile-view image of an expressionless face of the observed subject is processed. Under the assumption that input images are nonoccluded, scale- and orientation-invariant face images (e.g., Fig. 1), each subsequent image of the observed subject (acquired during the same monitoring session with the pertinent subject) is processed in the following manner. The face region is extracted from the input frontal-view face image. The face-profile region is extracted from the input profile-view face image. To do so, watershed segmentation with markers is applied on the morphological gradient of the input color image. For the frontal view, the segmented face region is subjected to a multidetector processing: per facial component (eyes, eyebrows, mouth), one or more spatial samples of its contour are generated. From each spatially sampled contour of a facial component, we extract a number of points. In total, we extract 19 different frontal-face feature points. For the profile-view, we extract ten feature points from the contour of the segmented face-profile region (i.e., face profile contour). By performing an intra-solution consistency

check, a certainty factor CF is assigned to each extracted point. A comparison of CFs assigned to frontal-face feature points leads to a selection of the most accurate of the redundantly extracted data. Subtle changes in the analyzed facial expression are measured next. Motivated by AUs of the FACS system, these changes are represented as a set of midlevel feature parameters describing the state and motion of the feature points and the shapes formed by certain feature points. Based on these feature parameters, a rule-based algorithm with the fast-direct chaining inference procedure interprets the extracted facial information in terms of 32 AUs occurring alone or in combination. With each scored AU, the utilized algorithm associates a factor denoting the certainty with which the pertinent AU has been scored.

Face and face-profile detection, facial feature extraction, parametric representation of the extracted information and AU coding are explained in Sections II–IV. Experimental results are presented in Section V.

II. Face and Face-Profile Detection

The first step in automatic facial expression analysis is to locate the face in the scene. Possible strategies for face detection vary a lot, depending on the type of input images [31]. We address this problem as a segmentation problem in two objects: the face and the background. For its low computational complexity and its good localization properties we chose the watershed segmentation with markers as the segmentation means. For each input face image (either in frontal or in profile view), the markers of the two objects are extracted as follows. First, a color-based segmentation extracts the skin region as the largest connected image component with hue, saturation, and value within the range $[\max(-0.7, \boldsymbol{H_{avg}} - 0.35), \min(\boldsymbol{H_{avg}} + 0.35, 1.05)]$, $[0, 0.7]$, and $[0.1, 0.9]$, respectively, where $\boldsymbol{H_{avg}}$ is the average hue in the horizontal middle of the image. Although people have different skin color, several studies have shown that the major difference lies in the intensity rather than in the chrominance [26], [30], [31]. After analyzing 360 dual-view face images of different people (see Section V-A), we have found out that the hue $\in [-\pi, \pi]$ of the human face color seldom exceeds the range $[-0.7, 1.05]$ and that the saturation $\in [0, 1]$, remains within the range [0, 0.7]. Similar results have been reported in [30], [26]. The experimentation also showed that the hue never deviates more than 0.35 from its average value for a given face image. Hence, under the assumption 1) that we do not deal with face detection in arbitrary scenes, 2) that the largest part of an input, true-color, PAL-camera acquired image is either a portrait or a profile view of a face (see Section V-A) and, in turn, 3) that calculating the average hue in the horizontal middle of the image will indeed identify the face-skin color of the observed person, we defined human face-skin color hue, saturation, and value domains as given above.

The skin region that results in this way is in general well localized but might suffer from small inaccuracies along the face boundaries. Moreover, the marker-based watershed segmentation requires that the objects' markers (for the face and for the background, respectively) are placed completely within the pertinent objects. In order to deal with the above, the Face marker is extracted by performing a binary erosion with a small square structuring element (3×3) on the skin region, an operation that, roughly speaking, "shrinks" the latter by 1–2 pixels. In the absence of a model for the color of the background, such as the skin color model is for the face, we base the extraction of the background marker also on the skin region. To do so, we consider a box, which is by T pixels larger than the bounding box of the skin region, as the background marker. The number T should be rather small (in our experiments five) since otherwise, in the presence of clutter in the background, strong background edges may appear between the two markers. Once the markers of the two objects are extracted, we apply the watershed segmentation algorithm [29] on the morphological gradient of the input color image. The gradient is estimated as the color difference between the morphological opening and closing operators, each of which is applied separately to each of the three components of the color image. We choose the Euclidian distance in the $\mathrm{Lu} * \mathrm{v}*$ color space as a metric of color difference, since the $\mathrm{Lu} * \mathrm{v}*$ space is perceptually uniform under this metric [25]. Fig. 2 outlines the employed algorithm. It yields a good localization of the face given that the most prominent color edge between the markers is indeed the face contour.

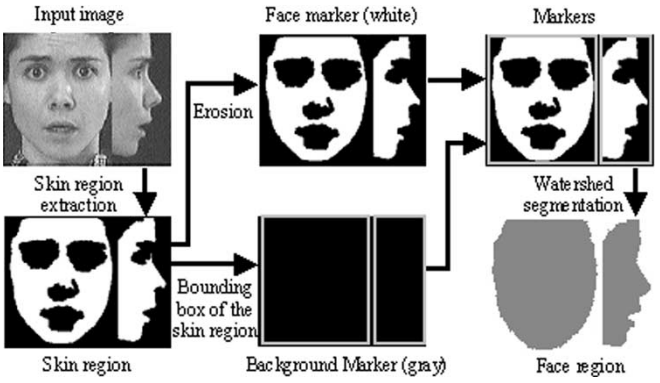


Fig. 2. Face region extraction by watershed segmentation with markers.

III. Facial Feature Extraction

Contractions of facial muscles change the appearance of permanent and transient facial features. Permanent facial features are facial components such as eyebrows, eyes, and mouth. Their shape and location can alter immensely with expressions (e.g., pursed lips versus delighted smile). Transient facial features are any facial lines and bulges that did not become permanent with age but appear with expressions. To reason about shown facial expression and the facial muscle actions that produced it, one must first detect facial features and their current appearance.

Our approach to facial features' detection in images of faces utilizes the face region and/or the face-profile region, extracted from an input face image as described above, and adopts the assumption that input images acquired during a single monitoring session with a subject are nonoccluded, scale and orientation invariant, with face-profile images in right profile view.

A. Profile Face Feature Extraction

The contour of the segmented face-profile region is treated as the face profile contour in further processing. To extract the fea-

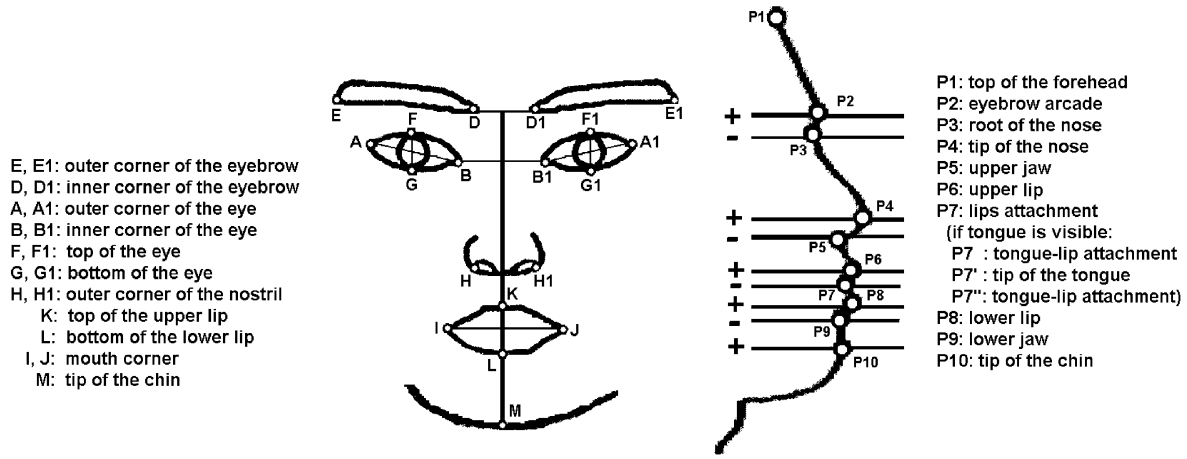


Fig. 3. Feature points (fiducials of the contours of the face components and of the profile contour).

Search window basis W_{PB}: width(W_{PB}) × height(W_{PB}), width(W_{PB}) = height(W_{PB}) = 3% length($P1_N P4_N$)
P1 (stable point, contractions of the facial muscles do not cause its physical displacement): uppermost point of *f*
P4 (stable point, contractions of the facial muscles do not cause its physical displacement): rightmost point of *f*
P2: maximum of *f* between P1 and P4; W_{P2}: 2 * width(W_{PB}) × 4 * height(W_{PB}); left-middle(W_{P2}) = $P2_N$
P3: minimum of *f* between P1 and P4; $W_{P3} = W_{PB}$; center(W_{P3}) = $P3_N$
P10: lowermost maximum of *f*; W_{P10}: 15 * width(W_{PB}) × 15 * height(W_{PB}); top-middle(W_{P10}) = $P10_N$
P5: the first minimum of *f* between P4 and P10; W_{P5}: width(W_{PB}) × 2 * height(W_{PB}); left-middle(W_{P5}) = $P5_N$
P9: the last minimum of *f* between P4 and P10; W_{P9}: 15 * width(W_{PB}) × 15 * height(W_{PB}); center(W_{P9}) = $P9_N$
P7: the first minimum of *f* between P5 and P9; W_{P7}: 5 * width(W_{PB}) × 5 * height(W_{PB}); center(W_{P7}) = $P7_N$
if P7 is the only minimum of *f* between P5 and P9, P7' and P7" do not exist, otherwise:
P7": the second minimum of *f* between P5 and P9; $W_{P7"} = W_{P7}$; center($W_{P7"}$) = $P7_N$
P7': maximum of *f* between P7 and P7"; $W_{P7'} = W_{P7}$; center($W_{P7'}$) = $P7_N$
P6: maximum of *f* between P5 and P7; W_{P6}: 4 * width(W_{PB}) × 8 * height(W_{PB}); center(W_{P6}) = $P6_N$
P8: maximum of *f* between P7 (P7" if exists) and P9; W_{P7}: 4 * width(W_{PB}) × 8 * height(W_{PB}); center(W_{P7}) = $P7_N$

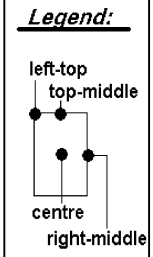


Fig. 4. Definitions of the profile-contour fiducial points P and the related search windows W_P given in the order in which the points are extracted from the profile contour function f defined for image I. The size of each search window W_P (width × height) is defined relative to the size of the *search window basis* W_{PB}, which is defined relative to the location of stable facial points P1 and P4. The positioning of each search window W_P is defined relative to the position of point P_N extracted from the face profile image N of a neutral expression of the observed subject.

ture points from the face profile contour, we move from image to function analysis and treat the right-hand side of the face profile contour as a profile contour function. We extract ten profile-contour fiducial points, illustrated in Fig. 3, as the extremities of this function.

The zero-crossings of the function's first-order derivative define extremities. Usually, many extremities are found. To handle the false positives and to ascertain correct extraction of the feature points illustrated in Fig. 3 in all situations (e.g., also when the tongue is visible and P7′ and P7″ exist), we proceed as follows. We exploit the knowledge about facial anatomy, the knowledge about spatial arrangement of the extreme points, and the information extracted from the image of a neutral facial expression of the observed subject, and we extract the feature points in a particular order (Fig. 4). After analyzing 240 dual-view face images of different people showing different facial expressions (see Section V.A), a standard "search" window $\boldsymbol{W_P}$ has been defined for each fiducial point $\boldsymbol{P}$ with respect to anatomically possible directions and magnitudes of the motion on the skin surface affecting the temporal location of $\boldsymbol{P}$. For instance, nonrigid movements of the eyebrows (raised and/or frowned eyebrows) cause upward and/or outward movement of P2, while the upward pull of the skin along the nose (i.e., wrinkled nose) causes outward and downward movement of P2. This and the fact that no facial muscle activity can push the eyebrow arcade either to the nasal bone or to the middle of the forehead, define the size and the positioning of the search window $\boldsymbol{W_{P2}}$ given in Fig. 4. Furthermore, each search window $\boldsymbol{W_P}$ has been defined relative to the "search window basis" $\boldsymbol{W_{PB}}$, and, in turn, relative to the referential, stable facial points P1 and P4, (Fig. 4). Given these referential points, the search window basis $\boldsymbol{W_{PB}}$ and the search windows $\boldsymbol{W_P}$ are defined in a scale- and person-invariant manner. Fiducial point $\boldsymbol{P}$ is determined eventually such that it represents a specific global extremity (Fig. 3) of the profile contour function within the search window $\boldsymbol{W_P}$, which is set around the location of $\boldsymbol{P_N}$ discerned for the face-profile image $\boldsymbol{N}$ of a neutral expression of the observed subject.

B. Frontal Face Feature Extraction

Each of the known techniques for facial feature detection in static face images (e.g., snake fitting, template matching, local spatial filtering) has circumstances under which it performs poorly and circumstances under which it performs extremely well. Introducing redundancy in the extracted facial expression data by employing multidetector processing and then selecting

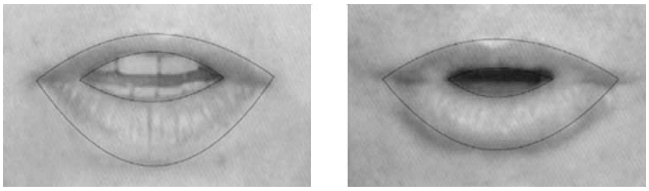


Fig. 5. Typical results of the *mouth template matching*.

the best of the acquired results could yield, therefore, a more accurate and complete set of detected facial features. Hence, we employ multidetector processing of the face region segmented from an input frontal-view face image to spatially sample the contours of the facial components. This "hybrid" facial feature detector is expected to result in a more robust performance than either a single detector used for all facial features or a set of different detectors, each of which used for one facial feature. Each facial feature detector that we exploit is an already existing facial feature detector. The utilized detectors have been chosen because they are simple and easy to implement. Yet, another set of detectors, which under the same conditions perform similarly (qua robustness and accuracy) to the detectors we are currently using to spatially sample the contours of the eyebrows, eyes, nostrils, and mouth from the facial portrait, could be chosen instead (e.g., the template-matching methods proposed in [27]). Since virtually all the employed facial feature detectors have been presented elsewhere, we provide in Appendix A just a short overview of the utilized methods.

After the multidetector processing of the face region, we proceed with the feature point extraction. For the cases where multiple detectors are used to localize the contour of a certain facial component, a relevant set of fiducial points is extracted from each spatially sampled contour of the pertinent facial component. For instance, from each localized mouth contour Mouth, we extract four feature points (see Fig. 5). In total, we extract 19 different feature points corresponding to the vertices and/or the apices of the contours of the facial components (Fig. 3).

C. Data Certainty Evaluation and Feature Selection

We assign the same certainty factor to each fiducial point that belongs to the same contour of a facial component spatially sampled by a certain detector. For example, we assign the same certainty factor $\mathrm{CF_A} = \mathrm{CF_F} = \mathrm{CF_G} = \mathrm{CF_B} \in [0,1]$ to all fiducial points of the right eye. To do so, we measure first the distance between the currently detected inner corner of the eye $\mathrm{B_{current}}$ and point $\mathrm{B_{neutral}}$ detected in the neutral expression image of the subject. Then we calculate the pertinent $\mathrm{CF_B}$ by using the functional form (1).

$$\mathrm{CF_B} = \mathbf{sigm}(\boldsymbol{d}(\mathrm{B_{current}}, \mathrm{B_{neutral}}); 7; 3.5) \quad (1)$$

$$\mathbf{sigm}(\mathrm{x}; \mu; \sigma) = 1/(1 + \exp((x - \mu)/\sigma)). \quad (2)$$

In (1), $\boldsymbol{d}(\mathrm{p1}, \mathrm{p2})$ is the distance between points p1 and p2 measured in pixels and $\mathbf{sigm}(\mathrm{x}; \mu; \sigma)$ is a Sigmoid function given in (2), whose parameters are determined under the assumption that there are 60 to 80 pixels across the width of the subject's eye (see Section V.A). This functional form implies that if $\boldsymbol{d}(\mathrm{B_{current}}, \mathrm{B_{neutral}}) \leq 1, \mathrm{CF_B} = 1$, if $\boldsymbol{d}(\mathrm{B_{current}}, \mathrm{B_{neutral}}) = 7, \mathrm{CF_B} = 0.5$, and if $\boldsymbol{d}(\mathrm{B_{current}}, \mathrm{B_{neutral}}) \geq 14, \mathrm{CF_B} = 0$ will be assigned to the fiducial points of the right eye (Fig. 6).

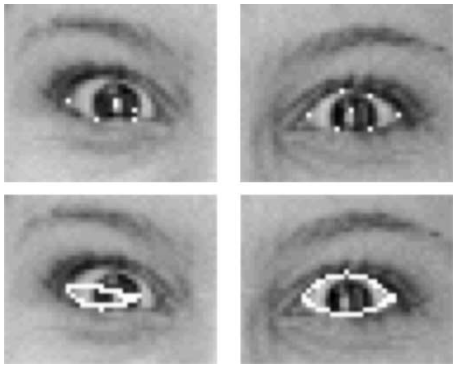


Fig. 6. Spatial sampling of the eye contour—measured error: $d(B_{\mathrm{current}}, B_{\mathrm{neutral}}) = 14$ (left) $d(B_{\mathrm{current}}, B_{\mathrm{neutral}}) = 1$ (right).

We use the inner corners of the eyes as the referential points for calculating CFs of the fiducial points of the eyes because of the stability of these points with respect to nonrigid facial movements: contractions of the facial muscles do not cause physical displacements of these points. Under the assumption that the images acquired during a single monitoring session with an observed subject are scale and orientation invariant (see the beginning of Section III), the location of the stable facial points such as the inner corners of the eyes should remain the same during the entire session. Hence, the certainty of spatial sampling of a facial feature obtained by a given detector can be estimated based upon the error made by the given detector while localizing the stable points belonging to the facial feature at issue; the larger the degree of the detection error, the lower the certainty of the data at issue.

The referential features used for calculating CFs of other fiducial points are the tip of the nose (point P4 of the profile contour, Fig. 3), the size of the eyebrow area, the inner corners of the nostrils, and the medial point of the mouth $\boldsymbol{M}$. Point $\boldsymbol{M}$ is calculated as the center of gravity of the distribution obtained from the mouth region-of-interest filtered to reveal colors in the vicinity of the pure red [18]. Independently of bilateral facial muscle actions that can affect the facial appearance of the mouth (e.g., mouth stretching, smile, etc.), the medial point of the mouth remains stable. However, this point does not remain stable if unilateral muscle actions occur, causing subtle facial changes in one mouth corner only. In such cases, the CF calculated based upon the referential point $\boldsymbol{M}$ will be decreased even though the relevant mouth contour could be spatially sampled with high precision.

The utilized intra-solution consistency check assumes indirectly that all fiducial points are accurately extracted from the reference expressionless-face image of the current subject. To ascertain this assumption, we inspect visually the fiducial points extracted automatically from a neutral expression image acquired at the beginning of each monitoring session with a subject and, if necessary, we mark them manually in the pertinent image. Other images acquired during a single monitoring session are processed in the entirely automatic manner described above.

Eventually, in order to select the best of sometimes redundantly available solutions (e.g., for the fiducial points belonging to the eyebrows), we perform an inter-solution check. We compare, namely, the CFs of the feature points extracted from the contours spatially sampled by different detectors of the same facial component. The feature points having the highest CF are used for further analysis of shown AUs.

In the case of the fiducial points of the mouth, we also perform an "inter-solution consistency check" as a part of the inter-solution check. It compares the outputs of the *vertical mouth classifier* and the *horizontal mouth classifier* (see Appendix A) with the properties of the mouth contour localized by a mouth detector. It proceeds as follows.

1) Let us designate the *curve fitting of the mouth* detector as $\mathbf{det}(1)$, the *mouth template matching* detector as $\mathbf{det}(2)$, the *vertical mouth classifier* as $\mathbf{Vdet}$, the *horizontal mouth classifier* as $\mathbf{Hdet}$, and the certainty factor assigned to the mouth features extracted from the mouth contour localized by $\mathbf{det}(\boldsymbol{i})$ as $\mathbf{CF{-}det}(\boldsymbol{i})$.
2) $(\forall \boldsymbol{i} \in \{1, 2\})$ try to fire the following rules:
   If result($\mathbf{Vdet}$) = "smile" AND ($\mathrm{y}(\mathrm{I}_{\mathbf{det}(\boldsymbol{i})}) < \mathrm{y}(\mathrm{I}_{\mathrm{neutral}})$ OR $\mathrm{y}(\mathrm{J}_{\mathbf{det}(\boldsymbol{i})}) < \mathrm{y}(\mathrm{J}_{\mathrm{neutral}})$) Then $\mathbf{CF{-}det}(\boldsymbol{i}) = ++10\%$,
   If result($\mathbf{Vdet}$) = "sad" AND ($\mathrm{y}(\mathrm{I}_{\mathbf{det}(\boldsymbol{i})}) > \mathrm{y}(\mathrm{I}_{\mathrm{neutral}})$ OR $\mathrm{y}(\mathrm{J}_{\mathbf{det}(\boldsymbol{i})}) > y(J_{\mathrm{neutral}})$) Then $\mathbf{CF{-}det}(\boldsymbol{i}) = ++10\%$,
   If result($\mathbf{Hdet}$) = "stretched" AND ($\mathrm{x}(\mathrm{I}_{\mathbf{det}(\boldsymbol{i})}) < \mathrm{x}(\mathrm{I}_{\mathrm{neutral}})$ OR $\mathrm{x}(\mathrm{J}_{\mathbf{det}(\boldsymbol{i})}) > \mathrm{x}(\mathrm{J}_{\mathrm{neutral}})$) Then $\mathbf{CF{-}det}(\boldsymbol{i}) = ++10\%$, and
   If result($\mathbf{Hdet}$) = "puckered" AND ($\mathrm{x}(\mathrm{I}_{\mathbf{det}(\boldsymbol{i})}) > \mathrm{x}(\mathrm{I}_{\mathrm{neutral}})$ OR $\mathrm{x}(\mathrm{J}_{\mathbf{det}(\boldsymbol{i})}) < \mathrm{x}(\mathrm{J}_{\mathrm{neutral}})$) Then $\mathbf{CF{-}det}(\boldsymbol{i}) = ++10\%$.
   The values of $\mathrm{y}(\mathrm{I}_{\mathbf{det}(\boldsymbol{i})})$ and $\mathrm{x}(\mathrm{I}_{\mathbf{det}(\boldsymbol{i})})$ correspond to the y-coordinate and x-coordinate of point I, respectively, which has been extracted from the mouth contour localized by the detector $\mathbf{det}(\boldsymbol{i})$. The values of $\mathrm{y}(\mathrm{I}_{\mathrm{neutral}})$ and $\mathrm{x}(\mathrm{I}_{\mathrm{neutral}})$ correspond to the y-coordinate and x-coordinate of point I, respectively, which has been localized in the neutral expression image of the currently observed subject. Although the $\mathbf{Vdet}$ and $\mathbf{Hdet}$ achieved a 100% average recognition rate when tested on a set of 100 full-face images, both methods use only some average properties of the image [18], which do not necessarily depict subtle differences between various mouth expressions. Therefore, when the detector $\mathbf{det}(\boldsymbol{i})$ passes this inter-solution consistency check successfully, we increase the associated certainty factor $\mathbf{CF{-}det}(\boldsymbol{i})$ for a mere 10%. This increase of 10% has been decided on based upon a number of visual observations conducted by several human observers which suggested that this increase would result neither in an overestimated nor in an underestimated certainty factor $\mathbf{CF{-}det}(\boldsymbol{i})$.
3) Select the mouth feature points having the highest CF for further analysis of shown AUs. If this process results in a draw, select the mouth feature points extracted from the mouth contour localized by the detector $\mathbf{det}(\boldsymbol{i})$ whose results have been selected more frequently during the current session with the observed subject.

Profile-contour fiducial points motion	
up/down (P) = $y(P_{neutral}) - y(P_{current})$ if ***up/down(P)*** > ε, ***P*** moves up	***in/out(P)*** = $x(P_{neutral}) - x(P_{current})$ if ***in/out(P)*** > ε, ***P*** moves inward
Profile-contour fiducial points absence	**Fiducial points state**
If there is no maximum of the profile contour function *f* between P5 and P7, ***absent(***P6***)***. Similarly for P7', P7", P8 and P9 (see Fig. 3 and Fig. 4).	***increase/decrease(PP')*** = $PP'_{neutral} - PP'_{current}$, where $PP' = \sqrt{\{(x_P - x_{P'})^2 + (y_P - y_{P'})^2\}}$ If ***increase/decrease(PP')*** < ε, distance ***PP'*** increases.
Shapes formed between certain profile contour fiducial points	
The physical meanings of ***angular(P6P8)*** = *true* and ***increased_curvature(P5P6)*** = *true* are shown below.	

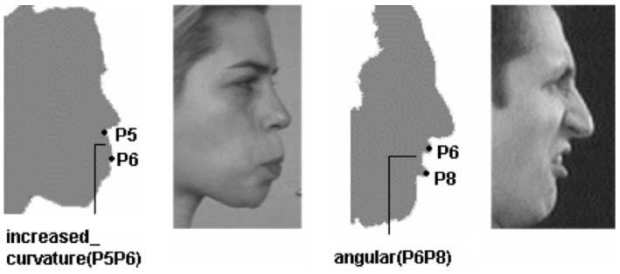


Fig. 7. Midlevel feature parameters for AU recognition: two describing the motion of the feature points (upper part of the table), two describing their state (middle part of the table), and two describing shapes formed between certain points (lower part of the table). The value of $y(P)$ and y_P corresponds to the y-coordinate of point P (similar for the x-coordinate); ε is 1 pixel.

IV. Facial Action Detection

Each AU of the FACS system is anatomically related to the contraction of one or more specific facial muscles [6]. Contractions of facial muscles alter the shape and location of the facial components. Some of these changes in facial expression are observable from the changes in the position of the feature points illustrated in Fig. 3. To classify detected changes in the position of the feature points in terms of AUs, the pertinent changes are represented first as a set of suitable midlevel feature parameters.

A. Parametric Feature Representation

Six midlevel feature parameters, defined in Fig. 7, describe the changes in the position of the fiducial points depicted in Fig. 3. They are calculated for various feature points, for each input image (see Tables II and III, Section IV-B), by comparing the currently extracted points with the relevant points extracted from the neutral expression image of the subject.

More specifically, two midlevel feature parameters describe the motion of the feature points: $\mathbf{up/down}(\boldsymbol{P})$ and $\mathbf{in/out}(\boldsymbol{P})$. These parameters are calculated only for profile contour fiducial points. The parameter $\mathbf{up/down}(\boldsymbol{P}) = y(P_{\mathrm{neutral}}) - y(P_{\mathrm{current}})$ describes the upward or downward movement of point P. If $y(P_{\mathrm{neutral}}) - y(P_{\mathrm{current}}) > \varepsilon$, point P moves up. If $y(P_{\mathrm{neutral}}) - y(P_{\mathrm{current}}) < \varepsilon$, point P moves down. P_{neutral} is point P localized in the neutral expression image of the currently observed subject. P_{current} is point P localized in the currently examined image of the observed subject. The values of $y(P)$ corresponds to the y-coordinate of point P

TABLE I
FACS RULES FOR RECOGNITION OF 32 AUS THAT OUR METHOD ENCODES AUTOMATICALLY IN DUAL-VIEW FACE IMAGES

AU1	Pulls the eyebrows' inner corners upward, causes the skin of the center forehead to wrinkle horizontally.
AU2	Pulls the eyebrows' outer corner(s) upward, causes the skin of the outer forehead to wrinkle horizontally.
AU4	Pulls the eyebrows closer together, produces a bulge between the eyebrows, lowers the eyebrows slightly.
AU5	Raises the upper eyelid(s), widens the eye opening.
AU6	Raises the cheek(s), pushes the skin surrounding the eye(s) towards the socket, narrows the eye opening.
AU7	Tightens the upper and lower eyelid(s), narrows the eye opening.
AU8	Pulls the lips towards each other, parts the lips.
AU9	Wrinkles the nose, lowers the brows, produces a bulge between the brows and the root of the nose, raises the upper lip.
AU10	Raises the upper lip, deepens the nasolabial furrow, does not wrinkle the nose.
AU12	Pulls the lip corners upward obliquely.
AU13	Pulls the lip corners sharply upward.
AU15	Pulls the corners of the lips downward, stretches the lips slightly, flattens the skin of the chin boss.
AU16	Pulls the lower lip downward laterally, causes the lower lip to protrude.
AU17	Pushes the chin boss and the lower lip upward and stretches the skin on the chin boss.
AU18	Pushes the mouth forward medially, de-elongates the mouth, causes the lips to protrude forwards (as by saying "fool").
AU19	Causes at least the tip of the tongue to be visible.
AU20	Pulls the lips backward laterally, flattens the skin of the lips and the chin boss.
AU23	Tightens the lips slightly, making the lips appear more narrow.
AU24	Presses the lips together, tightens and narrows the lips to a small extent.
AU25	Parts the lips, does not part the jaws.
AU26	Parts the lips, parts the jaws, does not stretch the mouth.
AU27	Stretches the mouth as lower jaw is pulled down.
AU28,t,b	AU28: lips sucked into the mouth. AU28t: Upper lip sucked into the mouth. AU28b: Bottom lip sucked into the mouth.
AU29	Pushes the jaw forward, causing the chin to stick out and the lower teeth to extend in front of upper teeth.
AU35	The cheeks are sucked into the mouth, producing an 8-like shape of the mouth and crevices in cheeks.
AU36t,b	AU36t: Pushes the tongue under the upper lip, causes a bulge above the upper lip. AU36b: Pushes the tongue under the lower lip, causes a bulge below the lower lip.
AU38	Flares out the nostril wings, widens the nostril openings.
AU39	Compresses the nostril wings, narrows the nostril openings.
AU41	Causes the upper eyelid(s) to drop down, appearing at "half-mast", reducing the eye opening.

and the value assigned to ε is 1 pixel. Midlevel feature parameter $\mathbf{in/out}(\boldsymbol{P}) = x(P_{\text{neutral}}) - x(P_{\text{current}})$ describes the inward or outward movement of point P. This parameter is calculated only for profile contour fiducial points. If $x(P_{\text{neutral}}) - x(P_{\text{current}}) < \varepsilon$, point P moves outward. If $x(P_{\text{neutral}}) - x(P_{\text{current}}) > \varepsilon$, point P moves inward. Two midlevel feature parameters describe the state of the feature points: $\mathbf{absent}(\boldsymbol{P})$ and $\mathbf{increase/decrease}(\boldsymbol{PP'})$. Midlevel feature parameter $\mathbf{absent}(\boldsymbol{P})$ denotes the absence of point P belonging to the profile contour function f. This parameter is calculated only for points P6, P7′, P7″, P8, and P9 (see Figs. 3 and 4). If there is no maximum of the profile contour function f between P5 and P7, then $\mathbf{absent}(\mathrm{P6})$. If there is no minimum of the profile contour function f between P8 and P10, then $\mathbf{absent}(\mathrm{P9})$. Similar rules are used for P7′, P7″, and P8. Feature parameter $\mathbf{increase/decrease}(\boldsymbol{PP'}) = PP'_{\text{neutral}} - PP'_{\text{current}}$ describes the increase or decrease of the distance between points P and P'. If $PP'_{\text{neutral}} - PP'_{\text{current}} < \varepsilon$, distance PP' increases. If $PP'_{\text{neutral}} - PP'_{\text{current}} > \varepsilon$, distance PP' decreases. The distance PP' between points P and P' is calculated as given in Fig. 7. Finally two midlevel feature parameters describe two specific shapes formed between certain feature points. Midlevel feature parameter $\mathbf{angular}(\mathbf{P6P8}) =$ true denotes the presence of an angular shape formed between the profile contour fiducial points P6 and P8 (see Fig. 7). This parameter is calculated only for points P6 and P8. Feature parameter $\mathbf{increased\_curvature}(\mathbf{P5P6}) =$ true denotes the presence of an increased curvature between the profile contour fiducial points P5 and P6 (see Fig. 7). This parameter is calculated only for points P5 and P6.

We assign a certainty factor $\mathrm{CF} \in [0, 1]$ to each calculated midlevel feature parameter. We do so based upon the CFs associated with the selected feature points (see Section III-C), whose state or motion is described by the pertinent midlevel feature parameter. For example: $\mathrm{CF}_{\text{increase/decrease(BD)}} = \mathbf{min}(\mathrm{CF_B}, \mathrm{CF_D})$, and $\mathrm{CF}_{\text{up/down(P6)}} = \mathrm{CF}_{\text{in/out(P6)}} = \mathrm{CF}_{\text{angular(P6P8)}} = \mathrm{CF}_{\text{increased\_curvature(P5P6)}} = \mathrm{CF_{P6}} (= \mathrm{CF_{P4}})$.

B. Action Unit Recognition

The last step of automatic facial action detection is to translate the extracted facial information (i.e., the calculated feature parameters) into an AU-coded description of the shown facial expression. To achieve this, we apply a rule-based method with the fast-direct chaining inference procedure to two separate sets of rules.

Motivated by the rules of the FACS system (Table I), each of the rules utilized for AU recognition is defined in terms of the predicates of the midlevel feature representation (Fig. 7) and each encodes a single AU in a unique way according to the relevant FACS rule. A set of 24 rules, say **set-1**, for encoding 24 AUs occurring alone or in combination in an input face-profile image is given in Table II. A set of 22 rules, say **set-2**, for encoding 22

TABLE II
RULES FOR RECOGNITION OF 24 AUS IN A FACE-PROFILE IMAGE

rule 1	IF *up/down*(P2) > ε THEN **AU1**
rule 2	IF *in/out*(P2) < ε AND *increase/decrease*(P2P3) ≤ ε THEN **AU4**
rule 3	IF NOT (AU9 OR AU12 OR AU15 OR AU17 OR AU18 OR AU20) AND *angular*(P6P8) = *true* THEN **AU8**
rule 4	IF *in/out*(P2) < ε AND *increase/decrease*(P2P3) > ε THEN **AU9**
rule 5	IF *increase/decrease*(P2P3) ≤ ε AND *increase/decrease*(P5P6) > ε AND *in/out*(P6) < ε THEN **AU10**
rule 6	IF *in/out*(P6) > ε AND *in/out*(P8) > ε AND *increase/decrease*(P5P6) ≥ ε AND *increase/decrease*(P6P8) < ε THEN **AU12**
rule 7	IF *in/out*(P6) > ε AND *in/out*(P8) > ε AND *increase/decrease*(P5P6) > ε AND *increase/decrease*(P6P8) ≥ ε THEN **AU13**
rule 8	IF *up/down*(P6) < ε AND *up/down*(P8) < ε AND *increased_curvature*(P5P6) = *false* THEN **AU15**
rule 9	IF *increase/decrease*(P8P10) > ε AND *up/down*(P8) < ε AND *in/out*(P8) < ε THEN **AU16**
rule 10	IF NOT (AU28 OR AU28t OR AU28b) AND *in/out*(P10) > ε THEN **AU17**
rule 11	IF *in/out*(P6) < ε AND *in/out*(P8) < ε AND *increase/decrease*(P5P6) ≤ ε AND *increase/decrease*(P8P10) ≤ ε AND *increase/decrease*(P6P8) < ε THEN **AU18**
rule 12	IF *absent*(P7') = *false* AND *absent*(P7") = *false* THEN **AU19**
rule 13	IF *in/out*(P6) > ε AND *in/out*(P8) > ε AND *increase/decrease*(P5P6) ≤ ε AND *increase/decrease*(P6P8) ≤ ε AND *increased_curvature*(P5P6) = *false* THEN **AU20**
rule 14	IF NOT (AU28 OR AU28t OR AU28b) AND *increase/decrease*(P6P8) > ε AND *increase/decrease*(P6P8) < *t1* THEN **AU23**
rule 15	IF NOT (AU28 OR AU28t OR AU28b) AND *increase/decrease*(P6P8) > ε AND *increase/decrease*(P6P8) ≥ *t1* THEN **AU24**
rule 16	IF *increase/decrease*(P6P8) < ε AND *increase/decrease*(P4P10) ≥ ε THEN **AU25**
rule 17	IF *increase/decrease*(P4P10) < ε AND *increase/decrease*(P4P10) ≤ *t2* THEN **AU26**
rule 18	IF *increase/decrease*(P4P10) < ε AND *increase/decrease*(P4P10) > *t2* THEN **AU27**
rules 19-21	IF *absent*(P6) AND *absent*(P8) THEN **AU28**. IF *absent*(P6) THEN **AU28t**. IF *absent*(P8) THEN **AU28b**.
rule 22	IF *in/out*(P10) < ε THEN **AU29**
rules 23-24	IF *increased_curvature*(P5P6) = *true* THEN **AU36t**. IF *absent*(P9) THEN **AU36b**.

AUs occurring alone or in combination in an input frontal-face image is given in Table III. If the input images are face-profile images, the method encodes 24 AUs occurring alone or in combination by utilizing the rules of **set-1**. If the input images are frontal-face images, the method encodes 22 AUs occurring alone or in combination by utilizing the rules of **set-2**. Yet, if the input images are dual-view face images, the method encodes 32 AUs occurring alone or in combination by utilizing the rules of both **set-1** and **set-2**.

The applied fast direct-chaining inference procedure takes advantage of both a relational representation of the knowledge and the depth-first search to find as many conclusions as possible within a single "pass" through the knowledge base [24]. 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 (**set-1** or **set-2**). Then it searches a linkage between the fired rule and the rule that it 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. As can be seen, the fast direct-chaining inference process is more efficient than both the forward chaining and the backward chaining because it tries to fire only the rules that may potentially contribute to the inference process.

To prevent firing of a rule more than once, we utilize a list of fired rules (LFR). Thus, if a rule has fired (i.e., the rule's premise $\boldsymbol{p}$ is true and $\mathrm{CF}_p \geq \mathrm{T}$), the rule number is added to the LFR. The value of the threshold T is set to $\mathrm{T} = 0.05$. We do so to enable potential encoding of all shown AUs, even if the reached conclusions might have low certainties (due to the propagation, and hence accumulation, of potentially low certainties of extracted facial data).

We assign an initial certainty factor $\mathrm{CF} = 0$ to each AU that can be scored. With each actually scored AU, we associate a factor $\mathrm{CF} \in [0, 1]$ denoting the certainty with which the pertinent AU has been scored. Its value equals the overall certainty factor CF_p of the premise $\boldsymbol{p}$ of the rule whose firing caused the AU in question to be scored.

The certainty factor $\mathrm{CF_p}$ of the premise $\boldsymbol{p}$ of a fired rule is calculated as follows:

1) if $\boldsymbol{p}$ contains a clause $\boldsymbol{c}$ of a kind "NOT AU$\boldsymbol{i}$," then $\mathrm{CF}_p = \mathrm{CF}_p$;
2) if $\boldsymbol{p}$ contains $\boldsymbol{c}1$ ***AND*** $\boldsymbol{c}2$, where $\boldsymbol{c}1$ and $\boldsymbol{c}2$ are the clauses of the premise $\boldsymbol{p}$, then $\mathrm{CF_p} = \mathbf{min}(\mathrm{CF}_{c1}, \mathrm{CF}_{c2})$;
3) if $\boldsymbol{p}$ contains $\boldsymbol{c}1$ ***OR*** $\boldsymbol{c}2$, where $\boldsymbol{c}1$ and $\boldsymbol{c}2$ are the clauses of the premise $\boldsymbol{p}$, then $\mathrm{CF_p} = \mathbf{max}(\mathrm{CF}_{c1}, \mathrm{CF}_{c2})$;
4) if $\boldsymbol{p}$ contains just clause $\boldsymbol{c}$, being of a different kind than "NOT AU$\boldsymbol{i}$," en $\mathrm{CF}_p = \mathrm{CF}_c$;
5) $(\forall \boldsymbol{c})\mathrm{CF}_c = \mathrm{CF_{fp}}$, where $\boldsymbol{fp}$ is the feature parameter to which clause $\boldsymbol{c}$ is related (see also Section IV-A).

In the case of dual-view input face images, some AUs could be scored twice (e.g., AU12, see Tables II and III). Hence, the last processing step of the utilized algorithm deals with those redundantly available scores. For each such pair of the redundantly inferred conclusions, it discards the one with which a lower CF has been associated.

V. EXPERIMENTAL EVALUATION

There are at least two crucial issues in evaluating the performance of an automated system. The first concerns the acquisition of a relevant test data set and the second is that of validation.

TABLE III
RULES FOR RECOGNITION OF 22 AUS IN A FRONTAL-VIEW FACE IMAGE

rule 1	IF *increase/decrease*(BD) < ε OR *increase/decrease*(B1D1) < ε THEN **AU1**
rule 2	IF *increase/decrease*(AE) < ε OR *increase/decrease*(A1E1) < ε THEN **AU2**
rule 3	IF *increase/decrease*(DD1) > ε THEN **AU4**
rule 4	IF *increase/decrease*(FG) < ε OR *increase/decrease*(F1G1) < ε THEN **AU5**
rule 5	IF AU12 OR AU13 THEN **AU6**
rule 6	IF NOT(AU12) AND ((FG > ε AND *increase/decrease*(GX) > ε) OR (F1G1 > ε AND *increase/decrease*(G1Y) > ε)) THEN **AU7**
rule 7	IF NOT(AU12 OR AU13 OR AU15 OR AU18 OR AU20 OR AU23 OR AU24 OR AU35) AND KL > ε AND *increase/decrease*(CK) < ε THEN **AU8**
rule 8	IF (*increase/decrease*(IB) > ε AND *increase/decrease*(CI) < ε) OR (*increase/decrease* (JB1) > ε AND *increase/decrease*(CJ) < ε) THEN **AU12**
rule 9	IF (*increase/decrease*(IB) > ε AND *increase/decrease*(CI) > ε) OR (*increase/decrease* (JB1) > ε AND *increase/decrease*(CJ) > ε) THEN **AU13**
rule 10	IF *increase/decrease*(IB) < ε OR *increase/decrease*(JB1) < ε THEN **AU15**
rule 11	IF NOT(AU28) AND IJ ≥ *t3* AND *increase/decrease*(IJ) > ε AND *increase/decrease*(KL) ≤ ε THEN **AU18**
rule 12	IF *increase/decrease*(IJ) < ε AND |*increase/decrease*(IB)| = ε AND |*increase/decrease*(JB1)| = ε THEN **AU20**
rule 13	IF KL > ε AND *increase/decrease*(KL) > ε AND *increase/decrease*(IJ) ≤ ε AND *increase/decrease*(JB1) ≥ ε AND *increase/decrease*(IB) ≥ ε THEN **AU23**
rule 14	IF NOT(AU12 OR AU13 OR AU15) AND KL > ε AND *increase/decrease*(KL) > ε AND IJ > *t3* AND *increase/decrease*(IJ) > ε THEN **AU24**
rule 15	IF *increase/decrease*(KL) < ε AND *increase/decrease*(CM) ≥ ε THEN **AU25**
rule 16	IF *increase/decrease*(CM) < ε AND CM ≤ *t4* THEN **AU26**
rule 17	IF CM > *t4* THEN **AU27**
rule 18	IF |KL| = ε THEN **AU28**
rule 19	IF IJ < *t3* THEN **AU35**
rule 20	IF NOT(AU8 OR AU12 OR AU13 OR AU18 OR AU24) AND *increase/decrease*(HH1) < ε THEN **AU38**
rule 21	IF NOT(AU8 OR AU15 OR AU18 OR AU24 OR AU28) AND *increase/decrease*(HH1) > ε THEN **AU39**
rule 22	IF NOT(AU7) AND ((FG > ε AND *increase/decrease*(FG) > ε AND *increase/decrease*(FX) > ε) OR (F1G1 > ε AND *increase/decrease*(F1G1) > ε AND *increase/decrease*(F1Y) > ε)) THEN **AU41**

A. Test Data Set

As already remarked by many researchers (e.g., [7], [16], [20]), no database of images exists that is shared by all diverse facial-expression-research communities. In general, only isolated pieces of such a facial database exist. An example is the unpublished database of Ekman-Hager AU Exemplars [8], which has been used by Bartlett *et al.* [1], Donato *et al.* [5], and Tian *et al.* [27] to train and test their methods for AU detection from face image sequences. Another example is the database of static full-face images, so called FACS Dictionary [10], which is also not publicly available. This, together with the inapplicability of the databases mentioned above for the purposes of testing our face-profile-based AU encoder, incited us to generate our own database of test images.

The following criteria have been defined for our database of static face images.

1) *Resolution*: The images should have standard PAL camera resolution, that is, when digitized, images should measure 720 × 576 pixels. The largest part of each image is either a portrait or a profile view of a face. In other words, there are at least 450 pixels across the width of the subject's face in the case of a portrait image and at least 300 pixels across the width of the face in the case of a profile-view face image. In turn, this implies that there are approximately 60 to 80 pixels across the width of the subject's eye (as assumed in (1)).
2) *Color*: The images should be true-color (24-bit) images.
3) *DB structure*: The images should belong to one of three database clusters:
   a) Portraits of faces (no in-plane or out-plane head rotations are present); include images scanned from photographs used as behavioral science research material (see Fig. 8, first row);
   b) Profile-view images of faces (no in-plane or out-plane head rotations are present; see Fig. 8, second row);
   c) Dual-view face images (i.e., combined portraits and profiles of faces; see Fig. 8, third row).
4) *Distribution*: the database is installed on our group's server and can be easily accessed by any group member.

Similarly to the method presented in this paper, most of the existing approaches to AU detection assume that the presence of the face in the input image is ensured [16]. However, in most real-life situations the location of the face in the scene is not known a priori. The presence of a face can be ensured either by using a method for automatic face detection in arbitrary scenes (see [31]) or by using a camera setting that ascertains the assumption at issue. The method proposed here does not perform face detection in an arbitrary scene; it operates on face images stored in our facial database, almost each of which has been acquired by a head-mounted CCD digital PAL camera device (Fig. 9). The pertinent device contains two cameras—the camera set in front of the face acquires portraits while the camera placed on the right side of the face acquires face-profile images. The utilized camera setting ascertains the

Fig. 8. Examples of facial database images. First row: portraits of faces. Second row: profile-view face images. Third row: dual-view face images.

Fig. 9. Head-mounted two-cameras device.

assumption that the images remain orientation and scale invariant during a monitoring session with a subject and that the face-profile-images are in right profile view (e.g., Fig. 8).

The database images represent a number of demographic variables including ethnic background, gender, and age, and provide, in principle, a basis for generality of research findings. Overall, the subjects were students and college personnel (in total 25 different persons) of both sexes, young but still ranging in age from 20 to 45, and of either European, African, Asian, or South American ethnic background. In order to avoid effects of the unique properties of particular people, each DB partition has been supplied with images of several individuals (e.g., the dual-views DB partition contains images of eight different subjects). The subjects were asked to display expressions that included single AUs and combinations of those. They were instructed by an expert (a certified FACS coder) on how to perform the required facial expressions. A total of 330 portraits (excluding some 60 images scanned from the photographs used as behavioral science research material), 240 profile-view images, and 560 dual-view images of subjects' faces were recorded during sessions, which began with displaying a neutral expression.

B. Validation Studies

Validation studies on the AU detection method proposed here address the question of whether the conclusions reached by our method are acceptable to human observers judging the same face images. The presented validation of the rule base and the overall method is based upon the dual-view face images contained in our database.

First, two experts (i.e., certified FACS coders) were asked to evaluate the available 560 dual-view face images in terms of displayed AUs. Inter-observer agreement as to the depicted AUs in the images was found for a total of 454 images. The pertinent observers' judgments of these 454 test images were further compared to those generated by our method. Overall results of this comparison are given in Table IV in the following terms.

TABLE IV
OUR METHOD'S PERFORMANCE IN AU CODING OF 454 TEST DUAL-VIEW STATIC FACE IMAGES MEASURED FOR AUS PER FACIAL FEATURE, FOR UPPER- AND LOWER-FACE AUS, AND OVERALL

	upper-face AUs		lower-face AUs		
	eyebrows	eyes	nose	mouth	chin
Correct	433	437	443	423	436
Partially correct	21	17	10	28	17
Incorrect	0	0	1	3	1
Recognition rate	**95.4%**	**96.3%**	**97.6%**	**93.2%**	**96.0%**
Correct	422		413		
Partially correct	32		37		
Incorrect	0		4		
Recognition rate	**93.0%**		**91.0%**		
Correct	392				
Partially correct	58				
Incorrect	4				
Recognition rate	**86.3%**				

1) *Correct* denotes that the AU codes generated by our method were completely identical to the AU codes scored by human observers judging the same images.
2) *Partially correct* denotes that AU-coded description obtained by the method is similar but not identical to the one given by human observers when interpreting the same image (e.g., some AU codes may be missing or may be recognized in addition to those recognized by human observers).
3) *Incorrect* denotes that none of the AU codes discerned by human observers were recognized by the method.
4) *Recognition rate* has been calculated as the ratio between the number of correctly recognized test images and the total number of test images. If more than one AU of a particular feature was misidentified in a test image, the pertinent image was counted once for the given feature. If several AUs of different features were misidentified in a test image, that image was counted for each of the pertinent

Fig. 10. Facial expression of AU7 + AU12 activation.

features. To calculate the percentage of agreement (i.e., the recognition rates), human FACS coders typically use the ratio between the number of correctly recognized AUs and the total number of AUs shown in the stimulus image. However, it is more appropriate to calculate the recognition rates based on the number of test images when one evaluates the performance of an automated system. This is because the system may score additional AUs besides those scored by human observers; such errors would not be taken into account if the recognition rates were measured based upon the number of correctly scored AUs and the total number of AUs shown in an image.

As can be seen from Table IV, in 86% of 454 test cases our method coded the analyzed facial expression using the same set of AU codes as the human observers. If we consider only the images in which the AUs were encoded with higher certainty factors (say CF $>$ 0.3; there are in total 423 such images), agreement between the system and the human observers was even 91%.

As far as misidentifications produced by our method are concerned, most of them arose from confusion between similar AUs (AU1 and AU2, AU6 and AU7, AU18, and AU35) and from subtle activations that remained unnoticed by human observers (e.g., AU26, AU38, AU39). The reason for the confusion between AU1 and AU2 (i.e., recognizing AU1 in addition to AU2) is that activation of AU2, which raises the outer portion of the eyebrow(s), tends to pull the inner eyebrow (AU1) as well. Although human observers also confuse AU6 and AU7 often [6], [27], in the case of our method, the reason for the confusion between AU6 and AU7 are the utilized rules for recognition of these AUs. Namely, if AU12 is present, AU6 will be scored (Table III) although this does not necessarily match the actually shown expression (Fig. 10). The confusion between AU18 and AU35 is also caused due to the utilized rules for encoding these AUs. Since inward pull of the cheeks is not detected by the system, only the width of the mouth distinguishes AU18 from AU35, causing misidentification of a weak AU35 (Table III). The reason for most of the mistaken identifications of AU26, AU38, and AU39 are subtle activations of these AUs, which remained unnoticed by the human observers. Actually, in most of such cases, our method coded the input images correctly, unlike the human observers. Yet such cases were addressed as misidentification.

Thus, comparing an automated system's performance to that of human judges is not enough. Human observers sometimes disagree in their judgments of AUs pictured in an analyzed image (e.g., that is why we reduced the initial set of 560 images to the test set of 454 images). They occasionally make mistakes and if the tested system does not produce the same mistakes, its performance measure is reduced. To estimate the performance of an automated system precisely, it is necessary to compare it to a validated standard. A better, readily accessible, standard set of face images objectively encoded in terms of displayed AUs is, therefore, necessary. Yet no effort in establishing such a benchmark database of test images has yet been reported (see also Section V.1).

VI. Conclusion

In this paper, we proposed a novel, automated method for detecting facial actions based upon changes in contours of facial components and/or face profile contour detected in a static frontal-view and/or profile-view face image.

The significance of this contribution is in the following.

1) The presented approach to automatic AU recognition extends the state of the art in automatic facial gesture analysis in several directions, including the kind of face images (static), the facial view (frontal, profile, and dual view), the number of AUs (32 in total), the difference in AUs, and the data certainty propagation handled. Namely, the previously reported automated AU detectors do not deal with static images, cannot handle more than one facial view at a time, do not assign certainty measures to the inferred conclusions (let alone varying them in accordance with the certainty of the input data), and, at best, can detect 16 to 20 AUs.
2) This paper provides a basic understanding of how to achieve automatic AU coding in both frontal-face and face-profile static images. It exemplifies how such knowledge can be used for devising procedures of greater flexibility and improved quality (e.g., inaccurate/partial data from one facial view can be substituted by data from the other view). This can form the basis of further research on AU analysis from multiple facial views.

Based upon the validation study explained in Section V-B, it can be concluded that the proposed method's performance in AU recognition from dual-view static images of faces exemplifies an acceptable level of expertise. The achieved results are similar to those reported for other automated FACS coders. The method achieves an average recognition rate of 86.3% for encoding 32 AU codes and their combinations in 454 test samples, while other automated FACS coders have (in the best case and for face image sequences) an average recognition rate of 88% for encoding 16 AU codes and their combinations in 113 test samples [27].

Though it is quite acceptable, the performance of the presented method can be improved in several respects.

1) The proposed algorithm cannot handle distractions like occlusions (e.g., by a hand), glasses, and facial hair. Hence, its analysis is limited to nonoccluded faces without a beard, moustache, and glasses.
2) It cannot deal with rigid head movements; the analyzed images have to be scale and orientation invariant with respect to the image of the expressionless face of the currently observed subject, as if they were acquired by a head-mounted camera device (such as the one illustrated in Fig. 9).
3) Head-mounted camera devices usually reduce the freedom with which the subject can move around and they are commonly perceived as being uncomfortable or even cumbersome.
4) The proposed method cannot encode the full range of facial behavior (i.e., all 44 AUs defined in FACS); it performs facial action coding in static frontal-view face images in terms of 22 AU codes, in profile-view face images in terms of 24 AU codes, and in dual-view face images in terms of 32 AU codes (Tables II and III).

Further efforts will be required if these limitations are to be addressed. In addition, it will be interesting to test the proposed method with a substantially large database.

APPENDIX A
FRONTAL FACE FEATURE DETECTORS

This appendix provides a short overview of the detectors we are using to spatially sample the contours of the eyebrows, eyes, nostrils, and mouth from an input frontal-view face image.

We apply a simple analysis of image histograms in a combination with various filter transformations to locate six regions of interest (ROIs) in the face region segmented from an input frontal-view face image: two eyebrows, two eyes, nose, and mouth. The details of this procedure can be found in [15], [18]. Then, to spatially sample the contour of a certain facial component, we apply one or more facial-feature detectors to the pertinent ROI.

A. Eyebrows

Two different detectors localize the contours of the eyebrows in the eyebrow ROIs. One applies the contour-following algorithm based on four-connected chain codes and the other fits a 2-D model of the eyebrow consisting of two second degree parabolas. The details of these algorithms are reported in [21].

B. Eyes

The contours of the eyes are localized in the eye ROIs by a single detector representing an adapted version of the method for hierarchical-perceptron feature localization [28]. The detector employs a set of $81 \times 4 \times 1$ back-propagation neural networks with a Sigmoid transfer function to locate the iris of the eye and the eye microfeatures illustrated in Fig. 5. The border between the eyelids and the eye on which the microfeatures lie is then approximated by two third-degree polynomials. The details of this algorithm can be found in [15].

C. Nostrils and Chin

The contours of the nostrils are localized in the nose ROI by applying a method that fits two 2-D small circular models onto the two small regions delimited as the nostril regions by a seed-fill algorithm. The seed-fill algorithm is also used to color eyes and mouth regions in the face region. An adapted version of the Vornoi-diagrams-based algorithm delimits the symmetry line between them. The tip of the chin is localized as the first peak after the third deepest valley (the mouth) of the brightness distribution along the symmetry line [15].

D. Mouth

We utilize two detectors to spatially sample the contour of the mouth in the mouth ROI. *Curve fitting of the mouth* applies a simple boundary-following algorithm to achieve a coarse estimation of the mouth contour and a second-order least-square model algorithm to fit four second degree parabolas on the coarse mouth contour. *Mouth template matching* detector localizes the contour of the mouth in the mouth ROI by fitting a 2-D model of the lips to the mouth (Fig. 6). We also employ two detectors to classify the horizontal and the vertical movements of the mouth. The *vertical mouth classifier* utilizes a set of back-propagation neural networks to classify the mouth movements into one of the categories "'smile," "neutral" and "sad." The *horizontal mouth classifier* employs rule-based reasoning to classify the mouth movements into one of the categories "stretched," "neutral," and "puckered." Details of these algorithms are reported in [18].

E. Computational Costs

For a standard frontal-view image (i.e., ± 500 pixels across the width of the face), the multidetector processing takes approximately 8 s on a Pentium 2, 0 GHz. This represents more than 75% of the total time spent on AU detection in a dual-view face image. Yet, none of the utilized detectors runs a computationally expensive algorithm. Hence, we feel that a careful reimplementation, aimed at obtaining optimal code of the employed detectors will greatly improve the performance.

ACKNOWLEDGMENT

The authors would like to thank I. Patras and J. Wojdel, of Delft University of Technology, as well as the anonymous reviewers for their helpful comments and suggestions.

REFERENCES

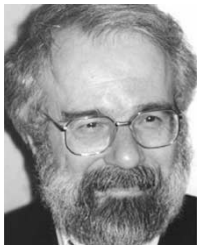
[1] M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, "Measuring facial expressions by computer image analysis," *Psychophysiolgy*, vol. 36, pp. 253–263, 1999.
[2] M. Black and Y. Yacoob, "Recognizing facial expressions in image sequences using local parameterized models of image motion," *Comput. Vis.*, vol. 25, no. 1, pp. 23–48, 1997.
[3] 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," *Psychophysiolgy*, vol. 36, pp. 35–43, 1999.
[4] C. Darwin, *The Expression of the Emotions in Man and Animals*. Chicago, IL: Univ. of Chicago Press, 1872, 1965.
[5] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, "Classifying facial actions," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 21, pp. 974–989, Oct. 1999.

[6] P. Ekman and W. Friesen, *Facial Action Coding System*. Palo Alto, CA: Consulting Psychol. Press, 1978.
[7] P. Ekman, T. S. Huang, T. J. Sejnowski, and J. C. Hager, Eds., "Final Report to NSF of the Planning Workshop on Facial Expression Understanding," Human Interaction Lab., Univ. California, San Francisco, 1993.
[8] P. Ekman, J. Hager, C. H. Methvin, and W. Irwin, "Ekman-Hager Facial Action Exemplars," Human Interaction Lab., Univ. California , San Francisco.
[9] I. Essa and A. Pentland, "Coding, analysis, interpretation and recognition of facial expressions," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, pp. 757–763, July 1997.
[10] W. V. Friesen and P. Ekman, "Dictionary—Interpretation of FACS Scoring," Human Interaction Laboratory, Univ. California, San Francisco.
[11] 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. Automatic Face Gesture Recognition*, 2002, pp. 272–278.
[12] *Handbook of Emotions*, M. Lewis and J. M. Haviland-Jones, Eds., Guilford Press, New York, 2000, pp. 236–249.
[13] K. Mase, "Recognition of facial expression from optical flow," *IEICE Trans.*, vol. E-74, no. 10, pp. 3474–3483, 1991.
[14] A. Ortony and T. J. Turner, "What is basic about basic emotions?," *Psychol. Rev.*, vol. 74, pp. 315–341, 1990.
[15] 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.
[16] ——, "Automatic analysis of facial expressions: The state of the art," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, pp. 1424–1445, Dec. 2000.
[17] ——, "Toward an affect-sensitive multimodal human-computer interaction," in *Proc. IEEE*, vol. 91, Sept. 2003, pp. 1370–1390.
[18] M. Pantic, M. Tomc, and L. J. M. Rothkrantz, "A hybrid approach to mouth features detection," *Proc. IEEE Int. Conf. Systems, Man, Cybernetics*, pp. 1188–1193, 2001.
[19] 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.
[20] A. Pentland, "Looking at people," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, pp. 107–119, Jan. 2000.
[21] B. Raducanu, M. Pantic, L. J. M. Rothkrantz, and M. Grana, "Automatic eyebrow tracking using boundary Chain code," in *Proc. Advanced School Computing Imaging Conf.*, 1999, pp. 137–143.
[22] J. Russell and J. Fernandez-Dols, *The Psychology of Facial Expression*. New York: Cambridge Univ. Press, 1997.
[23] *Handbook Methods in Non-Verbal Behavior Research*, K. R. Scherer and P. Ekman, Eds., Cambridge Univ. Press, Cambridge, MA, 1982.
[24] M. Schneider, A. Kandel, G. Langholz, and G. Chew, *Fuzzy Expert System Tools*. New York: Wiley, 1997.
[25] L. Shafarekno, M. Petrou, and J. Kittler, "Automatic watershed segmentation of randomly textured color images," *IEEE Trans. Image Processing*, vol. 6, pp. 1530–1544, Nov. 1997.
[26] K. Sobottka and I. Pitas, "A novel method for automatic face segmentation, facial feature extraction and tracking," *Signal Process. Image Commun.*, vol. 12, no. 3, pp. 263–281, 1998.
[27] Y. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, pp. 97–115, Jan. 2001.
[28] J. M. Vincent, D. J. Myers, and R. A. Hutchinson, "Image feature location in multi-resolution images using multi-layer perceptrons," in *Neural Networks for Vision, Speech & Natural Language*, R. Lingard, D. J. Myers, and C. Nightingale, Eds. London, U.K.: Chapman & Hall, 1992, pp. 13–29.
[29] L. Vincent and P. Soille, "Watersheds in digital spaces: An efficient algorithm based on immersion simulations," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 33, pp. 583–589, June 1991.
[30] J. Yang and A. Waibel, "A real-time face tracker," in *Proc. Workshop on Applications of Computer Vision*, 1996, pp. 142–147.
[31] M. H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, pp. 34–58, Jan. 2002.

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

She joined the Data and Knowledge Systems Group of the Mediamatics Department, Delft University of Technology as an Assistant Professor, in 2001. Her research interests pertain to the application of AI and computational intelligence techniques in the analysis of different aspects of human behavior for the realization of perceptual, context-sensitive, multimodal human-computer interfaces.

Dr. Pantic is a member of the ACM and the American Association of Artificial Intelligence.



Leon J. M. Rothkrantz received the M.Sc. degree in mathematics from the University of Utrecht, Utrecht, The Netherlands, the Ph.D. degree in mathematics from the University of Amsterdam, Amsterdam, the Netherlands, and the M.Sc. degree in psychology from the University of Leiden, Leiden, the Netherlands, in 1971, 1980, and 1990, respectively.

He joined the Data and Knowledge Systems group of the Mediamatics Department, Delft University of Technology, Delft, The Netherlands as an Associate Professor, in 1992. His long-range research goal is the design and development of natural, context-aware, multimodal man-machine interfaces. His current research focuses on a wide range of the paper's related issues including lip-reading, speech recognition and synthesis, facial expression analysis and synthesis, multimodal information fusion, natural dialogue management, and human affective feedback recognition.