Facial gesture recognition in face image sequences: A study on facial gestures typical for speech articulation

M. Pantic and L.J.M. Rothkrantz

Delft University of Technology

ITS / Mediamatics

P.O. Box 356, 2600 AJ Delft, The Netherlands

{M.Pantic, L.J.M.Rothkrantz}@cs.tudelft.nl

Abstract—*Automatic analysis of facial gestures is rapidly becoming an area of intense interest in computer science and human-computer interaction design communities. However, the basic goal of this area of research – translating detected facial changes into a human-like description of shown facial expression – is yet to be achieved. One of the main impediments to achieving this aim is the fact that human interpretations of a facial expression differ depending upon whether the observed person is speaking or not. A first step in tackling this problem is to achieve automatic detection of facial gestures that are typical for speech articulation. This paper presents our approach to seizing this step in the research on automatic facial expression analysis. It presents all: a robust and flexible method for recognition of 22 facial muscle actions from face image sequences, a method for automatic determination of whether the observed subject is speaking or not, and an experimental study on facial muscle actions typical for speech articulation.*

I. INTRODUCTION

Facial gestures (facial muscle actions) regulate our social interactions: they represent visible speech signals and they clarify whether our current focus of attention (e.g., a person or what has been said) is important, funny or unpleasant for us. They are direct, naturally preeminent means for humans to communicate their emotions [1, 2]. Automatic analyzers of subtle facial changes, therefore, seem to have a natural place in various vision systems including automated tools for psychological research, lip reading, bimodal speech analysis, affective computing, face and visual-speech synthesis, and perceptual user interfaces. Thus, in recent years, there has been a tremendous interest in automating facial gesture analysis.

Most approaches to automatic facial gesture analysis in face image sequences attempt to recognize a set of prototypic emotional facial expressions, i.e., happiness, sadness, fear, surprise, anger and disgust [3]. Yet, in everyday life such prototypic expressions occur rather infrequently; emotions are displayed more often by subtle changes in one or few discrete facial features, such as raising the eyebrows in surprise [1]. To detect such subtlety of human emotion, automatic recognition of facial gestures (i.e., fine-grained changes in facial expression) is needed.

From several methods for recognition of facial gestures based on visually observable facial muscular activity, the FACS system [4] 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 [3, 5]. 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 on automatic facial gesture recognition from face image sequences, the method presented in [6] performs the best in this aspect: it encodes 16 AUs occurring alone or in a combination in frontal-view face image sequences.

However, even if an automatic detector of all possible facial muscle actions would be at hand, emotional interpretation of facial cues would remain by no means a trivial task. This goal is made difficult by the rich shadings of affective/attitudinal states that humans recognize in a facial expression. Another major element of difficulty is that a shown facial gesture may be easily misinterpreted if the presence of visual speech data is not taken into account. For example, a frown may be displayed by the speaker to emphasize the difficulty of the currently discussed problem and it may be shown by the listener to denote that he did not understand the problem at issue. To date, however, automatic facial information analyzers do not perform usually user-profiled interpretation of sensed data and virtually all approaches to facial gesture analysis have largely avoided dealing with questions that involve whether the observed subject is speaking or not. The later is easy to do if one can limit the context. For example, if you know that except of the observed subject there is no other person in the area, then pursing the lips will probably represent a facial signal of being bored or being in a mode of thinking and not a visible speech signal. But, as we move towards more generally competent perceptual user interfaces, which should

facilitate videoconferences, virtual visits to Internet sites, etc., we will have to directly confront the problem of distinguishing the facial gestures that are typical for speech articulation from those signaling attitude or affect. Hence, both a reliable detector of whether the observed subject is speaking or not and the knowledge about facial gestures which form the typical visible speech signals (to be treated as noise in affect-sensitive analysis of visual speech data) are needed for an (user-profiled or not) emotional interpretation of facial cues.

Within our research on facial gesture analysis from frontal-view face image sequences, we investigated first whether and to which extent human facial gestures and speech onset/offset can be recognized automatically. Hereafter, we investigated which facial gestures form typical visual speech signals. This paper presents the preliminary results of our research. The devised method for rule-based recognition of 22 AUs from frontal-view face image sequences is presented in section 2. Section 3 gives an overview of a neural-network-based method for automatic determination of whether the observed subject is speaking or not. Experimental evaluations of the two methods and an experimental study on facial muscle actions typical for speech articulation are presented in section 4. Section 5 concludes the paper.

II. FACIAL GESTURE RECOGNITION

The problem of automatic facial gesture recognition from face image sequences is usually divided into three sub-problem areas (Fig. 1): detecting prominent facial features such as eyes and mouth, representing subtle changes in facial expression as a set of suitable mid-level feature parameters, and interpreting these data in terms of facial gestures such as the AUs of the FACS system.

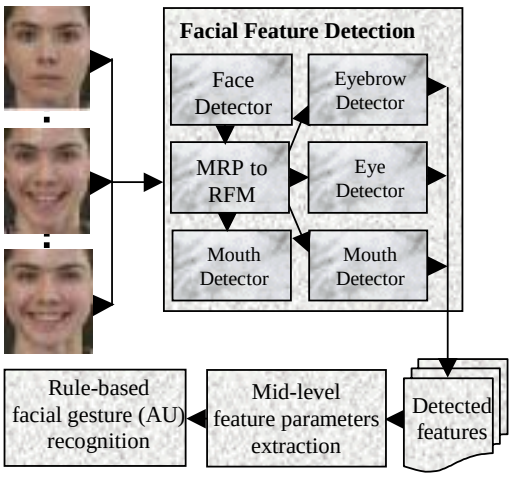

Fig. 1: Outline of the AU-recognition method

A. Facial Feature Detection

To reason about shown facial gestures, the face and its components (i.e., prominent facial features) should be detected first. In order to do so, we apply a multi-phase multi-detector processing of an input frontal-view face image sequence. The two phases of the proposed method for detection of prominent facial features are coarse detection and fine detection.

In the first phase, we apply a HSV color-based segmentation of the face ("Face Detector" in Fig. 1). The face region is segmented from an input frame 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 [7]. Then we use a simple analysis of image histograms ("MRP to RFM" in Fig. 1) to locate 7 regions of interest (ROI): two eyebrows, two eyes, nose, mouth and chin.

In the second phase, to spatially sample the contour of a certain permanent facial feature, we apply one or more facial-feature detectors to the pertinent ROI. For example, the contours of the eyes are localized in the ROIs of the eyes by using a single detector representing an adapted version of a hierarchical-perceptron feature-location method [7]. On the other hand, the contour of the mouth is localized in the mouth ROI by applying both a 4-parameters deformable template and a method that fits three 2^{nd} degree parabolas [8]. For further details about these and other detectors used to spatially sample the contours of the prominent facial features, readers are referred to [7, 8].

B. Parametric Feature Representation

The contours of the facial features, generated by the facial feature detection method (Fig. 1), are utilized for further analysis of shown facial gestures.

First, we carry out feature points' extraction under two assumptions: (1) the face images are non-occluded and in frontal view, and (2) the first frame is in a neutral expression. We extract 22 fiducial points: 19 are extracted as vertices or apices of the contours of the facial features (Fig. 2), 2 represent the centers of the eyes (points X and Y), and 1 represents the the middle point between the nostrils (point C). We assign a certainty factor to each of the extracted points, based on an "intra-solution consistency check". For example, the fiducial points of the right eye are assigned a certainty factor $CF \in [0, 1]$ based upon the calculated deviation of the actually detected inner corner $B_{current}$ from the pertinent point $B_{neutral}$ localized in the first frame of the input sequence. The functional form of this mapping is:

$$CF = \boldsymbol{sigm}(\boldsymbol{d}(B_{current}, B_{neutral}); 1, 4, 10)$$

where $\boldsymbol{d}(p1, p2)$ is the block distance between points p1 and p2 (i.e., maximal difference in x and y direction) while $\boldsymbol{sigm}(x; \alpha, \beta, \gamma)$ is a Sigmoid function. The major

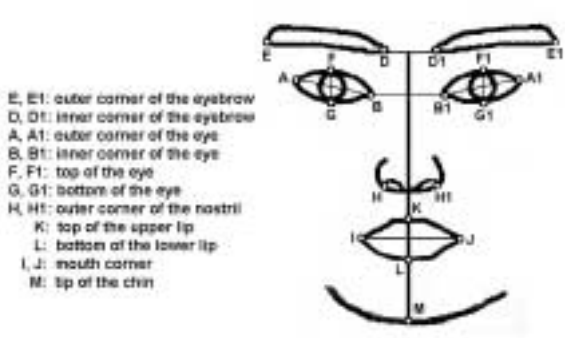

Fig. 2: Feature points (fiducials of the features' contours)

impulse for the usage of the inner corners of the eyes as the referential points for calculating CFs of the fiducial points of the eyes comes from the stability of these points with respect to non-rigid facial movements: facial muscles' contractions do not cause physical displacements of these points. For the same reason, the referential features used for calculating CFs of the fiducial points of the eyebrows, nose/chin and mouth are the size of the relevant eyebrow area, the inner corners of the nostrils and the medial point of the mouth respectively. Eventually, in order to select the best of sometimes redundantly available solutions (e.g., for the fiducial points belonging to the mouth), an inter-solution consistency check is performed by comparing the CFs of the points extracted by different detectors of the same facial feature.

AUs of the FACS system are anatomically related to contractions of facial muscles [4]. Contractions of facial muscles produce motion on the skin surface and changes in the shape and location of the prominent facial features. Some of these changes are observable from changes in the position of the fiducial points. To classify detected changes in the position of the fiducial points in terms of AUs, these changes should be represented first as a set of suitable feature parameters. Motivated by the FACS system, we represent these changes as a set of mid-level feature parameters describing the state and motion of the fiducial points. We defined a single mid-level feature parameter, which describes the state of the fiducials. This parameter, which is calculated for each frame for various fiducial points by comparing the currently extracted fiducial points with the relevant fiducial points extracted from the neutral frame, is defined as:

$$\textbf{\textit{inc/dec(AB)}} = \textbf{\textit{AB}}_{neutral} - \textbf{\textit{AB}}_{current}, \text{ where } \textbf{\textit{AB}} = \sqrt{\{(x_A - x_B)^2 + (y_A - y_B)^2\}}$$

If ***inc/dec(AB)*** < 0, distance ***AB*** increases.

C. Action Unit Recognition

The last step in automatic facial gesture analysis is to translate the extracted facial information (i.e., the calculated feature parameters) into a description of shown facial changes, e.g., into the AU codes.

Table 1: The description of 22 AUs to be recognized and the related rules for AU recognition

AU	AU description & the related rule
1	Raised inner portion of the eyebrow(s) IF ***inc/dec***(BD) < 0 OR ***inc/dec***(B1D1) < 0 THEN AU1
2	Raised outer portion of the eyebrow(s) IF ***inc/dec***(AE) < 0 OR ***inc/dec***(A1E1) < 0 THEN AU2
4	Eyebrows pulled closer together (frown) IF ***inc/dec***(DD1) > 0 THEN AU4
5	Raised upper eyelid(s) IF ***inc/dec***(FG) < 0 OR ***inc/dec***(F1G1) < 0 THEN AU5
6	Raised cheeks (smile); IF AU12 OR AU13 THEN AU6
7	Raised lower eyelid(s) IF ***not***(AU12) AND ((FG > 0 AND ***inc/dec***(GX) > 0) OR (F1G1 > 0 AND ***inc/dec***(G1Y) > 0)) THEN AU7
8	Lips pulled towards each other IF ***not***(AU12 OR AU13 OR AU15 OR AU18 OR AU20 OR AU23 OR AU24 OR AU35) AND KL > 0 AND ***inc/dec***(CK) < 0 THEN AU8
12	Mouth corner(s) pulled up IF (***inc/dec***(IB) > 0 AND ***inc/dec***(CI) < 0) OR (***inc/dec***(JB1) > 0 AND ***inc/dec***(CJ) < 0) THEN AU12
13	Mouth corner(s) pulled sharply up IF (***inc/dec***(IB) > 0 AND ***inc/dec***(CI) > 0) OR (***inc/dec***(JB1) > 0 AND ***inc/dec***(CJ) > 0) THEN AU13
15	Mouth corner(s) pulled down IF ***inc/dec***(IB) < 0 OR ***inc/dec***(JB1) < 0 THEN AU15
18	Mouth pushed medially forward (as when saying "fool") IF ***not***(AU28) AND IJ ≥ ***t1*** AND ***inc/dec***(IJ) > 0 AND ***inc/dec***(KL) ≤ 0 THEN AU18
20	Mouth stretched horizontally IF ***inc/dec***(IJ) < 0 AND ***inc/dec***(IB) = 0 AND ***inc/dec***(JB1) = 0 THEN AU20
23	Tightened lips IF KL > 0 AND ***inc/dec***(KL) > 0 AND ***inc/dec***(IJ) ≤ 0 ND ***inc/dec***(JB1) ≥ 0 AND ***inc/dec***(IB) ≥ 0 THEN AU23
24	Lips pressed together IF ***not***(AU12 OR AU13 OR AU15) AND KL > 0 AND ***inc/dec***(KL) > 0 AND IJ > ***t1*** AND ***inc/dec***(IJ) > 0 THEN AU24
25	Parted lips IF ***inc/dec***(KL) < 0 AND ***inc/dec***(CM) ≥ 0 THEN AU25
26	Parted jaws IF ***inc/dec***(CM) < 0 AND CM ≤ ***t2*** THEN AU26
27	Mouth stretched vertically; IF CM > ***t2*** THEN AU27
28	Lips sucked into the mouth; IF KL = 0 THEN AU28
35	Cheeks sucked into the mouth; IF IJ < ***t1*** THEN AU35
38	Widened nostrils IF ***not***(AU8 OR AU12 OR AU13 OR AU18 OR AU24) AND ***inc/dec***(HH1) < 0 THEN AU38
39	Compressed nostrils IF ***not***(AU8 OR AU15 OR AU18 OR AU24 OR AU28) AND ***inc/dec***(HH1) > 0 THEN AU39
41	Dropped upper eyelid(s) IF ***not***(AU7) AND ((FG > 0 AND ***inc/dec***(FG) > 0 AND ***inc/dec***(FX) > 0) OR (F1G1 > 0 AND ***inc/dec***(F1G1) > 0 AND ***inc/dec***(F1Y) > 0)) THEN AU41

To achieve this, we utilize a fast-direct-chaining rule-based method that encodes 22 AUs occurring alone or

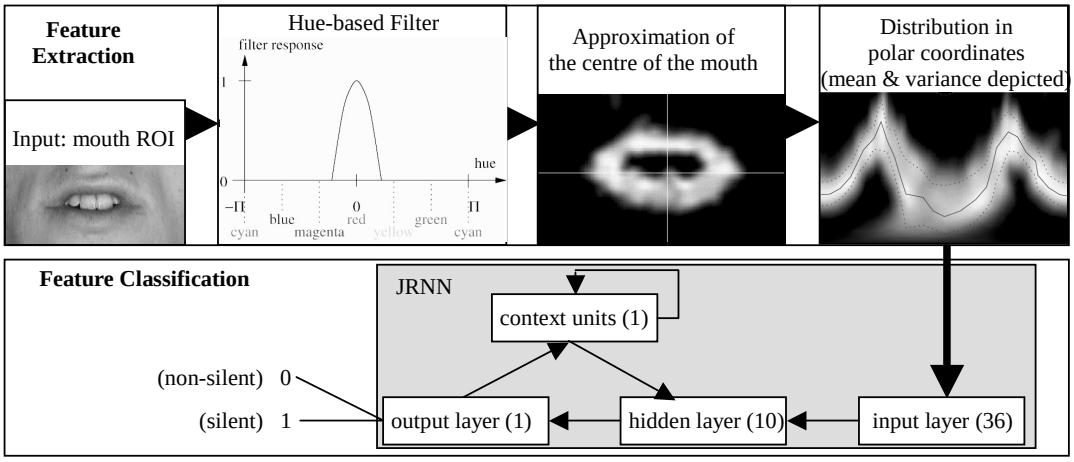

Fig. 3: The outline of the neural-network-based method for speech onset/offset detection

in a combination in the current frame of the input face-profile image sequence. A full list of the utilized rules is given in Table 1. Motivated by the FACS system [4], each of these rules is defined in terms of the predicate of the mid-level representation and each encodes a single AU in a unique way according to the relevant FACS rule.

III. SPEECH ONSET/OFFSET DETECTION

The human perception of speech is not restricted to the auditory part of a speech signal; even in ideal auditory conditions, the visual information provided by the face (i.e., lip motion) plays an important role in our speech-recognition process and, with the degradation of the auditory signal (e.g., due to hearing disorders or noisy environment), it becomes crucial [9]. To detect whether the observed subject is speaking or not directly from an input face image sequence, we used a neural-network-based method for visual speech onset/offset detection developed in the scope of our research on visual speech processing (i.e., lip-reading) [10]. The utilized method processes the mouth ROI extracted from an input frame of a face image sequence in two steps (Fig. 3): feature extraction and classification into one of the "silent" and "non-silent" categories.

A. Feature Extraction

Based on the observation that the lips appear in a face image as more reddish than the rest of the face, the mouth ROI of an input frame is first transformed into the HSV color space and then into the red domain so that only the lips are highlighted in the image. Image in red domain is obtained from the Hue component of the input frame by applying the following function:

$$f(\boldsymbol{h}) = \begin{cases} 1 - \dfrac{(\boldsymbol{h} - \boldsymbol{h}_0)^2}{\boldsymbol{w}^2} & |\boldsymbol{h} - \boldsymbol{h}_0| \le \boldsymbol{w} \\ 0 & |\boldsymbol{h} - \boldsymbol{h}_0| > \boldsymbol{w} \end{cases}$$

where $\boldsymbol{w} = 0.23$ and $\boldsymbol{h}_0 = 0$ gave a fair segmentation of the lips from the rest of the face for our camera setup.

The filtered image is transformed further into polar coordinates $\boldsymbol{J(\alpha, r)}$ around the center of the mouth. This center point is computed as the center of gravity of the distribution obtained from filtering the image. The mean $\boldsymbol{M(\alpha)}$ and variance $\boldsymbol{\sigma(\alpha)}$ of the distribution $\boldsymbol{J(\alpha, r)}$ for a given angle α relate directly to the distance of the lips from the center of the mouth and the thickness of the lips respectively:

$$M(\alpha) = \frac{\int_r J(\alpha, r) \cdot r}{\int_r J(\alpha, r)}, \quad \sigma(\alpha) = \frac{\int_r J(\alpha, r) \cdot (r - M(\alpha))^2}{\int_r J(\alpha, r)}$$

These values are used further to represent the mouth shape. A 36-dimensional feature vector, representing the mean and variance sampled in 18 uniformly distributed sample points, is extracted as the data vector for further processing.

B. Silence Detection

In the second step of the utilized algorithm, a trained Jordan Recurrent Neural Network (JRNN) classifies the input frame as either "silent" or "non-silent" (Fig. 3).

In a JRNN the time dependency is represented by the recurrent nature of the NN itself: while the input to the JRNN is a single frame feature vector, the time related information is preserved within the network by context neurons. The input to a context neuron is the output of the whole network and its own output from the previous time step. The single hidden layer of the network is fed with activation of both the input neurons and context neurons. In the case of the JRNN used for speech onset/offset detection, there is only one output neuron and hence only one context neuron.

IV. EXPERIMENTAL STUDIES

We conducted three experimental studies within our research on automatic facial gesture analysis. The first

was aimed at evaluating the performance of our method for AU recognition. The second pertained to evaluating the proposed method for speech onset/offset detection. The third was aimed at discerning the facial muscle actions that are typical for speech articulation.

A. Image Database

Most of the existing approaches to either facial gesture recognition or lip-reading assume that the presence of the face in the input image is ensured [3, 10]. However, in most of the real-life situations where such automated systems are to be employed (e.g., videoconferencing, human-computer interaction, etc.) the location of the face in the scene is not known a priori. The presence of a face can be ensured either by employing an existing method for automatic face detection in arbitrary scenes (e.g., see [11]) or by using a camera setting that will ascertain the assumption at issue. The two algorithms proposed here do not perform face detection in an arbitrary scene; they operate on frontal-view face image sequences acquired by a head-mounted CCD digital PAL camera (Fig. 4).

Fig. 4: Mounted-camera device and an example of an input frame

The face image sequences used in our experiments have been obtained with the help of six 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 (66% female) and age (20 to 35 years). Two datasets have been acquired:

- ***Dataset 1***: 48 image sequences of subjects displaying series of facial expressions including single AUs and combinations of those. The first frame is in a neutral expression and the length is from 95 to 250 frames. No movement of the lips due to a speech articulation is present.
- ***Dataset 2***: 6 image sequences of subjects speaking a set of 5 sentences while maintaining a neutral facial expression. The sentences are from the POLYPHONE corpus [12] and contain all of the phonemes used in the Dutch language. The length of sequences varies from 850 to 1050 frames.

B. AU Recognition

Dataset 1 has been used to evaluate the performance of the proposed method for AU recognition. Metadata were associated with the acquired test data in terms of AUs that were scored by 5 certified FACS coders other than the one displaying the judged facial expressions. As the actual test data set, we used 40 image sequences for which the overall inter-coders' agreement about displayed AUs was above 75%. 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, given in Table 2, show that in 93% of test cases, our method for AU recognition coded the analyzed facial expression using the same AU codes as the human observers.

Table 2: Recognition results for the upper face AUs (AU1, AU2, AU4, AU5, AU6, AU7, AU41), the AUs affecting the nose (AU38, AU39), the AUs affecting the jaw (AU26, AU27) and those affecting the mouth (AU8, AU12, AU13, AU15, AU18, AU20, AU23, AU24, AU25, AU28, AU35):

 # 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	54	50	4	0	92.6%
nose	13	12	0	1	92.3%
mouth	102	95	4	3	93.1%
jaw	23	21	1	1	91.3%
Total:	192	178	9	5	**92.7%**

C. Silence Detection

Dataset 2 has been used to evaluate the performance of our method for visually based speech onset/offset detection. First, the speech onset and offset points were labeled manually based upon the auditory signal only. Those boundary points were used to label all the frames in video sequences of dataset 2 as well as the pertinent extracted feature vectors as either 1 for silence or 0 for speech. These labels were further used as target output values for the silence recognition system.

About 10% of the feature vectors extracted from the full dataset 2 were used to form the test set; the rest was used to train the utilized JRNN. The training set was chosen to contain both intervals of beginning and of ending of silence and non-silent frames. The fitness of the network was measured with mean square error of its response for a single epoch. Fig. 5 depicts the response of the trained JRNN for a single video-sequence containing 5 sentences.

When compared with other NN architectures (feed forward and time delayed neural networks), JRNN gave by far the most superior results; it was trained much faster and it gave the smoothest and most accurate results. For further details about both the JRNN-based method for speech onset/offset recognition and the comparison of the performance of JRNN and those of other NN architectures, readers are referred to [10].

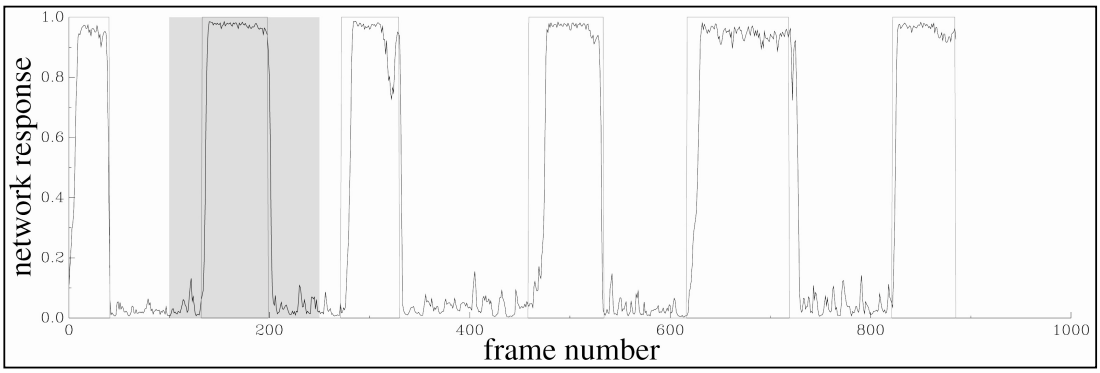

Fig. 5: The response of JRNN for a video sequence containing 5 sentences. The straight lines depict target outputs and the grey area represents the part of the sequence that was used as a test set.

D. AUs Typical for Speech Articulation

Dataset 2 has been used in our experimental study on facial muscle actions typical for speech articulation. The aim of this experiment was to discern the AUs that can be encoded by our method for AU recognition and, at the same time, form typical visible speech signals. Our major impulse to perform this experiment comes from our intention to build an adequate affect-sensitive analyzer of human facial interactive cues in the future. In order to achieve this aim by using the proposed rule-based method for AU recognition, we must be able first to discern which AUs are to be treated as noise (present as visible speech signals) by such an automatic affect-sensitive analyzer of human facial interactive cues.

30 intervals of variable length (70 to 110 frames) of the video sequences of dataset 2, which were labeled as non-silent by the JRNN, were AU-coded by our method for AU recognition. A set of 5 AUs (from a total of 22 AUs that can be encoded by the proposed AU coder) is found to be typical for speech articulation in the Dutch language: AU18, AU20, AU23, AU25 and AU26.

In 16% of test cases AU8 and AU24 were encoded as well. Yet, we did not label these AUs as being typical for speech articulation since their frequency of the occurrence is much lower than that of the AUs listed above (they occur in more than 56% of test cases). Also, in 7% of test cases AU1 and AU2 were encoded. However, due to the fact that AU1 and AU2 affect the eyebrows rather than the mouth, they were excluded from the set of AUs typical for speech articulation.

V. CONCLUSIONS

The presented method for automatic AU recognition extends the state of the art in automatic facial gesture analysis in face image sequences in terms of number of AUs handled. The significance of this contribution is also in the performed experimental studies that suggest: (i) that it is possible to determine whether the observed subject is speaking or not from visual data only, and (ii) that at least 5 AUs are typical for speech articulation and could be, therefore, treated as noise in affect-sensitive interpretation of visual speech data.

The presented algorithm for automatic AU coding of face image sequences does not take into account the temporal nature of facial gestures. Yet, the presented AU coder could greatly speed up the time-consuming (manual) process of acquiring AU-labeled data on which models that can capture the temporal nature of facial gestures (e.g., HMM) could be trained. Devising both a HMM-based AU coder and an affect-sensitive analyzer of AU-coded "silent" and "non-silent" facial data forms the main focus of our further research.

REFERENCES

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

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

[3] 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.

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

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

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

[7] 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.

[8] M. Pantic, et al., "A hybrid approach to mouth features detection", *Proc. IEEE Conf. SMC*, 2001, pp. 1188-1193.

[9] A. Adjoudani, et al., "A multimedia platform for audio-visual speech processing", *Proc. Eurospeech*, 1997, vol. 3, pp. 1671-1674.

[10] J. Wojdel and L. Rothkrantz, "Visually based speech onset/offset detection", *Proc. Euromedia*, 2000, pp. 156-160.

[11] R. Feraud, et al., "A fast and accurate face detector based on neural networks", *IEEE TPAMI*, vol. 23, no. 1, pp. 42-53, 2001.

[12] M. Damhuis, et al., "Creation and analysis of the Dutch polyphone corups", *Proc. ICSLP*, 1994, pp. 1803-1803.