Multimodal emotion recognition from expressive faces, body gestures and speech

George Caridakis , Ginevra Castellano , Loic Kessous , Amaryllis Raouzaiou , Lori Malatesta , Stelios Asteriadis and Kostas Karpouzis
Image, Video and Multimedia Systems Laboratory,
National Technical University of Athens
9, Heroon Politechniou str., 15780, Athens, Greece
{gcari, araouz, lori, stiast, kkarpou}@image.ece.ntua.gr
InfoMus Lab, DIST - University of Genova
Viale Causa 13, I-16145, Genova, Italy
Ginevra.Castellano@unige.it
Department of Speech, Language and Hearing,
University of Tel Aviv
Sheba Center, 52621, Tel Aviv, Israel
kessous@post.tau.ac.il

Abstract. In this paper we present a multimodal approach for the recognition of eight emotions that integrates information from facial expressions, body movement and gestures and speech. We trained and tested a model with a Bayesian classifier, using a multimodal corpus with eight emotions and ten subjects. First individual classifiers were trained for each modality. Then data were fused at the feature level and the decision level. Fusing multimodal data increased very much the recognition rates in comparison with the unimodal systems: the multimodal approach gave an improvement of more than 10% with respect to the most successful unimodal system. Further, the fusion performed at the feature level showed better results than the one performed at the decision level.

Keywords: Affective body language, Affective speech, Emotion recognition, Multimodal fusion

1 Introduction

In the last years, research in the human-computer interaction area increasingly addressed the communication aspect related to the "implicit channel", that is the

Please use the following format when citing this chapter:

Caridakis, G., Castellano, G., Kessous, L., Raouzaiou, A., Malatesta, L., Asteriadis, S., Karpouzis, K., 2007, in IFIP International Federation for Information Processing, Volume 247, Artificial Intelligence and Innovations 2007: From Theory to Applications, eds. Boukis, C., Pnevmatikakis, L., Polymenakos, L., (Boston: Springer), pp. 375-388.

channel through which the emotional domain interacts with the verbal aspect of the communication [1]. One of the challenging issues is to endow a machine with an emotional intelligence. Emotionally intelligent systems must be able to create an affective interaction with users: they must be endowed with the ability to perceive, interpret, express and regulate emotions [2]. Recognising users' emotional state is then one of the main requirements for computers to successfully interact with humans. Most of the works in the affective computing field do not combine different modalities into a single system for the analysis of human emotional behaviour: different channels of information (mainly facial expressions and speech) are considered independently to each other. Further, there are only a few attempts to integrate information from body movement and gestures. Nevertheless, Sebe et al. [3] and Pantic et al. [4] highlight that an ideal system for automatic analysis and recognition of human affective information should be multimodal, as the human sensory system is. Moreover, studies from the psychology show the need to consider the integration of different non-verbal behaviour modalities in the human-human communication [5].

In this paper we present a multimodal approach for the recognition of eight acted emotional states (anger, despair, interest, pleasure, sadness, irritation, joy and pride) that integrates information from facial expressions, body movement and gestures and speech. In our work we trained and tested a model with a Bayesian classifier, using a multimodal corpus with ten subjects collected during the Third Summer School of the HUMAINE EU-IST project, held in Genova in September 2006. In the following sections we describe the systems based on the analysis of the single modalities and compare different strategies to perform the data fusion for the multimodal emotion recognition.

2 Related work

Emotion recognition has been investigated with three main types of databases: acted emotions, natural spontaneous emotions and elicited emotions. The best results are generally obtained with acted emotion databases because they contain strong emotional expressions. Literature on speech (see for example Banse and Scherer [6]) shows that most part of the studies were conducted with emotional acted speech. Feature sets for acted and spontaneous speech have recently been compared by [7]. Generally, few acted-emotion speech databases included speakers with several different native languages. In the last years, some attempts to collect multimodal data were done: some examples of multimodal databases can be found in [8] [9] [10].

In the area of unimodal emotion recognition, there have been many studies using different, but single, modalities. Facial expressions [11] [12], vocal features [13] [14], body movements and postures [15] [16], physiological signals [17] have been used as inputs during these attempts, while multimodal emotion recognition is currently gaining ground [18] [19] [20]. Nevertheless, most of the works consider the integration of information from facial expressions and speech and there are only a few attempts to combine information from body movement and gestures in a multimodal framework. Gunes and Piccardi [21] for example fused at different levels facial expressions and body gestures information for bimodal emotion recognition. Further, el Kaliouby and Robinson [22] proposed a vision-based computational model to infer acted mental states from head movements and facial expressions.

A wide variety of machine learning techniques have been used in emotion recognition approaches [11] [1]. Especially in the multimodal case [4], they all employ a large number of audio, visual or physiological features, a fact which usually impedes the training process; therefore, it is necessary to find a way to reduce the number of utilized features by picking out only those related to emotion. One possibility in this direction is to use neural networks, since they enable us to pinpoint the most relevant features with respect to the output, usually by observing their weights. An interesting work in this area is the sensitivity analysis approach by Engelbrecht et al. [23]. Sebe et al. [3] highlight that probabilistic graphical models, such as Hidden Markov Models, Bayesian networks and Dynamic Bayesian networks are very well suited for fusing different sources of information in multimodal emotion recognition and can also handle noisy features and missing values of features all by probabilistic inference.

In this work we combine a wrapper feature selection approach to reduce the number of features and a Bayesian classifier both for the unimodal and the multimodal emotion recognition.

3 Collection of multimodal data

The corpus used in this study was collected during Third Summer School of the HUMAINE EU-IST project, held in Genova in September 2006. The overall recording procedure was based on the GEMEP corpus [10], a multimodal collection of portrayed emotional expressions: we recorded data on facial expressions, body movement and gestures and speech.

3.1 Subjects

Ten participants of the summer school distributed as evenly as possible concerning their gender (Figure 1) participated to the recordings. Subjects represented five different nationalities: French, German, Greek, Hebrew, Italian.

Fig. 1. The participants who took part to the recordings.

3.2 Technical set up

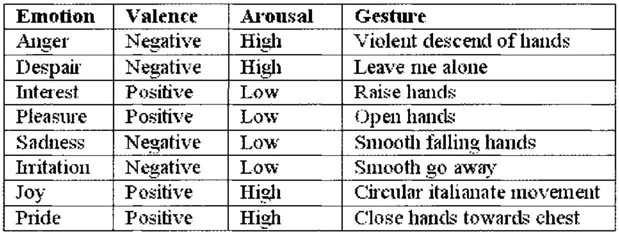
Two DV cameras (25 fps) recorded the actors from a frontal view. One camera recorded the actor's body and the other one was focused on the actor's face. We have chosen such a setup because the resolution required for facial features extraction is much larger than the one for body movement detection or hand gestures tracking. This could only be achieved if one camera zoomed in the actor's face. Video streams were synchronised manually after the recording process. We adopted some restrictions concerning the actor's behaviour and clothing. Long sleeves and covered neck were preferred since the majority of the hand and head detection algorithms are based on colour tracking. Further, uniform background was used to make the background subtraction process easier. As for the facial features extraction process we considered some prerequisites such as the lack of eyeglasses, beards, moustaches.

For the voice recordings we used a direct-to-disk computer-based system. The speech samples were directly recorded on the hard disk of the computer using sound editing software. We used an external sound card connected to the computer by IEEE 1394 High Speed Serial Bus (also known as FireWire or i.Link). A microphone mounted on the actors' shirt was connected to an HF emitter (wireless system emitter) and the receiver was connected to the sound card using a XLR connector (balanced audio connector for high quality microphones and connections between equipments). The external sound card included a preamplifier (for two XLR inputs) that was used in order to adjust the input gain and to minimize the impact of signal-to-noise ratio of the recording system. The sampling rate of the recording was 44.1 kHz and the quantization was 16 bit, mono.

3.3 Procedure

Participants were asked to act eight emotional states: anger, despair, interest, pleasure, sadness, irritation, joy and pride, equally distributed in the space valence-arousal (see Table 1). During the recording process one of the authors had the role of the director guiding the actors through the process. Participants were asked to perform specific gestures that exemplify each emotion. Selected gestures are shown in Table 1.

Table 1. The acted emotions and the *emotion-specific gestures*.



Emotion	Valence	Arousal	Gesture
Anger	Negative	High	Violent descend of hands
Despair	Negative	High	Leave me alone
Interest	Positive	Low	Raise hands
Pleasure	Positive	Low	Open hands
Sadness	Negative	Low	Smooth falling hands
Irritation	Negative	Low	Smooth go away
Joy	Positive	High	Circular italianate movement
Pride	Positive	High	Close hands towards chest

As in the GEMEP corpus [10], a pseudo-linguistic sentence was pronounced by the actors during acting the emotional states. The sentence "Toko, damato ma gali sa" was designed in order to fulfil different needs. First, as the different speakers have different native languages, using a specific language was not so adequate to this study. Then we wanted the sentence to include phonemes that exist in all the languages of all the speakers. Also, the words in the sentence are composed of simple diphones (‘ma' and 'sa'), two ('gali' 'toko') or three diphones ('damato'). Then, the vowels included ('o' , 'a' , 'i') are vowels that are relatively distant in a vowel space, for example the vowel triangle, and have a pronunciation mostly similar in all the languages of the group of speakers. We suggested the speakers a meaning for the sentence. 'Toko' is supposed to be the name of an 'agent', i.e., a real or artificial individual, who the speakers/users are interacting with. We chose for this word two stops consonants (also known as plosives or stop-plosives) /t/ and /k/ and two identical vowels /o/. This was done in order to allow the study of certain acoustic correlates. Then 'damato ma gali sa' is supposed to mean something like 'can you open it'. The word 'it' could correspond to a folder, a file, a box, a door or whatever.

Each emotion was acted three times by each actor, so that we collected 240 posed gestures, facial expressions and speech samples.

4 Feature extraction

4.1 Face feature extraction

An overview of the proposed methodology is illustrated in Figure 2. The face is first located, so that approximate facial feature locations can be estimated from the head position and rotation. Face roll rotation is estimated and corrected and the head is segmented focusing on the following facial areas: left eye/eyebrow, right eye/eyebrow, nose and mouth. Each of those areas, called feature-candidate areas, contains the features whose boundaries need to be extracted for our purposes. Inside the corresponding feature-candidate areas precise feature extraction is performed for each facial feature, i.e. eyes, eyebrows, mouth and nose, using a multi-cue approach, generating a small number of intermediate feature masks. Feature masks generated for each facial feature are fused together to produce the final mask for that feature. The mask fusion process uses anthropometric criteria [24] to perform validation and weight assignment on each intermediate mask; all the feature's weighted masks are then fused to produce a final mask along with confidence level estimation.

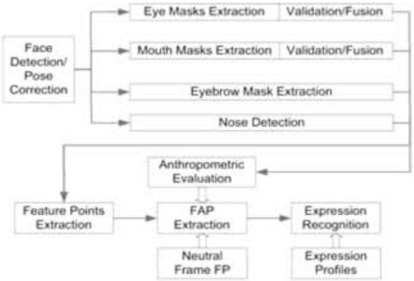


Fig. 2. High-level overview of the facial feature extraction process.

Since this procedure essentially locates and tracks points in the facial area, we chose to work with MPEG-4 FAPs (Facial Animation Parameters) and not Action

Units (AUs), since the former are explicitly defined to measure the deformation of these feature points. In addition to this, discrete points are easier to track in cases of extreme rotations and their position can be estimated based on anthropometry in cases of occlusion, whereas this is not usually the case with whole facial features. Another feature of FAPs which proved useful is their value (or magnitude), which is crucial in order to differentiate cases of varying activation of the same emotion (e.g. joy and exhilaration) [25] and exploit fuzziness in rule-based systems [12]. Measurement of FAPs requires the availability of a frame where the subject's expression is found to be neutral. This frame is called the *neutral frame* and is manually selected from video sequences to be analyzed or interactively provided to the system when initially brought into a specific user's ownership. The final feature masks are used to extract 19 Feature Points (FPs) [25]; Feature Points obtained from each frame are compared to FPs obtained from the neutral frame to estimate facial deformations and produce the FAPs. Confidence levels on FAP estimation are derived from the equivalent feature point confidence levels. The FAPs are used along with their confidence levels to provide the facial expression estimation.

In accordance with the other modalities, facial features need to be processed so as to have one vector of values per tune. FAPs originally correspond to every frame in the tune. Two approaches were reviewed. The first approach consisted of extracting the most prominent frame within a tune. During this process, a mean value is calculated for every FAP within the tune and next the frame with the highest variation is selected based on this set of values. On the other hand a way to imprint the temporal evolution of the FAP values is to calculate a set of statistical features over these values and their derivatives. The whole process was inspired by the equivalent process performed in the acoustic features. We have selected the latter since the recognition rate achieved was superior. More sophisticated techniques to extract the most prominent frame are included in our plans for future work.

4.2 Body feature extraction

Tracking of body and hands of the subjects was done using the EyesWeb platform [26]. Starting from the silhouette and the hands blobs of the actors, we extracted five main expressive motion cues, using the EyesWeb Expressive Gesture Processing Library [27]: quantity of motion and contraction index of the body, velocity, acceleration and fluidity of the hand's barycenter. The data were normalised according to the behaviour shown by each actor, considering the maximum and the minimum values of each motion cue in each actor, in order to compare data from all the subjects. Automatic extraction allows to obtain temporal series of the selected motion cues over time, depending on the video frame rate. For each profile of the motion cues we selected then a subset of features describing the dynamics of the cues over time. Based on the model proposed in [28] we extracted the following dynamic indicators of the motion cues temporal profile: initial and final slope, initial and final slope of the main peak, maximum value, ratio between the maximum value and the duration of the main peak, mean value, ratio between the mean and the maximum value, ratio between the absolute maximum and the biggest following relative maximum, centroid of energy, distance between maximum value and centroid of energy, symmetry index, shift index of the main peak, number of peaks, number of peaks preceding the main one, ratio between the main peak duration and the whole profile duration. This process was made for each motion cue of all the videos of the corpus, so that each gesture is characterised by a subset of 80 motion features.

4.3 Speech feature extraction

The set of features that we used contains features based on intensity, pitch, MFCC (Mel Frequency Cepstral Coefficient), Bark spectral bands, voiced segment characteristics and pause length. The full set contains 377 features. The features from the intensity contour and the pitch contour are extracted using a set of 32 statistical features. This set of features is applied both to the pitch and intensity contour and to their derivatives. Not any normalization has been applied before feature extraction. In particular, we didn't perform user or gender normalization for pitch contour as it is often done in order to remove difference between registers. We considered the following 32 features: maximum, mean and minimum values, sample mode (most frequently occurring value), interquartile range (difference between the 75th and 25th percentiles), kurtosis, the third central sample moment, first (slope) and second coefficients of linear regression, first, second and third coefficients of quadratic regression, percentiles at 2.5 %, 25 %, 50 %, 75 %, and 97.5 %, skewness, standard deviation, variance. Thus, we have 64 features based on the pitch contour and 64 features based on the intensity contour. This feature set was used originally for inspecting a contour such as a pitch contour or a loudness contour, but these features are also meaningful for inspecting evolution over time or spectral axis. Indeed, we also extracted similar features on the Bark spectral bands as done in [29]. We also extracted 13 MFCCs using time averaging on time windows. Features derived from pitch values and lengths of voiced segments were also extracted using a set of 35 features applied to both of them. We also extracted features based on pause (or silence) length and non-pauses lengths (35 each).

5 Uni-modal and multimodal emotion recognition

In order to compare the results of the unimodal and the multimodal systems, we used a common approach based on a Bayesian classifier (BayesNet) provided by the software Weka, a free toolbox containing a collection of machine learning algorithms for data mining tasks [30]. In Figure 3 we show an overview of the framework we propose:

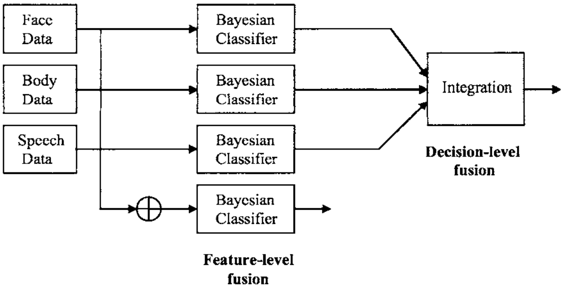


Fig.3. Overview of the framework.

As shown in the left part of the diagram, a separate Bayesian classifier was used for each modality (face, gestures, speech). All sets of data were normalized. Features discretisation based on Kononenko's MDL (minimum description length) criterion [31] was done to reduce the learning complexity. A wrapper approach to feature subset selection (which allows to evaluate the attribute sets by using a learning scheme) was used in order to reduce the number of inputs to the classifiers and find the features that maximize the performance of the classifier. A best-first search method in forward direction was used. Further, in all the systems, the corpus was trained and tested using the cross-validation method.

We evaluated two different models: (1) a model obtained training and testing all the 240 data samples, even when data from some modalities is missing in some samples; (2) a model obtained using only the data available for the three modalities. As shown in the results in the next section, the first model allows to manage also data with missing samples, but it is less precise; the second one is less flexible, but is more precise in the classification.

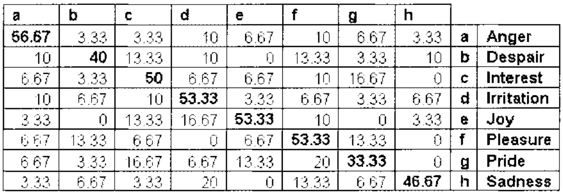
To fuse facial expressions, gestures and speech information, two different approaches were implemented (right of Figure 3): feature-level fusion, where a single classifier with features of the three modalities is used; and decision-level fusion, where a separate classifier is used for each modality and the outputs are combined a posteriori. In the second approach the output was computed combining the posterior probabilities of the unimodal systems. We made experiments using two different approaches for the decision-level fusion. The first approach consisted of selecting the emotion that received the best probability in the three modalities. The second approach consisted of selecting the emotion that corresponds to the majority of 'voting' from the three modalities; if a majority was not possible to define (for example when each unimodal system gives in output a different emotion), the emotion that received the best probability in the three modalities was selected.

6 Results

6.1 Emotion recognition from facial expressions

Table 2 shows the confusion matrix of the emotion recognition system based on facial expressions when all the samples are used (first model). The overall performance of this classifier was 48.3 % (it increases up to 59.6 % when only the samples available for all the modalities are used). The most recognised emotions were anger (56.67 %), irritation, joy and pleasure (53.33 %). Pride is misclassified with pleasure (20%), while sadness is misclassified with irritation (20 %), an emotion in the same valence-arousal quadrant.

Table 2: Confusion matrix of the emotion recognition system based on facial expressions.

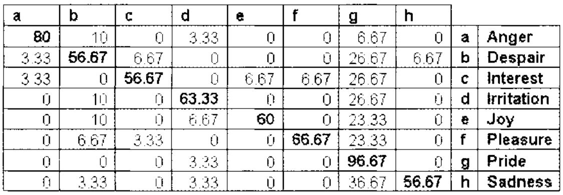


a	b	c	d	e	f	g	h		
56.67	3.33	3.33	10	6.67	10	6.67	3.33	a	Anger
10	**40**	13.33	10	0	13.33	3.33	10	b	Despair
6.67	3.33	**50**	6.67	6.67	10	16.67	0	c	Interest
10	6.67	10	**53.33**	3.33	6.67	3.33	6.67	d	Irritation
3.33	0	13.33	16.67	**53.33**	10	0	3.33	e	Joy
6.67	13.33	6.67	0	6.67	**53.33**	13.33	0	f	Pleasure
6.67	3.33	16.67	6.67	13.33	20	**33.33**	0	g	Pride
3.33	6.67	3.33	20	0	13.33	6.67	**46.67**	h	Sadness

6.2 Emotion recognition from gestures

Table 3 shows the performance of the emotion recognition system based on gestures when all the samples are used (first model). The overall performance of this classifier was 67.1 % (it increases up to 83.2 % when only the samples available for all the modalities are used). Anger and pride are recognised with very high accuracy (80 and 96.67 % respectively). Sadness was partly misclassified with pride (36.67 %), as well as the majority of the emotions, except for anger.

Table 3: Confusion matrix of the emotion recognition system based on gestures.

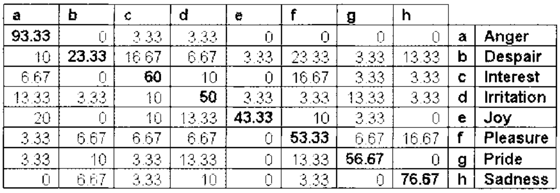


a	b	c	d	e	f	g	h		
80	10	0	3.33	0	0	6.67	0	a	Anger
3.33	**56.67**	6.67	0	0	0	26.67	6.67	b	Despair
3.33	0	**56.67**	0	6.67	6.67	26.67	0	c	Interest
0	10	0	**63.33**	0	0	26.67	0	d	Irritation
0	10	0	6.67	**60**	0	23.33	0	e	Joy
0	6.67	3.33	0	0	**66.67**	23.33	0	f	Pleasure
0	0	0	3.33	0	0	**96.67**	0	g	Pride
0	3.33	0	3.33	0	0	36.67	**56.67**	h	Sadness

6.3 Emotion recognition from speech

Table 4 displays the confusion matrix of the emotion recognition system based on speech when all the samples are used (first model). The overall performance of this classifier was 57.1 (it increases up to 70.8 % when only the samples available for all the modalities are used). Anger and sadness are classified with high accuracy (93.33 and 76.67% respectively). Despair obtained a very low recognition rate and was mainly confused with pleasure (23.33%).

Table 4: Confusion matrix of the emotion recognition system based on speech.

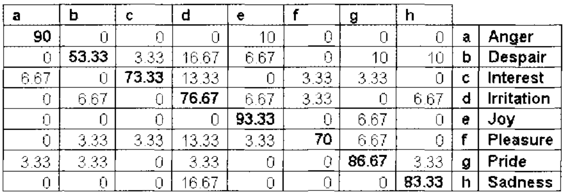


a	b	c	d	e	f	g	h		
93.33	0	3.33	3.33	0	0	0	0	a	Anger
10	**23.33**	16.67	6.67	3.33	23.33	3.33	13.33	b	Despair
6.67	0	**60**	10	0	16.67	3.33	3.33	c	Interest
13.33	3.33	10	**50**	3.33	3.33	13.33	3.33	d	Irritation
20	0	10	13.33	**43.33**	10	3.33	0	e	Joy
3.33	6.67	6.67	6.67	0	**53.33**	6.67	16.67	f	Pleasure
3.33	10	3.33	13.33	0	13.33	**56.67**	0	g	Pride
0	6.67	3.33	10	0	3.33	0	**76.67**	h	Sadness

6.4 Feature-level fusion

Table 5 displays the confusion matrix of the multimodal emotion recognition system when all the samples are used (first model). The overall performance of this classifier was 78.3 % (it increases up to 89.4 % when only the samples available for all the modalities are used), which is much higher than the performance obtained by the most successful unimodal system, the one based on gestures. The diagonal components reveal that all the emotions, apart from despair, can be recognised with more than the 70 % of accuracy. Anger was the emotion recognised with highest accuracy, as in all the unimodal systems.

Table 5: Confusion matrix of the multimodal emotion recognition system.



a	b	c	d	e	f	g	h		
90	0	0	0	10	0	0	0	a	Anger
0	**53.33**	3.33	16.67	6.67	0	10	10	b	Despair
6.67	0	**73.33**	13.33	0	3.33	3.33	0	c	Interest
0	6.67	0	**76.67**	6.67	3.33	0	6.67	d	Irritation
0	0	0	0	**93.33**	0	6.67	0	e	Joy
0	3.33	3.33	13.33	3.33	**70**	6.67	0	f	Pleasure
3.33	3.33	0	3.33	0	0	**86.67**	3.33	g	Pride
0	0	0	16.67	0	0	0	**83.33**	h	Sadness

6.5 Decision level fusion

The approach based on decision-level fusion obtained lower recognition rates than that based on feature-level fusion. The performance of the classifier was 74.6 %, both using the best probability and the majority voting plus best probability approach.

The performance of the classifier increases up to 85.1 % for the first approach and 88.20 % for the second approach when only the samples available for all the modalities are used.

7 Discussion and conclusions

We presented a multimodal framework for analysis and recognition of emotion starting from expressive faces, gestures and speech. We trained and tested a model with a Bayesian classifier, using a multimodal corpus with eight acted emotions and ten subjects of five different nationalities.

We experimented our approach on a dataset of 240 samples for each modality (face, body, speech), considering also instances with missing values. We also evaluated a model built disregarding instances with missing values. The first model obtained lower recognition rates for the eight emotions than the second one, both in the unimodal systems and in the multimodal system, but it allows to manage also data with missing values, condition close to a real situation. Considering the performances of the unimodal emotion recognition systems, the one based on gestures appears to be the most successful, followed by the one based on speech and the one based on facial expressions. We note that in this study we used *emotion-specific gestures*: these are gestures that are selected so as to express each specific emotion. An alterative approach which may also be of interest would be to recognise emotions from different expressivity of the same gesture (one not necessarily associated with any specific emotion) performed under different emotional conditions. This would allow good comparison with contemporary systems based on facial expressions and speech and will be considered in our future work. Fusing multimodal data increased very much the recognition rates in comparison with the unimodal systems: the multimodal approach gave an improvement of more than 10 % compared to the performance of the system based on gestures, when all the 240 samples are used. Further, the fusion performed at the feature level showed better performances than the one performed at the decision-level, highlighting that processing input data in a joint feature space is more successful.

We can conclude that using three different modalities highly increases the recognition performance of an automatic emotion recognition system. That is helpful also when some values for features of some modalities are missing. On the other hand, humans use more than one modality to recognise emotions and process signals in a complementary manner, so it is expected that an automatic system shows a similar behaviour.

Acknowledgments: The research work has been realised in the framework of the EU-IST Project HUMAINE (Human-Machine Interaction Network on Emotion), a Network of Excellence (NoE) in the EU 6th Framework Programme (2004-2007).

References

1. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G.: Emotion recognition in human-computer interaction, IEEE Signal Processing Magazine, January 2001.
2. Picard, R.: Affective computing, Boston, MA: MIT Press (1997).
3. Sebe, N., Cohen, I., Huang, T.S.: Multimodal Emotion Recognition, Handbook of Pattern Recognition and Computer Vision, World Scientific, ISBN 981-256-105-6, January 2005.
4. Pantic, M., Sebe, N., Cohn, J., Huang, T.S.: Affective Multimodal Human-Computer Interaction, ACM Multimedia, pp. 669 - 676, Singapore, November 2005.
5. Scherer, K. R., Wallbott, H. G.: Analysis of Nonverbal Behavior. HANDBOOK OF DISCOURSE: ANALYSIS, Vol. 2, Cap.11, Academic Press London (1985).
6. Banse, R., Scherer, K.R.: Acoustic Profiles in Vocal Emotion Expression. Journal of Personality and Social Psychology. 614-636 (1996).
7. Vogt, T., André, E.: Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. IEEE International Conference on Multimedia & Expo (ICME 2005).
8. Gunes H., Piccardi M.: A Bimodal Face and Body Gesture Database for Automatic Analysis of Human Nonverbal Affective Behavior, Proc. of ICPR 2006 the 18th International Conference on Pattern Recognition, 20-24 Aug. 2006, Hong Kong, China.
9. Bänziger, T., Pirker, H., Scherer, K.: Gemep - geneva multimodal emotion portrayals: a corpus for the study of multimodal emotional expressions. In L. Deviller et al. (Ed.), Proceedings of LREC'06 Workshop on Corpora for Research on Emotion and Affect (pp. 15-019). Genoa. Italy (2006).
10. Douglas-Cowie, E., Campbell, N., Cowie, R., Roach, P.: Emotional speech: towards a new generation of databases. Speech Communication, 40, 33–60 (2003).

11. Pantic, M., Rothkrantz, L.J.M.: Automatic analysis of facial expressions: The state of the art. IEEE Trans. on Pattern Analysis and Machine Intelligence, 22(12):1424–1445 (2000).
12. Ioannou, S., Raouzaiou, A., Tzouvaras, V., Mailis, T., Karpouzis, K., Kollias, S. : Emotion recognition through facial expression analysis based on a neurofuzzy network, Neural Networks, Elsevier, Vol. 18, Issue 4, May 2005, pp. 423-435.
13. Cowie, R., Douglas-Cowie, E.: Automatic statistical analysis of the signal and prosodic signs of emotion in speech. In Proc. International Conf. on Spoken Language Processing, pp. 1989–1992 (1996).
14. K.R. Scherer: Adding the affective dimension: A new look in speech analysis and synthesis, In Proc. International Conf. on Spoken Language Processing, pp. 1808–1811, (1996).
15. Camurri, A., Lagerlöf, I, Volpe, G.: Recognizing Emotion from Dance Movement: Comparison of Spectator Recognition and Automated Techniques, International Journal of Human-Computer Studies, 59(1-2), pp. 213-225, Elsevier Science, July 2003.
16. Bianchi-Berthouze, N., Kleinsmith, A. A categorical approach to affective gesture recognition, Connection Science, V. 15, N. 4, pp. 259-269. (2003).
17. Picard, R.W., Vyzas, E., Healey, J.: Toward machine emotional intelligence: Analysis of affective physiological state, IEEE Trans. on Pattern Analysis and Machine Intelligence, 23(10):1175–1191 (2001).
18. Pantic M., Rothkrantz, L.J.M.: Towards an Affect-sensitive Multimodal Human-Computer Interaction, Proceedings of the IEEE, vol. 91, no. 9, pp. 1370-1390 (2003).
19. Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C.M., Kazemzaeh, A., Lee, S., Neumann, U., Narayanan, S.: "Analysis of Emotion Recognition using Facial Expressions, Speech and Multimodal information," Proc. of ACM 6th int'l Conf. on Multimodal Interfaces (ICMI 2004), State College, PA, Oct. 2004. pp205-211.
20. Kim, J., André, E., Rehm, M., Vogt, T., Wagner, J.: Integrating information from speech and physiological signals to achieve emotional sensitivity. Proc. of the 9th European Conference on Speech Communication and Technology (2005).
21. Gunes H, Piccardi M.: Bi-modal emotion recognition from expressive face and body gestures. Journal of Network and Computer Applications (2006), doi:10.1016/j.jnca.2006.09.007
22. el Kaliouby, R., Robinson, P.: Generalization of a Vision-Based Computational Model of Mind-Reading. In Proceedings of First International Conference on Affective Computing and Intelligent Interfaces, pp 582-589 (2005).
23. Engelbrecht, A.P., Fletcher, L., Cloete, I.: Variance analysis of sensitivity information for pruning multilayer feedforward neural networks, Neural Networks, 1999. IJCNN '99. International Joint Conference on, Vol.3, Iss., 1999, pp:1829-1833 vol.3.
24. Young, J. W.: Head and Face Anthropometry of Adult U.S. Civilians, FAA Civil Aeromedical Institute, 1963-1993 (final report 1993)
25. Raouzaiou, A., Tsapatsoulis, N., Karpouzis, K., Kollias, S.: Parameterized facial expression synthesis based on MPEG-4, EURASIP Journal on Applied Signal Processing, Vol. 2002, No 10, 2002, pp. 1021-1038.
26. Camurri, A., Coletta, P., Massari, A., Mazzarino, B., Peri, M., Ricchetti, M., Ricci, A. and Volpe, G.:Toward real-time multimodal processing: EyesWeb 4.0, in Proc. AISB 2004 Convention: Motion, Emotion and Cognition, Leeds, UK, March 2004.
27. Camurri, A., Mazzarino, B., and Volpe, G.: Analysis of Expressive Gesture: The Eyesweb Expressive Gesture Processing Library, in A. Camurri, G.Volpe (Eds.), Gesture-based Communication in Human-Computer Interaction, LNAI 2915, Springer Verlag (2004).
28. Castellano, G., Camurri, A., Mazzarino, B., Volpe, G.: A mathematical model to analyse the dynamics of gesture expressivity, in Proc. of AISB 2007 Convention: Artificial and Ambient Intelligence, Newcastle upon Tyne, UK, April 2007.

29. Kessous, L., Amir, N.: Comparison of feature extraction approaches based on the Bark time/frequency representation for classification of expressive speechpaper submitted to Interspeech 2007.
30. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco (2005).
31. Kononenko, I.: On Biases in Estimating Multi-Valued Attributes. In: 14th International Joint Conference on Articial Intelligence, 1034-1040 (1995).