Learning spatio-temporal models of facial expressions

M. Pantic, I. Patras, M.F. Valstar
Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, The Netherlands

Abstract
The human face is used to regulate the conversation by gazing or nodding, to interpret what has been said by lip reading, and to communicate and understand somebody's affective state and intentions on the basis of the shown facial expression. Machine understanding of human facial behavior could revolutionize human-machine interaction technologies and fields as diverse as security, behavioral science, medicine, communication, and education. Yet development of an automated system that detects and interprets human facial signals is rather difficult. This article summarizes our research efforts in meeting this challenge. It presents two systems for machine recognition of facial muscle actions (i.e., Action Units, AUs) in face video and a case-based reasoning system capable of classifying facial expressions (coded in terms of AUs) into the emotion categories learned from the user.

Keywords
Facial expression analysis, temporal templates, particle filtering, case-based reasoning, emotion.

1 Introduction
The human face provides a number of signals essential for interpersonal communication in our social life. The face houses the speech production apparatus and is used to identify other members of the species, to regulate the conversation by gazing or nodding, and to interpret what has been said by lip reading. It is our direct and naturally preeminent means of communicating and understanding somebody's affective state and intentions on the basis of the shown facial expression [5].
Automating the analysis of facial behavior would be highly beneficial for fields as diverse as security, medicine and education. In security contexts, facial expressions play a crucial role in establishing or detracting from credibility. In medicine, facial expressions are the direct means to identify when specific mental processes are occurring. In education, pupils' facial expressions inform the teacher of the need to adjust the instructional message. As far as interfaces between humans and computers (PCs / robots / machines) are concerned, facial expressions provide a way to communicate information about needs and demands to the machine. Where the user is looking (gaze tracking) can be effectively used to free computer users from the classic keyboard and mouse. Also, certain facial signals (e.g. a wink) can be associated with certain commands (e.g. a mouse click) offering an alternative to traditional mouse and keyboard commands. The human ability to read emotions from someone's facial expressions is the basis of facial affect processing that can lead to expanding interfaces with emotional communication and, in turn, to obtaining a more flexible, adaptable, and natural interaction between humans and machines.

2 Facial action coding
Most approaches to automatic facial expression analysis attempt to recognize a small set of prototypic emotional facial expressions, i.e., fear, sadness, disgust, happiness, anger, and surprise (for an exhaustive survey of the past work on this research topic, the reader is referred to [10]). This practice may follow from the work of Darwin and more recently Ekman [5], 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 raising of the eyebrows in surprise. To detect such subtlety of human emotions and, in general, to make the information conveyed by facial expressions available for usage in various aforementioned applications, automatic recognition of facial muscle actions, such as the action units (AUs) of the FACS system [3], is needed. Facial Action Coding System (FACS) is designed for human observers to describe changes in facial expression in terms of observable facial muscle actions (AUs). FACS provides the rules for visual detection of 44 different AUs and their temporal segments (onset, apex, offset) in a face image sequence. Using these rules, a human coder decomposes a shown facial expression into the specific AUs that produced the expression.
Few approaches have been reported for automatic AU recognition in face images (for an exhaustive survey of the past work on this research topic, the reader is referred to [6]). These include automatic detection of 16 AUs from face video using lip tracking, template matching and neural networks [14], color and motion based detection of 20 AUs occurring alone or in combination in profile-view face video [8], and automatic detection of 18 AUs from face video using Gabor filters, AdaBoost and Support Vector Machines [1]. In contrast to these methods, which address mainly the problem of spatial modeling of facial expressions, the methods proposed in this article address the problem of temporal modeling of facial expressions as well. In other words, the methods proposed here are very suitable for encoding temporal activation patterns (onset → apex → offset) of AUs shown in an input face video.

2.1 AU detection using temporal templates
Figure 1 outlines our method for AU detection in face video using temporal templates. Temporal templates are

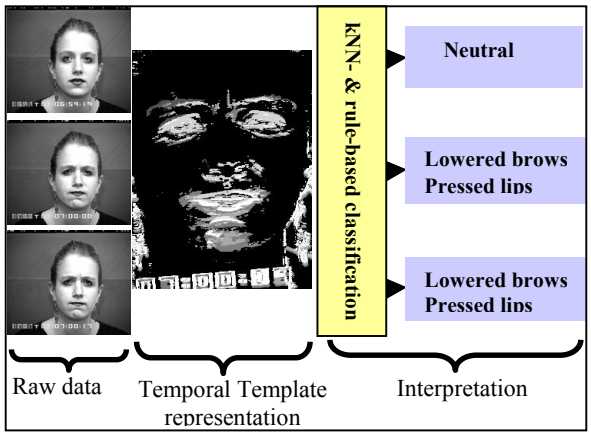

Figure 1. *AU detection using temporal templates*

2D images constructed from image sequences, which show motion history, that is, where and when motion in the input image sequence has occurred [2]. We employ Motion History Images (MHI), which in contrast to Motion Energy Images preserve not only spatial but also the temporal information. More specifically, the value of a pixel in an MHI image indicates where and when motion in the input image sequence has occurred. This value decays over time, so that a high intensity pixel denotes recent motion, a low intensity pixel denotes a motion that occurred earlier in time, and intensity zero denotes no motion at all at that specific location.

Before we can construct a MHI from an input video, the face present in the video needs to be registered in two ways. Intra registration removes all rigid head movements within the input video while the inter registration places the face at a predefined location in the scene. The inter registration process warps the face onto a predefined 'normal' face, eliminating inter-person variation of face shape and facilitating the comparison between the facial expression shown in the input video and template facial expressions. Under the assumption that each input image sequence begins and ends with a neutral facial expression, we downsample the number of frames to a fixed number of (n+1) frames. In this way our system becomes robust to the problem of varying duration of facial expressions.

After the registration and time warping of the input image sequence, the MHI is obtained as follows. Let *I(x, y, t)* be an image sequence of pixel intensities of *k* frames and let *D(x, y, t)* be the binary image that results from pixel intensity change detection, that is by thresholding $|I(x,y,t)-I(x,y,t-1)|>th$, where *x* and *y* are the spatial coordinates of picture elements and *th* is the minimal intensity difference between two images. In an MHI, say H_t, the pixel intensity is a function of the temporal history of motion at that point with *t* being a frame of the downsampled input video (with (n+1) frames). Using the known parameter *n*, H_t is defined as:

$$H_t(x,y,t)=\begin{cases} s*t & D(x,y,t)=1 \\ H_t(x,y,t-1) & otherwise \end{cases} \qquad (1)$$

where $s=(255/n)$ is the intensity step between two history levels and where $H_t(x,y,t)=0$ for $t\le 0$. The final MHI, say *H(x, y)*, is found by iteratively computing equation *(1)* for $t=1\ldots n+1$.

For automatic detection of AU from MHI-represented face image sequences, we employ a combined kNN/rule-based classifier. The utilized kNN algorithm is straightforward: for a test sample it uses a distance metric to compute which *k* (labeled) training samples are "nearest" to the sample in question and then casts a majority vote on the labels of the nearest neighbors to decide the class of the test sample. Parameters of interest are the distance metric being used and *k*, the number of neighbors to consider. The optimal parameters were experimentally determined to be the simple Euclidian distance measure and *k* = 3 [15]. Although it gives a good indication about the AUs shown in an input video, the kNN algorithm can confuse AUs that have partially the same MHI-representation. To address this drawback, we created a set of rules. With these rules we can correctly reclassify samples that the kNN algorithm misclassifies at first. For instance, the kNN classifier often confuses AU4 and AU1+AU4. Both produce activity in the same part of the MHI, but AU4 causes the eyebrows to move inward and downward, while AU1+AU4 first causes an upward movement followed by an inward and downward movement of the eyebrows. This results in high activation between the brows and relatively low activation above the inner corners of the brows in the case of AU4 activation. Hence, the rules used to resolve the confusion in question have been defined based upon this kind of knowledge about the facial muscle anatomy.

When tested on the Cohn-Kanade Facial Expression Database [4] and the MMI Facial Expression Database [9], the proposed method achieved a recognition rate of 68%, respectively 61%, when detecting 21 AUs occurring alone or in combination in an input face image sequence (for details about this method see [15]).

2.2 AU detection using temporal rules

Figure 2 outlines our method for AU detection in face video using temporal rules. The method processes an input face image sequence in four steps: Face Detection, Facial Fiducial Points Detection, Point Tracking and AU Coding. To detect the face region in the first frame of an input face video, we adopt a real-time face detector proposed in [1], which represents an adapted version of the original Viola-Jones face detector [16]. The Viola-Jones face detector consists of a cascade of classifiers trained by AdaBoost. Each classifier usess integral image filters, which remind of Haar Basis functions and can be computed very fast at any location and scale. For each stage in the cascade, a subset of features is chosen using a feature selection based on AdaBoost. The adapted version of the Viola-Jones face detector that we employ uses GentleBoost instead of AdaBoost and it uses a smart training procedure in which, after each single feature, the system can decide whether to test another feature or to make a decision. By this the system retains information about the continuous outputs of each feature detector rather than converting to binary decisions at each stage of the cascade.

The detected face region is then divided in 20 relevant Regions of Interest (ROIs), each one corresponding to one facial point to be detected. A combination of heuristic techniques based upon the analysis of the vertical and horizontal image histograms achieves this. The employed facial feature point detection method [17] uses individual feature patch templates to detect points in the relevant ROI. These feature models are 13×13 pixels GentleBoost templates built from both gray level intensities and Gabor wavelet features. In the training phase, the feature models are learned using a representative set of positive and negative examples, where the positive examples are image patches centered on a particular facial feature point and the negative examples are image patches randomly displaced a small distance from the same facial feature. In the testing phase, each ROI is filtered first by the same set

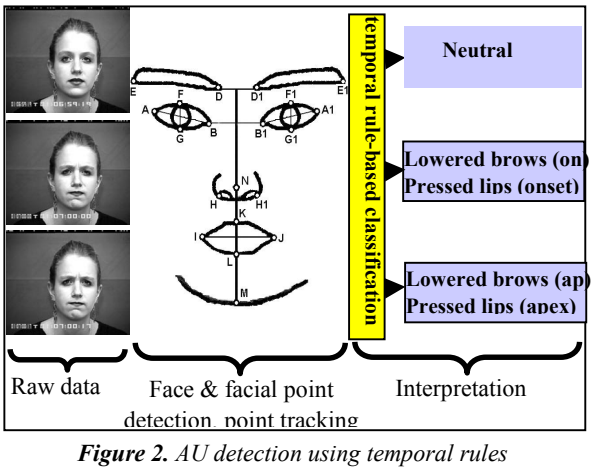

Figure 2. *AU detection using temporal rules*

of Gabor filters used in the training phase (in total, 48 Gabor filters are used). Then, for a certain facial point an input 13×13 pixels window (*sliding window*) is slid pixel by pixel across 49 representations of the relevant ROI (grayscale plus 48 Gabor filter representations). For each position of the sliding window, a GentleBoost classifier outputs a response depicting the similarity between the 49-dimensional representation of the sliding window and the learned feature point model. After scanning the entire ROI, the position with the highest response reveals the feature point in question.

After 20 fiducial points are localized in the first frame of the input face image sequence, windows 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 [12]. Particle filtering has become a 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 equation *(2)* below, 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 and the robust function that we use is the absolute value.

$$d = \sum_{i=1}^{M} \rho\left(\left\| \frac{c_i}{m_c} - \frac{o_i}{m_o} \right\|_1 \mu_c\right) / M \rightarrow (2)$$

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. We defined two 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 and parameter ***inc/dec(PP')*** = $PP'_{t1} - PP'_t$ describes the increase or decrease of the distance between points P and P'. 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 combination in the input face videos. For instance, to recognize the temporal segments of AU4, which pulls the eyebrows closer together, we exploit the following temporal rules:

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

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

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

When tested on the Cohn-Kanade Facial Expression Database [4] and the MMI Facial Expression Database [9], the proposed method achieved a recognition rate of 90% when detecting 27 AUs occurring alone or in combination in an input face image sequence (for details about this method see [7]).

3 User-profiled facial-afect recognition

As already noted above, virtually all systems for automatic facial affect analysis attempt to recognize a small set of universal/basic emotions [10]. However, pure expressions of "basic" emotions are seldom elicited; most of the time people show blends of emotional displays [5]. Hence, the classification of human non-verbal affective feedback into a single "basic"-emotion category is not realistic. Also, not all non-verbal affective cues can be classified as a combination of the "basic" emotion categories. Think for instance about the frustration, skepticism or boredom. Furthermore, it has been shown that the comprehension of a given emotion label and the ways of expressing the related affective state may differ from culture to culture and even from person to person [13]. Hence, pragmatic choices (user-profiled choices) must be made regarding the selection of affective states to be recognized by an automatic analyzer of human affective feedback.

The rest of this paper describes our case-based reasoning system that performs classification of AUs into the emotion categories learned from the user. The utilized case base is a dynamic, incrementally self-organizing event-content-addressable memory that allows fact retrieval and evaluation of encountered events based upon the user preferences and the generalizations formed from prior input. Each event (case) is one or more micro-events, each of which is a set of AUs. Micro-events related by the goal of communicating one specific affective state are grouped within the same dynamic memory chunk. In other words, each memory chunk represents a specific emotion category and contains all micro-events to which the user assigned the emotion label in question. The indexes associated with each dynamic memory chunk comprise individual AUs and AU combinations that are most characteristic for the emotion category in question. Finally, the micro-events of each dynamic memory chunk are hierarchically ordered according to their typicality: the larger the number of times a given micro-event occurred, the higher its hierarchical position within the given chunk. The initial endowment of the dynamic memory is achieved by asking the user to associate an interpretation (emotion) label to a set of 40 typical facial expressions (micro-events that might be hardwired to emotions according to [13]).

The classification of the AUs detected in an input face image into the emotion categories learned from the user is further accomplished by case-based reasoning about the content of the dynamic memory. To solve a new problem of classifying a set of input AUs into the user-defined interpretation categories, the following steps are taken:

1. Search the dynamic memory for similar cases, retrieve them, and interpret the input set of AUs using the interpretations suggested by the retrieved cases.
2. If the user is satisfied with the given interpretation, store the case in the dynamic memory. Otherwise, adapt the memory according to user-provided feedback on the interpretation he associates with the input facial expression.

The utilized retrieval and adaptation algorithms employ a pre-selection of cases that is based upon the clustered organization of the dynamic memory, the indexing structure of the memory, and the hierarchical organization of cases within the clusters/ chunks according to their typicality (for details about this method see [11]).

Two validation studies on a prototype system have been carried out. The question addressed by the 1st validation study was: How acceptable are the interpretations given by the system after it is trained to recognize 6 basic emotions? The question addressed by the 2nd validation study was: How acceptable are the interpretations given by the system, after it is trained to recognize an arbitrary number of user-defined interpretation categories? In the first case, a human FACS coder was asked to train the system. In the second case, a lay expert, without formal

training in emotion signals recognition, was asked to train the system. The same expert used to train the system was used to evaluate its performance, that is, to judge the acceptability of interpretations returned by the system. For basic emotions, in 100% of test cases the expert approved of the interpretations generated by the system. For user-defined interpretation categories, in 83% of test cases the lay expert approved entirely of the interpretations and in 14% of test cases the expert approved of most but not of all the interpretation labels generated by the system for the pertinent cases.

4 Conclusion

In this paper, we presented two methods for AU detection in a nearly frontal view face video and a facial expression recognition system that performs classification of AUs into the emotion categories learned from the user.

The presented approaches extend the state of the art in automatic AU detection from face image sequences in two ways including temporal modeling of facial expressions and the number of AUs (21 and 27 AUs in total) handled. Namely, the automated systems for AU detection from face video that have been reported so far address mainly the problem of spatial modeling of facial expressions and, at best, can detect 16 to 18 AUs (from in total 44 AUs). Our methods also improve other aspects of automated AU detection compared to earlier works. The performance of both proposed methods is invariant to occlusions like glasses and facial hair as long as these do not entirely occlude facial points that are tracked (this is of importance for the second proposed AU detector). Also, the methods perform well independently of changes in the illumination intensity. As far as our method for automatic facial affect interpretation is concerned and given that the previously reported facial expression analyzers are able to classify facial displays only in one of the 6 basic emotion categories, the proposed method extends the state of the art in the field by enabling facial expression interpretation in a user-adaptive manner.

However, the proposed methods cannot recognize the full range of facial behavior (i.e. all 44 AUs defined in FACS). Furthermore, they assume that the input data are facial displays which are isolated or pre-segmented, showing a single temporal pattern (onset → apex → offset) of an expression that begins and ends with a neutral state. In reality, such segmentation is not available; human facial behavior is more complex and transitions from a facial (emotional) expression to another do not have to involve intermediate neutral state. Hence, our facial behavior analyzers cannot deal with spontaneously occurring facial displays. Further research efforts are necessary if the full range of human (spontaneous and posed) facial behavior is to be coded in an automatic way.

The work of M. Pantic is supported by the Netherlands Organization for Scientific Research Grant EW-639.021.202. The work of I. Patras and M.F. Valstar is supported by the Netherlands BSIK-MultimediaN-N2 Interaction project.

References:

1. Bartlett, M.S.; Littlewort, G.; Lainscsek, C.; Fasel, I.; Movellan, J. (2004). Machine learning methods for fully automatic recognition of facial expressions and facial actions. *Proc. Int'l Conf. Systems, Man and Cybernetics (SMC'04)*, 592-597.
2. Bobick, A.F.; Davis, J.W. (2001). The recognition of human movement using temporal templates. *IEEE Trans. Pattern Analysis & Machine Intelligence*, **23**, 257-267.
3. Ekman, P.; Friesen, W.V. (1978). *FACS Manual*. Consulting Psychologist Press, Palo Alto, USA.
4. Kanade, T.; Cohn, J.; Tian, Y. (2000). Comprehensive database for facial expression analysis. In *Proc. Int'l Conf. Face and Gesture Recognition (FGR'00)*, 46-53.
5. Keltner, D.; Ekman, P. (2000). Facial expression of emotion. *Handbook of Emotions*. Guilford Press, New York, USA, 236-249.
6. Pantic, M. (2005). Face for Interface. *The Encyclopedia of Multimedia Technology and Networking*. Idea Group Publishing, Hershey, USA.
7. Pantic, M.; Patras, I. (2005). Detecting facial actions and their temporal segments in nearly frontal view face image sequences. *Proc. Int'l Conf. System, Man and Cybernetics (SMC'05)*.
8. Pantic, M.; Patras, I.; Rothkrantz, L.J.M. (2002). Facial action recognition in face profile image sequences. *Proc. Int'l Conf. Multimedia and Expo (ICME'02)*, 37-40.
9. Pantic, M.; Valstar, M.F.; Rademaker, R.; Maat, L. (2005). Web-based database for facial expression analysis. In *Proc. Int'l Conf. Multimedia and Expo (ICME'05)*.
10. Pantic, M.; Rothkrantz, L.J.M. (2003). Toward an Affect-Sensitive Multimodal Human-Computer Interaction. *Proceedings of the IEEE,* **91**, 1370-1390.
11. Pantic, M.; Rothkrantz, L.J.M. (2004). Case-based reasoning for user-profiled recognition of emotions from face images. In *Proc. Int'l Conf. Multimedia and Expo (ICME'04)*, 391-394.
12. Pitt, M.K.; Shephard, N. (1999). Filtering via simulation: auxiliary particle filtering", *J. Amer. Stat. Assoc.*, **94**, 590-599.
13. Russell J.; Fernandez-Dols, J. (1997). *The psychology of facial expression*, Cambridge University Press, Cambridge, USA.
14. Tian, Y.; Kanade, T.; Cohn, J.F. (2001). Recognizing action units for facial expression analysis. *IEEE Trans. Pattern Analysis & Machine Intelligence*, **23**, 97-115.
15. Valstar, M.F.; Pantic, M.; Patras, I. (2004). Motion History for Facial Action Detection from Face Video. *Proc. Int'l Conf. System, Man and Cybernetics (SMC'04)*, 635-640.
16. Viola, P.; Jones, M. (2001). Robust real-time object detection. *ICCV Workshop on Statistical and Computation Theories of Vision*.
17. Vukadinovic, D.; Pantic, M. (2005). Fully automatic facial feature point detection using Gabor feature based boosted classifiers. *Proc. Int'l Conf. System, Man and Cybernetics (SMC'05)*.