The use of Active Appearance Model for facial expression recognition in crisis environments

D. Datcu
D.Datcu@ewi.tudelft.nl

L.J.M. Rothkrantz
L.J.M.Rothkrantz@ewi.tudelft.nl

Man-Machine Interaction Group
Delft University of Technology
2628 CD, Delft,
The Netherlands

ABSTRACT

In the past a crisis event was notified by local witnesses that use to make phone calls to the special services. They reported by speech according to their observation on the crisis site. The recent improvements in the area of human computer interfaces make possible the development of context-aware systems for crisis management that support people in escaping a crisis even before external help is available at site. Apart from collecting the people's reports on the crisis, these systems are assumed to automatically extract useful clues during typical human computer interaction sessions. The novelty of the current research resides in the attempt to involve computer vision techniques for performing an automatic evaluation of facial expressions during human-computer interaction sessions with a crisis management system. The current paper details an approach for an automatic facial expression recognition module that may be included in crisis-oriented applications. The algorithm uses Active Appearance Model for facial shape extraction and SVM classifier for Action Units detection and facial expression recognition.

Keywords

Facial expression recognition, Active Appearance Models, face detection, crisis management systems.

INTRODUCTION

For the past decades, many projects have been started with the purpose of learning the machines to recognize human faces and facial expressions. The need to extract information from images is enormous. Computer vision has become one of the most challenging subjects nowadays. In the same time, human computer interfaces tend to become more and more user oriented. The recent improvements in this area make possible the development of context-aware systems for crisis management. These systems are assumed to automatically extract useful clues during typical human computer interaction sessions. Among other communication channels, the visual channel plays an important role in the attempt to assess the emotional state of the user in the crisis environment. Because the system is designed to be working in crisis situations, there is a requirement to have efficient mechanisms to detect certain emotional user states as stress, anger, fear. Such information can help other system components to increase the confidence of their results/decisions given certain user and context conditions. The need for automatic mechanisms that collect emotion-oriented information related to the actors at a crisis site comes from the difficulty on analyzing the rather chaotic and unpredictable behavior of the people during major crisis events and by the lack of their ability to make clear statements on their observations. It is of a great importance to extract as much information as possible and to try to disambiguate the partial reports coming from human actors at the crisis environments.

Face analysis as a computer-vision task has many applications and has direct relevance to the face-recognition and facial expression recognition problem. Such techniques work at the lowest level of data abstraction and the quality of the information provided by them is essential for further information processing components working within a specialized system. Beside other types of information provided by other system components, the results contribute at the general awareness of the system over the crisis event. More exactly it helps at filtering the incoming information provided by end-user devices such as PDAs, PCs, automatic cameras and microphones.

The current paper details a robust approach for an automatic facial expression recognition module for video sequences that may be included in crisis-oriented applications. The solution splits the system into three components: the Viola&Jones face detector (Viola and Jones 2001), the Facial Characteristic Point - FCP extractor based on Active Appearance Model (Cootes et al. 1998), and the tracker and classifier of temporal emotional patters. Additionally an intermediate procedure for the detection of Action Units (Ekman and Friesen 1978), links the FCP based model to the recognition of facial expressions. The methods presented in the paper are optimized and can be integrated as real-time data processing components in an automatic emotion recognition framework for crisis management applications. Finally, we report recognition rates for the six universal facial expressions based on a range of experiments.

RELATED WORK

The online Facial Expression Dictionary (FED) is an ongoing project at the Man-Machine Interaction group of the TU Delft (de Jongh 2002). The goal of the project is to develop a non-verbal dictionary which contains information about non-verbal communication of people. Resembling a verbal dictionary, instead of words the FED contains facial expressions. An improved, automatic, version of FED is presented in (Datcu and Rothkrantz, 2007). One method for the analysis is the internal representation of facial expressions based on collections of Action Units (AU) as defined in Facial Action Coding System (FACS) by (Ekman and Friesen 1978). It is one of the most efficient and commonly used methodologies to handle facial expressions. Recent researches making use of this technique are (Bartlett et al. 2004) and (Datcu and Rothkrantz, 2004). Some attempts to automatically detect the salient facial features implied computing descriptors such as scale-normalized Gaussian derivatives at each pixel of the facial image and performing some linear combinations on their values. It was found that a single cluster of Gaussian derivative responses leads to a high robustness of detection given the pose, illumination and identity (Gourier et al. 2004). A representation based on topological labels is proposed in (Yin et al. 2004). The paper of (Saatci and Town 2006) presents an approach to determine the gender and expression of faces by using Active Appearance Model. (Wilhelm et al. 2005) presents a comparison between the use of ICA and AAM with different classifiers for the recognition of facial expressions, age, gender and identity. Four emotional states were employed for the analysis that was realized by using SVM classifier. The work shows an improvement of the recognition results by involving a first classification of the gender. The work of (Zhou et al. 2003) proposes a Bayesian inference solution based on tangent shape approximation constructed in the form of Bayesian Tangent Shape Model. The work of (Lee and Elgammal 2006) presents a novel nonlinear generative model using conceptual manifold embedding and empirical kernel maps for facial expressions. The algorithm deals with the complex nonlinear deformations of the shape and appearance in facial expressions and provides accurate emotional based synthesis. (Liebelt et al. 2006) develop an interactive multi-level algorithm for AAM fitting to 2D images and 3D shape alignment to disparity data. (Antonini et al. 2006) makes use of the active appearance model to derive a set of high level features defined as expression descriptive units (EDU) and compare the performance of different classifiers for the recognition of the six universal facial expressions.

OVERVIEW OF THE SYSTEM

In the case of recognizing the facial expressions in a video sequence, a fixed length, non-overlapping time window forms the basis of input data for the subsequent processing steps. Initially the analysis window is positioned at the beginning of the video sequence. The Viola&Jones face detector component is run for each video frame from the given time window. The next Active Appearance Model based component aims at the retrieval of the shape from all the detected faces. The Face Characteristic Point FCP detection component tries to identify the locations of specific landmarks on the face area. Certain parameters as distances among these points are used as features for recognizing human emotions. The information concerning the behavior of the facial movements is connected with Facial Action Coding System – FACS. It is well known in the area of psychology and in the current research it is used for modeling different emotions through the activation of different facial muscles. The last component uses Support Vector Machines as classifier for the recognition of facial expressions (Figure 1).

The development of the facial expression recognition system has two steps. The first step assumes the training of the classifiers involved in the processes of detecting faces in video frames, localizing FCPs on the face area and of recognizing the facial expressions. A special attention is carried out on the selection of data samples for training each classifier. The second step is the testing of the classifiers for performing specific tasks. This corresponds with the functionality of the final system in a real life application.

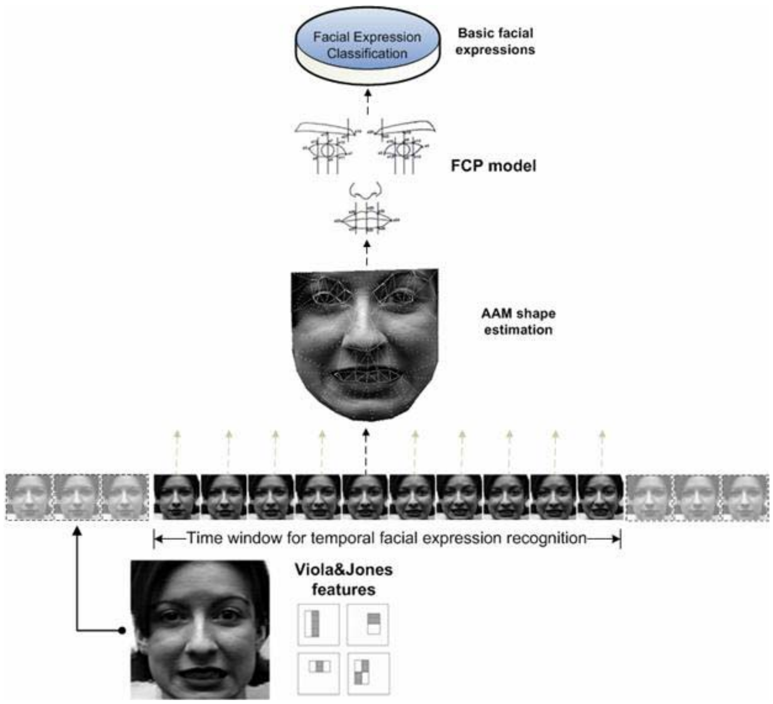

Figure 1 The processing steps in temporal facial expression recognition

VIOLA&JONES FACE DETECTOR

In order to detect a face in a frame of a video sequence we use AdaBoost classifier in combination with Viola&Jones features. During the training stage, the classifier works as a feature selection method. The most relevant features that are identified present a high power of discriminating between face versus non-face samples. The features are computed as difference of sums of pixel intensities from rectangular areas on the input image. Depending on the location of the areas, there are three types of features. The first is the *two-rectangle feature* and it is computed as the difference between the sums of pixel intensities in two horizontally or vertically adjacent rectangular areas (Figure 2: type 1 and type 2). The second type is a *three-rectangle feature* and is computed as the difference between the sums of pixel intensities in two horizontally or vertically outside rectangular areas and the middle area (Figure 2: type 3 and type 4). The third type is a *four-rectangle feature* is computed as difference between diagonal opposite rectangular areas (Figure 2: type 5). The Viola&Jones features are optimally computed by using an intermediate representation of the input image known as the integral image. Each type of Viola&Jones feature can be determined by using a few references on the integral image representing the cumulated sum of pixel intensities on every point of the original image.

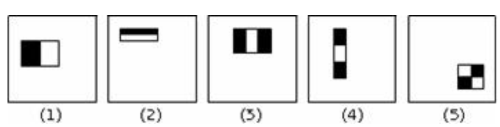

Figure 2 The five basic types features used in our approach

The learning process implies a selection of the best features in a sequence of steps. At each step a weak learning algorithm is employed in the attempt to classify the samples given each feature. After the training of a weak classifier, a re-weighting procedure assigns higher priorities to misclassified samples for the next learning steps. The final face detector consists in a weighted linear combination of classification functions. The face detection process makes use of a fixed size sliding window and a multi-resolution pyramid to determine all the rectangular areas to be analyzed.

ACTIVE APPEARANCE MODEL

The Active Appearance Model (AAM) is employed to extract the shape and texture information on specific faces from each frame of the video sequence. The use of a face detection algorithm as a prior step has the advantage of speeding up the search for the shape parameters during AAM based processing. This enhancement makes possible for a real-time implementation of the algorithm. AAM creates a statistical model of shape variation and a model of texture variation. The average shape is determined considering the training set of face shape samples. The shape samples are aligned using a Generalized Procrustes Analysis. Each face sample is then warped so that the control points match the ones of the mean shape (Figure 3).

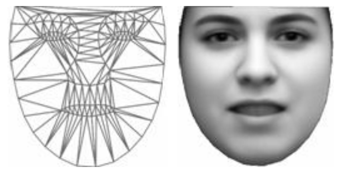

Figure 3 The mean face shape (left) and the mean face texture aligned to the mean shape (right)

According to the separate PCA analysis on shape and texture, a synthesized face sample can be written in terms of shape s and texture t as: $s = \bar{s} + \Phi_s b_s$ and $t = \bar{t} + \Phi_t b_t$, where the values of $\bar{s}$ and $\bar{t}$ represent the mean face shape and the mean face texture. The matrices Φ_s and Φ_t contain the eigenvectors of the shape and texture variations. The final combined model contains information regarding both the shape and texture and is written as:

$$b = \begin{bmatrix} W_s b_s \\ b_t \end{bmatrix} = \begin{bmatrix} W_s \Phi_s^T (s - \bar{s}) \\ \Phi_t^T (t - \bar{t}) \end{bmatrix}$$

W_s is the diagonal matrix that introduces the weighting between units of intensities and units of distances. The third PCA: $b = \Phi_c c$ models the correlation between the shape and texture by taking into account the previous two PC transforms. A new sample is generated by: $s = \bar{s} + \Phi_s W_s^{-1} \Phi_{c,s} c$, $t = \bar{t} + \Phi_t W_t^{-1} \Phi_{c,t} c$, where:

$$\Phi_c = \begin{bmatrix} \Phi_{c,s} \\ \Phi_{c,t} \end{bmatrix}$$

In accordance with the model, each face is represented by a set of parameters defining the model and pose parameters of the face object in the image. A fit function is chosen to represent to probability that the model parameters reflect the target face object. An iterative scheme for correcting the model face parameters is assumed to lead to the optimal solution.

FCP MODEL

The shape information extracted by the AAM from a face image is used to compute a set of suitable parameters that describe well the appearance of the facial features. The first step is the selection of the optimal key points on the face area from the shape data. The key points P_i are defined as Facial Characteristic Points (FCPs) and the FCP-set (Figure 4) is derived from Kobayashi & Hara model (Kobayashi and Hara 1972). In the second step a transform converts the FCP-set to some parameters v_i of an intermediate model. The parameterization has the advantage of providing the classifier with data that encode the most important aspects of the facial expressions. Furthermore, it acts as a dimensionality reduction procedure since the dimension of the feature space is lower than the dimension of the image space. An advantage of the model is that it also can handle certain degree of asymmetry by using some parameters for both left and right sides of the face. The feature parameters are computed as the values of certain Euclidean distances between key points. The symmetry of the model is assumed to make the recognition process of facial expressions robust to occlusion or poor illumination i.e. if the left eye area is not directly visible do not use the related information. The assumption that the facial expressions of a face captured in such conditions can still be evaluated, is based on the supposition that the face detection procedure is also robust enough in such working conditions so as to be able to detect the face first. The parameters v_i model the variability of facial expressions in terms of distances among several pairs of FCPs. The complete list of such parameters is given in **Table 1**.

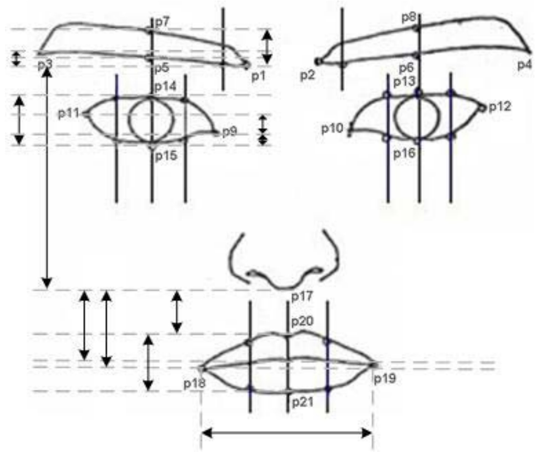

Figure 4 The Facial Characteristic Point – FCP model

v_i	meaning	Visual feature	v_i	meaning	Visual feature	v_i	meaning	Visual feature
v_1	$(P_1, P_7)_y$	Left eyebrow	v_7	$(P_{14}, P_{15})_y$	Left eye	v_{13}	$(P_{17}, P_{20})_y$	Mouth
v_2	$(P_1, P_3)_y$	Left eyebrow	v_8	$(P_9, P_{11})_y$	Left eye	v_{14}	$(P_{20}, P_{21})_y$	Mouth
v_3	$(P_2, P_8)_y$	Right eyebrow	v_9	$(P_9, P_{15})_y$	Left eye	v_{15}	$(P_{18}, P_{19})_y$	Mouth
v_4	$(P_2, P_4)_y$	Right Eyebrow	v_{10}	$(P_{13}, P_{16})_y$	Right eye	v_{16}	$(P_{17}, P_{18})_y$	Mouth
v_5	$(P_1, P_{17})_y$	Left Eyebrow	v_{11}	$(P_{10}, P_{12})_y$	Right eye	v_{17}	$(P_{17}, P_{19})_x$	Mouth
v_6	$(P_2, P_{17})_y$	Right eyebrow	v_{12}	$(P_{10}, P_{16})_y$	Right eye			

Table 1. The set of visual feature parameters

TEMPORAL CLASSIFICATION OF FACIAL EXPRESSIONS

The variances in the parametric model describe emotional patterns for each facial expression. These capture the subtle changes in the shapes of facial features during the video sequence. The method used to encode the emotional patterns take into account the total variance for each parameter of the same face in all the frames in the temporal analysis window. If $V_T = (v_1, v_2, ..., v_m)$ where m=17 is the vector of parameters extracted from the video frame at time T, then $PV_{T'} = \Delta V$ is the variance of the parameters according to the last and the first frame in the sequence starting at time T'.

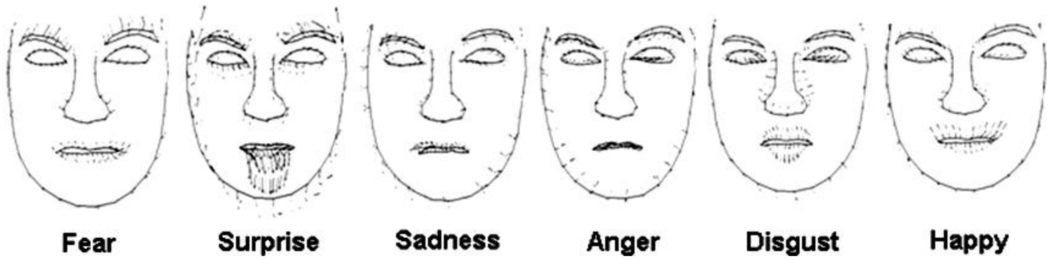

Figure 5 Shape deformation templates for the basic six facial expressions

Figure 5 illustrates an example of temporal variances for the reduced set of FCP distance-based parameters for each of the six basic emotions. It is easy to notice, for instance, the increase of v_{15} in the case of the first facial expression (Fear). This variance contributes to the realization of Action Unit AU20 'Lip stretcher'. The decrease of v_{14} determines the activation of AU25 'Lips part'. The onset of facial expression surprise is associated with the decrease of v_{15} and v_1, the increase of v_{14}, v_7 and v_5. The increase of v_{14} contributes at the activation of AU27 and the increase of v_5 contributes at the realization of AU5. The decrease of v_5 during the onset of facial expression "sadness" contributes at the activation of AU4 'Brow Lowerer'. Figure 6 plots the small changes for the parameters described in **Table 1**, for each facial expression.

RESULTS

The database used for analyzing the temporal changes of emotional patterns consists of a selection from Cohn-Kanade AU-Coded Facial Expression Database (Kanade et al. 2000). Our database contains 474 video sequences having the structure as shown in Table 2. Each video sequence consists of a certain number of frames, depending on the available number of frames that were recorded initially. For the experiments the frame data were extrapolated to a fixed number of frames per video sequence. In addition to this, we used video samples showing faces with different degrees of rotation, occlusion and illumination changes that were recorded during specific crisis simulation sessions.

Expression	Fear	Surprise	Sadness	Anger	Disgust	Happy
#samples	84	105	92	30	56	107

Table 2. The structure of the database for the analysis

For the research, Support Vector Machine is used as classification method for both Action Unit detection and facial expression recognition. Non-linearity that is specific to facial expression representation is handled through kernel methods (non-linear SVM) that first preprocess the data by a non-linear mapping and then apply the linear algorithm in the input feature space. The constraints aim at determining the model parameters that fit the training data and minimize the complexity of the decision function in the same time. The method for evaluating the result is 2 fold Cross Validation. If $PV_{T'} = (PV_{(1)}, PV_{(2)}, ..., PV_{(n)})$ denotes the vector of parameter variances associated with the changes in the face shape during the all video sequence, $PV_{(k)}$ is the column vector corresponding to the kth fixed time-window, and T the space of output variables i.e. the facial expression labels, then f is the associated deterministic function and $t_n = f(x_n) + \varepsilon_n$ represents the possible overlapping target facial expressions.

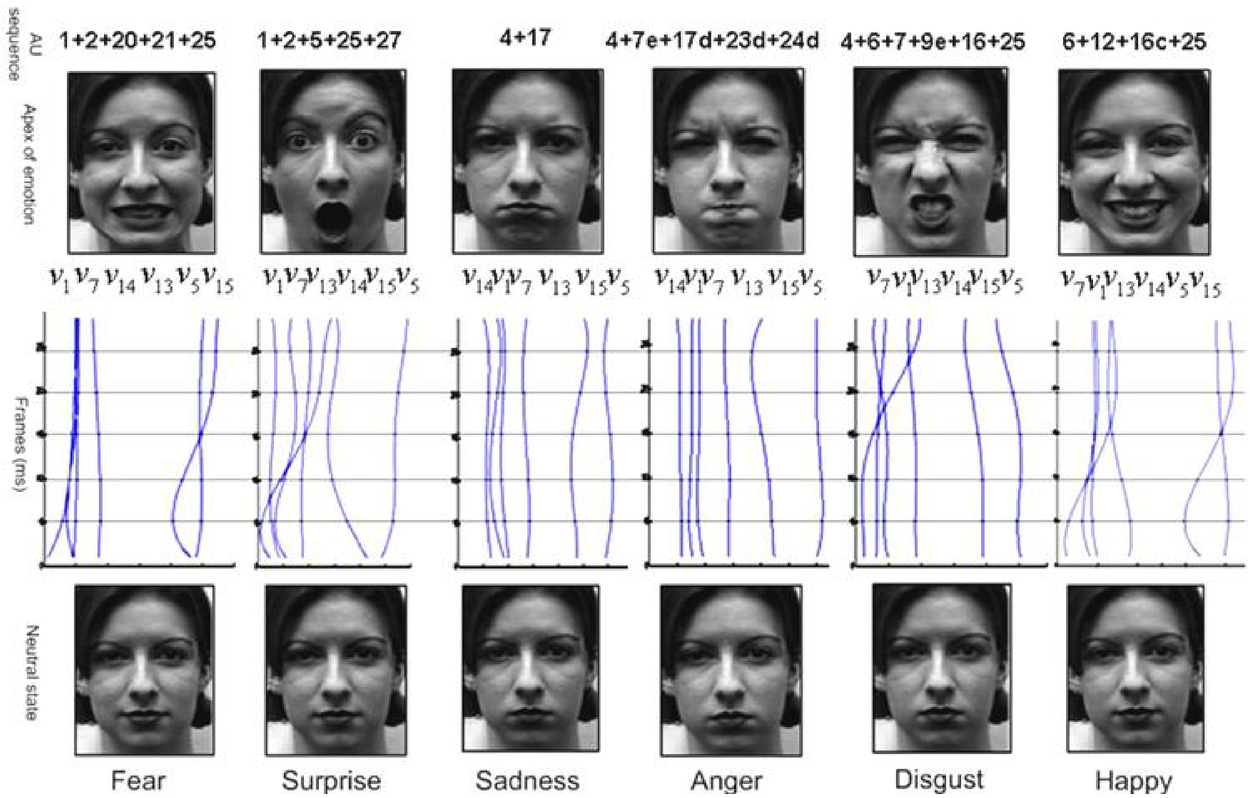

Figure 6 The temporal variances of a reduced set of parameters in six emotion enhanced video sequences

The result is a classifier with a certain level of robustness to over-fitting. In the current research we used the polynomial kernel having the degree 3. The results of facial expression recognition are shown in Figure 7 and Table 3. The results for the recognition of individual Action Units is given in Table 4.

(%)	Fear	Surprise	Sadness	Anger	Disgust	Happy
Fear	**88.09**	2.38	4.76	3.57	1.19	0
Surprise	0	**88.67**	2.83	8.49	0	0
Sadness	5.43	2.17	**85.86**	2.17	1.08	3.26
Anger	10.71	0	3.57	**85.71**	0	0
Disgust	5.35	5.35	3.57	1.78	**82.14**	1.78
Happy	4.62	0	7.40	2.77	5.55	**79.62**

Table 3. The confusion matrix (%) for the facial expression recognition using SVM (polynomial kernel of degree 3)

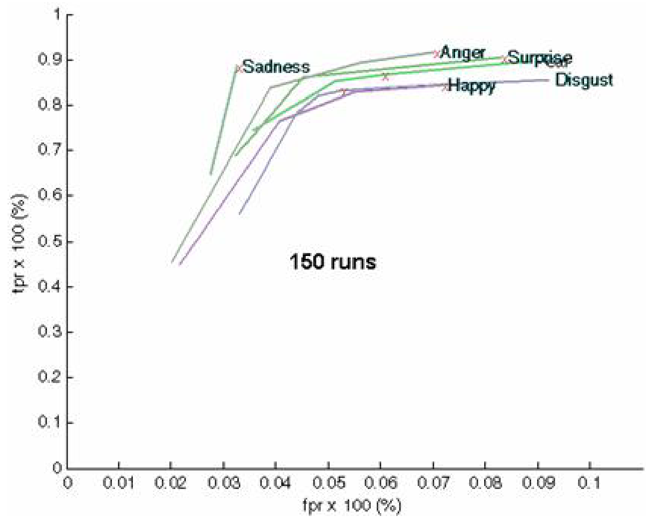

Figure 7 The ROC graph showing the performance of the SVM classifier (polynomial kernel)

AU	true positive rate	accuracy	AU	true positive rate	accuracy
AU1	89.76%±5.11%	82.70%±6.56%	*AU15*	96.58%±1.16%	95.15%±0.30%
AU2	94.84%±3.52%	88.61%±5.37%	*AU16*	87.03%±1.86%	83.54%±0.60%
AU4	89.11%±1.40%	85.44%±0.90%	*AU17*	97.89%±2.98%	97.05%±1.79%
AU5	71.70%±15.46%	73.00%±8.95%	*AU18*	99.15%±1.20%	98.73%±1.19%
AU6	84.86%±0.23%	79.96%±0.30%	*AU20*	89.51%±3.10%	86.08%±2.98%
AU7	94.58%±2.33%	91.98%±2.98%	*AU22*	81.46%±0.22%	78.06%±2.98%
AU9	90.55%±2.98%	87.13%±0.30%	*AU23*	98.23%±0.39%	96.62%±1.19%
AU10	94.24%±0.54%	91.14%±2.39%	*AU24*	96.84%±4.48%	96.41%±3.88%
AU11	92.50%±6.38%	91.35%±6.27%	*AU25*	99.37%±0.90%	98.95%±0.30%
AU12	68.60%±20.60%	71.97%±17.05%	*AU26*	97.88%±2.40%	97.47%±2.39%
AU13	90.28%±7.52%	87.76%±6.56%	*AU27*	97.45%±0.00%	96.62%±0.00%
AU14	81.69%±17.06%	79.80%±4.11%	*AU28*	98.10%±2.69%	97.68%±2.09%

Table 4. The results of Action Unit detection using SVM (polynomial kernel of degree 3)

CONCLUSION

The recent improvements in the area of human computer interfaces make possible the development of context-aware systems for crisis management. These systems are assumed to automatically extract useful clues during typical human computer interaction sessions. Among other communication channels, the visual channel plays an important role in the attempt to assess the emotional state of the user in the crisis environment. The algorithms described in the current paper are to be incorporated in a crisis management system and focus on the recognition of the six prototypic human facial expressions. Automatic facial expression recognition stands for an important component of a system designed to provide support in crisis situations. The value of the automatic mechanisms that collect emotion-oriented information related to the actors at a crisis site is given by the difficulty on analyzing the rather chaotic and unpredictable behavior of the people during major crisis events and by the lack of their ability to make clear statements on their observations. It is of a great importance to extract as much information as possible and to try to disambiguate the partial reports coming from human actors at the crisis environments.

The algorithms used in the research employ the Active Appearance Model for computing the facial shape, and SVM classifier for detecting the Action Units and for recognizing the facial expressions. The constructed classifiers show good computational and classification performance. For training the classifiers we used data samples selected from Cohn-Kanade database and other sources showing faces in good recording sessions and also faces in different pose, occlusion and illumination conditions collected during specific crisis simulation sessions.

The described methods allow for a real-time implementation of the facial recognition system. The use of a Viola&Jones face detector before the use of the AAM FCPs extractor provides shorter search times for the determination of the face and facial features shapes. The employment of temporal emotional patterns represents an efficient solution for the tracking and classification of facial expressions based on facial movements in video sequences. As an alternative, the recognition of facial expressions implied a prior detection of Action Units. The variant offers the means for handling the dynamics of the facial expressions through the set of AUs defined in Facial Action Coding System FACS.

ACKNOWLEDGEMENTS

The research reported here is part of the Interactive Collaborative Information Systems (ICIS) project, supported by the Dutch Ministry of Economic Affairs, grant nr: BSIK03024.

REFERENCES

1. Antonini, G., Sorci, M., Bierlaire, M., Thiran, J.-P., Discrete Choice Models for Static Facial Expression Recognition, ACIVS'06, 710-721, 2006.
2. Cootes, T.F., Edwards, G.J., Taylor, C.J., Active Appearance Models, In H.Burkhardt and B.Neumann, editors, 5th European Conference on Computer Vision, Vol.2, 484-498, Springer, 1998.
3. Datcu, D. and Rothkrantz, L.J.M, Multimodal Web based system for human emotion recognition, Euromedia 2007, Delft, The Netherlands, 2007.
4. Datcu, D. and Rothkrantz, L.J.M., Automatic recognition of facial expressions using Bayesian Belief Networks, Proceedings of IEEE SMC 2004, ISBN 0-7803-8567-5, 2209-2214, October 2004.
5. Ekman P. and Friesen W., Facial Action Coding System, Consulting Psychologists Press, Inc., Palo Alto California, USA, 1978.
6. Freund Y. and Schapire R.E., A decision-theoretic generalization of on-line learning and an application to boosting, In 2nd European Conference on Computational Learning Theory, 1995.
7. Harris C.G. and Stephens M., A combined corner and edge detector, Proceedings 4th Alvey Vision Conference, Manchester, 189-192, 1988.
8. Jongh de E.J., FED: An online facial expression dictionary as a first step in the creation of a complete nonverbal dictionary, TU Delft, 2002.
9. Kanade T, Cohn J. and Tian Y., Comprehensive database for facial expression analysis, Proc. IEEE Int'l Conf. on Automatic Face and Gesture Recognition. 46–53, 2000.
10. Kearney G.D. and McKenzie S., Machine interpretation of emotion: design of a memory-based expert system for interpreting facial expressions in terms of signaled emotions. (JANUS), Cognitive Science 17, Vol. 4, 589–622, 1993.
11. Kobayashi H. and Hara F., Facial Interaction Between Animated 3D Face Robot and Human Beings, IEEE Computer Society Press, 3732-3737, 1997.
12. Lee C.S. and Elgammal A, Nonlinear Shape and Appearance Models for Facial Expression Analysis and Synthesis, Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06) - Volume 01, 497 – 502.
13. Liebelt J., Xiao J., Yang J., Robust AAM Fitting by Fusion of Images and Disparity Data, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, June 17-22, Vol. 2, 2483-2490, 2006.
14. Morishima S. and Harashima H., Emotion Space for Analysis and Synthesis of Facial Expression, IEEE International Workshop on Robot and Human Communication, 674-680, 1993.
15. Rothkrantz L.J.M. and Pantic M., Expert Systems for Automatic Analysis of Facial Expressions, Elsevier, Image and Computing, Vol. 18, 881-905, 2000.

16. Saatci Y., Town C., Cascaded Classification of Gender and Facial Expression using Active Appearance Models, The 7th Conference on Automatic Face and Gesture Recognition FGR'06, 2006.
17. Sojka E, A New Approach to Detecting Corners in Digital Images, IEEE ICIP, 2003.
18. Tipping M.E., The Relevance Vector Machine, Advances in Neural Information Processing Systems, Vol. 12, 652-658, 2000.
19. Treptow A. and Zell A., Combining Adaboost Learning and Evolutionary Search to Select Features for Real-time Object Detection, University of Tuebingen, Department of Computer Science, Germany, 2003.
20. Vapnik V., The nature of statistical learning, Springer, New York, 1995.
21. Viola P. and Jones M., Robust Real-time Object Detection, Second International Workshop on Statistical and Computational Theories of Vision-Modeling, Learning, Computing, and Sampling, 2001.
22. Wilhelm, T., Bohne, H.-J., Gross, H.-M., Classification of Face Images for Gender, Age, Facial Expression, and Identity, ICANN 2005, LNCS 3696, 569-574, 2005.
23. Zhao J. and Kearney G, Classifying facial emotions by back-propagation neural networks with fuzzy inputs, International Conference on Neural Information Processing, Vol. 1, 454–457, 1996.
24. Zhou Z.-H. and Geng X., Projection Functions for Eye Detection, State Key Labaratory for Novel Software Technology, NU, China, 2002.
25. Zhou Y., Gu L., Zhang H-J, Bayesian tangent shape model: estimating shape and pose parameters via Bayesian inference, ISBN: 0-7695-1900-8, IEEE Computer Vision and Pattern Recognition, Proceedings. 2003, 2003.