Proceedings of the 2004 IEEE International Workshop on
Robot and Human Interactive Communication
Kurashiki, Okayama Japan September 20-22, 2004

FACIAL ACTION UNIT RECOGNITION USING TEMPORAL TEMPLATES

Michel Valstar, Ioannis Patras and Maja Pantic
Delft University of Technology
EEMCS / Mediamatics Dept.
Delft, the Netherlands
{M.F.Valstar, I.Patras, M.Pantic}@ewi.tudelft.nl

Abstract

Automatic recognition of human facial expressions is a challenging problem with many applications in human-computer interaction. Most of the existing facial expression analyzers succeed only in recognizing a few emotional facial expressions, such as anger or happiness. Instead of being another approach to automatic detection of prototypic facial expressions of emotion, this work attempts to measure a large range of facial behavior by recognizing facial action units (AUs, i.e. atomic facial signals) that produce expressions. The proposed system performs AU recognition using temporal templates as input data. Temporal templates are 2D images, constructed from image sequences, which show where and when motion in the image sequence has occurred. A two-stage learning machine, combining a k-Nearest-Neighbor (kNN) algorithm and a rule-based system, performs the recognition of 15 AUs occurring alone or in combination in an input face image sequence. Each rule utilized for recognition of a given AU (or a given AU combination) is based on the presence of a specific temporal template in a particular facial region, in which the presence of facial muscle activity characterizes the AU (or AU combination) in question. When trained and tested on the Cohn-Kanade face image database, the proposed method achieved an average recognition rate of 76.2%.

1 INTRODUCTION

Humans interact with each other far more naturally than they do with machines. This is why face-to-face interaction cannot be still substituted by human-computer interaction in spite of the theoretical feasibility of such a substitution in numerous professional areas including education and certain medical branches. In fact, existing man-machine interfaces are perceived by a broad user audience as the bottleneck in the effective utilization of the available information flow [1]. Hence, to improve man-machine interaction one should emulate the way in which humans communicate with each other.

Although speech alone is often sufficient for communicating with another person (e.g., in a phone call), considerable research in social psychology has shown that non-verbal communicative cues are essential to synchronize the dialogue, to signal comprehension or disagreement and to let the dialogue run smoother and with less interruptions [2]. The terms 'face-to-face' and 'interface' indicate that the human face has a significant role in interpersonal interactions. The face is the means to identify other members of the species, to clarify and stress what is said, to signal comprehension, disagreement and intentions [3]. Logically, automatic analysis of faces and facial expressions has numerous applications in human-computer interaction and has attracted, therefore, the interest of many AI researchers.

The majority of the existing approaches to automatic facial expression analysis focus at the recognition of few prototypic emotional facial expressions (e.g. sadness, anger or happiness) produced on command [4]. Yet such prototypic facial expressions occur relatively rarely in everyday life; emotions and attitudinal states are displayed more often by subtle changes in one or few discrete facial features such as raising the eyebrows in disbelief. Instead of being another approach to automatic detection of prototypic facial expressions of emotions, this work attempts to recognize a large range of facial behavior by recognizing facial actions (i.e. atomic facial signals) that produce expressions.

0-7803-8570-5/04/$20.00 ©2004 IEEE

Table 1. Description of Action Units as defined in FACS. The first column lists the AUs detected in our experiments, the second gives a description of the AUs.

Action Unit	Description
AU1	Raised inner eyebrow
AU2	Raised outer eyebrow
AU4	Eyebrows drawn together, lowered eyebrows
AU6	Raised cheek, compressed eyelid
AU7	Tightened eyelid
AU9	Wrinkled nose
AU11	Deepened nasolabial furrow
AU12	Lip corners pulled up
AU15	Lip corners depressed
AU17	Chin raised
AU20	Mouth stretched horizontally
AU25	Lips parted (jaws on each other)
AU26	Jaw dropped
AU27	Mouth stretched vertically (mouth wide open)

The proposed method is based upon the Facial Action Coding System (FACS) [5]. This is the best known and the most commonly used system developed for human observers to measure facial movement in terms of visually observable muscle actions. With FACS, a human observer decomposes an observed facial expression into one or more of 44 FACS-defined Action Units (AUs) that produced the expression in question.

Few efforts were reported towards automatic AU detection from face image sequences. Tian et al. [6] presented a system based upon lip tracking and template matching that recognizes 15 AUs occurring alone or in a combination in a frontal-view face image sequence. Bartlett et al. [7] reported on automatic detection of 3 AUs using Gabor filters, support vector machines and Hidden Markov Models to analyze a frontal-view face image sequence. Pantic et al. [8] reported on efforts to detect 20 AUs occurring alone or in a combination in profile-view face image sequences.

In contrast to these existing AU detectors, which can detect a certain set of AUs and none other, the method for AU detection presented here seems to represent a more general solution to automatic AU detection. Namely, if presented with a suitable data set, the proposed method could be trained to detect any arbitrary facial expression (i.e. activation of an individual AU or a set of AUs). When trained and tested on the Cohn-Kanade image database [11], used by Tian et al. [6] as well, the proposed method accomplished an average recognition rate of 76.2% for 15 AUs occurring alone or in combination in an input frontal-view face image sequence. In Table 1 a description of the detected AUs is given.

The proposed method performs AU recognition using temporal templates as input data. Section 2 elaborates on temporal templates and how we construct them from image sequences. A two-stage learning machine, combining a kNN algorithm and a rule-based system, performs the actual recognition of AUs. To score a specific AU (AU combination), it verifies whether a particular temporal template is present in a particular facial region. Section 3 provides the details of the pertinent two-stage learning machine. Finally, experimental results and concluding remarks are summarized in sections 4 and 5.

2 TEMPORAL REPRESENTATION

Bobick and Davis first introduced temporal templates [9]. They are 2D images constructed from image sequences, effectively reducing a 3D spatio-temporal space to a 2D representation. They eliminate one dimension while retaining the temporal information; the locations where movement occurred in an input image sequence are depicted in the related 2D image.

To be able to construct temporal templates we either need the background to be static or the motion of the object of interest to be separable from the background. If the temporal template is constructed

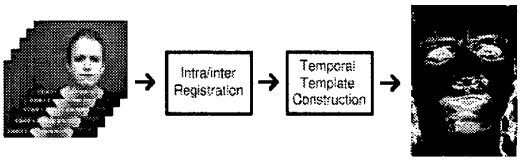

Figure 1. Creating MHI from an image sequence

without preserving the information about the time when the movement occurred, we refer to it as a Motion Energy Image (MEI). If instead we preserve the temporal motion history information by assigning different intensities to different moments of the movement, we refer to it as a Motion History Image (MHI) (Fig. 1). In our system we will use the MHIs, because we are interested in the motion history.

2.1 Face Image Sequence Registration

As already mentioned above, useful temporal templates can be constructed only if the observed background is static or if the motion of the object of interest is separable from the background.

Furthermore, to be able to compare separate temporal tem-plates, the faces in the image sequences must have the same position and orientation. Hence, to construct useful comparable temporal templates, we

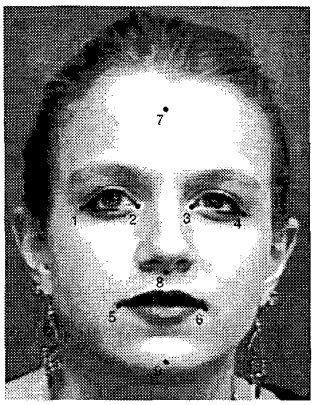

Figure 2. Manually selected facial points

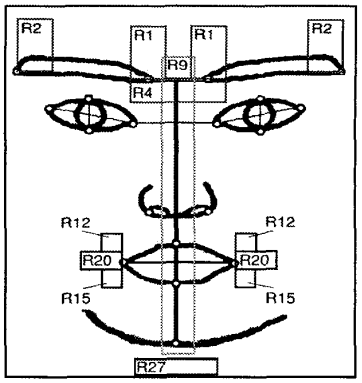

Figure 3. Facial regions for measurement of temporal template

need the input face image sequences to be registered in two ways. First, all rigid head movements within one image sequence must be eliminated. Second, all utilized image sequences must have the faces in the same position and on the same scale.

To achieve the first registration, we first select by hand 9 facial points from the first frame of the image sequence (Fig. 2). These points are then tracked in all subsequent frames using a condensation based template tracking technique [10]. The size of the template being used has an impact on the tracking performance. Experimental trials revealed that for a large window (100 x 100 pixels) the best performance is achieved. For registration of each frame with respect to the first frame we apply an affine transformation. This transformation uses facial points whose spatial position remains the same even if a facial muscle contraction occurs. Otherwise we cannot be sure whether the movement of a point is due to unwanted rigid head motion or due to the activation of AUs. We call this process intra-registration.

As already mentioned above, all image sequences must be registered with respect to a predefined set of facial points, otherwise faces in different image sequences could have different position as well as variations in size. This inter-registration process is also carried out by an affine transformation. Under the assumption that all image sequences begin with a neutral facial expression, the transformation matrix is computed by comparing the neutral position of the facial points defined for the current image sequence with the predefined position of these facial points.

2.2 Temporal Template Construction

Once properly registered, the available image sequences are used to construct temporal templates. Since we do not employ MEIs in the further AU recognition process, we are only interested in the construction of MHIs. Let I(x, y, t) be an image sequence of k frames and let D(x, y, t) be a binary image sequence indicating regions of motion, where x and y are the spatial coordinates of picture elements.

In an MHI, say H_τ, the pixel intensity is a function of the temporal history of motion at that point with τ being the period of time to be considered. Bobick and Davis' implementation of the MHI is as follows [7]:

$$H_\tau(x,y,t)=\begin{cases}\tau & D(x,y,t)=1\\ \max((H_\tau(x,y,t-1)-1),0) & otherwise\end{cases} \quad (1)$$

Bobick and Davis studied spontaneous body gestures. In their problem definition it is not known when the movement of interest begins or ends. Therefore they need to vary the observed period τ and try to classify all resulting MHIs. Because we assume that the beginning and end of a facial expression are known and coincide with the duration of an image sequence, we don't need to vary τ. Therefore we are able to normalize the temporal behaviour by distributing the grey values in the MHI over the available range (0-255, assuming that we are using 8 bit greylevel images). Thus, variations in display duration of an AU are canceled out, which makes it possible to compare facial expressions that have a different period but are otherwise identical.

Initially the image sequences may have a different number of frames. So, while the MHIs are temporally normalized, the number of history levels in them may still differ from one image sequence to another. To be able to compare the sequences properly, we want to create all MHIs having a fixed number of history

levels n. Therefore the image sequence is sampled to $n+1$ frames. The number of history levels is experimentally determined to be the number that, when used to construct the MHIs, results in the highest recognition rate. Using the known parameter n we modified the MHI operator into:

$$H(x,y,t)=\begin{cases} s*t & D(x,y,t)=1 \\ H(x,y,t-1) & otherwise \end{cases} \quad (2)$$

where $s=(255/n)$ is the intensity step between two history levels and $H(x,y,t)=0$ for $t<0$

.

3 TWO STAGE LEARNING MACHINE

Initially we used just a simple kNN learning machine to classify an input image sequence into one of m facial expression classes, each of which corresponds either to an individual AU or to an AU combination. The employed 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. In our tests we tried the Manhattan, Euclidian, Tanimoto and Minkowski distances. Experimental evaluation showed that the simple Euclidian distance measure *dist* performs the best:

$$dist=\sqrt{\sum_{i=1}^{d}(x_i-x'_i)^2} \quad (3)$$

where x is the test sample, x' is a training sample and d is the dimensionality of our sample space.

Unfortunately, applying kNN only resulted in recognition rates that were lower than what we expected (see Tables 2 and 3). Inspection of the test results revealed that some of the mistakes that the classifier made are deterministic. To exploit these deterministic mistakes we created a set of rules based on the knowledge of a human FACS coder. We defined facial regions for which the presence of motion characterizes a certain AU. For example, activation in region R2 is characteristic for the activation of AU 2 (see Fig. 3). We calculate this activation in region R_i as:

$$act(Ri)=\sum_{x,y\in Ri}H(x,y,n) \quad (4)$$

where H is the MHI operator defined in (2) and n is

Table 2. Expert rules. The first column lists the AU prediction made by the kNN learner, the second lists the rules and the third column lists the AU that is assigned to the sample if the rule fires. The R_i are the facial regions where activation is characteristic for certain AUs. The th_i are thresholds for each (part of a) rule.

KNN pred.	Rule	New value
AU4	$\frac{act(R1)}{(act(R2)+act(R4))} > th_1 \wedge act(R4) < th_2$	AU1
AU4	$act(R6) > act(R4)$	AU6
AU4	$\frac{act(R1)}{(act(R2)+act(R4))} > th_3 \wedge act(R4) > th_4$	AU1+ AU4
AU4	$act(R6)+act(R9)-act(R4) > th_5$	AU4+ AU7+ AU9
AU6	$\frac{act(R4)}{act(R6)} > th_6$	AU1+ AU4
AU6	$act(R9) > th_6$	AU4+ AU7+ AU9
AU1 +AU4	$\frac{act(R4)}{act(R1)} > th_7 \wedge act(R6)+act(R9)-act(R4) < th_8$	AU4
AU1 +AU4	$\frac{act(R4)}{act(R1)} > th_9 \wedge act(R6)+act(R9)-act(R4) > th_{10}$	AU4+ AU7+ AU9
AU25	$act(R12) > th_{11}$	AU12
AU25	$act(R15) > th_{12}$	AU12+ AU25
AU26	$\frac{act(R27)}{act(R25)} > th_{13}$	AU27
AU26	$\frac{act(R12)}{act(R27)} > th_{14}$	AU12+ AU25
AU27	$act(R27) < th_{15}$	AU26
AU12	$\frac{act(R25)}{act(R12)} > th_{16} \wedge \frac{act(R12)}{act(R15)} > th_{17}$	AU12+ AU25
AU12	$\frac{act(R25)}{act(R12)} > th_{18} \wedge \frac{act(R12)}{act(R15)} < th_{19}$	AU20+ AU25
AU12 +AU25	$act(R15) > act(R12) \wedge act(R20) > th_{20}$	AU15+ AU17
AU12 +AU25	$act(R15) > act(R12) \wedge act(R20) < th_{21}$	AU20+ AU25

the number of history levels in each MHI. Note that the activation measure assigns higher values to recent facial motion than to motion that occurs further in the past. The regions are positioned relative to the same facial points we used for the registration of the image sequences. Using these regions we were able to construct a set of rules, which are based on the activation values of facial regions typical for a certain AU. With these rules we can correctly reclassify test

Table 3. Confusion matrix for lower-face AUs. Numbers in parentheses are results for kNN only.

Real label	AUs	Predicted label 25	26	27	12	17	15+17	12+25	20+25	11+20+25
	25	**7(7)**	3(3)	0(0)	0(1)	0(0)	0(0)	1(1)	1(0)	0(0)
	26	2(2)	**11(11)**	4(4)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
	27	0(0)	4(5)	**15(14)**	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
	12	0(1)	0(0)	0(0)	**6(6)**	0(0)	0(0)	4(3)	0(0)	0(0)
	17	1(1)	0(0)	0(0)	0(1)	**5(5)**	1(1)	0(0)	1(0)	2(2)
	15+17	1(3)	1(1)	0(0)	0(0)	1(1)	**14(12)**	0(0)	0(0)	0(0)
	12+25	0(0)	1(1)	0(0)	1(2)	0(0)	0(0)	**20(19)**	0(0)	0(0)
	20+25	0(0)	1(1)	1(1)	0(2)	0(0)	1(1)	1(1)	**5(3)**	0(0)
	11+20+25	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	2(2)	0(0)	**1(1)**
Recognition rate: 70.6% (65.5%)										

Table 4. Confusion matrix for upper-face AUs. Numbers in parentheses are results for kNN only.

Real label	AUs	Predicted label 1	4	6	1+2	1+4	4+7+9
	1	**2(0)**	0(2)	0(0)	0(0)	1(1)	0(0)
	4	1(0)	**5(4)**	1(1)	0(0)	1(3)	0(0)
	6	0(0)	1(1)	**24(24)**	1(1)	0(0)	0(0)
	1+2	0(0)	1(1)	0(0)	**6(6)**	1(2)	0(0)
	1+4	0(0)	0(0)	0(2)	1(1)	**5(3)**	0(0)
	4+7+9	0(0)	1(0)	0(0)	0(0)	0(2)	**3(2)**
Recognition rate: 81.8% (69.6%)							

samples that were at first misclassified by the kNN learner. For example, the kNN learning machine 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 of the eyebrows followed by an inward and downward movement. This results in high activation between the brows and relatively low activation above the inner corners of the brows. Fig. 3 shows the defined facial regions and Table 2 lists the rules we applied to our system.

For each data set, the values of the thresholds th_i are determined automatically during the training phase. Their values are set to the maximum or minimum activation of the training samples of the kNN-predicted class (depending on the sign: >, respectively <, see Table 2). This reduces the probability that test samples that were correctly classified by the kNN classifier, get misclassified in the second stage. For example, suppose we know that AUx and y are often confused and that high activation in region R_j indicates AUy. If the kNN-classifier decides AUx for test sample i, we will only reclassify i if the activation of region R_j is greater than the maximum activation of the training samples labeled as AUx.

4 EXPERIMENTAL EVALUATION

The database used in experimental studies on our system is the Cohn-Kanade AU-Coded Facial Expression Image Database [11]. The database consists of video's of facial expressions, made by 138 subjects. Each recording contains one combination of Aus. We used the pertinent imagery to recognize 9 lower-face AU combinations (Table 3) and 6 upper-face AUs (Table 4). We did so by training two different learning machines: one for the upper-face AUs and one for the lower-face AUs.

The parameter k of the kNN algorithm is an important parameter affecting the recognition rate. Setting k = 4 for the upper-face AU recognition and k = 5 for the lower-face AU recognition resulted in the highest recognition rates.

Tables 2 and 3 show the confusion matrices of upper and lower face AU detection. As can be seen, the algorithm using kNN only confuses the class containing AU4 with AU1+AU4. The second stage correctly reclassifies two out of three of these confusions.

However, we are not able to solve all confusions using the rule-base technique. Table 3 shows that our system confusesAUs 25, 26 and 27. For AU 25, the lips must be parted. For AU26 the jaw must be slightly dropped. For AU27 the jaw is dropped low and the mouth is stretched vertically. However, sometimes the difference is difficult to see and even human FACS coders have trouble distinguishing between these AUs.

Also, multiple demonstrations of our system have been held. Using a simple webcam and no alterations to the lighting condition, in all occasions the system performed as expected, although no recognition rates have been recorded.

5 CONCLUSIONS

This paper presents a method for the automatic recognition of facial action units (AUs) using temporal

templates. It proposes a two-stage classifier, which at the first stage consists of a general kNN classification scheme and at the second stage uses domain specific knowledge in a rule-based system. We have applied our method to real image sequences from the Cohn-Kanade database and obtained a recognition rate of 70.6% for lower face AUs and a recognition rate of 81.8% for upper face AUs.

For future research, we would consider representations of the image sequences by features that can be extracted from the temporal templates. In particular, we will investigate on features that can describe the motion density and motion direction. Furthermore, special consideration should be given to the appropriate modeling of the temporal dynamics of the extracted features and their interdependencies. To this direction, further research with. Hidden Markov Models or Dynamic Bayesian Networks is needed.

Another approach is to further exploit the temporal dynamics of MHIs by introducing Multilevel Motion History Images (MMHIs), which overcome the problem of self-occlusion inherent to normal MHIs. This would give a better representation of the order and speed in which the facial motion occurs and would also allow us to use a better definition for the facial region activation value (equation 4).

Finally, another issue is the limitations imposed by the absence of a sufficient number of training samples for each AU (or for each combination of AUs). Training and testing in larger databases and addressing the issues related to combinations of AUs are therefore directions that we should consider.

ACKNOWLEDGMENT

The authors would like to thank Jeffrey Cohn of the University of Pittsburgh for providing the Cohn-Kanade database. This work has been supported by the Netherlands Organization for Scientific Research (NWO) Grant EW-639.021.202.

REFERENCES

[1] B. Schneiderman, "Universal Usability", *Communications of the ACM*, Vol. 43, No. 5, pp. 85-91, 2000

[2] E. Boyle, A.H. Anderson and A.Newlands, "The effect of visibility on dialogue and performance in a co-operative problem solving task", *Language and speech*, Vol. 37, no. 1, pp 1-20, 1994

[3] J.A. Russell and J.M. Fernez-Dols, Eds., *The Psychology of Facial Expression*, vol. 9, no. 3, pp. 185-211, 1990

[4] M. Pantic and L.J.M. Rothkrantz, "Toward an Affect-Sensitive Multimodal Human-Computer Interaction", IEEE proceedings vol. 91, no. 9, pp. 1370-1390, 2003

[5] P.Ekman and W.V. Friesen, "The Facial Action Coding System: A Technique for the Measurement of Facial Movement", San Francisco: Consulting Psychologist, 1976

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

[7] M. Bartlett et al., "Measuring Facial Expressions by Computer Image Analysis", *Psychophysiology,* vol 36, pp253-264, 1999.

[8] M. Pantic, I. Patras and L.J.M. Rothkrantz, 'Facial action recognition in face profile image sequences', in *Proc. IEEE Int'l Conf. on Multimedia and Expo*, vol. 1, pp. 37-40, Lausanne, Switzerland, August 2002

[9] A.F. Bobick and J.W. Davis, "The Recognition of Human Movement Using Temporal Templates", *IEEE trans. on pattern analysis and machine intelligence vol 23*, 2001.

[10] M. Isard and A. Blake, "Condensation – Conditional Density Propagation for Visual Tracking", *Int. J. Computer Vision,* 1998.

[11] T. Kanade, J. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis", *Proc. IEEE Int'l Conf. Face and Gesture Recognition*, pp. 46-53, 2000.