The eNTERFACE'05 Audio-Visual Emotion Database

O. Martin [(1)], I. Kotsia [(2)], B. Macq[(1)], I. Pitas[(2)]

(1): Université catholique de Louvain,
Laboratoire de Télécommunications et de Télédétection,
2, Place du Levant, 1348 Louvain-la-Neuve, Belgium
Tel: +32 10 47 80 75
Fax: +32 10 47 20 89
Email: {martin, macq}@tele.ucl.ac.be

(2): Aristotle University of Thessaloniki
Department of Informatics, Box 451,
54124 Thessaloniki, Greece
Tel: +30 231 099 63 04
Fax: +30 231 099 63 04
E-mail: {ekotsia, pitas}@auth.csd.edu.gr

Abstract

This paper presents an audio-visual emotion database that can be used as a reference database for testing and evaluating video, audio or joint audio-visual emotion recognition algorithms. Additional uses may include the evaluation of algorithms performing other multimodal signal processing tasks, such as multimodal person identification or audio-visual speech recognition. This paper presents the difficulties involved in the construction of such a multimodal emotion database and the different protocols that have been used to cope with these difficulties. It describes the experimental setup used for the experiments and includes a section related to the segmentation and selection of the video samples, in such a way that the database contains only video sequences carrying the desired affective information. This database is made publicly available for scientific research purposes.

1. Introduction

During the past decade, research on automatic emotion recognition has attracted the interest of an ever-growing community of researchers. Numerous systems achieving emotion recognition from visual or prosodic features have been developed. However, it remains very difficult to compare the relative performances of the existing prototypes due to the lack of common databases and protocols.

In the past few years, the Cohn-Kanade facial database [1] imposed itself as the main benchmark database for facial expression recognition algorithms. It includes over 2000 image sequences from over 200 different subjects, expressing up to six different emotions. A relatively large number of other facial expression databases are also widely employed. To name just a few, the Japanese Female Facial Expression (JAFFE) database [2] contains 213 images of 7 facial expressions, posed by 10 Japanese female models. The AR Face Database [3] is a collection of over 4000 high-resolution color images of faces with different facial expressions, illumination conditions, and occlusions.

For emotion recognition systems based on the analysis of user's prosody, an even-wider variety of databases are used. It becomes very difficult to assess the performances of a specific method, as those performances are strongly related to the database used for the experiments. Among the variety of databases used for the analysis of emotions in the speech signal, the database described by Amir [4] distinguishes from the other available databases as it includes only naturally occurring expressions of emotions. The database contains 30 subjects recalling an emotional event in which they participated. Another interesting database in the field is the one collected by the Reading-

Leeds project, which consists of radio broadcasts of material such as interviews; in which emotions arise spontaneously from the content of the interaction [5].

For multimodal emotion recognition, the need for a common database is even greater: speech and image signal processing communities are often working independently from each other and relatively few joint audio-visual studies of emotions have been conducted so far. The lack for an existing multimodal affective database is a crucial problem. A detailed analysis of the state of the art, coupled with an attempt to fill the need for a multimodal emotion database has recently been made by Douglas-Cowie et. al [6], whose approach focuses on the generation of genuine emotions. Although their result is interesting, a database containing the 6 archetypal emotions defined by Ekman et al. [7] is still needed, as most of existing system aims at recognizing this set of archetypal emotions.

To fill this need for a multimodal affective database containing the six archetypal emotions, 46 subjects were invited to react to six different situations, each of them eliciting one of the following emotions: happiness, sadness, surprise, anger, disgust and fear.

Two human experts decided whether or not the subject had expressed itself in such a way that an untrained human observer could without ambiguity recognize the emotion present in the reaction, for each of the emotions to be elicited. In a post-processing step, samples in which the emotion was not clearly recognized were discarded, so that the database would only contain video samples carrying relevant affective information. In this post-processing step, decision was made to remove 4 subjects whose none of the video samples carried a believable affective message.

The paper is structured as follows. After the present introduction, a second section depicts an overview of the database. A third section presents the different protocols that have been followed in the creation of the database. A fourth section focuses on the technical aspects related to the recordings, while a fifth and last section details the post-processing step that led to a relevant multimodal emotion database.

2. Database Overview

The final version of the database thus contains 42 subjects, coming from 14 different nationalities. Table 1 shows the geographic distribution of the subjects who participated in the database.

Country	Number of subjects	Country	Number of subjects
Belgium	9	Cuba	1
Turkey	7	Slovakia	1
France	7	Brazil	1
Spain	6	U.S.A.	1
Greece	4	Croatia	1
Italy	1	Canada	1
Austria	1	Russia	1

Table 1. Geographic distribution of the subjects who participated in the database

Among the 42 subjects, a percentage of 81% were men, while the remaining 19% were women. A percentage of 31% of the total set wore glasses, while 17% of the subjects had a beard.

The recordings lasted for two weeks. All the experiments were driven in English. Each subject was told to listen to six successive short stories, each of them eliciting a particular emotion. They had then to react to each of the situations and two human experts judged whether the reaction expressed the emotion in an unambiguous way. If this was the case, the sample was added to the database. If not, it was discarded. For a more complete description of the recording protocol, the reader is invited to refer to the following section.

3. Protocol Description

As extensively described in [8], an emotion can only be expressed genuinely if the stimulus is natural, i.e. if the reaction is spontaneous. When asked to simulate an emotion, a subject may try to reproduce the facial expression and prosody that relates to the expression of that emotion, but the result is always incomplete: some of the muscles involved in the expression of the emotion will not be activated when the emotion is simulated, while other muscles would be activated to produce what is believed to be the expression of the emotion to be elicited, as perceived by the subject [8].

The aspects evoked in the first paragraph indicate that the database should ideally contain only genuine expressions of emotions. However, as the database should also consist of high-quality video samples (with constant illumination, background, head pose, etc…) to be useful for practical applications, the choice that was made was to get as close as possible to spontaneous emotions, while keeping at the same time a fully controlled recording environment.

To achieve this goal, the first idea was to immerge the subject into a situation that evokes a specific emotion, and

then to let the subject react to the situation in his own language. The only indication that would be given to the subject would be to be as 'emotive' as possible, as the presence of a camera often inhibits people. Unfortunately, allowing the subject to react in its own language has a main drawback: the prosodic features (such as pitch variations, speaking rate, etc…) largely depend on the language itself. To illustrate by an example, the speaking rate is typically higher for an Italian than for a French-speaking Swiss subject. As one of the goals of the database is to have prosodic features that depend only on the emotion that is expressed, the choice to conduct all experiments in English had to be made.

The idea behind the second version of the protocol was thus to let the subjects freely react to the proposed situations, but only in English. Unfortunately, the results were not very convincing either. As most subjects were not English-native speakers, the reactions were not spontaneous at all. The fact that the subjects had to think about what their reactions should be and then translate those reactions into English led to totally unnatural utterances. Therefore, it was chosen to use pre-defined answers for each situation. The remaining of this section presents this last version of the protocol thoroughly, along with the situations and reactions corresponding to each emotion to be elicited.

For each emotion, the protocol is the following. First, the subject is asked to listen carefully to a short story and to 'immerge' himself into the situation. Once he is ready, the subject may read, memorize and pronounce (one at the time) the five proposed utterances, which constitute five different reactions to the given situation. The subjects are asked to put as much expressiveness as possible, producing a message that contains only the emotion to be elicited. If the result is satisfying, the procedure may continue with the next emotion. In the opposite case, the subject is asked to repeat the attempt. When the subject obviously didn't know how an emotion should be expressed, the experimenters decided to suggest a way of expressing it, based on their knowledge[1] of the way the emotion is generally expressed. In other cases, the experimenters chose not to repeat the experience when they believed no satisfying results could be achieved with the subject under consideration.

In the remaining of this section, the stories corresponding to each of the emotions will be presented, along with the five different reactions. Pictures depicting each of the emotions illustrate the discussion. As can be seen on the figures, the facial expressions are quite different from the ones that can be found in most of the existing facial expression databases. The reason is that the subject is speaking while expressing the emotion, thus producing a mouth shape that corresponds to a mix between the influence of both the pronounced phoneme and the internal emotional state.

Figure 1 shows the subject in its neutral state. Tables 2 to 7 correspond to the stories and reactions to the six emotions considered, while figures 2 to 7 show instances of the facial expressions related to the considered emotions.

Figure 1. A subject depicting the neutral state

Situation to elicit "anger"
"You are in a foreign city. A city that contains only one bank, which is open today until 4pm. You need to get 200$ from the bank, in order to buy a flight ticket to go home. You absolutely need your money today. There is no ATM cash machine and you don't know anyone else in the city. You arrive at the bank at 3pm and see a big queue. After 45 minutes of queuing, when you finally arrive at the counter, the employee tells you to come back the day after because he wants to have a coffee before leaving the bank. You tell him that you need the money today and that the bank should be open for 15 more minutes, but he is just repeating that he does not care about anything else than his coffee…"
Reactions
R1: What??? No, no, no, listen! I need this money!
R2: I don't care about your coffee! Please serve me!
R3: I can have you fired you know!
R4: Is your coffee more important than my money?
R5: You're getting paid to work, not drink coffee!

Table 2. Situation and reactions to elicit "anger"

[1] Such knowledge is extracted from psychological studies such as found in [8].

Figure 2. A subject depicting the expression of "anger"

Situation to elicit "disgust"
"You are in a restaurant. You are already a bit sick and the restaurant looks quite dirty, but it is the only restaurant in the village, so you don't really have the choice… When you finally receive your plate, which is a sort of noodle soup, you take your spoon, ready to eat. Although you are very hungry, the soup does not taste very good. It seems that it is not very fresh…Suddenly you see a huge cockroach swimming in your plate! You're first surprised and you jump back out of your chair. Then, you look again at your plate, really disgusted. "
Reactions
R1: That's horrible! I'll never eat noodles again.
R2: Something is moving inside my plate
R3: Aaaaah a cockroach!!!
R4: Eeeek, this is disgusting!!!
R5: That's gross!

Table 3. Situation and reactions to elicit "disgust"

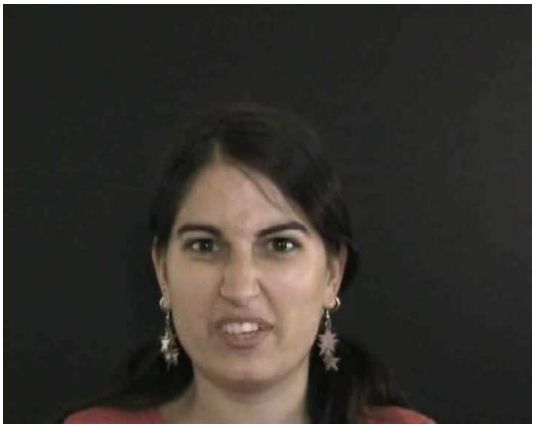

Figure 3. A subject depicting the expression of "disgust"

Situation to elicit "fear"
"You are alone in your bedroom at night, in your bed. You cannot sleep because you are nervous. Your bedroom is located on the second floor of your house. You are the only person living there. Suddenly, you start hearing some noise downstairs. You go on listening and realize that there is definitely someone in the house, probably a thief…or maybe even a murderer! He's now climbing up the stairs, you are really scared."
Reactions
R1: Oh my god, there is someone in the house!
R2: Someone is climbing up the stairs
R3: Please don't kill me...
R4: I'm not alone! Go away!
R5: I have nothing to give you! Please don't hurt me!

Table 4. Situation and reactions to elicit "fear"

Figure 4. A subject depicting the expression of "fear"

Situation to elicit "happiness"
"You learned this morning that you won the big prize of 5.000.000€ at the lottery! You're in a very happy mood of course, because you realize that some of your dreams will now become true! After the surprise to learn that you have won, comes the happy state of mind when you start dreaming about your new projects. You are in a restaurant, inviting your friends for a good meal, and telling them how happy you feel."
Reactions
R1: That's great, I'm rich now!!!
R2: I won: this is great! I'm so happy!!
R3: Wahoo... This is so great.
R4: I'm so lucky!
R5: I'm so excited!

Table 5. Situation and reactions to elicit "happiness"

Figure 5. A subject depicting the expression of "happiness"

Situation to elicit "sadness"
"You just came back from an exhausting day at work. You are in a neutral state of mind when suddenly the telephone rings. You take the phone call and realize that it is your boy (girl) friend. He (she) announces you that he (she) doesn't want to go on the relationship with you. You first don't believe it, but after a while you start realizing what just happened. When you think about all the good moments you spent with your boy (girl) friend, and associate these memories with the fact that the relationship just finished, you start feeling really sad"
Reactions
R1: Life won't be the same now
R2: Oh no, tell me this is not true, please!
R3: Everything was so perfect! I just don't understand!
R4: I still loved him (her)
R5: He (she) was my life

Table 6. Situation and reactions to elicit "sadness"

Figure 6. A subject depicting the expression of "sadness"

Situation to elicit "surprise"
"Your best friend invites you for a drink after your day at work. You join him on the Grand Place of Mons[2], for a beer. Then, he suddenly tells you that he's actually gay! You are very surprised about it, you really didn't expect that!"
Reactions
R1: You have never told me that!
R2: I didn't expect that!
R3: Wahoo, I would never have believed this!
R4: I never saw that coming!
R5: Oh my God, that's so weird!

Table 7. Situation and reactions to elicit "surprise"

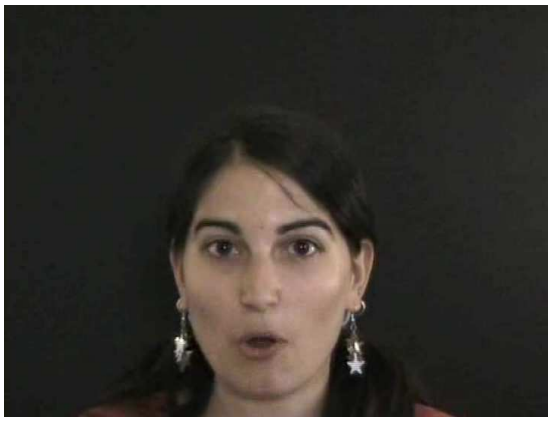

Figure 6. A subject depicting the expression of "surprise"

4. Technical Aspects

The database was recorded using a standard mini-DV digital video camera. The resolution of the camera was 800.000 pixels. The recording of the speech signal was realized through the use of a high-quality microphone, specially conceived for speech recordings. The microphone was situated roughly 30cm below the subject's mouth, outside of the camera field.

The background consists of a monochromatic dark gray panel that covered the entire area behind the subject, to allow easier face detection and tracking.

Illumination was made constant through the use of a set of occultation panels, placed in front of every window. Lighting material consisted of a strong spotlight (500 watts), situated right behind the camera, facing the user. The spotlight was covered with a semi-transparent plastic film to soften the light, decrease the shadows and protect

[2] Mons is the place where the recordings took place.

the subject from the very intense source of light. Two additional directional spots were situated between the subject and the background panel, so as to cancel shadows produced on the background panel by the main spotlight. The two additional spots were also covered with a semi-transparent plastic film. Figure 8 shows the recording setup.

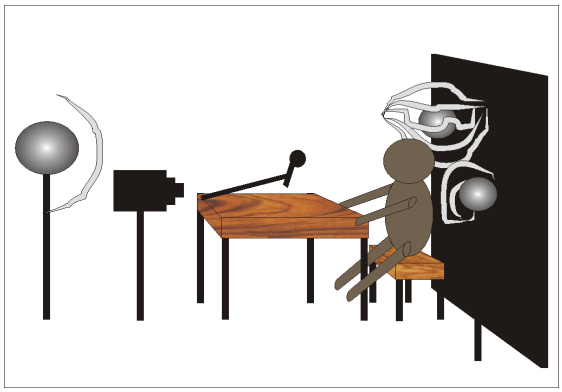

Figure 8. The recording setup.

The recording room was small (around ten square meters) and furnished with electronic equipment. The doors remained closed at all time to prevent external sound to interfere with the experiments.

Two PhD students, experts in emotion recognition, conducted the experiments. One of the experimenter was placed on the left of the camera, while the other sat on a chair at the right of the camera's tripod. Both experimenters stayed outside of the main spotlight field at all time.

5. Post-Processing Operation

Expressing emotions 'on demand' is a very difficult task. In addition, none of the subjects was an actor: they were participants of a European workshop on multimodal interfaces [9], mostly as research engineers. When asked to express themselves emotively, some subjects performed pretty well while other totally failed to express the requested emotions. Other subjects manage to perform in a satisfying way, but only after several trials. To take this variety of outcomes into account and still end up with a high-quality database, it is necessary to apply a post-processing step to the set of original recordings: human experts have to examine carefully each recorded sample. Their role is to decide whether or not each considered sample is expressing the requested emotion in an unambiguous way. If this is the case, the sample is added to the database. If not, it is discarded.

As described in the third section of the present paper, each subject was asked to express 6 different emotions. For each emotion, five reactions were simulated. The database then consists of a list of folders, each of whom being dedicated to one subject and being named by the keyword 'subject', followed by the subject index (subject 1, subject 2 … subject 42). A folder of the type 'subject *i*' contains seven subfolders: one for the neutral sample and the six remaining for the six emotions considered. These subfolders were named in accordance with the emotion they represented ('anger', 'disgust', 'fear', 'happiness', 'sadness', 'surprise' and 'neutral'). Each of these subfolders contains themselves five subfolders, corresponding to the five reactions and named with the keyword 'sentence' followed by the sentence index (sentence 1, sentence 2...sentence 5).

The video sequences were processed using a 720x576 Microsoft AVI format. The frame rate was equal to 25 frames per second, while pixel aspect ration was D1/DV PAL (1.067). The video was compressed using a DivX 5.0.5 Codec, to ensure easy portability. The audio sample rate was 48000 Hz, in an uncompressed stereo 16-bit format.

Eventually, the database consists of a total of 1166 video sequences. Out of these 1166 video sequences, 264 concern women recordings (23%) and 902 men recordings (77%).

From the 42 people participating in the database, 25 of them (60%) performed satisfactory in all 6 emotions, generating 5 believable reactions to each of the proposed situations. The remaining 17 subjects (40%) did not manage to reveal the right emotion in all of their reactions, thus produced unqualified video samples that had to be discarded. Table 8 depicts the number of reactions that were kept for each subject, and each of the considered emotions.

S.	Ang.	Dis.	Fear	Hap.	Sad.	Sur.
1	5	5	5	5	5	5
2	5	5	1	5	5	5
3	5	4	2	5	5	0
4	5	5	5	5	5	5
5	5	5	5	5	5	5
6	5	3	5	5	5	5
7	5	5	5	5	5	5
8	5	5	5	5	5	5

9	5	5	5	5	5	5
10	5	5	4	5	5	5
11	5	5	5	5	5	5
12	0	5	5	5	0	1
13	5	5	5	5	5	1
14	5	5	5	5	5	5
15	5	5	5	5	5	5
16	5	5	5	5	5	5
17	5	5	5	5	5	5
18	0	3	0	0	0	0
19	5	5	1	5	0	5
20	5	1	4	5	5	5
21	5	5	5	5	5	5
22	5	5	5	5	5	5
23	5	5	4	5	5	5
24	5	5	4	5	5	5
25	5	3	5	5	5	5
26	5	3	5	5	5	5
27	5	4	5	5	5	5
28	5	2	4	5	5	5
29	5	5	5	5	5	5
30	5	5	5	5	5	5
31	5	5	5	5	5	5
32	5	5	5	5	5	5
33	5	4	5	5	5	5
34	5	5	5	5	5	5
35	5	5	5	5	5	5
36	5	2	5	5	5	5
37	5	5	5	5	5	5
38	5	5	5	5	5	5
39	5	5	5	5	5	5
40	5	5	5	5	5	5
41	5	5	5	5	5	5
42	5	5	5	5	5	5

Table 8. Number of video sequences retained (per subject and per emotion)

Table 9 shows the total number of video sequences that were retained for each of the considered emotion.

Emotion	Number of video sequences
Anger	200
Disgust	189
Fear	187
Happiness	205
Sadness	195
Surprise	190

Table 9. Number of video sequences in the database, per emotion

6. Conclusion

As announced in the introduction, the need for a good benchmark database is an important limiting factor in the evaluation of multimodal emotion recognition algorithms. This paper described how such a benchmark database was created, using a protocol that immerges the subject in a situation evoking the emotion to be elicited. The entire process that led to a relevant multimodal affective database is exposed, along with the technical aspects involved in such a work. The database is made publicly available for scientific research purposes, through a website given in the acknowledgments section.

7. Acknowledgments

Olivier Martin is funded through a FIRST fellowship from the Walloon region, Belgium (contract FIRST n° EPH3310300R0312/215286).

The database is available, free of charge, for research purposes only. To obtain a copy of the database, the interested reader is invited to refer to the “results” section of the website http://www.enterface.net/enterface05. The database is available via the link “Project #2 database”.

The authors would like to thank the eNTERFACE’05 (http://www.enterface.net) organization committee for their support in the recording of the database. The authors would also like to thank all participants of the eNTERFACE’05 workshop for their participation in this database. Special thanks go to the team of the project “Multimodal Caricatural Mirror” [10] for their helpful contribution in the creation of the protocols presented in this paper.

8. References

[1] T. Kanade, J. F. Cohn, and Y. L. Tian: “Comprehensive database for facial expression analysis”. Proc. 4th IEEE International Conference on Automatic Face and Gesture Recognition (FG'00), pages 46--53, 2000.

[2] http://www.irc.atr.jp/~mlyons/jaffe.html

[3] A. M. Martinez and R. Benavente: "The AR Face Database," tech. rep., CVC #24, 1998.

[4] N. Amir, S. Ron, and N. Laor: “Analysis of an emotional speech corpus in Hebrew based on objective criteria”. Proceedings of the ISCA Workshop on Speech and Emotion (pp. 29–33), 2000

[5] P. Greasley, J. Setter, M. Waterman, C. Sherrard, P. Roach, S. Arnfield, and D. Horton: "Representation of prosodic and emotional features in a spoken language database". Proceedings of the 13th International Congress of Phonetic Sciences. Stockholm. 242-245, 1995.

[6] E. Douglas-Cowie, R. Cowie, and M. Schröder: “A New Emotion Database: Considerations, Sources and Scope”. Proceedings of the ISCA Workshop on Speech and Emotion (pp. 39–44), 2000.

[7] P. Ekman and W. V. Friesen: “Facial Action Coding System”, Consulting Psychologist Press, 1977

[8] P. Ekman and W. V. Friesen: “Unmasking the Face”, Palo Alto: Consulting Psychologists Press, 1984.

[9] http://www.enterface.net/enterface05

[10] O. Martin, I. Kotsia, A. Savran, J. Adell, A. Huerta, R. Sebbe, B. Macq and I. Pitas: “A Multimodal Caricatural Mirror”, Proceedings of the eNTERFACE Summer Workshop on Multimodal Interfaces, pp.13—20, Mons, Belgium, July-August 2005.