Semantic Audio-Visual Data Fusion for Automatic Emotion Recognition

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

KEYWORDS
Data fusion, automatic emotion recognition, speech analysis, face detection, facial feature extraction, facial characteristic point extraction, Active Appearance Models, Support Vector Machines.

ABSTRACT

The paper describes a novel technique for the recognition of emotions from multimodal data. We focus on the recognition of the six prototypic emotions. The results from the facial expression recognition and from the emotion recognition from speech are combined using a bi-modal multimodal semantic data fusion model that determines the most probable emotion of the subject. Two types of models based on geometric face features for facial expression recognition are being used, depending on the presence or absence of speech. In our approach we define an algorithm that is robust to changes of face shape that occur during regular speech. The influence of phoneme generation on the face shape during speech is removed by using features that are only related to the eyes and the eyebrows. The paper includes results from testing the presented models.

INTRODUCTION

The ability to replicate the human competence in naturally processing emotional clues from different channels during interpersonal communication has achieved an even higher role for the modern society nowadays. Smart human computer interfaces, affect sensitive robots or systems to support and coordinate people's daily activities are all about to incorporate knowledge on how emotions are to be perceived and interpreted in a similar way they are sensed by human beings.
While the study on human emotion recognition using unimodal information has considerably maturated for the last decade, the research on multimodal emotion understanding is still at the preliminary phase (Pantic and Rothkrantz, 2003). Sustained efforts attempt answering the question of what is the role and how information from various modalities can support or attenuate each other so as to get the smooth determination of human emotions. Recent research works have pointed to the advantage of using combinations of facial expressions and speech for correctly determining the subject's emotion (Busso et al., 2004; Zeng et al., 2007).
In the current paper we investigate the creation of a bimodal emotion recognition algorithm that incorporates facial expression recognition and emotion extraction from speech. Mostly we are interested on the design of a multimodal emotion data fusion model that works at the high, semantic level and that takes account of the dynamics in facial expressions and speech. Following recent comparable studies, we aim at obtaining higher performance rates for our method when compared to the unimodal approaches. The algorithms we use derive representative and robust feature sets for emotion classification model. The current research is a continuation of our previous work on facial expression recognition (Datcu and Rothkrantz, 2007; Datcu and Rothkrantz, 2005) and emotion extraction from speech signals (Datcu and Rothkrantz, 2006).

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.

RELATED WORK

The paper of (Wimmer et al., 2008) studies early feature fusion models based on statistically analyzing multivariate time-series for combining the processing of video based and audio based low-level descriptors (LLDs).
The work of (Hoch et al., 2005) presents an algorithm for bimodal emotion recognition in automotive environment. The fusion of results from unimodal acoustic and visual emotion recognizers is realized at abstract decision level.
For the analysis, the authors used a database of 840 audiovisual samples that contain recordings from seven different speakers showing three emotions. By using a fusion model based on a weighted linear combination, the performance gain becomes nearly 4% compared to the results in the case of unimodal emotion recognition.
(Song et al., 2004) presents a emotion recognition method based on Active Appearance Models – AAM for facial feature tracking. Facial Animation Parameters – FAPs are extracted from video data and are used together with low level audio features as input for a HMM to classify the human emotions.
The paper of (Paleari and Lisetti, 2006) presents a multimodal fusion framework for emotion recognition that relies on MAUI - Multimodal Affective User Interface paradigm. The approach is based on the Scherer's theory

Component Process Theory (CPT) for the definition of the user model and to simulate the agent emotion generation.
(Sebe et al., 2006) proposes a Bayesian network topology for recognizing emotions from audio and facial expressions. The database they used includes recordings of 38 subjects which show 11 classes of affects. According to the authors, the achieved performance results pointed to around 90% for bimodal classification of emotions from speech and facial expressions compared to 56% for the face-only classifier and about 45% for the prosody-only classifier.
(Zeng et al., 2007) conducted a series of experiments related to the multimodal recognition of spontaneous emotions in a realistic setup for Adult Attachment Interview. They use Facial Action Coding System – FACS (Ekman and Friesen, 1978) to label the emotion samples. Their bimodal fusion model combines facial texture and prosody in a framework of Adaboost multi-stream hidden Markov model (AdaMHMM).
(Joo et al., 2007) investigates the use of S-type membership functions for creating bimodal fusion models for the recognition of five emotions from speech signal and facial expressions. The achieved recognition rate of the fusion model was 70.4% whereas the performance of the audio-based analysis was 63% and the performance of the face-based analysis was 53.4%. (Go et al., 2003) uses Z-type membership functions to compute the membership degree of each of the six emotions based on the facial expression and the speech data. The facial expression recognition algorithm uses multi-resolution analysis based on discrete wavelets. An initial gender classification is done by the pitch of the speech signal criterium. The authors report final emotion recognition results of 95% in case of male and 98.3% for female subjects. (Fellenz et al., 2000) uses a hybrid classification procedure organized in a two-stages architecture to select and fuse the features extracted from face and speech to perform the recognition of emotions. In the first stage, a multi-layered perceptron (MLP) is trained with the backpropagation of error procedure. The second symbolic stage involves the use of PAC learning paradigm for Boolean functions.
(Meng et al., 2007) presents a speech-emotion recognizer that works in combination with an automatic speech recognition system. The algorithm uses Hidden Markov Model – HMM as a classifier. The features considered for the experiments consisted of 39 MFCCs plus pitch, intensity and three formants, including some of their statistical derivates.
(Busso et al., 2004) explores the properties of both unimodal and multimodal systems for emotion recognition in case of four emotion classes. In this study, the multimodal fusion is realized separately at the semantic level and at the feature level. The overall performance of the classifier based on feature level fusion is 89.1% which is close to the performance of the semantic fusion based classifier when the product-combining criterion is used.

MULTIMODAL APPROACH

In our approach, the emotion recognition algorithm works for the prototypic emotions (Ekman and Friesen, 1978) and is based on semantic fusion of audio and video data. We have based our single modality data processing methods on previous work (Datcu and Rothkrantz, 2006; Datcu and Rothkrantz, 2007) we have conducted for the recognition of emotions from human faces and speech.

Facial Expression Recognition

In the case of video data processing, we have developed automatic systems for the recognition of facial expressions for both still pictures and video sequences. The recognition was done by using Viola&Jones features and boosting techniques for face detection (Viola and Jones, 2001), Active Appearance Model – AAM for the extraction of face shape and Support Vector Machines –SVM (Vapnik 1995; Vapnik 1998) for the classification of feature patterns in one of the prototypic facial expressions. For training and testing the systems we have used Cohn-Kanade database (Kanade et al., 2000) by creating a subset of relevant data for each facial expression. The structure of the final dataset is presented in Table 1.

Table 1: The structure of the Cohn-Kanade subset for facial expression recognition.

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

The Active Appearance Model – AAM (Cootes et al., 1998) makes sure the shapes of the face and of the facial features are correctly extracted from each detected face.
Starting with the samples we have collected from the Cohn-Kanade database, we have determined the average face shape and texture (Figure 1).

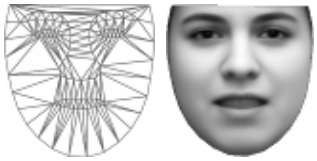

Figure 1: The mean face shape (left) and the mean face texture aligned to the mean shape (right).

According to the AAM model, the shape and texture can be represented as depicted in Equation 1, 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.

Equation 1

$$s = \bar{s} + \Phi_s b_s$$

$$t = \bar{t} + \Phi_t b_t$$

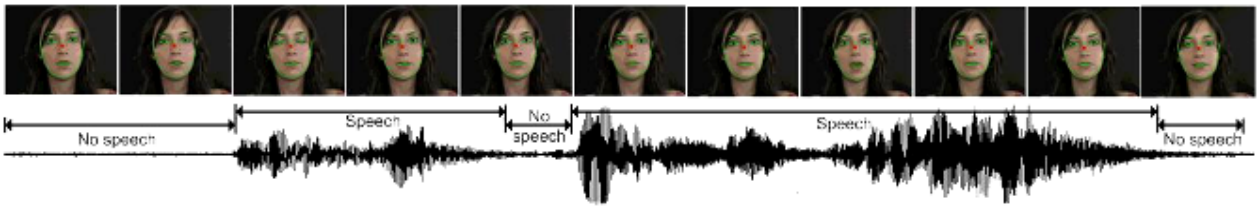

Figure 2: The silence/non-silence detection using multimodal data.
Sample taken from eNTERFACE 2005 database: "Oh my god, there is someone in the house!"

The final combined model contains information regarding both the shape and texture and is written as in Equation 2. The term W_s is a diagonal matrix that introduces the weighting between units of intensities and units of distances.

Equation 2

$$b = \begin{matrix} W_s b_s \\ b_t \end{matrix}$$

Based on the AAM face shape, the facial expression recognition algorithm generates a set of features to be used further on during the emotion classification stage. The features stand for geometric parameters as distances computed between specific Facial Characteristic Points – FCPs (Figure 3).
For the recognition of expressions in still pictures, the distances determined from one face form a representative set of features to reflect the emotion at a certain moment of time. In the case of recognition of facial expressions in video sequences, the features are determined as the variation of the same distances between FCPs as observed during several consecutive frames.

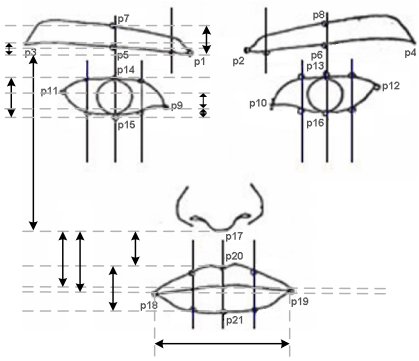

Figure 3: The Facial Characteristic Point FCP model.

Our algorithm for the recognition of emotions in videos implies an initial processing of the multimodal data. Firstly, the audio-video input data is rescaled by conversion to a specific frame-rate (Figure 4). This process may imply downscaling by skipping some video and audio frames. Secondly, the audio data is processed in order to determine the silence and non-silence segments. The resulting segments are correlated to the correspondent audio data and constitute the major data for the analysis.

In the case of facial expression recognition, within each segment an overlaping sliding window (Figure 5) groups together adjacent video frames. Based on the set of video frames, the recognition of facial expressions determines the most probable facial expression using a voting algorithm and a classifier trained on still pictures.

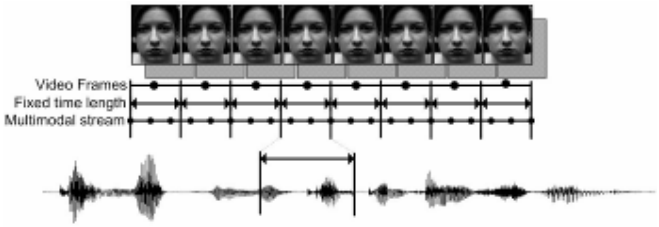

Figure 4: Multimodal frame rescaling algorithm.

For the video oriented classifier, the most probable facial expression is determined by taking into account the variation of the features extracted from all the video frames in the group.

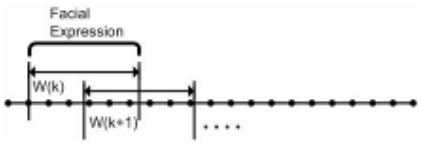

Figure 5: Video frame selection algorithm.

The indentification of silence and non-silence segments is realized by using both acoustic and video information (Figure 2). Apart from running acoustic analysis of the data, speech can be detected by tracking features from a simple FCPs based model that includes data from the mouth area.
The recognition of emotions is realized differently for silence segments and non-silence segments (Figure 6). For silence segments, the emotion is represented by the facial expression as detemined by the facial expression classification algorithm.
For the non-silence segments, the emotion recognition is based on the multimodal semantic fusion of the results of the emotion classification on single modalities. Additionally, the facial expression classification algorithm for non-silence segments determines the most probable facial expression by considering a different set of geometric features. The input features in this case relate to only FCPs from the upper part of the face.

Table 2: The geometric feature set for facial expression recognition for silence data segments.

		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			

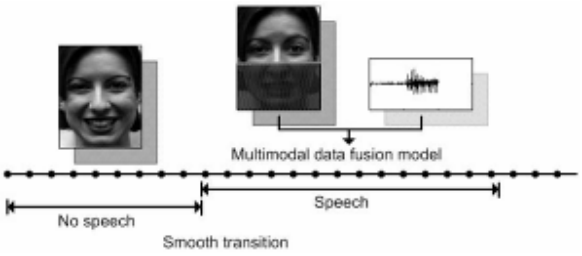

Figure 6: Emotion recognition regarding the transition between two adjacent segments.

The reason for not considering the FCPs of the mouth is explained by the natural influence of the phoneme generation on the mouth shape during the process of speaking.
The geometric features used for the recognition of facial expressions are illustrated in Table 2 for non-silence data segments and in Table 3 for silence data segments.

Table 3: The geometric feature set for facial expression recognition for speech-containing enhanced data segments.

		Feature			Feature
v_1	$(P_1, P_7)_y$	Left eyebrow	v_7	$(P_{14}, P_{15})_y$	Left eye
v_2	$(P_1, P_3)_y$	Left eyebrow	v_8	$(P_9, P_{11})_y$	Left eye
v_3	$(P_2, P_8)_y$	Right eyebrow	v_9	$(P_9, P_{15})_y$	Left eye
v_4	$(P_2, P_4)_y$	Right eyebrow	v_{10}	$(P_{13}, P_{16})_y$	Right eye
v_5	$(P_1, P_9)_y$	Left eyebrow	v_{11}	$(P_{10}, P_{12})_y$	Right eye
v_6	$(P_2, P_{10})_y$	Right eyebrow	v_{12}	$(P_{10}, P_{16})_y$	Right eye

All the FCPs are adjusted for correcting against the head rotation prior to computing the values of the geometric features used for the facial expression classification.

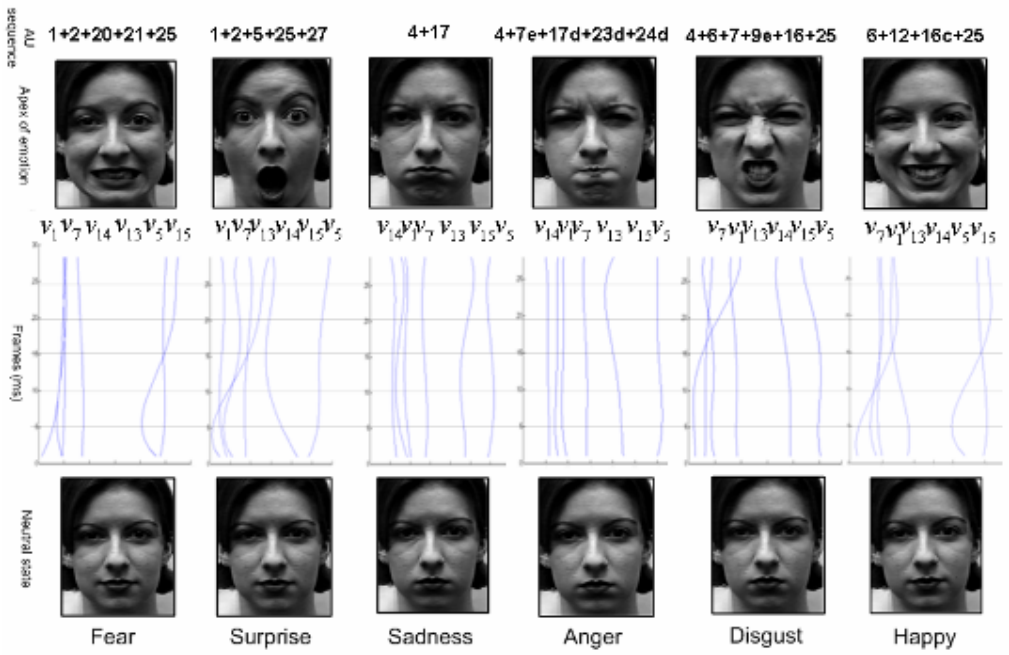

Figure 7: The dependency of temporal changes on emotion featured sequences (reduced parameter set).

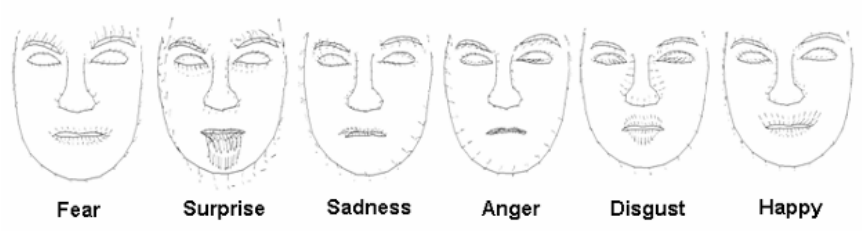

Figure 8: The emotion facial shape deformation patterns for the six prototypic emotion classes.

Moreover, another adjustement of the FCPs applies a correction against the varience of the distance between the subject and the camera. This is realized by scaling all the distance-oriented feature values by the distance between the inner corners of the eyes. The models that use the feature sets in Table 2 and Table 3 allow for the independent consideration of features from both sides of the face. The advantage of a facial expression recognition system that makes use of such a set of features is the ability to still offer good results for limited degrees of occlusion. For such cases, the features computed from the side that is not occluded can be mirrored to the features from the occluded side of the face.

The values of the geometric features over time may be plot for each facial expression (Figure 7).

An alternative to the previously described set of features is to take into account the dynamics of the features presented in Table 2 so as to determine the emotion given the relative deformation of the facial features in time (Figure 8).

Emotion recognition from speech

In the case of emotion recognition from speech, the analysis is handled separately for different number of frames per speech segment (Datcu and Rothkrantz, 2006). In the current approach there are five types of split methods applied on the initial audio data. Each type of split produces a number of data sets, according to all the frame combinations in one segment.

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 pro speaker. 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).

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.

The GentleBoost strong classifier 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.

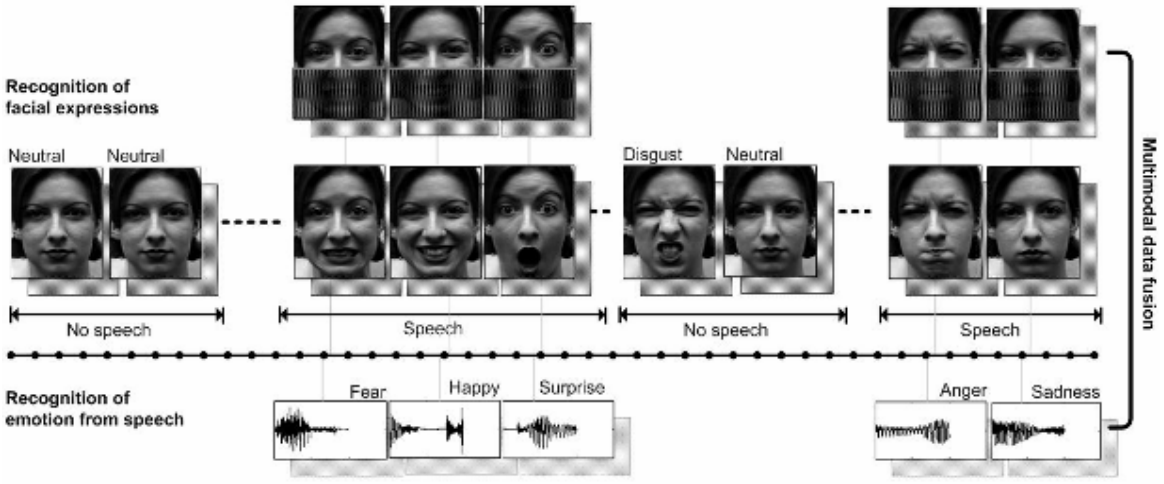

Figure 9: The sequencial recognition of human emotions from audio and video data.

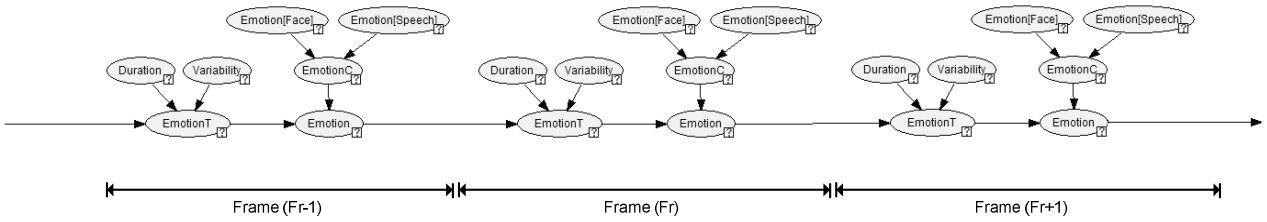

Figure 10: The DBN model for multimodal emotion recognition.

Bimodal emotion data fusion model

Figure 9 illustrates an example of the multimodal emotion recognition algorithm. High-level data fusion works only during the analyses of speech-enhanced multimodal segments.

The fusion model aims at determining the most probable emotion of the subject given the emotions determined in the previous frames. A data window contains the current and the previous *n* frames for the analysis.
Figure 10 depicts the Dynamic Bayesian Network - DBN for the emotion data fusion.
The variable *Duration* represents the stability of last determined emotion in consecutive frames in the analysis window. It has three states: *short*, *normal* and *long*. Each of the three possible situations are assumed to hold differently for each facial expression. Accordingly, it is assumed that for instance an emotion transition is likely to happen from one emotion to another after the former emotion has been shown during a number of consecutive frames.
The variable *Variability* represents the knowledge on the variability of previously determined emotions in the analysis window. It has three states: *low*, *medium* and *high*. Presumably the probability of emotion transition should be higher when the subject has shown rapid changes of emotions during the previous time.
The variable *EmotionT* represents the most probable emotion taking into account only the emotion of the subject determined at the previous frame in the analysis window and the probability of showing another emotion.
The variable *EmotionC* is the emotion of the subject as it is computed by the facial expression recognition and the emotion extraction from speech at the current frame. The variable *Emotion* is the emotion of the subject at the current frame to be determined.

RESULTS

For the classification of facial expressions, different models have been taken into account. In our experiments we have used 2-fold Cross Validation method for testing the performance of the models. For training, we have used Cohn-Kanade database for experiments on facial expression recognition and Berlin database for emotion extraction from speech.
We have partly used the eNTERFACE'05 audio-visual emotion database (Martin et al., 2006) for testing our multimodal algorithms for emotion recognition.

The partial results presented in the paper show the performance achieved by our algorithms for facial expression recognition in processing silence (Table 4 and Table 5) and speech segments (Table 6 and Table 7).
Table 5 shows the results of algorithms that use the dynamic behaviour shown by geometric features as input for the emotion classification process. Addionally we show the results in the case of emotion recognition from speech (Table 8).
Ongoing work is set to test the multimodal fusion model by using eNTERFACE'05 data set.
The results of the facial expression recognition clearly show that a higher performance is obtained by the models that make use of features computed from the entire face shape in comparison to the model that uses information regarding only the eyes and eyebrows.

Table 4: The results for facial expression recognition using SVM (polynomial kernel of degree 3) for still pictures.

(%)	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 5: The results for facial expression recognition using SVM (polynomial kernel of degree 3) for sequence of frames.

(%)	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 6: The results for facial expression recognition using SVM (polynomial kernel of degree 3) for still pictures using only the eyes and eyebrows information.

(%)	Fear	Surprise	Sadness	Anger	Disgust	Happy
Fear	**66.67**	6.67	13.33	0	13.33	0
Surprise	0	**63.64**	0	36.36	0	0
Sadness	0	0	**64.71**	0	35.29	0
Anger	20.00	20.00	0	**60.00**	0	0
Disgust	0	0	25.00	12.50	**62.50**	0
Happy	39.13	0	0	0	0	**60.87**

Table 7: The results for facial expression recognition using SVM (polynomial kernel of degree 3) for sequence of frames using only the eyes and eyebrows information.

(%)	Fear	Surprise	Sadness	Anger	Disgust	Happy
Fear	**70.59**	0	0	0	29.41	0
Surprise	15.00	**70.00**	15.00	0	0	0
Sadness	15.79	15.79	**63.16**	0	5.26	0
Anger	16.67	16.66	0	**66.67**	0	0
Disgust	0	21.22	2	11.11	**65.67**	0
Happy	0	36.36	0	0	0	**63.64**

Table 8: The optimal classifier for each emotion class, Berlin data set.

(%)	**ac (%)**	**tpr (%)**	**fpr (%)**
Anger	0.83±0.03	0.72±0.16	0.13±0.06
Boredom	0.84±0.07	0.49±0.18	0.09±0.09
Disgust	0.92±0.05	0.24±0.43	0.00±0.00
Fear	0.87±0.03	0.38±0.15	0.05±0.04
Happy	0.81±0.06	0.54±0.41	0.14±0.13
Sadness	0.91±0.05	0.83±0.06	0.08±0.06

IMPLEMENTATION

Figure 11 shows a snap shot of our software implementation (Datcu and Rothkrantz, Software Demo 2007) for the bimodal human emotion recognition system. Our system runs on Windows machines. For the detection of faces we have mainly used the implementation of Viola&Jones method from Intel's Open Source Computer Vision Library – OpenCV. We have used AAM-API (Stegmann, 2003) libraries for the implementation of Active Appearance Models. For the speech processing part we have built Tcl/Tk scripts in combination with Snack Sound Toolkit, a public domain toolkit developed at KTH.

Finally, we have built our facial feature extraction routines, the facial expression recognition system and the emotion extraction from speech in C++ programming language. For the classification component, we have used LIBSVM (Chang and Lin, 2001). On an Intel Core 2 CPU @2.00 GHz, 2.00 GB of RAM our software implementation works at speed of about 5 fps.

The AAM module of our initial facial expression recognition system requires the detection of faces for each frame in the incoming video sequence. We have obtained a considerable improvement in terms of speed by nearly doubling the frame rate with an algorithm that uses information regarding the face shape of one subject at the current frame as initial location for the AAM fitting procedure at the next frame in the video sequence (Figure 12). In this way the face detection is run only at certain time intervals comparing to the case when it is run for all the frames.

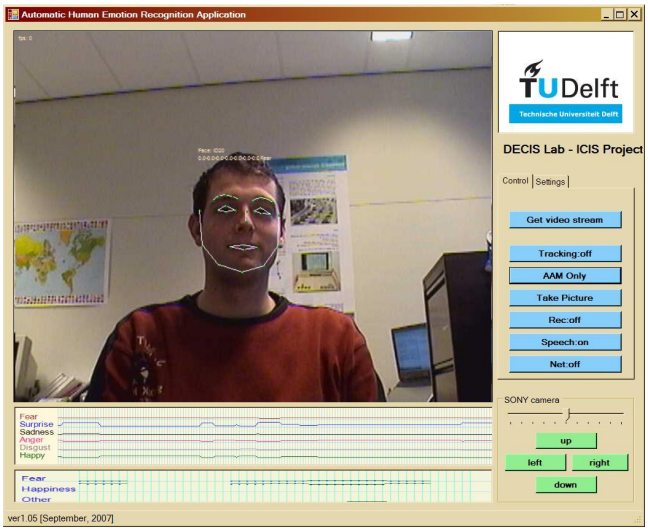

Figure 11: A snap shot of our software implementation of the bimodal emotion recognition algorithm.

The disadvantage of this algorithm is the higher probability of generating faulty results for the extraction of the face shape.

Moreover, the faulty such cases attract definitive errorneous results for the rest of the frames in the sequence. This happens because of the possibly sharp moves of the head, rather low speed of the implementation and because of the incapacity of the AAM algorithm to match model face shapes to image face shapes when the translation, rotation or scalation effects present high magnitudes. The case can be overcome by using an effective face tracking algorithm to anticipate the move of the head in the sequence of frames.

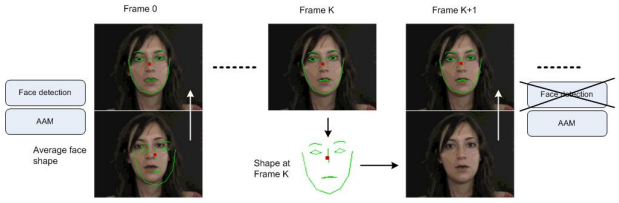

Figure 12: The shape at frame K is used as initial shape estimation for the AAM shape matching algorithm at frame K+1; the face detection is run only once every n frames instead of once for every frame.

CONCLUSION

In our experiments we have used Cohn-Kanade database for training and testing our facial expression recognition models. For the emotion recognition from speech we have used Berlin database that contains utterances in German language. A better approach is to use a unique multimodal database for running the full set of experiments on the algorithms detailed in the paper.

The use of additional semantic information regarding, for instance the emotion level from text or gestures - would greately increase the performance of the multimodal emotion recognition system. In such a situation, more advanced multimodal data fusion models may be developed.

Eventually, the fusion technique described in the paper focuses on information from only the upper part of the face. Instead, an efficient alternative would be to filter out the influence of phonemes and to run the same type of facial expression recognition models also for the speech-enhanced multimodal segments.

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.

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

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

Busso, C., Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, S. Narayanan, "Analysis of Emotion Recognition using Facial Expressions, Speech and Multimodal Information", Proceedings of ACM 6th International Conference on Mutlmodal Interfaces (ICMI 2004), ISBN: 1-58113-890-3, State College, PA, 2004.

Chang, C. C., C. J. Lin, "LIBSVM: a library for support vector machines", http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.

Cootes, T. F., G. J. Edwards, C. J. Taylor, "Active appearance models", Lecture Notes in Computer Science, vol. 1407, pp. 484–98, 1998.

Datcu, D., L. J. M. Rothkrantz, "Multimodal workbench for human emotion recognition", Software Demo at IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'07, Minneapolis, Minnesota, USA, 2007.

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 in still pictures and videos using Active Appearance Models. A comparison approach.", CompSysTech'07, ISBN 978-954-9641-50-9, pp. VI.13-1-VI.13-6, Rousse, Bugaria, June 2007.

Datcu, D., L. J. M. Rothkrantz, "The use of Active Appearance Model for facial expression recognition in crisis environments", Proceedings ISCRAM2007, ISBN 9789054874171, pp. 515-524, 2007.

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

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

Fellenz, W. A., J. G. Taylor, R. Cowie, E. Douglas-Cowie, F. Piat, S. Kollias, C. Orovas, B. Apolloni, "On emotion recognition of faces and of speech using neuralnetworks, fuzzy logic and the ASSESS system", Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks, IJCNN 2000, ISBN: 0-7695-0619-4, pp. 93-98, 2000.

Go, H. J., K. C. Kwak, D. J. Lee, "Emotion recognition from the facial image and speech signal," SICE Annual Conference, Japan, pp. 2890-2895, 2003.

Hoch, S., F. Althoff, G. McGlaun, G. Rigoll, "Bimodal fusion of emotional data in an automotive environment", IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP '05, ISBN: 0-7803-8874-7, Vol.2, pp. 1085-1088, 2005.

Joo, J. T., S. W. Seo, K. E. Ko, K. B. Sim, "Emotion Recognition Method Based on Multimodal Sensor Fusion Algorithm", 8th International Symposium on Advanced Intelligent Systems ISIS'07, 2007.

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

Martin, O., I. Kotsia, B. Macq, I. Pitas, "The eNTERFACE'05 Audio-Visual Emotion Database", Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06), ISBN:0-7695-2571-7, 2006.

Meng, H., J. Pittermann, A. Pittermann, W. Minker, "Combined speech-emotion recognition for spoken human-computer interfaces", IEEE International Conference on Signal Processing and Communications, Dubai (United Emirates), 2007.

OpenCV: Open Source Computer Vision Library; http://www.intel.com/technology/computing/opencv/.

Pantic, M., L. J. M. Rothkrantz, "Toward an affect-sensitive multimodal human-computer interaction", Proceedings of the IEEE, Special Issue on human-computer multimodal interface, 91(9):1370–1390, 2003.

Paleari, M., C. L. Lisetti, "Toward Multimodal Fusion of Affective Cues", HCM'06, Santa Barbara, California, USA, pp. 99-108, October 27, 2006.

Sebe N., I. Cohen, T. Gevers, T. S. Huang, "Emotion Recognition Based on Joint Visual and Audio Cues", Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), 2006.

Song, M., J.J. Bu, C. Chen, N. Li, "Audio-visual based emotion recognition-a new approach", Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004 - CVPR 2004, ISBN: 0-7695-2158-4, pp. 1020-1025, 2004.

Snack Sound Toolkit; http://www.speech.kth.se/snack/.

Stegmann, M. B., "The AAM-API: An Open Source Active Appearance Model Implementation", Medical Image Computing and Computer-Assisted Intervention - MICCAI 2003, 6th Int. Conference, Montréal, Canada, Springer, pp. 951-952, 2003.

Vapnik, V., "The Nature of Statistical Learning Theory", Springer-Verlag, New York, 1995.

Vapnik, V., "Statistical Learning Theory", John Wiley and Sons, Inc., New York, 1998.

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

Zeng, Z., Y. Hu, G. I. Roisman, Z. Wen, Y. Fu, T. S. Huang, "Audio-Visual Spontaneous Emotion Recognition", Artificial Intelligence for Human Computing, ISBN: 978-3-540-72346-2, Springer Berlin/Heidelberg, pp.72-90, 2007.

Wimmer, M., B. Schuller, D. Arsic, G. Rigoll, B. Radig, "LOW-LEVEL FUSION OF AUDIO AND VIDEO FEATURE FOR MULTI-MODAL EMOTION RECOGNITION", In 3rd International Conference on Computer Vision Theory and Applications (VISAPP), Madeira, Portugal, 2008.