Multimodal emotion recognition from expressive faces, body gestures and speech

Ginevra Castellano[1], Loic Kessous[2], and George Caridakis[3]

[1] InfoMus Lab, DIST - University of Genova
Viale Causa 13, I-16145, Genova, Italy
Ginevra.Castellano@unige.it
[2] Department of Speech, Language and Hearing, University of Tel Aviv
Sheba Center, 52621, Tel Aviv, Israel
kessous@post.tau.ac.il
[3] Image, Video and Multimedia Systems Laboratory, National Technical University of Athens
9, Heroon Politechniou str., 15780, Athens, Greece
gcari@image.ece.ntua.gr

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 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] highlight that an ideal system for automatic analysis and recognition of human affective information should be multimodal. Moreover, studies from the psychology show the need to consider the integration of different behaviour modalities in the human-human communication [4].

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

In the area of unimodal emotion recognition, there have been many studies using different, but single, modalities. Facial expressions [5], vocal features [6], body movements [7] have been used as inputs during these attempts, while multimodal emotion recognition is currently gaining ground [8]. 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 [9] for example fused at different levels facial expressions and body gestures information for bimodal emotion recognition.

A wide variety of machine learning techniques have been used in emotion recognition approaches. Especially in the multimodal case, 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 used 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. [10].

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 [11], a multimodal collection

of portrayed emotional expressions: we simultaneously recorded data on facial expressions, body movement and gestures and speech.

3.1 Subjects

Ten participants of the summer school (6 male and 7 female) participated to the recordings. Subjects represented five different nationalities: French, German, Greek, Hebrew, Italian.

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.
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. 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. The external sound card included a preamplifier (for two XLR inputs) that was used in order to adjust the input gain and to minimise 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. All instructions were provided based on the procedure used during the collection of the GEMEP corpus [11]. For selecting the emotion-specific gestures we have borrowed ideas from a figure animation research area dealing with posturing of a figure [12] and came up with the gestures 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 [11], 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. We suggested the speakers a meaning for the sentence. 'Toko' is supposed to be the name of a person, who the speakers/users are interacting with. Then 'damato ma gali sa' is supposed to mean something like 'can you open it'.

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

As first step the face was located, so that approximate facial feature locations could be estimated from the head position and rotation. The head was 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 was 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 were fused together to produce the final mask for that feature. 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. 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 analysed or interactively provided to the system when initially brought into a specific user's ownership. The final feature masks were used to extract 19 Feature Points (FPs) [13]**;** Feature Points obtained from each frame were compared to FPs obtained from the neutral frame to estimate facial deformations and produce the FAPs. Confidence levels on FAP estimation were derived from the equivalent feature point confidence levels. The FAPs were used along with their confidence levels to provide the facial expression estimation.

In accordance with the other modalities, facial features needed to be processed so as to have one vector of values per sentence. FAPs originally correspond to every frame in the sentence. A way to imprint the temporal evolution of the FAP values was 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.

4.2 Body feature extraction

Tracking of body and hands of the subjects was done using the EyesWeb platform [14]. 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: quantity of motion and contraction index of the body, velocity, acceleration and fluidity of the hand's barycenter. All data were normalised. 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 a subset of features describing the dynamics of the cues over time: 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. 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. 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 [15].

A separate Bayesian classifier was used for each modality (face, gestures, speech). All sets of data were normalised. Features discretisation was done to reduce the learning complexity. A wrapper approach to feature subset selection was used in order to reduce the number of inputs to the classifiers and find the features that

maximise 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.

To fuse faces, gestures and speech information, two different approaches were implemented: 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: the output was computed combining the posterior probabilities of the unimodal systems.

6 Results

6.1 Emotion recognition from facial expressions

Table 2 shows the confusion matrix of the emotion recognition system based on facial expressions. The overall performance of this classifier was 48.3 %. 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.. The overall performance of this classifier was 67.1 %. Anger and pride are recognised with very high accuracy (80 and 96.67 % respectively). Sadness was partly misclassified with pride (36.67 %)

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. The overall performance of this classifier was 57.1. 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 Multimodal emotion recognition

The overall performance of the classifier based on feature-level fusion was 78.3 %, which is much higher than the one obtained by the most successful unimodal system, the one based on gestures. The approach based on decision-level fusion instead obtained lower recognition rates: the performance of the classifier was 74.6 %.

7 Discussion and conclusions

We presented a multimodal framework for analysis and recognition of emotions starting from expressive faces, gestures and speech.

We experimented our approach on a dataset of 240 samples for each modality. 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. This study considered a restricted set of data, collected from a relatively small group of subjects.

Nevertheless, it represents a first attempt to fuse together three different synchronised modalities, which is still uncommon in current research. Future work will consider new multimodal recordings with a larger and more representative set of subjects; further, we will investigate the mutual relationship between audio-visual information.

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. Scherer, K.R. and Ellgring, H.: Multimodal Expression of Emotion: Affect Programs or Componential Appraisal Patterns? Emotion 7(1).
5. 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).
6. Scherer, K.R.: Adding the affective dimension: A new look in speech analysis and synthesis, In Proc. International Conf. on Spoken Language Processing, pp. 1808–1811, (1996).
7. 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.
8. 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).
9. 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.
10. 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.
11. 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).
12. Densley, D.J., Willis, P.J. Emotional posturing: a method towards achieving emotional figure animation, ca, p. 8, Computer Animation 1997, (1997)
13. 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.
14. 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.
15. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco (2005).