Multimodal Web based system for human emotion recognition

Dragos Datcu, L.J.M. Rothkrantz
Man-Machine Interaction Group
Delft University of Technology
2628 CD, Delft,
The Netherlands
E-mail: {D.Datcu ; L.J.M.Rothkrantz}@ewi.tudelft.nl

KEYWORDS
Automatic emotion recognition, speech analysis, face detection, facial feature extraction, facial characteristic point extraction, AAM, SVM.

ABSTRACT

The system being described in the paper presents a Web interface for a fully automatic audio-video human emotion recognition. The analysis is focused on the set of six basic emotions plus the neutral type. Different classifiers are involved in the process of face detection (AdaBoost), facial expression recognition (SVM and other models) and emotion recognition from speech (GentleBoost). The Active Appearance Model – AAM is used to get the information related to the shapes of the faces to be analyzed. The facial expression recognition is frame based and no temporal patterns of emotions are managed. The emotion recognition from movies is done separately on sound and video frames. The algorithm does not handle the dependencies between audio and video during the analysis. The methodologies for data processing are explained and specific performance measures for the emotion recognition are presented.

INTRODUCTION

Nonverbal communication plays an important role in everyday life of the people. As people have better understanding of the emotions related mechanisms and the correlation with human behaviour, it became even more necessary to design systems to automatically detect the human's emotional state. Nowadays it has been a challenging task to realize algorithms to identify meaningful clues and to learn machines to analyze human faces and utterances for extracting emotions.
In the current paper we present a Web based system that performs automatic recognition of emotions (Figure 1) from speech and video data. The users can upload audio and video files and can run full emotion analysis remotely.
According to Ekman et al. (Ekman and Friesen 1978) people are born with the ability to generate and interpret only six facial expressions: happiness, anger, disgust, fear, surprise and sadness. All other facial expressions have to be learned from the environment the person grows up.
Our Web system provides the necessary algorithms for the analysis of the six basic emotions, given the face of the subject or the audio speech signal. For the case of facial expressions, the system is robust to the typical differences of the age, gender and race and culture. In the case of speech analysis, the emotions are restricted to the german language. This limitation was set by the availability of data sets used for training the emotion from speech classifier.

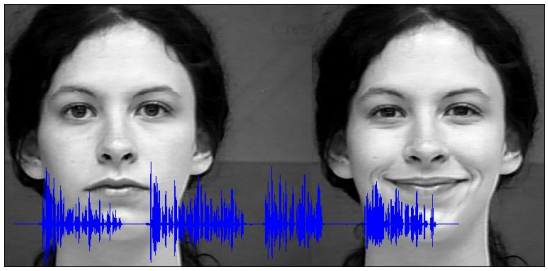

Figure 1. The Web based emotion recognition system performs analysis on both audio and video data.

RELATED WORK

The paper of (Saatci and Town, 2006) presents an approach to determine the gender and expression of faces by using Active Appearance Model. 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) propose 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. The recognition of facial expressions was recently tackled in different ways involving the use of Viola&Jones features on input image data (Wong et al., 2006), static and temporal relations on facial characteristic points and by using different classifiers as BBN (Shan et al., 2006), (Datcu and Rothkrantz, 2004), HMMs (Aleksic and Katsaggelos, 2006), RVM and SVM (Datcu and Rothkrantz, 2005).

Acknowledgments. 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.

(Rothkrantz and Pantic, 2000) proposed a point-based face model composed of two 2D facial views, namely the frontal- and the side view. Based on a point-based face model, expression-classification rules can be converted straight-forwardly into the rules of an automatic classifier.
The Facial Expression Dictionary (FED) (de Jongh, 2002) project aims at developing a non-verbal dictionary that contains information about non-verbal communication of people. The facial expression related analysis involves the user to select the region of the face and select the characteristic points of the face. The facial key points are predefined and the system helps to rapidly identify them on a new face sample. The face model used is that of Kobayashi and Hara (Kobayashi and Hara 1997). After the specification of all the FCPs, the system is able to run the emotion recognition process on the input face.
Face related analysis has been already incorporated in applications targetting e-learning (Loh et al., 2006), (Ben ammar and Neji, 2006) and smart meeting and conversation tracking systems (Zeng et al., 2006). The use of Active Appearance Model for extracting face shape information and the tracking of emotions in video sequences based on this shape data were recently researched in (Datcu and Rothkrantz, 2007).
For classifying human emotions in speech various attempts involved various algorithms. (Neiberg et al., 2006) have used a GMM to recognize emotions in spontaneous speech.
(Yu and al., 2004) applied a multilevel structure based on coupled hidden Markov models to estimate engagement levels in continuous natural speech. (Datcu and Rothkrantz, 2006) used the GentleBoost classifier to determine the optimal utterance segmentation for emotion recognition. The continuous speech signal is segmented into spoken utterances and the acoustic features are computed from each utterance portion. The extracted non-linguistic information is used for predicting the emotional states such as discrete emotion types or arousal/valence levels by employing SVM-based classifiers. The HMM uses the previous information to model the user's emotional state and engagement in conversation as a dynamic, continuous process. (Chateau et al., 2002) presents a study of the perception, the analysis and the modelling of styles or the 'emotional quality' of speech. The speech emotional quality is evaluated in terms of the emotional content that describes the listener's global impressions as elicited by their audition. criteria for evaluating the emotional quality are used to generate perceptive portraits of the speech. The evaluation is carried by using linear models to connect the perceptive portraits to physical data derived from signal analysis. Some work has also been focused on using additional information regarding speech.
The paper of (Lee and Narayanan, 2005) uses three sources of information - acoustic, lexical and discourse - for recognizing emotions. Linear discriminant and k-nearest neighbourhood classifiers are used to classify acoustical information to anger and frustration - as negative emotions and to neutral or positive emotions. The different features are extracted by using certain portions of the signal. A noticeable approach stands for multimodal analysis that aims at improving the recognition rates for the emotional state by fusing the results on separate modalities. The advantage of such methods relates to the overcoming the limited information that can be gathered from each single modality.
The work of (Busso et al, 2004) analysis the strengths and the limitations of systems based on the fusion of facial expression and acoustical information analysis at the decision level and in the case of feature level integration. (Kwon et al., 2003) provides a comparison on the emotion recognition performance of various classifiers. They obtained SVM and HMM based classifiers with significantly better results on SUSAS database from the previous approaches.
A recent research of (Rothkrantz et al., 2004) focuses on studying the effect of the workload on speech production by making use of a psychological experimental setup. A full analysis on each acoustic feature is conducted in order to create efficient models for stress detection.

MULTIMODAL APPROACH

The system provides the users with the capability of uploading audio and visual data for emotion analysis. For the audio data, the system accepts the processing of standard audio files. For the visual data there are two types of files that can be handled. The user can specify both static pictures and video sequences. In the case of processing a short video sequence, the user has a fixed limit for each video file to be processed.

Face detection

The detection of faces in both photos and video sequences is realized by using an implementation of an algorithm based on Viola&Jones features (Viola and Jones, 2001). These are visual features computed following simple addition and substraction operations on pixel intensities from rectangular areas (Figure 2). A strong classifier implementing Adaboost technique selects the most relevant Viola&Jones features that provide the best face detection rates. A combination of a fixed size sliding window and an multiresolution phyramid algorithm ensure the analysis of all possible areas for detecting faces in the frames. The input set of features for the classifier are selected from the fixed size of Viola&Jones features given the size of the sliding window.

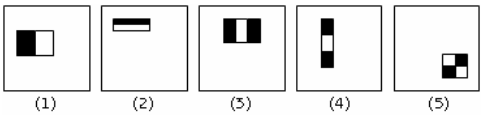

Figure 2. The five basic types of Haar-like features used for face detection. The summation operation is done on all pixel intensities from areas having the same color. The substraction is done between the two computed sums.

FCP model

The data set used for training the facial expression recognizer was Cohn-Kanade database (Kanade et al., 2000). The database contains a set of video sequences of recordings of several subjects acting on multiple scenarious. Each video sequence includes a subject showing a specific facial

expression from the neutral state to the apex of the emotion. In the original database only the last frame of each sequence is labelled using Auction Units AU codification. The process of creating the data set for training implied the selection of the last frame from each video sequence. The final set of selected samples has the structure as illustrated in
Table 1. 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 3) is derived from Kobayashi & Hara model (Kobayashi and Hara, 1972).

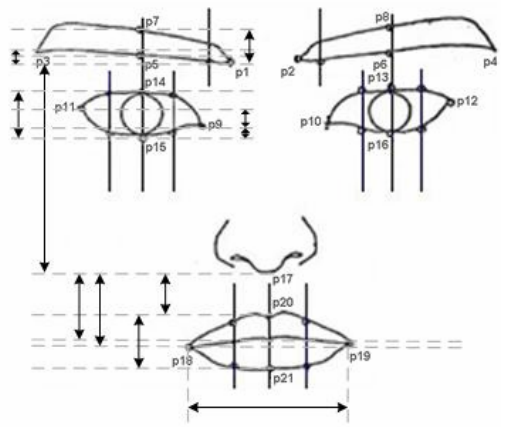

Figure 3:The Facial Characteristic Point FCP model.

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.

Table 1. The structure of the data set for facial expression recognition

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

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 related information. The assumption 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.
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 2.

Table 2: The set of visual feature parameters

		Visual feature			**Visual feature**			**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			

CLASSIFICATION OF FACIAL EXPRESSIONS

The method used to encode the emotional patterns takes into account the subtle changes of the face shapes in different emotional postures. The vector $V = (v_1, v_2, ..., v_m)$ where m=17, encodes the set of parameters extracted from the Facial Characteristic Point - FCP model and accordingly has associated a certain emotional label.
For the classification of facial expressions, different classifiers have been taken into account. The method used for determining the results is 2-fold Cross Validation.

Table 4 illustrates the performance of a SVM classifier with polynomial kernel that works as a detector of Action Units AUs (Ekman and Friesen, 1978). The main algorithm for recognizing emotions is based on the classification done using the distance oriented model between FCPs. The second approach is the technique of detecting the AUs first and then to recognize facial expressions from the AU sequences.

Table 3 presents the results in the case of the recognition of facial expressions by using Support Vector Machines – SVM as classifier.

The method used for determining the results is 2-fold Cross Validation.
Table 4 illustrates the performance of a SVM classifier with polynomial kernel that works as a detector of Action Units AUs (Ekman and Friesen, 1978). The main algorithm for recognizing emotions is based on the classification done using the distance oriented model between FCPs. The second approach is the technique of detecting the AUs first and then to recognize facial expressions from the AU sequences.

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

(%)	Fear	Surprise	Sadness	Anger	Disgust	Happy
Fear	**84.70**	3.52	3.52	4.70	1.17	2.35
Surprise	12.38	**83.80**	0.95	0	0	2.85
Sadness	6.45	3.22	**82.79**	1.07	3.22	3.22
Anger	3.44	6.89	6.89	**75.86**	6.89	0
Disgust	0	0	7.14	10.71	**80.35**	1.78
Happy	7.54	8.49	2.83	3.77	4.71	**72.64**

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

AU1	88.89%±2.08%	80.59%±0.00%
AU2	94.53%±0.25%	89.87%±2.39%
AU4	93.67%±0.47%	87.76%±1.19%
AU5	91.42%±1.13%	82.91%±2.09%
AU6	84.94%±5.93%	75.95%±3.58%
AU7	96.73%±1.92%	91.56%±1.79%
AU9	94.60%±1.88%	91.77%±2.09%
AU10	95.38%±1.82%	92.62%±2.69%
AU11	99.36%±0.30%	98.10%±0.30%
AU12	95.75%±0.41%	90.08%±2.09%
AU13	96.05%±2.44%	92.41%±1.79%
AU14	91.79%±1.51%	86.92%±5.37%
AU15	98.72%±1.20%	97.47%±0.60%
AU16	93.66%±0.69%	85.23%±1.19%
AU17	99.37%±0.90%	98.73%±0.00%
AU18	100.0%±0.00%	99.79%±0.30%
AU20	93.04%±2.60%	87.13%±2.09%
AU22	93.16%±3.18%	87.34%±2.98%
AU23	99.24%±0.38%	96.84%±1.49%
AU24	99.58%±0.60%	99.37%±0.30%
AU25	100.0%±0.00%	99.79%±0.30%
AU26	99.79%±0.30%	99.58%±0.00%
AU27	99.37%±0.90%	98.73%±0.00%
AU28	100.0%±0.00%	99.79%±0.30%

EMOTION ANALYSIS FROM SPEECH

The classifier chosen to model the emotion characteristics in speech is based on Gentle AdaBoost method for a maximum 200 training steps. The optimal classifiers are determined by employing ROC graphs to show the trade-off between the hit and the false positive rates. One important issue for the recognition of emotions in speech represents the segmentation of the speech signal. The way this process is done dramatically affects the subsequent results of the recognition of emotions.
One research question that rised was what is the optimal utterance segmentation method that gives the best results. Given the set of prosodic features, we determined the segmentation type and the utterance frame structure that leads to good recognition of emotions.
The data set used for emotion analysis from speech is Berlin (Burkhardt et al., 2005) – a database of German emotional speech. The database contains utterances of both male and female speakers, two sentences. The emotions were simulated by ten native German actors (five female and five male). The result consists of ten utterances (five short and five long sentences). The length of the utterance samples ranges from 1.2255 seconds to 8.9782 seconds. The recording frequency is 16kHz.
The final speech data set contains the utterances for which the associated emotional class was recognized by at least 80% of the listeners. Following a speech sample selection, an initial data set was generated comprising 456 samples and six basic emotions (*anger*: 127 samples, *boredom*: 81 samples, *disgust*: 46 samples, *anxiety/fear*: 69 samples, *happiness*: 71 samples and *sadness*: 62 samples).
In the case of emotion recognition from speech, the analysis is handled separately for different number of frames per utterance. In the current approach there are five types of splitting methods performed on initial data. Each type of splitting produces a number of data sets, according to all the frame combinations in one utterance.
The Praat (Boersma and Weenink, 2005) tool was used for extracting the features from each sample from all generated data sets. According to each data set frame configuration, the parameters **mean**, **standard deviation**, **minimum** and **maximum** of the following acoustic features were computed: Fundamental frequency (pitch), Intensity, F1, F2, F3, F4 and Bandwidth.
All these parameters form the input for separate GentleBoost classifiers according to data sets with distinct segmentation characteristics.

Results of emotion recognition from speech

The GentleBoost *committee* is trained for a maximum number of 200 stages. Separate data sets containing male, female and both male and female utterances are considered for training and testing the classifier models. The performance of each classifier is evaluated with the 5-fold cross validation. Depending on the number of sub-frames per speech frame, the different data sets are used to generate sets of classifiers. One curve on the graph stands for the set of representative GentleBoost strong classifiers generated by using the specific data set, associated with a certain split configuration. Each node on one curve relates to one classifier in the set. The ROC graph in Figure 4 shows the tradeoff between the hit and the false-positive rates for all the GentleBoost classifiers generated from Berlin data set.

Each point on the figure stands for one GentleBoost classifier that is selected using the highest true-positive rate criterion. For each emotion class, a total number of 200 points is taken into account and only the ones with the highest scores are displayed on the same emotion curve. By analyzing each emotion curve separately, the final strong committee to be chosen is the one that is the closest to the north-west corner of the figure. In other words, the classifier in question is the one that has the highest true positive rate (*tpr*) while the false positive rate (*fpr*) is the lowest in the set of classifiers on the same curve.

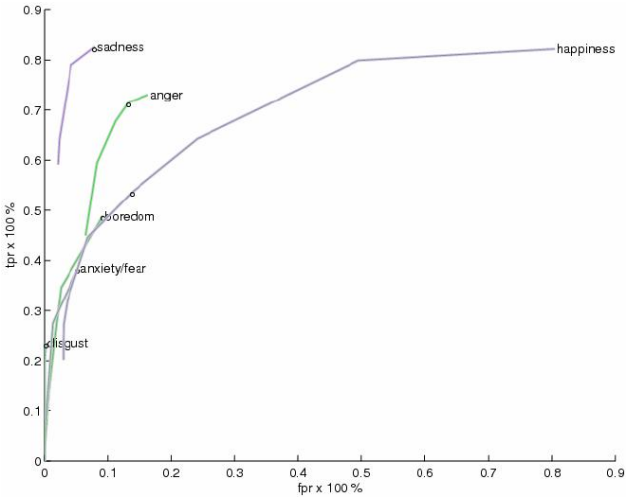

Figure 4. ROC graph that show the committees with the highest true positive rates for each emotion class.

Table 5 depicts the characteristics of each strong classifier that is selected for each emotion curve separately. The column *nr.stages* shows the number of stages required to train the associated strong committee.
An additional field (*ac*) in each table shows the accuracy rate achieved by the classifiers.
Each classifier is identified by the structure of the frames into the utterance sample (column *frames*). A digit from one binary sequence specifies that the correspondent frame contributes ('1') or not ('0') with features at the classification process.

Table 5: The results for the recognition of emotions using GentleBoost classifier.

emotion	nf	frames	nr. stages	ac (%)	tpr (%)	fpr (%)
anger	10	1101000001	5	0.83±0.03	0.72±0.16	0.13±0.06
boredom	2	10	58	0.84±0.07	0.49±0.18	0.09±0.09
disgust	10	0100001000	21	0.92±0.05	0.24±0.43	0.00±0.00
anxiety/fear	10	1110000011	86	0.87±0.03	0.38±0.15	0.05±0.04
happiness	10	1111010100	40	0.81±0.06	0.54±0.41	0.14±0.13
sadness	10	1011111101	13	0.91±0.05	0.83±0.06	0.08±0.06

An observation on the table proves that the majority of the strong classifiers lying on the emotion curves in the ROC graph clearly express the efficiency of using a ten frames per utterance configuration for the segmentation. The information presented in Table 6 is independent on the emotion class.

Table 6: The dependency of emotion recognition results on the number of frames per utterance for Berlin data set.

nf	ac (%)	tpr (%)	fpr (%)
1	0.85±0.11	0.36±0.63	0.07±0.17
2	0.83±0.31	0.44±0.67	0.10±0.41
3	0.84±0.17	0.46±0.62	0.09±0.23
5	0.84±0.13	0.50±0.63	0.10±0.22
10	0.77±0.33	0.58±0.64	0.20±0.45

One difference should be noted on the analysis methods used for choosing the best classifiers for Table 5 and Table 6. While for the first the criterion was to choose the classifiers with the best trade-off between hit rate and false positive rate, the last involved the choice for the classifiers with the highest true positive rate.

THE WEB BASED INTERFACE

The user can upload photos, audio and video files into the system. For each type of data files there is a size limit imposed for restricting the amount of processing necessary for running all the analysis. The numer of the users that can work remotely at the same time is also limited. From the functional point of view, the Web based emotion recognition system consists of a set of CGI applications that run behind the HTTP server. The Web interface is created using HTML and JAVA applets on the client side and MySQL database system, PHP scripting and CGI modules on the server side.
The files uploaded by the users are stored in a database on the server.
For facial expression recognition, the system first calls a CGI module that converts all photos to a common visual format. That makes sure that the other CGI components will be able to read the picture files without problems.
In case of video files, the video is decomposed so as to extract frame by frame and to store the collection of frames in a temporary location for further analysis. In case of videos, the frame sequence is scaled to a fixed rate of frames per second. In most of the cases this operation involves a downscaling to a lower rate so as to allow for a smaller amount of processing per whole sequence. The next step is the detection of faces into each frame. This is carried on by a face detection CGI application that is called from PHP environment. The results are passed back to the PHP script and written as parameters of the JAVA applet. In this way the user will be able to go through each frame, one for pictures, and to visualize the location of the detected faces in video or still picture. An additional step of the face detection program is the extraction of the areas of the faces and the storing of these as separate image files. The operation is useful for the next step that involves the running of the facial expression recognition CGI modules. Each face is analyzed several times by different CGI applications, one for each implementation of each facial expression classifier.
There are several classifiers implemented in this way and each of them uses the same image files containing the face to be analyzed. The results are again, passed back to the script that formats the HTML response back to the client so as to

include the information regarding the facial expression recognition data.
In the case of audio data, the uploaded files are analyzed first by a CGI script for extracting all the speech parameters such as pitch, frequencies, intensities and bandwidth. Another CGI application computes the certain parameters derived from those mentioned above, namely the mean, minimum, maximum and standard deviation. The last CGI application implementing the Gentle AdaBoost classifier, performs the recognition of emotions given the input data associated with the initial utterance file. The results are passed back in the same way as in the case of facial expression recognition.
The Web browser receives the PHP generated HTML response and the audio and video files back from the server. The user can visualize the initial audio-video data and the new information related to the recognition of emotions in speech and visual content. The mechanism implies that all the processing is done on the server and the client is no longer bothered with possible restrictions that may rise from the necessity of running data analysis on the local machine. Some experimental versions of the emotion recognition interface are illustrated in Figure 5 (for speech) and in Figure 6 (for video).
Some optimizations have been made on the server side in order to speed up the process of going though all the computations.
The implementation of the facial expression recognition and the emotion recognition from speech modules present a very high speed of frame and audio files processed per second. Practically, these can be integrated in a potential real time system.
Despite the fact that each CGI component – usually the ones implementing classification algorithms – needs very small time for processing the data, when a CGI is run, it requires a big amount of time for the initialization. Commonly, this comes from the need to load big data files of matrices required in the process of classification.
The improvement consists of an implementation of applications that are loaded in the memory and running continuously on the server or possible on other, more powerfull, computers, on the same network with that of the HTTP server. These kind of applications are idle during the time periods when the Web System is not accessed by users. When there are users attempting data analysis, the PHP script running on the HTTP server calls the CGI modules and takes back the provided results. Each CGI module keeps its own waiting list for handling the tasks according to the user requests.

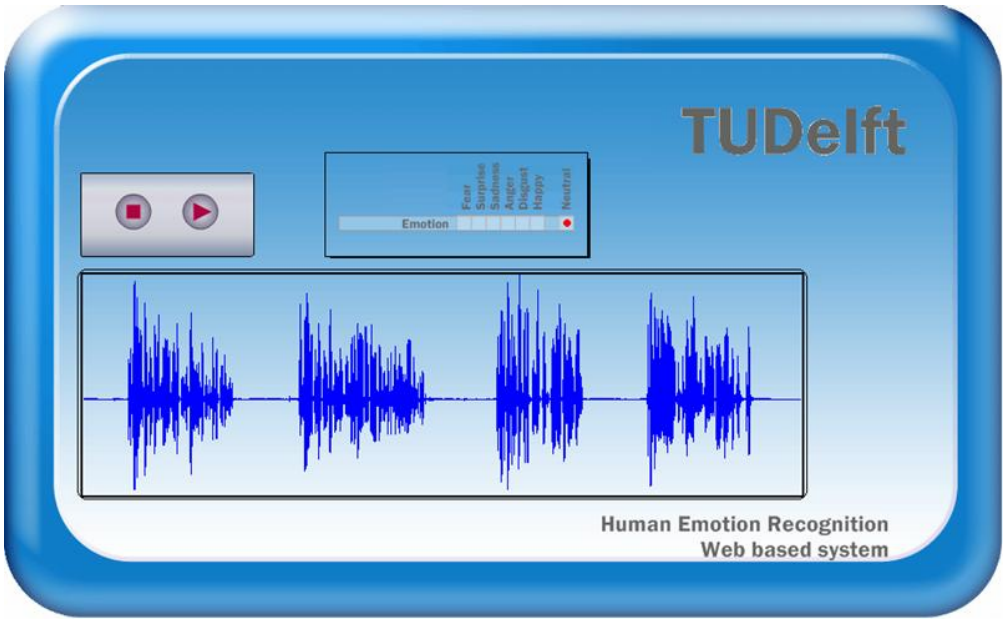

Figure 5. The Web interface for emotion recognition from speech

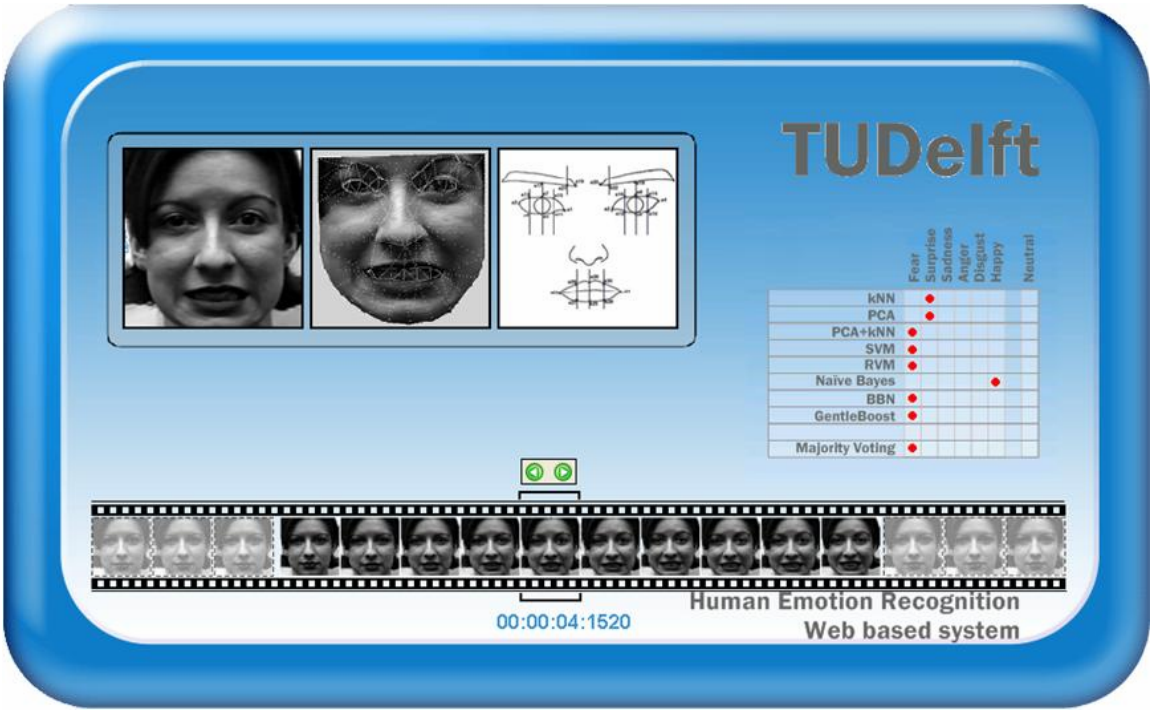

Figure 6. The Web interface for facial expression recognition

CONCLUSION

The system being described in the paper is able to perform automatic human emotion recognition. Beside the advantage of accessing a Web based interface for distance processing, the disadvantage of the system resides in the big amount of processing and time needed for running the complete human emotion analysis. In the case of multiple accesses the effects are considerable and it may lead to low performance. One more aspect is that the facial expression recognition is frame based and no temporal emotional patterns are used during the analysis of video sequences. Furthermore, a major drawback of the system is the lack of algorithms to handle the recognition of emotions in videos. If the user uploads a movie sequence, the system performs either audio or visual based emotion recognition. The emotion classification models do not take the combination of the two modalities into account at the same time In the case of emotion recognition from speech, the ability to run analysis only on german spoken utterances stands for a major restriction for the moment. In the research for the development of an algorithm for emotional speech analysis, we have conducted a set of analysis on different types of utterance segmentation.
As a base technique, we used the GentleBoost classifier with a maximum of 200 training stages. The optimal strong classifier has been selected by making use of ROC graphs. Although the initial research included also separate analysis for male and female voices, the chosen solution is only on mixed male-female voices due to the limited amount of space.
The administration side of the website is enhanced with sections that provide statistics on the use of the system, to track the user processing tasks and to analyse the system performance. After a prior consent from the users, the multimodal database of the system is enriched with new audio and video samples as they are provided for analysis. In the current version, the emotion classification is restricted to the set of six basic emotions plus the neutral. Ongoing work is carried for extending the set of emotions that are to be recognized.

REFERENCES

Aleksic, P. S., A. K. Katsaggelos, “Automatic Facial Expression Recognition using Facial Animation Parameters and Multi-Stream HMMs”, ISSN: 1556-6013, in IEEE Transactions on Information Forensics and Security, 2006.

Ben ammar, M. and Neji, M. (2006), “Affective e-Learning Framework”, In T. Reeves & S. Yamashita (Eds.), Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2006 (pp. 1595-1602). Chesapeake, 2006.

Boersma, P., Weenink, D., “Praat: doing phonetics by computer (Version 4.3.14)” [Computer program]. 2005.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B.: A Database of German Emotional Speech. Proceedings Interspeech, Lissabon, Portugal 2005.

Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C., M., Kazemzadeh, A., Lee, S., Neumann, U., Narayanan, S., “Analysis of emotion recognition using facial expressions, speech and multimodal information”, *ICMI*, State College, Pennsylvania, 2004.

Chateau, N., Maffiolo, V., Ehrette, T., s’Alessandro, C, “Modelling the emotional quality of speech in a telecommunication context”, *Proceedings of the International Conference on Auditory Display*, Kyoto, Japan 2002.

Datcu, D., Rothkrantz L.J.M., "Automatic recognition of facial expressions using Bayesian Belief Networks", *Proceedings of*

IEEE SMC 2004, ISBN 0-7803-8567-5, pp. 2209-2214, October 2004.

Datcu, D., L.J.M. Rothkrantz, "Machine learning techniques for face analysis", *Euromedia 2005*, ISBN 90-77381-17-1, pp. 105-109, April 2005.

Datcu, D., L.J.M. Rothkrantz, "The recognition of emotions from speech using GentleBoost Classifier", CompSysTech'06, June 2006.

Datcu, D., L.J.M. Rothkrantz, "Facial expression recognition using Active Appearance Model in crisis environments", ISCRAM'07, Delft, The Netherlands, May 2007.

Ekman, P. and W. Friesen. 1978. "Facial Action Coding System." Consulting Psychologists Press, Inc., Palo Alto California, USA.

Freund, Y. and R.E. Schapire. 1995. "A decision-theoretic generalization of on-line learning and an application to boosting." In *2nd European Conference on Computational Learning Theory*.

Harris, C.G. and M. Stephens. 1988. "A combined corner and edge detector". Proceedings 4th Alvey Vision Conference, Manchester, 189-192.

Jongh de, E.J. 2002. "FED: An online facial expression dictionary as a first step in the creation of a complete nonverbal dictionary." TU Delft.

Kanade, T., J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis,". *Proc. of the 4th IEEE Int. Con. on Automatic Face and Gestures Reco.*, France, 2000.

Kearney, G.D. and S. McKenzie. 1993. "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.

Kobayashi, H. and F. Hara. 1997. "Facial Interaction Between Animated 3D Face Robot and Human Beings". IEEE Computer Society Press, 3732-3737.

Kwon, Oh-Wook, Chan, K., Hao, J., Lee, Te-Won, "Emotion recognition by speech signals", *EUROSPEECH'03*, Geneva, pp. 125–128, 2003.

Lee, C., M., Narayanan, S., S., "Toward detecting emotions in spoken dialogs", IEEE Transactions on speech and audion processing 13(2), pp. 293–303, 2005.

Lee, C.S., A, Elgammal, "Nonlinear Shape and Appearance Models for Facial Expression Analysis and Synthesis", *Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06)*, volume 01, pp. 497 – 502, 2006.

Loh, May-Ping, Y.P. Wong, C.O. Wong, "Facial Expression Recognition for E-learning Systems using Gabor Wavelet & Neural Network", *Sixth IEEE International Conference on Advanced Learning Technologies (ICALT'06*), pp. 523-525, 2006.

Morishima, S. and H. Harashima. 1993. "Emotion Space for Analysis and Synthesis of Facial Expression". IEEE International Workshop on Robot and Human Communication, 674-680.

Neiberg, D., K. Elenius, K. Laskowski, "Emotion Recognition in Spontaneous Speech Using GMMs", INTERSPEECH 2006 – ICSLP.

Rothkrantz, L.J.M. and M. Pantic. 2000. "Expert Systems for Automatic Analysis of Facial Expressions". Elsevier, Image and Computing, Vol. 18, 881-905.

Rothkrantz, L., J., M., Wiggers, P., van Wees, J., W., A., van Vark, R., J., "Voice stress analysis", *Proceedings of Text, Speech and Dialogues*, 2004.

Saatci, Y., Town, C. 2006, "Cascaded Classification of Gender and Facial Expression using Active Appearance Models", *The 7th Conference on Automatic Face and Gesture Recognition*, FGR'06.

Shan, C., S. Gong, P. W. McOwan, "Dynamic Facial Expression Recognition Using A Bayesian Temporal Manifold Model", BMVC06, 2006.

Tipping, M.E. 2000. "The Relevance Vector Machine. Advances in Neural Information Processing Systems". Vol. 12, 652-658.

Treptow, A. and A. Zell. 2003. "Combining Adaboost Learning and Evolutionary Search to Select Features for Real-time Object Detection". University of Tuebingen, Department of Computer Science, Germany.

Viola, P. and M. Jones. 2001. "Robust Real-time Object Detection." Second International Workshop on Statistical and Computational Theories of Vision-Modeling, Learning, Computing, and Sampling.

Zhao J. and G. Kearney. 1996. "Classifying facial emotions by back-propagation neural networks with fuzzy inputs". International Conference on Neural Information Processing, Vol. 1, 454–457.

Zhou, Z.-H. and X. Geng. 2002. "Projection Functions for Eye Detection". State Key Labaratory for Novel Software Technology, NU, China.

Yu, C., Aoki, P. M., Woodruff, A., "Detecting user engagement in everyday conversations". *8th International Conference on Spoken Language Processing (ICSLP'04)*. 2004.

Zeng, Z., Y. Hu, Y. Fu, T. S. Huang, G. I. Roisman, Z. Wen, „Audio-visual emotion recognition in adult attachment interview", in Proceedings of the 8th international conference on Multimodal interfaces, ISBN:1-59593-541-X, pp. 139-145, 2006.

Zhou, Y., Gu, L., Zhang, H-J. 2003, "Bayesian tangent shape model: estimating shape and pose parameters via Bayesian inference", ISBN: 0-7695-1900-8, *Computer Vision and Pattern Recognition, Proceedings*, 2003.

Wong, W.S., W.Chan, D.Datcu, L.J.M. Rothkrantz, "Using a sparse learning relevance vector machine in facial expression recognition", *Euromedia'2006*, ISBN 90-77381-25-2, pp. 33-37, University of Ghent, April 2006.