2005 IEEE International Conference on Systems, Man and Cybernetics
Waikoloa, Hawaii October 10-12, 2005

Detecting Facial Actions and their Temporal Segments in Nearly Frontal-View Face Image Sequences

Maja Pantic and Ioannis Patras
Delft University of Technology
Electrical Engineering, Mathematics and Computer Science
Delft, The Netherlands
{M.Pantic,I.Patras}@ewi.tudelft.nl

Abstract – *The recognition of facial expressions in image sequences is a difficult problem with many applications in human-machine interaction. Facial expression analyzers achieve good recognition rates, but virtually all of them deal only with prototypic facial expressions of emotions and cannot handle temporal dynamics of facial displays. The method presented here attempts to handle a large range of human facial behavior by recognizing facial action units (AUs) and their temporal segments (i.e., onset, apex, offset) that produce expressions. We exploit particle filtering to track 20 facial points in an input face video and we introduce AU-dynamics recognition using temporal rules. When tested on Cohn-Kanade and MMI facial expression databases, the proposed method achieved a recognition rate of 90% when detecting 27 AUs occurring alone or in a combination in an input face image sequence.*

Keywords: particle filtering, tracking, facial expression analysis, temporal rules.

1 Introduction

The human face is used to regulate the conversation by gazing or nodding, to interpret what has been said by means of lip reading, and to communicate and understand somebody's affective state and intentions on the basis of the shown facial expression [1, 2]. Machine understanding of facial expressions could revolutionize human-machine interaction technologies and fields as diverse as security, behavioral science, medicine, and education [1]. Consequently, computer-based recognition of facial expressions has become an active research area.

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 (for an exhaustive survey of the past work on this research topic, see [3]). This practice may follow from the work of Darwin and more recently Ekman [2], who suggested that basic emotions have related prototypic expressions. In everyday life, however, such prototypic expressions occur relatively rarely; emotions are shown more often by subtle changes in facial display, such as flashing the eyebrows in surprise. Instead of classifying facial expressions into few basic emotion categories, this work attempts to measure a large range of facial behavior by recognizing action units (AUs, i.e., atomic facial signals) that produce expressions. As described in [4], all visually distinguishable facial activity can be described on the basis of 44 AUs. Hence, if a computer system would be able to detect these 44 AUs automatically, it will be able to identify each and every facial expression that the human face can possibly display. Few approaches have been reported for automatic recognition of AUs in face images [1]. 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., [5], [6]). Only recently there has been an emergence of efforts in automating AU coding of face images. Tian et al. used lip tracking, template matching and neural networks to identify 16 AUs occurring alone or in combination in frontal-view face image sequences [7]. They reported an 87.9% average recognition rate attained by their method for videos of the Cohn-Kanade Facial Expression database [8]. Bartlett et al. reported on accurate automatic recognition of 18AUs (95% average recognition rate) from full-face videos using Gabor filters and Support Vector Machines [9]. Valstar et al. used temporal templates (i.e., motion history images) and a combined k-Nearest-Neighbor and rule-based classifier to recognize 15 AUs from full-face image sequences with an average recognition rate of 65% [10].

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 instance, Schmidt and Cohn [11] 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 at the same time as AU12 or follow AU12 within 1 second. 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 [4], it is obvious that automated tools for the detection of AUs and their temporal dynamics would be highly beneficial. Nevertheless only a single effort in automating the detection of temporal segments of AUs in face image sequences has been reported so far. The work in question was aimed at automatic recognition of 23 AUs and their temporal segments in an input face-profile video [12]. In contrast to this previous work, we present here an

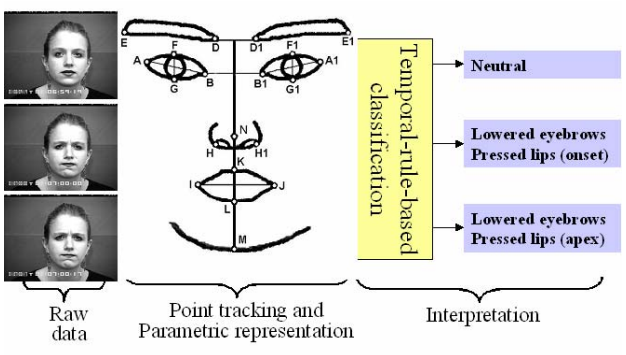

Fig. 1. Outline of the proposed method for recognition of AUs and their temporal segments from full-face video

automated system for recognition of 27 AUs and their temporal segments in input frontal-view face videos.

Fig. 1 outlines our method. After 20 fiducial points are initialized in the first frame of the input face image sequence, we exploit particle filtering to track these points automatically for the rest of the sequence. Based upon the changes in the position of the fiducial points, we measure changes in facial expression. Changes in the position of the fiducial points are transformed first into a set of mid-level parameters for AU recognition. Based upon the temporal consistency of mid-level parameters, a rule-based method encodes temporal segments (onset, apex, offset) of 27 AUs occurring alone or in a combination in the input face videos. Fiducial-point tracking, parametric representation and AU coding are explained in sections 2, 3 and 4. The evaluation of the proposed method using two benchmark databases, the Cohn-Kanade [8] and the MMI facial expression database [13], is given in section 5.

2 Facial Point Tracking

Facial muscle activity produces changes in the appearance of the facial components (eyes, nose, lips, etc.); their shape and location can alter immensely with facial expressions (e.g., pursed lips vs. jaw dropped). To reason about the shown facial expression and about the facial muscle actions that produced it, we track a set of 20 facial fiducial points (Fig. 1), the location of which alters during the facial expressions. At the first frame of the sequence, a number of windows that are interactively positioned around each of the facial points, define a number of color templates. Let us denote such a color template with $\boldsymbol{o} = \{\boldsymbol{o_i}\}$ where $\boldsymbol{i}$ is the pixel subscript. We subsequently track each color template for the rest of the image sequence with the auxiliary particle filter that was introduced by Pitt and Shepard [14]. Particle filtering has become the dominant tracking paradigm due to its ability to deal successfully with noise, occlusion and clutter. In order to adapt it for the problem of color-based template tracking, we define an observation model that is based on a robust color-based distance between the color template $o = \{o_i \mid i = 1\ldots M\}$ and a color template $c = \{c_i \mid i = 1\ldots M\}$ at the current frame. We attempt to deal with shadows by compensating for the global intensity changes. We use the distance function d, see *(1),* where M is the number of pixels in each template, m_c (and m_o) is the average intensity of template $c = \{c_i\}$ (and, respectively, of template $o = \{o_i\}$), i is the pixel index, $\|\cdot\|_1$ is the L_1 norm and ρ is the absolute value.

$$d = \sum_{i=1}^{M} \rho\left(\left\| \frac{c_i}{m_c} - \frac{o_i}{m_o} \right\|_1 \mu_c \right) / M \rightarrow (1)$$

We proceed under 2 assumptions (as defined for the face image sequences of both the Cohn-Kanade [8] and the MMI facial expression database [13]): (1) the input image sequence is non-occluded nearly frontal-view of the face, and (2) the first frame shows a neutral expression and no head rotations. To handle possible (small) head rotations and variations in scale of the observed face, we register each frame of the input image sequence with the first frame based on three referential points (Fig. 1): the tip of the nose (N) and the inner corners of the eyes (B and B1). We use these points as the referential points because of their stability with respect to non-rigid facial movements: facial muscle actions do not cause physical displacements of these points. Each frame is registered with the first frame by applying an affine transformation. Except of N, B and B1, which are tracked in unregistered input video sequences, other facial fiducial points are tracked in the registered input image sequence. Typical results are shown in Fig. 2.

3 Parameters for AU Recognition

Facial muscle actions 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 tracked points. To classify the tracked changes in

Fig. 2. Results of the facial point tracking

Table 1: Mid-level parameters for AU recognition

	Parameters		Parameters
AU1	up/down(D) > ε OR up/down(D1) > ε	AU2	up/down(E) > ε OR up/down(E1) > ε
AU4	inc/dec(DD1) > ε	AU5	up/down(F) > ε OR up/down(F1) > ε
AU6	(T1 > inc/dec(FG) > ε AND up/down(G) > ε AND up/down(I) > ε) OR (T1 > inc/dec(F1G1) > ε AND up/down(G1) > ε AND up/down(J) > ε)	AU7	(T1 > inc/dec(FG) > ε AND up/down(G) > ε AND |up/down(I)| ≤ ε) OR (T1 > inc/dec(F1G1) > ε AND up/down(G1) > ε AND |up/down(J)| ≤ ε)
AU10	inc/dec(KN) > ε	AU15	up/down(I) < -ε OR up/down(J) < -ε
AU12	(up/down(I) > ε AND inc/dec(NI) < -ε) OR (up/down(J) > ε AND inc/dec(NJ) < -ε)	AU13	(up/down(I) > ε AND inc/dec(NI) ≥ ε) OR (up/down(J) > ε AND inc/dec(NJ) ≥ ε)
AU16	inc/dec(LM) > ε	AU18	T2 > inc/dec(IJ) > ε AND inc/dec(KL) ≤ ε
AU20	inc/dec(IJ) < -ε AND |up/down(I)| ≤ ε AND |up/down(J)| ≤ ε	AU23	inc/dec(KL) > ε AND inc/dec(IJ) ≤ ε
AU24	inc/dec(KL) > ε AND T2 > inc/dec(IJ) > ε	AU25	inc/dec(KL) < -ε AND |inc/dec(NM)| ≤ ε
AU26	T3 < inc/dec(NM) < -ε	AU27	inc/dec(NM) < T3
AU28	|KL| ≤ ε	AU35	inc/dec(IJ) > T2
AU38	inc/dec(HH1) < -ε	AU39	inc/dec(HH1) > ε
AU41	(inc/dec(FG) > T1 AND |up/down(G)| ≤ ε) OR (inc/dec(F1G1)> T1 AND |up/down(G1)| ≤ ε)	AU44	(inc/dec(FG) > T1 AND up/down(G) > ε) OR (inc/dec(F1G1) > T1 AND up/down(G1) > ε)
AU43	|FG| ≤ ε AND |F1G1| ≤ ε	AU45	|FG| ≤ ε AND |F1G1| ≤ ε
AU46	|FG| ≤ ε OR |F1G1| ≤ ε		

terms of AUs and their temporal segments, these changes are transformed first into a set of mid-level parameters.

We defined two mid-level parameters: ***up/down(P)*** and ***inc/dec(PP')***. Parameter ***up/down(P)*** = $y(P_{t1}) - y(P_t)$ describes upward and downward movements of point P. If $y(P_{t1}) - y(P_t) > \varepsilon$, point P moves up. Otherwise, it moves down. P_{t1} is point P localized in the first frame of the input image sequence. P_t is point P tracked in frame t. The value of $y(P)$ is the y-coordinate of P. Parameter ***inc/dec(PP')*** = $PP'_{t1} - PP'_t$ describes the increase or decrease of the distance between points P and P'. If $PP'_{t1} - PP'_t < \varepsilon$, distance PP' increases. Otherwise, it decreases. Distance PP' is calculated as the Euclidian distance between points P and P'. The two parameters are calculated for various points (see Table 1), for each input frame.

4 Temporal Rules for AU Recognition

We transform the calculated mid-level parameters into a set of AUs describing the facial expression captured in the input video. We use a set of temporal rules to code 27 AUs occurring alone or in combination in an input face video. To minimize the effects of noise and inaccuracies in facial point tracking and to enable the recognition of the temporal dynamics of displayed AUs, we consider the temporal consistency of the mid-level parameters.

We divide activation of each AU into three temporal segments: the onset (beginning), apex, and offset (ending). Each temporal rule utilized for AU recognition is further defined in terms of the mid-level parameters (for the full list of mid-level parameters used to discriminate 27 different AUs, see Table 1) and each encodes a specific temporal segment of a single AU in a unique way. For instance, to recognize the temporal segments of AU4, which pulls the eyebrows closer together, we exploit the following temporal rules (ε is ± 2 pixels, i.e., 3% $DD1_{t1}$):

IF ($[inc/dec(DD1)]_t > [inc/dec(DD1)]_{t-1} + \varepsilon$)
AND $inc/dec(DD1) > \varepsilon$ THEN ***AU4-onset***

IF $|\, [inc/dec(DD1)]_t - [inc/dec(DD1)]_{t-1} \,| \le \varepsilon$
AND $inc/dec(DD1) > \varepsilon$ THEN ***AU4-apex***

IF ($[inc/dec(DD1)]_t < [inc/dec(DD1)]_{t-1} - \varepsilon$)
AND $inc/dec(DD1) > \varepsilon$ THEN ***AU4-offset***

Fig. 3 illustrates the meaning of these rules for the video from the Cohn-Kanade database shown in the 2nd row of Fig 2. The horizontal axis represents the time dimension (i.e., the frame number) and the vertical axis represents the value that the distance DD1 takes. As suggested by Fig. 3, distance DD1 should decrease and its value should be less than its neutral-expression value to label a frame as an "AU4 onset". The decrease of the value that the distance DD1 takes should terminate, resulting in a (relatively) stable temporal value of parameter $inc/dec(DD1)$, for a frame to be labeled as "AU4 apex". Eventually, distance DD1 should increase toward its neutral-expression value to label a frame as an "AU4 offset". Since each and every video of the Cohn-Kanade database depicts only the onset and the apex of the recorded expression, the offset of AU4 could not be depicted in Fig. 3.

Generally, for each and every AU, it must be possible to detect a temporal segment (an onset, apex, or offset) continuously over at least 5 consecutive frames for the facial action in question to be scored. Incited by the research findings that suggested that temporal changes in neuromuscular facial activity last from ¼ of a second (e.g., a blink) to several minutes (e.g., a jaw clench) [4], the temporal duration has been determined empirically based

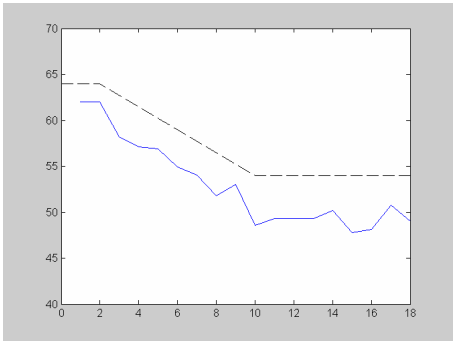

Fig. 3. Onset and apex of AU4 indicated by the ground truth (dashed line) and those detected by the method (full line) based upon the value that distance DD1 takes in the case of the Cohn-Kanade database sample shown in the 2nd row of Fig. 2.

Table 2: Rules for resolving temporal conflicts. The rules are used for both frames and temporal segments that are unlabeled or labeled incorrectly. UL ↔ "unlabeled".

	Previous frame	*Current frame (old label)*	*Next frame*	*Current (new label)*
R1	Onset	UL / Apex / Offset	Onset	Onset
R2	Onset	UL	Apex	Apex
R3	Apex	UL / Onset / Offset	Apex	Apex
R4	Apex	UL	Offset	Apex
R5	Offset	UL / Onset / Apex	Offset	Offset

on a video frame rate of 24 frames/second (i.e., 5 frames have a duration of less than ¼ of a second). However, since the samples of the Cohn-Kanade database have been acquired at an extremely low frame rate (12 or even less frames/second), the temporal duration was set to 3 frames for the samples from the Cohn-Kanade database.

Both inaccuracies in facial fiducial point tracking and occurrences of non-prototypic facial activity may result in temporal segments that are unlabeled (i.e., neither the onset, nor the apex, nor the offset) or in frames and temporal segments that are labeled incorrectly. The latter may arise, for instance, when an apex frame or an apex temporal segment of an AU is detected either between two onset segments or between two offset segments of that AU. 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 next frame / segment and re-label the current frame / segment according to the ruled-based system summarized in Table 2. For instance, any unlabeled temporal segment and/or any apex segment and/or any offset segment of an AU that has been detected between two onset segments of that AU are re-labeled as "onset". Finally, an AU should be recognized, in general, only when the full temporal model of that AU is observed (onset → apex→ offset). Yet, in order to deal with samples coming from the Cohn-Kanade database, we score AUs even if the relevant offsets are missing.

5 Experimental Results

5.1 Test Data Sets

In the last decade, the research on automatic analysis of facial expressions has become a central topic in machine vision research. Nonetheless, there is a glaring lack of a comprehensive, easily accessible reference set of images of facial expressions that could be used as a basis for benchmarks for efforts in the field [3].

To date, the Cohn-Kanade facial expression database [8] is the most comprehensive and the most commonly used database in research on automated facial expression analysis. It contains over 2000 AU-coded gray-scale image sequences of facial expressions in nearly frontal view shown by 210 adults (69% female, 81% Caucasian) being 18 to 50 years old. From those some 480 samples were made publicly available. The main drawbacks of this data set are as follows. First, each recording ends at the apex of the shown facial display. This makes research of facial expression temporal activation patterns (onset → apex → offset) less feasible using this data set. Further, many recordings contain the date/time stamp recorded over the chin of the subject. This makes changes in the appearance of the chin less visible and motions of the chin obscured by changes of the time/date stamp. Also, the database does not contain images of all possible single-AU activations; it contains mainly recordings of facial displays of emotions. Besides, the image sequences have been acquired at an extremely low frame rate (12 or even less frames/second). This makes fast neuromuscular facial activity (e.g., blink that lasts approximately ¼ of a second) difficult to observe and to track. Finally, the database is neither easily accessible nor easily searchable. Once permission for usage is issued, large, unstructured files of material are sent. In spite of these drawbacks we used the Cohn-Kanade facial expression data set to evaluate the performance of our method. We did so in order to make the results achieved by our method comparable to those accomplished by the previously reported facial expression analyzers that have been tested using this database (e.g., [7], [9], [10]).

The lack of easily accessible, suitable, and common training and testing material forms the major impediment to comparing, resolving, and extending the issues concerned with automated facial expression analysis from face images. It is this critical issue that we tried to address by building a novel face image database, which we call the MMI Facial Expression Database [13]. The MMI Database has been developed to address most (if not all) the issues mentioned above. It contains some 850 image sequences and 750 static images of faces in frontal and in profile view displaying various facial expressions of emotion, single AU activation, and multiple AU activation. The samples are all true color (24-bit) images which, when digitized, measure 720×576 pixels. They picture 52 different faces, ranging in age from 19 to 62, having either a European, Asian, African, or South American ethnic background. The image sequences are of variable length (40 – 520 frames), taken at 24 Hz frame rate, picturing one or more neutral-expressive-neutral facial behavior patterns. To date, approximately two thirds of the samples have been FACS coded for target AUs. Of these, 169 image sequences have been FACS coded per frame for temporal segments of target actions. Finally, the database has been developed as a web-based direct-manipulation application, allowing easy access and easy search of the available images [13].

5.2 Validation Studies

In order to evaluate the performance of our method, we used 90 image sequences of the Cohn-Kanade database picturing the activation of the following AUs: 1, 2, 4-7, 9, 10, 12, 15, 16, 20, 23-27, 38, 39, 44 and 45. To include into this test set more samples of facial displays of AUs 10, 13, 18, 28, 35, 41, 43, 45 and 46 we also used 45 image sequences of the MMI Facial Expression Database

picturing the facial expressions in question. The metadata (i.e., the human judgments about displayed AUs) associated with these 135 image sequences represent the ground truth with which we compared the judgments generated by our method. 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 for the Cohn-Kanade-database samples are summarized in Table 3 and these for the MMI-database samples are given in Table 4. Overall, we achieved an average recognition rate of 90% sample-wise for 27 different AUs occurring alone or in a combination in an input video sample.

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 fast blinks (i.e., AU45 having a duration of less than n frames in either onset or offset, where $n = 3$ for the Cohn-Kanade-database samples and $n = 5$ for the MMI-database samples). 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

Table 3: AU recognition results for 90 samples from the Cohn-Kanade database.

Legend**:** Upper face AUs: 1, 2, 4, 5-7, 9, 44, 45. AUs affecting the nose: 38, 39. AUs affecting the mouth: 10, 12, 15, 16, 20, 23-25. AUs affecting the jaw: 26, 27. # denotes the number of samples. C denotes correctly recognized samples. MA denotes the number of samples in which some AUs were missed or they were scored in addition to those depicted by human experts.

	#	C	MA	Rate
upper face	73	69	4	94.5%
nose	10	10	0	100%
mouth	82	76	6	92.7%
jaw	43	41	2	95.3%
all samples	90	84	6	**93.3%**

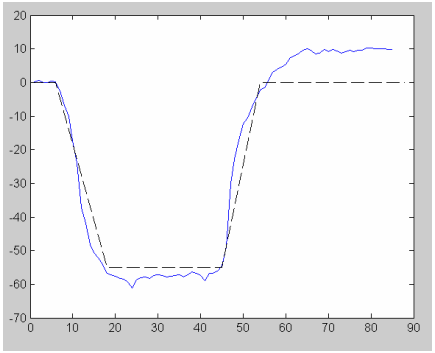

Fig. 4. Onset, apex and offset of AU27 indicated by the ground truth (dashed line) and those detected by the method (full line). Horizontal axis: the frame number. Vertical axis: the value that parameter *inc/dec*(NM) takes for a AU27-activation sample out of the MMI database.

Table 4: AU recognition results for 45 samples from the MMI Facial Expression database.

Legend**:** Upper face AUs: 41, 43, 45, 46. AUs affecting the mouth: 10, 13, 18, 28, 35. # denotes the number of samples. C denotes correctly recognized samples. MA denotes the number of samples in which some AUs were missed or they were scored in addition to those depicted by human experts.

	#	C	MA	Rate
upper face	30	26	4	86.7%
mouth	25	23	2	92%
all samples	45	39	6	**86.7%**

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 AU23 in cases when the mouth corners are tracked more toward the center of the mouth (rather than toward the cheeks). Note that AU23 and AU24 are also often confused by human FACS coders [4] and by other automated AU analyzers (e.g., [9]). In addition, note that the temporal pattern of feature motion in AU23 activation is very similar to the one occurring in AU24 activation. The same is the case with AU41 vs. AU43 activation. Hence, the distinction between these two pairs of AUs may be more amenable to appearance-based analysis than to feature motion analysis.

As can be seen from Fig. 3 and 4, the temporal segments of the AUs indicated by the ground truth varied slightly from those detected by our method. For most of AUs, in general, the boundaries of temporal segments of AUs were detected either at the same moment or a little bit later than prescribed by the ground truth. The measured delays take up to 1-2 frames on average for Cohn-Kanade-database samples, that is, up to $^1/_6$ of a second. However, in the case of AUs whose activation becomes apparent from the movement of the mouth corner (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 3 frames for Cohn-Kanade-database samples (up to ¼ of a second). The reason for these delays is the temporal rules used for AU recognition. 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 for 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.

Finally, upon a close inspection of the temporal rule used to recognize AU27 activation (see Table 1), 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. 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. 4).

6 Conclusions

Automating the analysis of facial actions (i.e., AUs) is important to advance the studies on human emotion, to design multimodal human-machine interfaces, and to boost various applications in fields such 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 nearly frontal view face video.

The presented approach extends the state of the art in automatic AU detection from face image sequences in two ways including the temporal segments of AUs (onset, apex, offset) and the number of AUs (27 in total) handled. Based upon the presented validation studies, 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 [7], our method achieves an average recognition rate of 90% for encoding of 27 AU codes and their combinations in 135 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 by Bartlett et al. [9], our method achieves an average recognition rate of more than 93.5% AU-wise (see Tables 3 and 4) 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 number of AUs and temporal dynamics handled, our method also improves other aspects of automated AU detection compared to earlier works. 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., point M in the case of a long beard). Finally, due to the usage of the color-based observation model, 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 nearly frontal-view face image sequence. Further research efforts are necessary if the full range of human facial behavior is to be coded in an automatic way.

Acknowledgments

The authors would like to thank Jeffrey Cohn of the University of Pittsburgh for providing the Cohn-Kanade database. This work of M. Pantic is supported by the Netherlands Organization for Scientific Research Grant EW-639.021.202. The work of I. Patras is supported by the Netherlands BSIK-MultimediaN-N2 Interaction project.

References

[1] M. Pantic, "Face for Interface", *Encyclopedia of Multimedia Technology and Networking*, vol. 1, 2005.
[2] D. Keltner and P. Ekman, "Facial expression of emotion", *Handbook of Emotions*, pp. 236-249, 2000.
[3] M. Pantic and L.J.M. Rothkrantz, "Toward an affect-sensitive multimodal HCI", *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1370-1390, 2003.
[4] P. Ekman and W. Friesen, *Facial Action Coding System*, Consulting Psychologist Press, 1978.
[5] K. Mase, "Recognition of facial expression from optical flow", *IEICE Transactions*, vol. E74, no. 10, pp. 3474-3483, 1991.
[6] I. Essa, and A. Pentland, "Coding, analysis, interpretation and recognition of facial expressions", *IEEE Trans. PAMI*, vol. 19, no. 7, pp. 757-763, 1997.
[7] Y. Tian, T. Kanade and J.F. Cohn, "Recognizing action units for facial expression analysis", *IEEE Trans. PAMI*, vol. 23, no. 2, pp. 97-115, 2001.
[8] T. Kanade, J. Cohn and Y. Tian, "Comprehensive database for facial expression analysis", *Proc. IEEE Conf. FGR*, pp. 46-53, 2000.
[9] M.S. Bartlett, et al., "Machine Learning Methods for Fully Automatic Recognition of Facial Expressions and Actions", *Proc. IEEE Conf. SMC*, pp. 592-597, 2004.
[10] M.F. Valstar, M. Pantic and I. Patras, "Motion history for facial action detection from face video", *Proc. IEEE Conf. SMC*, pp. 635-640, 2004.
[11] K.L Schmidt and J.F Cohn, "Dynamics of facial expression: Normative characteristics and individual differences", *Proc. IEEE Conf. Multimedia and Expo*, pp. 547-550, 2001.
[12] M. Pantic and I. Patras, "Temporal modeling of facial actions from face profile image sequences", *Proc. IEEE Conf. Multimedia and Expo*, pp. 49-52, 2004.
[13] M. Pantic, et al., "Web-based facial expression database", *Proc. IEEE Conf. Multimedia and Expo*, 2005. (www.mmifacedb.com)
[14] M.K. Pitt and N. Shephard, "Filtering via simulation: auxiliary particle filtering", *J. Amer. Stat. Assoc.*, vol. 94, pp. 590-599, 1999.