Bimodal Emotion Recognition

Marco Paleari[1,2], Ryad Chellali[1] and Benoit Huet[2]

[1] Italian Institute of Technology, TERA Department, Genova, Italy,
name.surname@iit.it
[2] EURECOM, Multimedia Department, Sophia Antipolis, France,
name.surname@eurecom.fr

Abstract. When interacting with robots we show a plethora of affective reactions typical of natural communications. Indeed, emotions are embedded on our communications and represent a predominant communication channel to convey relevant, high impact, information. In recent years more and more researchers have tried to exploit this channel for human robot (HRI) and human computer interactions (HCI). Two key abilities are needed for this purpose: the ability to display emotions and the ability to automatically recognize them. In this work we present our system for the computer based automatic recognition of emotions and the new results we obtained on a small dataset of quasi unconstrained emotional videos extracted from TV series and movies. The results are encouraging showing a recognition rate of about 74%.

Keywords: Emotion recognition; facial expressions; vocal expressions; prosody; affective computing; HRI;

1 Introduction

The abilities to recognize, process, and display emotions are well known to be central to human intelligence, in particular influencing abilities such as communications, decision making, memory, and perception [3]. In recent years more and more researchers in the human computer (HCI) and human robot interactions (HRI) societies have been investigating ways to replicate such a kind of functions with computer software [6, 13, 18]. In our domain, emotions could be used in many ways but two in particular are more relevant: 1) emotional communications for HRI [14], and 2) decision making for autonomous robots [5].

One of the key abilities of these systems is the ability to recognize emotions. The state of the art is rich with systems performing this task analyzing people's facial expressions and/or vocal prosody (see [18] for a thorough review). One of the main limitations of most of existing technologies is that they only have been tested on very constrained environments with acted emotions.

In this work we want to present our last results toward the development of a multimodal, person independent, emotion recognition software of this kind. We have tested our system on less constrained data in the form of movies and TV series video excerpts. The results we present are very promising and show that even in these almost unconstrained conditions, our system could perform well allowing to correctly identify as much as 74% of the presented emotions.

2 Multimodal Approach

In our approach we are targeting the identification of seven different emotions[3] by fusing information coming from both the visual and the auditory modalities.

The idea of using more than one modality arises from two main observations: 1) when one, or the other, modality is not available (e.g. the subject is silent or hidden from the camera) the system will still be able to return an emotional estimation thanks to the other one and 2) when both modalities are available, the diversity and complementarity of the information, should couple with an improvement on the general performances of the system.

Facial Expression Features We have developed a system performing real time, user independent, emotional facial expression recognition from video sequences and still pictures [10, 12]. In order to satisfy the computational time constraints required for real–time we developed a feature point tracking technology based on very efficient algorithms.

In a first phase, the face of the subjects in the video is automatically detected thanks to a slightly modified Viola-Jones face detector [17]. When the face is detected twelve regions are identified thanks to an anthropometric two dimensional mask similarly to what it is done by Sohail and Bhattacharya in [15]. Then, for each region of the face, we apply the Lucas–Kanade [16] algorithm to track a cloud of keypoints. Finally, the positions of these points are averaged to find one single center of mass per each region (see figure 1(a)). We call the set of the x and y coordinates of these 12 points *coordinates* feature set.

(a) Feature Points (b) Distances

Fig. 1. Video Features

As a second step we have extracted a more compacted feature set in a similar way to the one adopted by MPEG-4 Face Definition Parameters (FDPs) and Face Animation Parameters (FAPs). This process resulted in 11 features defined as distances and alignments $distance(j)$ from the keypoints in the *coordinates*

[3] the six "universal" emotions listed by Ekman and Friesen [4] (i.e. anger, disgust, fear, happiness, sadness, and fear) and the neutral state

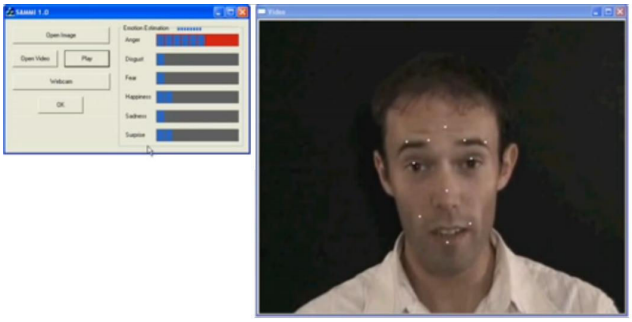


Fig. 2. Emotion Recognition System Interface

feature set (see figure 1(b)). Additionally we explicitly keep track, in this feature set, of the x and y displacement of the face and of a zooming factor (which is proportional to the z displacement). We refer to this set of distances, alignments, and displacements as to the *distances* feature set.

Prosodic Expression Features Our system for speech emotion recognition, takes deep inspiration from the work of Noble [8]. From the audio signal we extract: the fundamental frequency (pitch), the energy, the first three formants, the harmonicity (a.k.a. harmonics to noise ratio), the first 10 linear predictive coding coefficients (LPC), and the first ten mel–frequency cepstral coefficients (MFCC).

These 26 features are collected with the use of PRAAT[4] [1] and downsampled to 25 samples per second to help synchronization with the video features.

3 Emotion Recognition System

In the former section, we overviewed the modality of extraction of audio and video features. In this section, we detail the other procedures defining our emotion recognition system (see figure 2).

To evaluate this system we employ three measures: the recognition rate of the positive samples $CR_i^+ = \frac{well\_tagged\_samples\_of\_emo_i}{samples\_of\_emo_i}$, the average recognition rate $m(CR^+) = \frac{\sum well\_tagged\_samples\_of\_emo_i}{samples}$, and the weighted standard deviation $wstd(CR^+) = \frac{std(CR^+)}{m(CR^+)}$ [5]. The objective of our recognition system would be to maximize the $m(CR^+)$ while also minimizing the weighted standard deviation $wstd(CR^+)$.

For this experiment we have trained three different neural networks per each one of the six universal emotions using data from the audio, the coordinates, and the distances feature sets respectively.

[4] PRAAT is a C++ toolkit written by P. Boersma and D. Weenink to record, process, and save audio signals and parameters. See [1]

[5] *wstd* will be low if all emotions are recognized with the same likelihood and vice versa if some emotions are much better recognized than others, it will be high

It is important to notice that not all of the *audio*, *coordinates*, and *distances* features are used for all emotions. In [12] we presented a work in which we compare singularly each one of the $64 = 24 + 14 + 26$ features we have presented in sections 2 for the recognition of each one of the six "universal" emotions. As a result of this study we were able to select the best features and processing for recognizing each one of the selected emotions.

In table 1 we list the features which have been selected for this study.

Emotion	Audio features	Coordinate features	Distances features
Anger	Energy, Pitch, & HNR	Eye Region	Head Displacements
Disgust	LPC Coefficients	Eye Region	Eye Region
Fear	MFCC Coefficients	Eye Region	Head Displacements
Happiness	Energy, Pitch, & HNR	Mouth Region	Mouth Region & x Displacement
Sadness	LPC Coefficients	Mouth Region	Mouth Region
Surprise	Formants	Mouth Region	Mouth Region

Table 1. Selected features for the different emotions

In a first phase we have evaluated this setup on a publicly available multimodal database We have employed neural–networks with one hidden layer composed of 50 neurons which have been trained on a training set composed of 40 randomly selected subjects from the eNTERFACE'05 database [7]. The extracted data was fed to the networks for a maximum of 50 times (epochs). The remaining 4 subjects were used for test (the database contains videos of 44 subjects acting the 6 universal emotions). We have repeated these operations 3 times (as in an incomplete 11–fold cross validation) using different subjects for test and training and averaged the results.

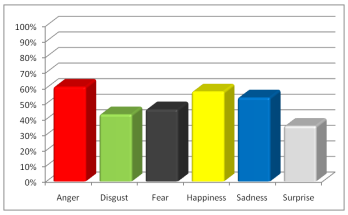
Then, the outputs of the 18 resulting neural–networks have been filtered with a 25 frames low–pass filter to reduce the speed in which the output can change; indeed, emotions do not change at a speed of 25 frames per second. This filtering shall also improve the results as discussed in [11].

For each emotion, we have employed a Bayesian approach to extract a single multimodal emotion estimate per frame o_{emo}. The Bayesian approach has been preferred to other simple decision level fusion approaches and more complex ones such as the NNET approach [9] as one returning very good results without requiring any training. The resulting system could recognize an average of 45.3% of the samples, $wstd(CR^+) = 0.73$.

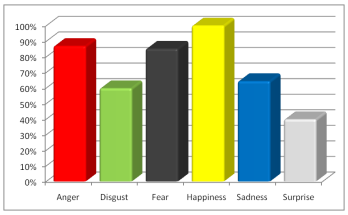
The reasons why the *wstd* is so high is because of the statistics of the outputs for the six Bayesian emotional detectors are very different. Therefore, we computed the minimum, maximum, average, and standard deviation values for each one of the detector outputs and proceeded to normalize the outputs to have a minimum estimate equal to 0 and a similar average value.

Performing this operation raise the $m(CR^+)$ to 50.3% while decreasing the $wstd(CR^+)$ to 0.19. In figure 3(a) we can see the CR^+ for the six different emotions after this phase of normalization.

To further boost the results we apply a double thresholding strategy to these results. Firstly, we define a *threshold* below which results are not accepted because they are evaluated as being not reliable enough.



(a) After outputs normalization



(b) With profiling (see table 2)

Fig. 3. CR^+ results

Secondly, we apply a function which we called inverse thresholding. In this case, we select more than one estimates for the same audio–video frame in the case in which two (or more) detector outputs are both above a certain $threshold^{-1}$. This operation is somehow similar to using a K–best approach but in this case more estimates are selected only when they are "needed".

Thresholds are defined as a function of the output mean and standard deviation values making the assumption that the distributions of the outputs of the detectors are Gaussians. We call the phase of choosing an appropriate couple of thresholds *profiling*. By choosing different profiles the system act differently and its behavior can be dynamically adapted to its specific needs.

It is interesting to note that infinite profiles can be defined which returns about the same number of estimations. Indeed, increasing the threshold or decreasing the inverse threshold have opposite influences on the number of estimations.

In table 2, we compare two possible profiling setting together with the originated results.

#	Recall	Thresholding Value	Inverse Thresholding Value	$m(CR^+)$	$wstd(CR^+)$
0	100%	0	1	50.3%	0.19
1	49.7%	$m(o_{emo}) + 1.2 * std(o_{emo})$	$m(o_{emo}) + 2.0 * std(o_{emo})$	61.1%	0.29
2	12.9%	$m(o_{emo}) + 3.0 * std(o_{emo})$	$m(o_{emo}) + 5.0 * std(o_{emo})$	74.9%	0.29

Table 2. Selected features for the different emotions

As expected, the two systems maintain low weighted standard deviation values while improving the mean recognition rate of the positive samples.

4 Relaxing Constraints

In the former sections we have introduced the topic of emotion recognition for human machine interactions (HMI) and overviewed our multimodal, person independent system. In this section we aim at relaxing the constraints to see how the system behaves in more realistic conditions.

To perform this task we have collected 107 short (4.2 ± 2.6 seconds) DivX quality excerpts from three TV series, namely "The Fringe", "How I met your

mother", and "The OC" and the Joe Wright's 2007 movie "Atonement" (see figure 4). The video sequences were selected to represent character(s) continuously in a shot longer than 1 second. It was required for at least one character to roughly face the camera along the whole video.

Fig. 4. Screenshots from the excerpts database

The result is a set of videos with very heterogeneous characteristics; for the visual modality we observe:

- more than one subject on the same video
- different ethnic groups
- different illumination conditions: non uniform lightening, dark images ...
- different gaze directions
- presence of camera effects: zoom, pan, fade ...

Also the auditory modality presents lesser constraints and in particular we have samples with:

- different languages (i.e. Italian and English)
- presence of ambient noise
- presence of ambient music
- presence of off–camera speech

4.1 Evaluation

Each one of these video is being evaluated thanks to an online survey on YouTube[6] We asked several subjects to tag the excerpts in the database with one (or more) of our 6 emotional categories; the neutral tag was added to this short list allowing

[6] `http://www.youtube.com/view_play_list?p=4924EA44ABD59031`

people to tag non emotional relevant excerpts. We currently have collected about 530 tags (4.99 ± 1.52 tags per video); each video segment has been evaluated by a minimum of 3 different subjects.

Few subjects decided to tag the videos only using audio or video but most exploited both modalities trying to understand what the emotional meaning of the characters in the video was. In average, every video was tagged with 2.2 different emotional tags but usually a tag is identifiable which was hit by over 70% of the subjects of our online survey. In 10 cases agreement on a single tag representing an excerpt could not pass the 50% threshold; in 8 of these cases neutral is among the emotions that are most indexed by our online survey, justifying the confusion. The remaining segments are tagged as representing two different emotions: a first one is represented by anger and surprise, the second by sadness and disgust. It is interesting to notice that, the emotions belonging to both couples have adjacent positioning on the Valence Arousal plane thus justifying, in part, the confusion among the two.

Figure 5 reports the distribution of the tags. As it can be observed the emotion neutral is predominant to the others representing about 40% of the tags that the subjects of our survey employed.

Fig. 5. Distribution of human tags

in \ out	ANG	DIS	FEA	HAP	SAD	SUR
Anger	13%	17%	10%	20%	17%	13%
Disgust	28%	0%	22%	6%	28%	17%
Fear	10%	13%	3%	26%	22%	26%
Happiness	17%	3%	23%	6%	20%	31%
Sadness	20%	12%	17%	17%	12%	22%
Surprise	11%	9%	23%	31%	26%	0%

Fig. 6. Correlation matrix of human tags

Sadness is the most common emotion in our database (with 16% tags), disgust is the emotion which is less identified by our online survey: only 3% of the tags human gave belong to this emotion.

Table 6 report the correlation matrix of the human tag. Each cell in the tab contains the percentage of videos of the emotion identified by the row which are also tagged as belonging to the emotion in column. As it appears in table 6 the emotions presented in the videos may be easily confused with each other. We identified 6 main reasons which can justify this result:

1. in films and TV series emotions tend to be complex mixes of emotions;
2. the excerpts are, for their very nature, extrapolated from the context; without it people are not always able to correctly recognize the expression;
3. the emotion presented could not always fit well into one of our categories;
4. in most cases the presented emotions are not characterized by high intensity, thus being confused with neutral states and similar emotions;
5. in some cases social norms makes character hide their emotional state possibly distorting or hiding the emotional message;

6. in some cases the intention of the director is to convey an emotion different to the one of the character being depicted: this emotion may be transferred by other means such as music, colors, etc. and influence the human perception.

4.2 Results

As it was pointed out in the former section, our online survey led most video excerpts to present two or more emotional tags.

Given the different characteristics of the train and test database (specifically the fact of presenting or not multiple emotional tags per video) a new metric needed to be defined. We decided that if an emotion is tagged by someone than it is reasonable to say that when a computer returned the same tag it did not make an error. With this idea in mind, without modifying the system described in section 3, and by applying the second profile from table 2, we analyzed audio and video of the multimedia excerpts of the newly designed emotional database.

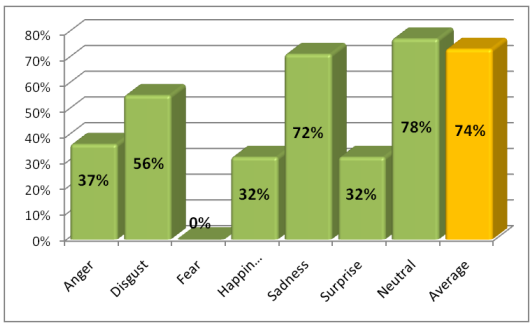


Fig. 7. Recognition rate on real videos

Figure 4.2 reports the result obtained by this system. The resulting average recognition rate on six emotions is of about 44% but it is boosted to 74% ($wstd = 0.36$) if neutral is considered as a seventh emotion. Please note that the number of frames tagged by our online survey as being neutral is about 6 times higher than the number of frames belonging to all the other emotions. Please also note that also considering the emotion neutral in the metric brings the recall rate back to 1: all frames are evaluated as belonging to one emotion or neutral.

Given the relatively small size of the employed database it may be normal for some emotions to be worse recognized than average (please note fear has only 5 samples). Nevertheless, it is important to comment the disappointing result obtained for the emotion "fear" and the very good one returned for "sadness".

Our analysis of the data suggested that the result obtained for "fear" may be explained with the differences underlying the emotional excerpts of this real–video database and and our original train base. Analyzing the videos we noticed that the videos of the eNTERFACE database depicted some kind of surprisedly

scared emotion while in our new database the emotion depicted is often similar to some kind of cognitive and cold fear. In other words, it is our conclusion that while both the emotion represented in the eNTERFACE database and the one represented in our test database are definable as fear, those two kind of emotions are different, e.g. they arise from different appraisals, and therefore have different expressions.

A similar behavior might as well have deteriorated the performances of the emotion anger; we know, indeed, that there are at least two kind of anger, namely "hot" and "cold".

Nevertheless, it is important to notice that, as a whole, the average recognition result clearly shows that without any modification or adaptation the system described here can work for emotion recognition of multimedia excerpts and it is likely to work on real scenarios too.

5 Concluding Remarks

In this paper, we have discussed the topic of multimodal emotion recognition and, in particular, a system performing bimodal audio–visual emotion recognition has been presented. Many different scenarios for human–robot interaction and human-centered computing will profit from such ability.

Our emotion recognition system has been presented and we have discussed the idea of thresholding, inverse thresholding, and profiling. The system is able to recognize about 75% of the emotions presented by the eNTERFACE'05 database at an average rate of more than 3 estimates per second.

Finally, we have shown the results obtained by this system on quasi unconstrained video conditions. For this study, an experimental database of 107 real video sequences from three TV series and a movie were extracted. The results on this small dataset confirm that our system works for the detection of emotions in real video sequences. In particular, we have showed that with the current setup the system could correctly tag as much as 74% of the frames (when considering neutral as a seventh emotion).

Because of the size of the database and number of tags, the metric we applied can be considered good, but different metrics shall be considered in the case in which many more tags were to be available; in particular we selected two: the first one only considers the most common human tag as the corrected one, the second weights the correctness of the computer outputs by the percentage of given human tags. With these two metrics the system performs 55% and 39% respectively.

Ongoing work consists in increasing the size of this database to extract more results. Future work will focus on the idea, developed in [2], of separating the frames of the video shots into two classes of silence/non silence frames to apply different processing; furthermore, we are trying to extend this idea by introducing a third and a fourth classes representing music frames and frames in which the voice does not belong to the depicted characters.

References

1. P. Boersma and D. Weenink. Praat: doing phonetics by computer, January 2008. [http://www.praat.org/].
2. D. Datcu and L. Rothkrantz. Semantic audio-visual data fusion for automatic emotion recognition. In *Euromedia' 2008*, Porto, 2008.
3. R. Davidson, K. Scherer, and H. Goldsmith. *The Handbook of Affective Science*. Oxford University Press, March 2002.
4. P. Ekman and W. V. Friesen. A new pan cultural facial expression of emotion. *Motivation and Emotion*, 10(2):159–168, 1986.
5. C.-H. J. Lee, K. Kim, C. Breazeal, and R. Picard. Shybot: friend-stranger interaction for children living with autism. In *CHI '08: CHI '08 extended abstracts on Human factors in computing systems*, pages 3375–3380, Florence, Italy, 2008. ACM.
6. S. Marsella and J. Gratch. Ema: A process model of appraisal dynamics. *Cognitive Systems Research*, 10(1):70–90, March 2009.
7. O. Martin, I. Kotsia, B. Macq, and I. Pitas. The eNTERFACE05 Audio-Visual Emotion Database. In *Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06)*. IEEE, 2006.
8. J. Noble. Spoken emotion recognition with support vector machines. *PhD Thesis*, 2003.
9. M. Paleari, R. Benmokhtar, and B. Huet. Evidence theory based multimodal emotion recognition. In *MMM '09* 15^{th} *Intl Conference on MultiMedia Modeling*, Sophia Antipolis, France, January 2009.
10. M. Paleari, R. Chellali, and B. Huet. Features for multimodal emotion recognition: An extensive study. In *Proceedings of IEEE CIS'10 Intl. Conf. on Cybernetics and Intelligence Systems*, Singapore, June 2010.
11. M. Paleari and B. Huet. Toward Emotion Indexing of Multimedia Excerpts. In *CBMI '08 Sixth International Workshop on Content-Based Multimedia Indexing*, London, June 2008. IEEE.
12. M. Paleari, B. Huet, and R. Chellali. Towards multimodal emotion recognition: A new approach. In *Proceedings of ACM CIVR'10 Intl. Conf. Image and Video Retrieval*, Xi'An, China, July 2010.
13. I. Poggi, C. Pelachaud, F. de Rosis, V. Carofiglio, and B. de Carolis. *GRETA. A Believable Embodied Conversational Agent*, pages 27–45. Kluwer, 2005.
14. C. Sapient Nitro. Share happy, project webpage. http://www.sapient.com/en-us/SapientNitro/Work.html#/?project=157, June 2010.
15. A. Sohail and P. Bhattacharya. *Signal Processing for Image Enhancement and Multimedia Processing*, volume 31, chapter Detection of Facial Feature Points Using Anthropometric Face Model, pages 189–200. Springer US, 2007.
16. C. Tomasi and T. Kanade. Detection and tracking of point features, April 1991. CMU-CS-91-132.
17. P. Viola and M. Jones. Robust real-time object detection. *International Journal of Computer Vision*, 2001.
18. Z.Zeng, M. Pantic, G. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 31(1):39–58, January 2009.