FACIAL GESTURE RECOGNITION IN FACE PROFILE IMAGE SEQUENCES

Maja Pantic
Delft University of Technology
ITS / Mediamatics Dept.
Delft, the Netherlands
M.Pantic@cs.tudelft.nl

Ioannis Patras
University of Amsterdam
Computer Science Dept.
Amsterdam, the Netherlands
yiannis@science.uva.nl

Leon Rothkrantz
Delft University of Technology
ITS / Mediamatics Dept.
Delft, the Netherlands
L.J.M.Rothkrantz@cs.tudelft.nl

ABSTRACT

Automatic analysis of facial gestures is an area of intense interest in the human-computer interaction design community. A robust way to discern facial gestures in images of faces, insensitive to scale, pose, and occlusion, is still the key research challenge in the automatic facial-expression analysis domain. A practical method recognized as the most promising one for addressing this problem is through a facial-gesture analysis of multiple views of the face. Yet, current systems for automatic facial-gesture analysis utilize mainly portraits or nearly frontal-views of faces. To advance the existing technological framework upon which research on automatic facial-gesture analysis from multiple facial views can be based, we developed an automatic system as to analyze subtle changes in facial expressions based on profile-contour fiducial points in a profile-view video. A probabilistic classification method based on statistical modeling of the color and motion properties of the profile in the scene is proposed for tracking the profile face. From the segmented profile face, we extract the profile contour and from it, we extract 10 profile-contour fiducial points. Based on these, 20 individual facial muscle actions occurring alone or in a combination are recognized by a rule-based method. A recognition rate of 85% is achieved.

1. INTRODUCTION

A long-term goal in human-computer interaction (HCI) research is to approach the naturalness of human-human interaction [1]. This means integrating "natural" means that humans employ to interact with each other into HCI. With this motivation, automatic speech recognition has been a topic of research for decades. Recently, also other human interactive modalities such as gaze, body and facial gestures have gained intense interest as potential modes of HCI [1].

As a step towards a multimodal HCI design, the main focus of our current research is whether and how facial gestures could be included as a new mode of HCI. The major impulse to investigate facial-gestures human communicative modality for inclusion into HCI comes from the significance of this modality within human-human interaction. Facial gestures (underlying a facial expression) regulate our social interactions [2]: they clarify whether our current focus of attention (a person, an object or what has been said) is important, funny or unpleasant for us. They are the most powerful, natural and immediate means for humans to communicate their emotions [2, 3]. Within our research, we first investigated whether and to which extent human facial gestures could be recognized automatically. This paper presents a part of our research concerning automatic recognition of facial gestures from face-profile images.

Most approaches to automated facial gesture analysis attempt to recognize a small set of prototypic emotional facial expressions, i.e., fear, sadness, disgust, anger, surprise and happiness [4]. This practice may follow from the work of Darwin and more recently Ekman [3], who suggested that basic emotions have corresponding prototypic expression. In everyday life, however, such prototypic expressions occur relatively infrequently; emotions are displayed more often by subtle changes in one or few discrete facial features, such as raising the eyebrows in surprise [2]. To detect such subtlety of human emotion, automatic recognition of facial gestures (i.e., fine-grained changes in facial expression) is needed.

Facial gestures are anatomically related to contractions of facial muscles [5]. Contractions of facial muscles produce changes in both the direction and magnitude of the motion on the skin surface and in the shape and location of the permanent facial features (eyes, mouth, etc.). To reason about shown facial gestures, the face, its features and their current appearance should be detected first. A problematic issue here is that of scale, pose, and occlusion: rigid head and body movements of the observed person usually cause changes in the viewing angle and the visibility of the tracked face and its features. As noted in [7], perhaps the most promising method for addressing this problem is through the use of multiple cameras yielding multiple views of the face and its features. To date, nonetheless, the works on automatic facial gestures analysis have avoided dealing with facial views other than a frontal one: portraits (e.g., [6, 8]) or nearly-frontal views of faces (e.g., [9, 10]) constitute the input data processed by the existing systems. For exhaustive reviews on the past attempts to address the problems of automatic facial gesture recognition in frontal and nearly-frontal views of faces, readers are referred to [4, 6].

From several methods for recognition of facial gestures based on visually observable facial muscular activity, the FACS system [5] is the most commonly used in the psychological research. Following this trend, all of the existing methods for automatic facial gesture analysis, including the method proposed here, interpret the facial display information in terms of the facial action units (AUs) of the FACS system [4, 6]. Yet none automatic system is capable of encoding the full range of facial mimics, i.e., none is capable of recognizing all 44 AUs that account for the changes in facial display. From the previous works, the automatic facial mimics analyzers presented in [10] and [8] perform the best in this aspect: they code 16 and, respectively, 27 AUs occurring alone or in a combination in frontal-view face images.

The research reported here addresses the problem of automatic AU coding from face profile image sequences. It was undertaken with two motivations:

1. In a frontal view of the face, facial gestures such as showing the tongue (AU19) or pushing the jaw forwards (AU29) represent

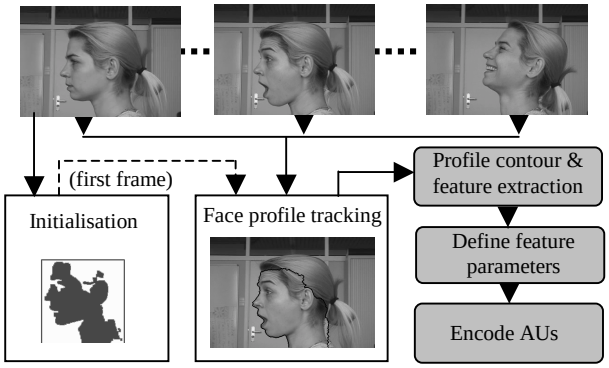

Fig. 1: Outline of the profile-based AU recognition

out-plane non-rigid facial movements which are difficult to detect [8, 9, 10]. Such facial gestures are clearly observable in a profile-view of the face.

2. A basic understanding of how to achieve automatic facial gesture analysis from human face profiles is necessary for the establishment of a technological framework for automatic facial gestures analysis from multiple facial views. Eventually, within such a framework, procedures of greater flexibility and improved performance can evolve.

Fig. 1 outlines the proposed method. For the first frame of the input face profile image sequence, a description of the local statistical properties of the tracked face profile is built based on a label field that is initiated by utilizing a HSV color-based segmentation of the face. A probabilistic classification method based on statistical modeling of the color and motion properties of the label field is used to track the label field in the rest of the sequence. The contour of the tracked face-profile region of the label field is extracted as the face profile contour. Under the assumption that the face images are in (nearly) left profile view, the left-hand-side points of the extracted profile contour that have a large curvature are extracted as feature points (profile-contour fiducials). Subtle changes in the tracked face profile are measured next. Motivated by AUs of the FACS system, these changes are represented as a set of mid-level feature parameters describing the state and motion of the feature points and the shapes formed between certain feature points. Based on these feature parameters, a rule-based algorithm interprets the extracted facial information in terms of 20 AUs occurring alone or in a combination. Face profile tracking, profile contour and feature extraction, parametric representation, AU coding and experimental evaluation are explained in sections 2, 3, 4, 5 and 6 respectively.

2. FACE PROFILE TRACKING

The first step in automatic facial gesture analysis is to locate the face in the scene. In order to do so, we adapted a semi-automatic method for object-based segmentation of complex-scene image sequences [11] for the purpose of human face tracking (Fig. 2).

Face region tracking is addressed as a segmentation problem in two objects: the Face and the Background. For the first frame of the sequence, markers of the two objects are extracted as follows. For the face region the marker is extracted as the largest connected image component with Hue, Saturation and Value within the range [5, 35], [0, 0.7] and [0.1, 0.9] respectively [8]. In the absence of a similar model for the Background, its marker is extracted as the bounding box of the Face marker. A color-based version of a watershed segmentation algorithm provides the final segmentation for the first frame [11]. The color-based watershed segmentation yields good localization of the face given that the most prominent color edge between the Background and the Face markers is indeed the face contour. Under the assumption that in the first frame of the sequence the face is seen in a nearly vertical position, this is usually the case. However, for the rest of the image sequence, rotations of the face (see Fig. 1) can result in bounding boxes that contain large parts of the background with probable strong edges. Therefore, for the rest of the image sequence, we perform the segmentation based on tracking of the local statistical color and motion properties of the Face and the Background.

The method operates at three levels. At *Level 1* (pixel level) a feature vector is estimated for each pixel in the current frame. At *Level 2* (region level) a watershed color segmentation method decomposes the current frame in a number of color regions. The statistical properties of the color regions are estimated subsequently under the assumption that the same process, which is modeled as a multivariate Gaussian, generates the feature vectors at pixels inside the same region. At *Level 3* (object level) a labeling based on probabilistic classification of the color regions takes place. Each color region is projected in the previous frame where an estimation of the label field is available. For each object present in a window surrounding the area of projection, we estimate the parameters of the model (a multivariate Gaussian) describing the objects color and motion properties. Then, each color region is assigned the object label such that the joint probability of the label field and the observed color and motion features is maximized. Once each color region is labeled, the local models are re-estimated for the current frame. An iterative region-classification / model-estimation routine is performed until no color region changes its label. Typical results of this method are illustrated in Fig. 3. For further details about this method, readers are referred to [11].

3. PROFILE CONTOUR & FEATURE EXTRACTION

The contour of the face profile region (referred to as "face profile contour" in the text below), generated by the face profile tracking method (Fig. 3), is utilized for further analysis of shown facial gestures. We proceed with feature points' extraction (Fig. 4) under two assumptions: (1) the face images are non-occluded nearly left profile view with possible in-plane head rotations, and (2) the first frame is in a neutral expression. After initializing the feature points in the first frame based on the face region of the initial label field, they are automatically extracted from the tracked face profile contour for the rest of the sequence.

To account for possible in-plane head rotations and variations in scale of the tracked face profile, the face profile contour is normalized in each frame based on two referential points (Fig. 4): the tip of the nose (P4) and the top of the forehead (P1). The major impulse to the usage of these referential points comes from their stability with respect to non-rigid facial features' movements: facial

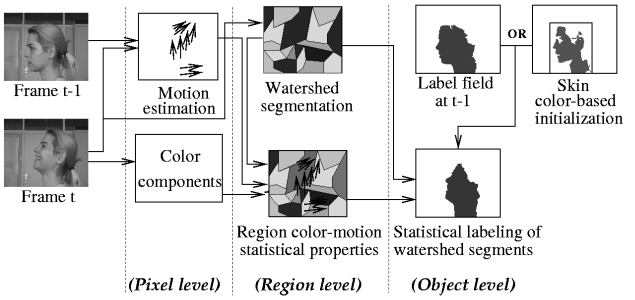

Fig. 2: Outline of the face profile tracking method

Fig. 3: Results for the "pleasant surprise" sequence. 1st and 3rd columns: Superposition of the contour of the face profile region on the original frames. 2nd and 4th columns: Face profile region (i.e., the label field for segmentation in two objects). Results are shown for frames 1, 36, 42, 91, 105, and 115.

muscles' contractions do not cause physical displacements of these points [12]. The tip of the nose and the top of the forehead are extracted as the leftmost and, respectively, the uppermost leftmost point of the generated contour. To handle possible inaccuracies in detection of the referential points caused by inaccuracies in the segmentation of the face profile region, we exploit all: information from the previous frames, the knowledge about temporal dynamics of rigid head movements (usually they occur gradually in time) and the knowledge about the facial stability of the referential points. A small window $\boldsymbol{W_B}$ (its height and width set to 3% of the length of P1P4 measured in frame ***t-1***) centered at the loci of a referential point extracted from frame ***t-1*** is searched for the pertinent point in frame ***t***. If a referential point cannot be defined such that it belongs to the face profile contour determined for frame ***t***, the relevant referential point determined for frame ***t-1*** is used instead. Finally, the face profile contour is normalized by carrying out affine transformations of it such that the line P1P4 between the referential points discerned for the current frame is of the same length and orientation as the line P1P4 determined for the first frame.

To extract the feature points from the normalized face profile contour, we move from image to function analysis and treat the left-hand side of the normalized face profile contour (up to the determined referential point P1) as the profile contour function. We extract the extremities of this function (i.e., the zero-crossings of the function's 1st order derivative). Given the *a priori* knowledge on where the convexities and concavities of a left face profile are, we analyze the extracted extremities to find out where the function is arched. The maximums and minimums of the function's 2nd order derivative are extracted as the feature points (Fig. 4). To ascertain correct extraction of the feature points when the tongue is visible (P7' and P7'' exist), we extract the feature points in the particular order (i.e., P1, P4, P2, P3, P10, P5, P9, P7 or P7' and P7", P6, P8). To handle inaccuracies in feature points' detection (e.g., frame 91, Fig. 3) and to remove false positives and negatives, we exploit both the knowledge about facial anatomy and geometric characteristics of the extreme points and the information from the previous frames. Similarly to the case of $\boldsymbol{W_B}$ defined for referential points P1 and P4, a standard "search" window $\boldsymbol{W_P}$ has been defined for each feature point ***P*** with respect to anatomically possible directions and magnitudes of the motion on the skin surface affecting the temporal location of ***P***. The feature point $\boldsymbol{P_t}$ is determined further for frame ***t*** such that it represents a specific zero crossing (Fig. 4) of the 1st order derivative of the profile contour function defined for frame ***t*** and belongs to the $\boldsymbol{W_P}$ set around the location of $\boldsymbol{P_{t1}}$ discerned for frame ***t1***. If $\boldsymbol{P_t}$ cannot be defined, $\boldsymbol{P_{t-1}}$ is used instead.

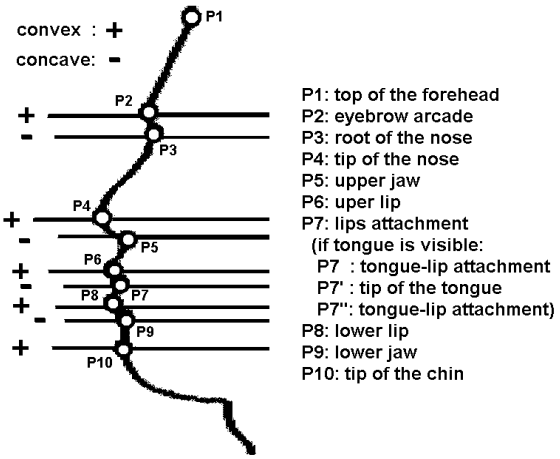

Fig. 4: Feature points (profile contour fiducials)

4. PARAMETRIC FEATURE REPRESENTATION

Each AU of the FACS system is anatomically related to contraction of a specific facial muscle [5]. Contractions of facial muscles produce motion in the skin surface and deform the shape and location of the facial features (eyebrows, mouth, chin, etc.). Some of these changes in facial expression are observable from the changes in the tracked face profile contour and the related feature points. To classify detected changes of the face profile contour in terms of facial muscle activity (i.e., in terms of AUs of the FACS system), these changes should be represented first as a set of suitable feature parameters.

We defined six mid-level feature parameters in total: two describing the motion of the feature points, two describing their state, and two describing shapes formed between certain feature points. The definitions of the parameters, which are calculated for each frame, are given in Fig. 5.

Feature points motion	
up/down (P) = $y_{Pt1} - y_{Pt}$ If ***up/down(P)*** < 0, ***P*** moves up.	***in/out(P)*** = $x_{Pt1} - x_{Pt}$ If ***in/out(P)*** > 0, P moves outward.
Feature points state	
If P9 equals P7, ***absent(***P9***)***. If there is no maximum of f" between P5 and P7, ***absent(***P6***)***. Similarly for P7', P7" and P8 (see Fig. 4).	***increase/decrease(AB)*** = $\boldsymbol{AB_{t1} - AB_t}$, where $\boldsymbol{AB} = \sqrt{\{(x_A - x_B)^2 + (y_A - y_B)^2\}}$ If ***increase/decrease(AB)*** < 0, distance ***AB*** increases.
Shapes formed by feature points	
The physic meanings of ***angular(P6P8)*** = *true* and ***increased_curvature(P5P6)*** are shown below.	

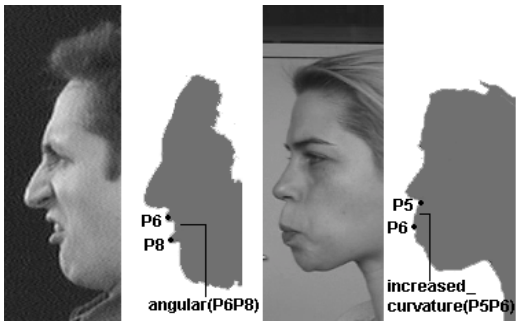

Fig. 5: Parametric representation of face-profile-contour features for AU recognition

5. ACTION UNIT RECOGNITION

The last step in automatic facial mimics analysis is to translate the extracted facial expression information (i.e., the calculated feature parameters) into a description of shown facial changes such as the AU-coded description of shown facial expression. To achieve this, we utilize a fast-direct-chaining rule-based method that encodes 20 AUs occurring alone or in a combination in the current frame of the input face-profile image sequence. A full list of the utilized rules is given in [13]. Motivated by the FACS system, each of these rules is defined in terms of the predicates of the mid-level representation (Fig. 5) and each encodes a single AU in a unique way according to the relevant FACS rule. For example, the rule utilized for coding AU12, which is described in the FACS system as an oblique upward pull of the lip corners (i.e., smile), is the following:

IF ***in/out***(P6) < 0 AND ***in/out***(P8) < 0 AND ***increase/decrease***(P5P6) ≤ 0 AND ***increase/decrease***(P6P8) ≤ 0 AND ***increased_curvature***(P5P6) = *false* THEN AU20.

6. EXPERIMENTAL EVALUATION

Though AU-coded facial expression image databases are available in general, these databases contain portraits or nearly frontal-views of human faces. Since these data are not suitable for testing our face-profile-based AU encoder, we generated our own test data.

The test data set has been created in office environments (e.g., Fig. 3, Fig. 5) with the help of 5 certified FACS coders drawn from college personnel. The acquired test images represent a number of demographic variables including ethnic background (European, Asian and South American), gender (60% female) and age (20 to 35 years). The subjects were asked to display series of expressions that included single AUs and combinations of those. Forty image sequences of variable length (110 to 240 frames) of nearly left-profile view of subjects' faces were recorded by utilizing a CCD digital PAL camera. The size of the face region in each frame was at least 135×175 pixels. Sequences began with a neutral expression with no head rotation. Metadata were associated with the acquired test data given in terms of AUs scored by 4 certified FACS coders. As the actual test data set, we used 32 image sequences for which the overall inter-coders' agreement about displayed AUs was above 75%. The AU-coded descriptions of shown expressions obtained by human FACS coders were compared further to those produced by our method. The results of this comparison are given in Table 1.

Table 1: Recognition results for the upper face AUs (AU1, AU4, AU9), the AUs affecting the jaw (AU17, AU26, AU27, AU29) and those affecting the mouth (AU8, AU10, AU12, AU15, AU16, AU18, AU19, AU20, AU23, AU24, AU25, AU28, AU36):

 # denotes the number of AUs' occurrences,
C denotes correctly recognized AUs' occurrences,
M denotes missed AUs' occurrences,
IC denotes incorrectly recognized AUs' occurrences.

	#	C	M	IC	Rate
upper face	18	15	1	2	83.3%
mouth	82	66	5	11	80.5%
jaw	36	33	1	2	90.9%
Total:	136	114	7	15	**84.9%**

7. CONCLUSIONS

In this paper, we introduced an automatic system for analyzing subtle changes in facial expression based on changes in face profile contour tracked in a nearly left-profile-view image sequence. 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 number of AUs, the difference in AUs and the facial view handled.
2. The proposed method for AU recognition provides a basic understanding of how to achieve automatic AU coding in face profile image sequences. Hereupon further research on facial gesture analysis from multiple facial views can be based. For example, as a first step, the proposed method could be combined with a method for AU recognition from frontal-view face image sequences to achieve AU recognition from dual-views of the face (as suggested in [8] for static face images).

In addition, the algorithm explained here could greatly speed up the time-consuming (manual) process of acquiring AU-labeled data on which generative probability models for AU recognition in face-profile image sequences could be trained (e.g., HMM for AU recognition – the main focus of our further research on this topic). Nonetheless, before this algorithm could be actually deployed for such a purpose, it needs some refinements. For instance, once the referential points have been located, they can be used for stabilizing the face profile region yielding a more robust face profile tracking. Also, although the proposed method demonstrates concurrent validity with manual FACS coding of test data set, additional field trials and quantitative validation studies are necessary to confirm this finding.

REFERENCES

[1] R. Sharma, et al., "Toward multimodal human-computer interaction", *Proc. of the IEEE*, vol. 86, no. 5, pp. 853-869, 1998.

[2] J. Russell and J. Fernandez-Dols, *The psychology of facial expression*, Cambridge University Press, 1997.

[3] D. Keltner and P. Ekman, "Facial expression of emotion", *Handbook of Emotions*, Guilford Press, pp. 236-249, 2000.

[4] M. Pantic and L.J.M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art", *IEEE TPAMI*, vol. 22, no. 12, pp. 1424-1445, 2000.

[5] P. Ekman and W. Friesen, *Facial Action Coding System*, Consulting Psychologist Press, 1978.

[6] G. Donato, et al., "Classifying facial actions", *IEEE TPAMI*, vol. 21, no. 10, pp. 974-989, 1999.

[7] A. Pentland, "Looking at people", *IEEE TPAMI*, vol. 22, no. 1, pp. 107-119, 2000.

[8] M. Pantic and L.J.M. Rothkrantz, "Expert system for automatic analysis of facial expressions", *Image and Vision Computing*, vol. 18, no. 11, pp. 881-905, 2000.

[9] M. Black and Y. Yacoob, "Recognizing facial expressions in image sequences using local parameterized models of image motion", *Computer Vision*, vol. 25, no. 1, pp. 23-48, 1997.

[10] Y. Tian, et al., "Recognizing action units for facial expression analysis", *IEEE TPAMI*, vol. 23, no. 2, pp. 97-115, 2001.

[11] I. Patras, et al., "A semi-automatic method for segmentation of image sequences with local region-based classification", *Int'l Conf. Signal and Image Processing*, pp. 205-210, 2000.

[12] L. Harmon, et al., "Identification of human face profiles by computer", *Pattern Recognition*, vol. 10, pp. 301-312, 1978.

[13] M. Pantic et al., *Facial mimics recognition from face profile image sequences*. Technical Report #TR-DKS-02-01, DKS group (http://www.kbs.twi.tudelft.nl/Publications/Report/), Delft University of Technology, Delft, 2002.