Motion History for Facial Action Detection in Video

Michel Valstar, Maja Pantic and Ioannis Patras
Delft University of Technology
Electrical Engineering, Mathematics and Computer Science
Man-Machine Interaction Group, Delft, the Netherlands
{M.F.Valstar,M.Pantic,I.Patras}@ewi.tudelft.nl

Abstract - *Enabling computer systems to recognize human facial expressions is a challenging research problem with many applications in behavioral science, medicine, security, and human-machine interaction. Instead of being another approach to automatic detection of prototypic facial expressions of emotion, this work attempts to analyze subtle changes in facial behavior by recognizing facial action units (AUs, i.e. atomic facial signals) that produce expressions. This paper proposes AU recognition based upon multilevel motion history images (MMHIs), which can be seen as an extension to temporal templates introduced by Bobick and Davis. By recording motion history at multiple time intervals (i.e., multilevel MHIs) instead of recording it once for the entire image sequence, we overcome the problem of self-occlusion which is inherent to temporal templates original definition. For automatic classification of an input MMHI-represented face video in terms of 21 AU classes, two approaches are compared: a Sparse Network of Winnows (SNoW) and a standard k-Nearest Neighbour (kNN) classifier. The system was tested on two different databases, the MMI-Face-DB developed by the authors and the Cohn-Kanade face database***.***

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 effectively, 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]. Of all different non-verbal communication means (body gesture, posture, touch), the facial expression is the most important means for interpersonal communication. It is the means to clarify what is said by means of lip-reading, to stress the importance of the spoken message by means of conversational signals like raising the eyebrows and to signal comprehension, disagreement, boredom etc. [3].

The majority of the existing approaches to automatic facial expression analysis focuses at the recognition of few prototypic emotional facial expressions such as happiness, surprise or anger [4], [5]. Yet such prototypic facial displays do not occur frequently in interpersonal communication [3]. To facilitate automatic analysis of subtler facial expressions, automatic detection of atomic facial signals (facial muscle actions), which when combined produce facial expressions, should be enabled first.

The method proposed here is based on the Facial Action Coding System (FACS) [6]. This is the best known and the most commonly used system developed for human observers to describe facial activity in terms of visually observable facial muscle actions (i.e., facial action units, AUs). With FACS, a human observer decomposes a shown facial expression into one or more of in total 44 AUs that produced the expression in question.

Few efforts were reported toward automatic AU detection from face image sequences. Tian et al. [4] presented a system based upon lip tracking and template matching that recognizes 15 AUs occurring alone or in 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 color and motion based detection of 20 AUs occurring alone or in combination in profile-view face video. Our previous work [9] reports on detecting 15 AUs and AU combinations by using temporal templates [10] generated from input face video and a two-stage classifier combining a kNN-based and a rule-based classifier.

Temporal templates are 2D images, constructed from image sequences, which show motion history, that is, where and when motion in the image sequence has occurred. A drawback innate to temporal templates proposed originally by Bobick and Davis [10] is the problem of motion self-occlusion due to overwriting. Let us explain this problem by giving an example. Let us denote an upward movement of the eyebrows as action A_1 and a downward movement of the eyebrows back to the neutral position as action A_2. Both actions produce apparent motion in the facial region above

0-7803-8566-7/04/$20.00 © 2004 IEEE.

Figure 1. Motion History Images (MHIs) of activated Action Units AU1+AU2 (left), AU16 (mid) and AU36L (right)

the neutral position of the eyebrows (Fig. 1, left). If A_2 follows A_1 in time and if the motion history of both actions is recorded within a single Motion History Image (MHI), then the motion history of action A_2 overwrites the motion history of A_1; the information about the motion history of action A_1 is lost. To overcome this problem, we propose in this paper to record the motion history at multiple time intervals and to construct Multilevel Motion History Image (MMHI), instead of recording the motion history once for the entire image sequence and constructing a single MHI.

In this paper, we examine further whether and to which extent are temporal templates applicable for AU detection in face video. For automatic detection of 21 AUs from an input (M)MHI-represented face video, we compare two approaches: a Sparse Network of Winnows (SNoW) classifier and a two-stage classifier combining a kNN-based and a rule-based classifier. The evaluation of these two approaches on two different databases, the MMI-Face-DB developed by the authors and the Cohn-Kanade face database [11], suggests that the approaches perform as well as humans in AU detection tasks.

2. Temporal Templates

Temporal templates are 2D images constructed from image sequences, effectively reducing a 3D spatio-temporal space to a 2D representation [10]. 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 camera and the background to be static or the motion of the object of interest to be separable from the motion induced by camera- and by background movements. If a temporal template is constructed without preserving the information about the time instance in which the movement occurred, we refer to it as to a Motion Energy Image (MEI). If, instead, we preserve the temporal information (motion history) by assigning different intensities to different moments of the movement, we refer to the resulting temporal template as to a Motion History Image (Fig. 1).

A drawback inherent to the originally proposed temporal template approach [10] is the problem of motion self-occlusion due to overwriting. As already explained above, if a motion on a location χ occurs at time instance *t1* and at time instance *t2* > *t1*, the recent motion (time instance *t2*) will overwrite the previously encountered motion (time instance *t1*). A way of dealing with this problem is to construct Multilevel Motion History Image (MMHI). Namely, instead of recording the motion history once for the entire image sequence (single MHI), the motion history is to be recorded at multiple time intervals (multilevel MHI).

2.1 Face Image Sequence Registration

In order to enable the construction of meaningful temporal templates that visualize only the motion of interest and that are comparable with each other, the faces in input image sequences must have the same position and orientation. In other words, input face image sequences must 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 manually 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 [12]. 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 (i.e., points 2, 3, and 8 illustrated in Fig. 2). For other points we cannot resolve whether the encountered movement of a point is due to a rigid head motion that we want to eliminate or due to the activation of an AU which we want to recognize. We call this process intra-registration.

As already mentioned above, all utilized image sequences must have the faces at the same position and on the same scale. This inter-registration process is also carried out by an affine transformation. The transformation matrix is computed by comparing the neutral position of the 9 facial points defined for the current image sequence with the positions of the same 9 facial points on a selected 'normal' face. This way all faces are normalized to the scale and position of this base face.

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 will explain only the construction process of MHIs and MMHIs. 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 to images for change detection. 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.

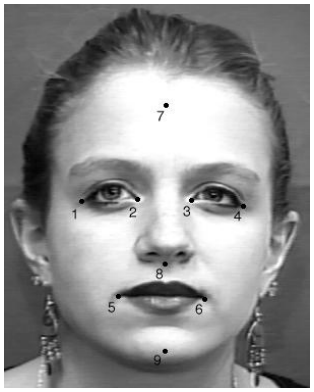

Figure 2. Manually selected points

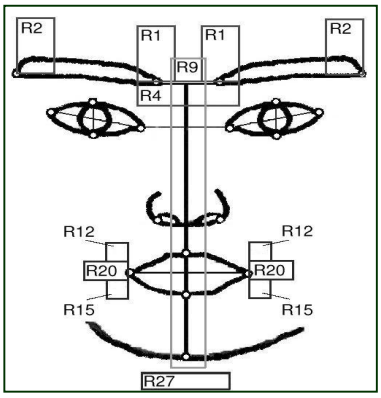

Figure 3. Facial regions for determining temporal templates activity

The implementation of the MHI is as follows [10]:

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

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 needed to vary the observed period τ and to try to classify all resulting MHIs. Because we assume that the beginning and the end of a facial expression are known and that they coincide with the duration of an image sequence, we do not need to vary τ; we are able to normalize the temporal behavior by distributing the gray values in the MHI over the available range (0-255, assuming that we are using 8 bit gray level images). In turn, we are able to cancel out variations in display duration of an AU which makes it possible to compare facial expressions that have a different period but are otherwise identical.

Initially, the input image sequences may have different numbers 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 such that they have a fixed number of history levels *n*. Therefore each image sequence is sampled to *n+1* frames. Using the known parameter *n* we modify 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 $t\le 0$. We have varied *n* between 3 and 9 and *th* between 0.15 and 0.20. In section IV we present our results with the values of *n* and *th* optimized for achieving the highest possible recognition rates.

With an MMHI, we want to encode motion occurring at different time instances on the same location such that it is uniquely decodable later on. To do so, we use a simple bit-wise coding scheme. If motion occurs at time instance *t* at position *(x, y)*, we add 2 to the power of (*t*-1) to the old value of the MMHI:

$$M(x,y,t)=M(x,y,t-1)+D(x,y,t)\cdot 2^{t-1} \quad (3)$$

with $M(x,y,t)=0$ for $t\le 0$. Because of the bitwise coding scheme, we are able to separate multiple motions occurring at the same position in the classification stage.

3 Classification Schemes

For AU automatic detection from (M)MHI-represented face image sequences, we compare two classification schemes: (i) a two-stage classifier combining a kNN-based and a rule-based classifier, and (ii) a SNoW classifier. A preliminary version of the first classification scheme has been presented in [9]. The second classification method has been originally proposed in [13]. It is a sparse network of linear functions over a pre-defined or incrementally learned feature space. It is a multi-class classifier specifically tailored to large-scale classification problems with a very large number of features which seems, therefore, suitable for our problem.

3.1 Combined kNN/rule-based classifier

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 the case of MHI-based data representation, both *k* and the distance measure were experimentally determined [9]. The distance measure $dist_{MHI}$ that performed the best was the simple Euclidian distance measure:

$$dist_{mhi}(x',x)=\sqrt{\sum_{i=1}^{d}\left(x_i-x'_i\right)^2} \quad (4)$$

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

In the case of MMHI-based data representation, the distance measure $dist_{MMHI}$ has been calculated as follows. Let us denote the current input sample with *S'* and the sample with which it is compared with *S*. Lets denote

further the *j*-th feature (pixel) of sample *S'* with S'_j, its corresponding MMHI representation as $M(S'_j)$, the *j*-th feature (pixel) of sample *S* with S_j, and its MMHI representation as $M(S_j)$, where

$$M(S'_j) = 2^{a_1} + 2^{a_2} + ... + 2^{a_l},$$
$$M(S_j) = 2^{b_1} + 2^{b_2} + ... + 2^{b_m}$$

such that $A_j = \{a_1, a_2, \ldots, a_n\}$ is the set of active history levels of the *j*-th feature of sample *S'* and $B_j = \{b_1, b_2, \ldots, b_n\}$ is the set of active history levels of the *j*-th feature of sample *S*. The distance measure $dist_{MMHI}$ is defined as:

$$dist_{mmhi}(S', S) = \frac{1}{2d} * \sum_{j=1}^{d} \left[\sum_{i=1}^{l} \min \left| a_{ij} - b_{kj} \right| + \sum_{i=1}^{m} \min \left| b_{ij} - a_{kj} \right| \right] \quad (5)$$

where *d* is the dimensionality of the feature space.

Though it gives a good indication about the AUs shown in a given sample, the kNN algorithm confuses commonly AUs that have partially the same (M)MHI. To address this drawback, we created a set of rules based on the knowledge of a human FACS coder. We defined facial regions in which the presence of motion characterizes certain AU activation. For example, the presence of motion in region R2 (Fig. 3) is characteristic for the activation of AU2. We calculate this activity in facial region *Ri* as follows:

$$act(Ri) = \begin{cases} \frac{1}{N} \sum_{x,y \in Ri} \left\lceil \frac{H(x,y,n)}{255} \right\rceil & MHI - data \\ \frac{1}{N} \sum_{j \in Ri} \left\lceil \frac{|A_j|}{n} \right\rceil & MMHI - data \end{cases} \quad (6)$$

where *H* is the MHI operator defined in *(2)*, *n* is the number of history levels in each (M)MHI, $|A_j|$ the cardinality of active history levels of the *j*-th pixel in a (M)MHI and *N* the number of pixels in the facial region *Ri*. The facial regions are positioned relative to the same facial points that we used for the registration of image sequences. Using these regions we built a set of rules, which are based on the activation values in facial regions being typical for certain AU activation. With these rules we can correctly reclassify samples that the kNN algorithm misclassified at first. For example, the kNN classifier often confuses AU4 and AU1+AU4. Both produce activity in the same part of the MHI (in regions R1 and R4 illustrated in Fig. 3), 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. Hence, the rules used to resolve the confusion in question are defined as follows. If the kNN classifier encodes AU4 and it is true that

$$\frac{act(R1)}{(act(R2) + act(R4))} > th_3 \wedge act(R4) > th_4,$$

where *act(Ri)* is the activity in facial region *Ri* defined in *(6)* and th_j are thresholds that are automatically defined during the training phase [9], then AU1+AU4 will be the final classification of the pertinent input sample. Otherwise, AU4 will be the final classification of the pertinent input sample. Similarly, if the kNN classifier encodes AU1+AU4 and it is true that

$$\frac{act(R4)}{act(R1)} > th_7 \wedge act(R6) + act(R9) - act(R4) < th_8,$$

then AU4 will be the final classification of the input sample in question. Otherwise, AU1+AU4 will be the final output of the two-stage classifier explained here. For a complete list of the utilized rules, the reader is referred to [9].

3.2 SNoW classifier

A Sparse Network of Winnows (SNoW) is an information processing structure that consists of an input layer of nodes and an output layer of target nodes. It learns a sparse network of linear functions in which the target concepts (class labels) are represented as linear functions over a common feature space [13].

The SNOW classifier that we employ for AU detection uses one target node for each AU to be detected. The data fed to the input layer of the utilized SNOW is a set of features extracted from the specific facial regions (see Fig. 3) of an input (M)MHI. These features have binary values (active or inactive). Namely, if the pixel at position (*x,y*) is the *i*-th out of *d* pixels that form the facial regions depicted in Fig. 3, and if it has active history levels $A_j = \{a_1, a_2, \ldots, a_l\}$, then the set of active features for the pixel in question can be defined as:

$$\mathcal{F}(x,y) = \{(i-1)*n + a_1, (i-1)*n + a_2, ..., (i-1)*n + a_l\} \quad (7)$$

where *n* is the number of history levels of the input (M)MHI. The total set of features forming the input to the utilized SNoW is the union of the active features for all the pixels of the facial regions depicted in Fig. 3. Target nodes are linked further via weighted connections to (some of the) input nodes. If $L_t = \{i, i, \ldots, i_m\}$ is the set of active input features that are linked to target node *t*, then *t* is *active* if and only if $\sum_{i \in Lt} w_i^t > \theta_t$, where θ_t is a threshold and w_i^t is the weight associated with the connection of the *i*-th feature to target node *t*. Different update rules, Winnow, Perceptron, and naive Bayes, can be used within a SNoW. The SNoW learning architecture inherits its generalization properties from the particular update rule that is being used. When using Winnow, it is a feature efficient learning algorithm, in that it scales linearly with the number of relevant features, and linearly with the number of features active in the domain. The SNOW classifier that we employ for AU detection uses Winnow update rule which, except

Table 1. Recognition rates for the MMI-Face-DB. The second column shows the number of samples with the specified action unit the database contains.

Action Units	nr	MHI		MMHI	
		Rec. rate	False positive	Rec. rate	False positive
1 + 2	10	0.90	3	0.90	3
2	6	0.50	0	0.50	2
1 + 4	6	0.50	6	0.50	3
4	12	0.67	6	0.75	10
6	10	0.70	1	0.70	1
9	11	0.82	0	0.45	0
8 + 25	10	0.60	4	0.40	3
10 + 25	10	0.90	1	0.80	1
11 + 25	10	0.70	1	0.70	1
12 + 25	10	1.00	10	0.80	9
14	11	0.27	4	0.18	3
15	8	0.37	9	0.75	28
16 + 25	10	0.60	8	0.50	16
17	10	0.60	3	0.70	6
18	10	0.70	4	0.60	5
20	10	0.60	3	0.60	1
22 + 25	10	0.60	3	0.40	0
25	9	0.33	10	0.33	5
25 + 26	11	0.36	11	0.46	11
27	10	0.80	0	0.70	0
26 + 30L	9	0.33	1	0.33	1
26 + 30R	9	0.67	3	0.67	2
26 + 36T	12	0.50	1	0.42	2
26+36B	10	0.60	2	0.30	1
26+36L	9	0.56	6	0.44	5
26+36R	10	0.40	6	0.60	2
Total:	**253**	**0.61**	**106**	**0.56**	**121**

of the threshold θ_t, utilizes two other update parameters: a promotion parameter $\alpha > 1$ and a demotion parameter $0 < \beta < 1$. These are being used to adjust the weights of the links when a mistake is made. If the system produces a false negative prediction during the training, all active weights are promoted by multiplication with α. Similarly, when the system produces a false positive prediction, all active weights are demoted by multiplication with β.

4 Experimental results

We evaluated the performance of the two AU recognition schemes described above using two different databases: the Cohn-Kanade face database [11] and the MMI-Face-DB developed by the authors.

4.1 Cohn-Kanade face database

The Cohn-Kanade face database contains over 2000 videos of the facial displays produced by 210 adults being 18 to 50 years old, 69% female, 81% Caucasian, 13% African and 6% from other ethnic groups. All facial displays were made on command and the recordings were made under constant lighting conditions. Only real expressions were recorded, which means that many AUs never occur alone. Many recordings contain the date/time stamp recorded over parts of the face. This occurrence is unwanted, for it causes (M)MHI activation which is of course unwanted.

Table 2. Recognition rates for the Cohn-Kanade database. The second column shows the number of samples with the specified action unit the database contains.

Action Units	nr	MHI		MMHI	
		Rec. rate	False positive	Rec. rate	False positive
1 + 2	21	0.52	2	0.45	3
1 + 4	21	0.61	12	0.67	22
4	21	0.38	10	0.29	7
6	61	1.00	8	0.89	8
12 + 25	59	0.63	10	0.31	4
12	27	0.78	26	0.54	32
15	4	0.25	10	0	2
17	26	0.54	5	0.46	16
20 + 25	18	0.11	1	0.11	5
25	42	0.64	10	0.88	37
27	42	0.93	5	0.78	3
Total:	**344**	**0.68**	**110**	**0.58**	**139**

4.2 MMI-Face-DB

Mainly due to the lack of facial displays depicting the activation of individual AUs in the Cohn-Kanade database, the authors of this paper decided to develop a new face database: the MMI-Face-DB. It consists of over 4000 videos and over 600 static images depicting the facial displays of 31 adults being 19 to 35 years old; 50% female, 81% being Caucasian, 14% Asian and 5% African. All facial displays were made on command and the recordings were made under constant lighting conditions. All but seven facial expression videos were recorded in profile and frontal view simultaneously (using a mirror). Two FACS experts coded the database. When in doubt, decisions were made by consensus.

The database contains a large amount of videos where the activation of individual AUs has been recorded. In the cases where this was not possible, expressions produced by the activation of the least possible number of AUs were recorded. For example, in AU16 (lower lip depressed): depressing the lower lip automatically parts the lips causing AU25 (lips parted) to be activated as well.

4.3 Evaluation results

The experiments using the kNN rule-based classifier have been applied to both MHI and MMHI represented data constructed from image sequences that were scaled down by a factor 4 in both the height and the width. This scaling has been done to increase the detection speed. The optimal kNN parameter k and (M)MHI constructor threshold th were experimentally determined to be $k = 3$ and $th = 0.19$. Table 1 shows the results of tests performed using the MMI-Face-DB. Results for MMHI are, overall, lower than they are for MHI. This is because of the definition of MMHI. The distance measure used makes it difficult to find the desired nearest neighbor of a sample when multiple levels are activated on the same position in a MMHI. Furthermore, as we do not have any examples displaying possible confusion caused by motion self-occlusion, we were not able to show the increased resolution of MMHI

Table 3. SNoW detection results. Column 2 lists the number of positive samples in the dataset.

Action Unit	nr	Recognition rate positive samples	Recognition rate negative samples
1	34	0.85	0.85
2	22	0.64	0.97
4	37	0.76	0.85
6	18	0.5	0.89
12	21	0.67	0.94
17	36	0.33	0.85
25	49	0.94	0.23
26	4	0.25	0.94
27	7	0.29	1.00
Total:	**228**	**0.58**	**0.84**

with respect to MHI in occasions where confusion caused by motion self-occlusion does occur.

The test results for classification using the kNN rule-based algorithm on the Cohn-Kanade database are shown in Table 2. Unfortunately, as already mentioned above, the Cohn-Kanade database has fewer AUs occurring alone, resulting in fewer AUs that can be recognized. The recognition rates for this database are somewhat higher. This is probably due to a larger number of samples per AU and fewer target classes providing less confusion possibilities.

For both datasets, all confusions of AU 11 (upper lip raised and deepened nasolabial furrows) were made with AU10 (upper lip raised). This is no surprise, however, as the facial changes produced by these AUs are very similar. There are many combinations of AUs with AU25 (for example AU16 + AU25), causing rather low recognition rates for this AU.

Finally, in Table 3 we show the result for detecting selected AUs in MHI data using the SNoW classifier. For each AU to detect a binary SNoW net is trained using an equal number of positive as negative samples. Each trained SNoW net is evaluated on the whole dataset on a leave-one-out basis. The MHI data was constructed from the Cohn-Kanade database; samples containing individual AUs as well as samples containing more than one AU were allowed in the training and test sets. The number of examples the SNoW algorithm requires to learn a linear function grows linearly with the number of relevant features. But although we downscaled our images and applied feature selection as described in section 3.2, the results suggest that there still are to many relevant features present in our sample MHIs for the number of samples in the dataset; AUs with little positive samples do not have good recognition rates for the detection of positively labeled samples.

5. Conclusions

This paper presented an evaluation of the use of Motion History Images in the field of Facial Action detection and suggested the use of Multilevel Motion History Images.

The test results show clearly that (M)MHIs are very suitable for detecting various AUs. Especially the AUs AU1+AU2 (eyebrows raised), AU10+AU25 (raised upper lip), AU12+AU25 (smile with lips parted) and AU27 (mouth stretched vertically) are easily recognized.

Though classification using MMHIs resulted in lower recognition rates than classification using conventional MHIs, we believe that MMHI data representation can offer many benefits in applications where confusions caused by motion self-occlusion are common (e.g., in hand gesture recognition, waving the hand is often confused with moving the hand from left to right only).

In order to gain better results using neural network classifiers such as the SNoW architecture, we plan to combine the Cohn-Kanade database and the MMI-face-DB. Together with a feature selection mechanism this should produce higher classification rates.

In future work we want to examine how sensitive the system is for registration errors. We also want to compare temporal templates with other features, such as optical flow.

Acknowledgements

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 effects 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] Y. Tian, T. Kanade, and J.F. Cohn, "Recognizing action units for facial expression analysis", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 97-115, 2001.

[5] M. Pantic and L.J.M. Rothkrantz, "Toward an Affect-Sensitive Multimodal Human-computer Interaction", *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1370-1390, 2003.

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

[7] M. Bartlett, J.C. Hager, P. Ekman, and T.J. Sejnowski, "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", *Proc. IEEE Int'l Conf. Multimedia and Expo*, vol. 1, pp. 37-40, Lausanne, Switzerland, August 2002.

[9] M.F. Valstar, I. Patras, and M. Pantic, "Facial action unit recognition using temporal templates", *Proc. IEEE Int'l Workshop on Human Robot Interactive Communication*, September 2004, accepted for publication.

[10] A.F. Bobick and J.W. Davis, "The Recognition of Human Movement Using Temporal Templates", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257-267, 2001.

[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.

[12] M. Isard and A. Blake, "Condensation – Conditional Density Propagation for Visual Tracking", *Int. J. Computer Vision*, vol 29, pp 5-28, 1998.

[13] M-H. Yang, D. Roth and N. Ahuja, "A SNoW-Based Face Detector", *Advances in Neural information Processing Systems*, vol. 12, pp. 855-861, 2000