J Multimodal User Interfaces
DOI 10.1007/s12193-012-0112-x

ORIGINAL PAPER

A cross-cultural, multimodal, affective corpus for gesture expressivity analysis

G. Caridakis · J. Wagner · A. Raouzaiou · F. Lingenfelser · K. Karpouzis · E. Andre

Received: 5 March 2012 / Accepted: 15 September 2012

Abstract A multimodal, cross-cultural corpus of affective behavior is presented in this research work. The corpus construction process, including issues related to the design and implementation of an experiment, is discussed along with resulting acoustic prosody, facial expressions and gesture expressivity features. However, research work presented here focuses more on the cross-cultural aspect of gestural behavior defining a common corpus construction protocol aiming to identify cultural patterns within non-verbal behavior across cultures i.e. German, Greek and Italian. Culture specific findings regarding gesture expressivity are derived from the affective analysis performed. Additionally, the multimodal aspect, including prosody and facial expressions, is researched in terms of fusion techniques. Finally, a release plan of the corpus to the public domain is discussed aiming to establish the current corpus as a benchmark multimodal, cross-cultural standard and reference point.

G. Caridakis (✉) · A. Raouzaiou · K. Karpouzis
Image, Video and Multimedia Systems Laboratory,
National Technical University of Athens, Athens, Greece
e-mail: gcari@image.ntua.gr

A. Raouzaiou
e-mail: araouz@image.ntua.gr

K. Karpouzis
e-mail: kkarpou@image.ntua.gr

J. Wagner · F. Lingenfelser · E. Andre
Multimedia Concepts and their Applications Laboratory,
Augsburg University, Augsburg, Germany
e-mail: johannes.wagner@informatik.uni-augsburg.de

F. Lingenfelser
e-mail: florian.lingenfelser@informatik.uni-augsburg.de

E. Andre
e-mail: andre@informatik.uni-augsburg.de

Keywords Affective computing · Multimodal corpora · Cross-cultural corpora · Gesture expressivity

1 Introduction

There is a wide variety of psychological researches concerning cultural differences in emotion expression and recognition of human beings. A well known psychological research is the one of P. Ekman, about cultural differences in facial expressions of emotions [18]. Together with successive studies occurs the impression that basic emotions can be recognized across different cultures by investigating facial cues-even if observed cultures do not share many similarities or relations. However, these studies acknowledge cross-cultural differences leading to major misinterpretations, especially between people from Caucasian and Asian backgrounds. So most studies take an intermediate stance concerning the universality of facial emotion expression across cultures [19].

As we approach the problem of cross-cultural emotion recognition from an engineering point of view, i.e. automatic, multimodal emotion recognition, we are interested in the possibility of establishing an architecture that is able to deal with cross-cultural recognition problems and embed into a universal emotion recognition framework. An important basis for such a framework is the collection of emotional corpora for a variety of cultures. Such corpora are presented in the next section. However, since previous studies used different experimental settings, the results are difficult to be compared and interpreted. For example, it is hard to identify culture-specific patterns of emotional expressions if emotions from TV shows are recorded for one culture and emotions from man-machine interactions for another. In particular, it is hard to say whether differences in emotional

Published online: 18 October 2012

expression result from cultural differences or differences in the experimental setting. The objective of our work was therefore to define a common protocol, annotation and evaluation scheme for recording and analyzing corpora containing emotional human behaviors across cultures. The resulting multimodal corpus includes human emotional data from three different cultures, i.e. German, Greek and Italian. Furthermore, it does not only contain synchronized multiple modalities for the three investigated cultures, i.e. speech, facial expressions and gestures, but also multiple human behavior capturing techniques, i.e. video, Wii mote and data gloves. In our paper, we present the annotation scheme we used across all three cultures, the multimodal features extracted from the three culture-specific corpora as well as first results from our recognition analysis. In particular, we present the results of a cross-cultural expressivity analysis for bare hand video-based gestures, a modality which has been scarcely explored in previous research work.

2 Related work

Designing, recording and labeling human affective expressions is a prerequisite in designing affective aware systems. Many aspects are included in the above mentioned processes involved in creating an affective corpus. Behavior spontaneity, recorded modalities, labeling are merely a few of the aspects that have to be taken under consideration when creating multimodal, affectively enriched corpora aiming to be used for affective analysis. Naturalistic behavior is considered ideal for validating real life affective analysis systems, although such behavior is relatively rare, filled with subtle context-based changes and difficult to be recorded with non-intrusive methods, while a large number of issues and internal processes of the subject involved in the affective elicitation methods influence the final result. Finally, the adopted emotion representation, annotation and labeling scheme should be predefined since these decisions are extremely important to both automatic affect recognition and user perception tests.

Despite the above mentioned difficulties, the necessity for creating reusable databases consisting of affectively enriched human behavior has resulted in a number of attempts for creating multimodal corpora. The importance of each corpus is determined by the effort and reasoning for each decision involved in the database creation as well as the research work performed from the automatic analysis view using the specific corpus. The Belfast database [14] mainly consists of sedentary interactions, from chat shows, religious programs and discussions between old acquaintances. The FeelTrace [12] tool was used for labeling the corpus recording the perceived emotional state via dimensional rating. The EmoTV corpus [1] is another corpus, which is in French and also draws material from TV interviews, but uses episodes with a wider range of body postures and more monologue, such as interviews on the street with people in the news. EmoTV uses ANVIL [23] as a platform and the coding scheme uses both verbal categorical labels and dimensional labels (intensity, activation, self-control and valence). A corpus construction attempt [22] was also performed within the HUMAINE EU-IST project framework during its Third Summer School held in Genoa in 2006. While the previous corpora consisted of real life interviews, the Genoa corpus included acted human behavior induced using a process similar to the one adopted in the GEMEP corpus [4]. The GEMEP (Geneva Multimodal Emotion Portrayals) corpus constitutes a repository of portrayed emotional expressions. A pseudo-linguistic sentence was pronounced by the participants while acting through eight emotional states uniformly distributed in valence-arousal space (two emotional states per quadrant).

Another interesting corpus is the Inter-ACT (INTEracting with Robots-Affect Context Task) [11], an affective and contextually rich multimodal video corpus including affective expressions of children playing chess with the Philips iCat robot [25]. Currently the corpus is protected for privacy reasons, due to the presence of children in the recordings. In the Activity Data and Spaghetti Data sets [15], volunteers' emotions were recorded during outdoor activities, while in EmoTaboo [35] pairs of people play the game Taboo while their faces, upper bodies, and voices are recorded. One of the subjects is a confederate, making sure that enough emotional reactions are observed in the other person.

The Interactive Emotional Dyadic Motion Capture database (IEMOCAP) [7] contains interactions of ten actors in dyadic sessions with markers on them. The scenarios elicited emotions such as happiness, anger, sadness, frustration and neutral state. An interesting multimodal corpus is the 3-D Audio-Visual Corpus of Affective Communication [20],a new audio-visual corpus for speech and facial expression in the form of dense dynamic 3D face geometries. The corpus consists of 14 native English speakers uttering 1109 sequences, with 11 suggested emotional states: "Negative", "Anger", "Sadness", "Stress", "Contempt", "Fear", "Surprise", "Excitement", "Confidence", "Happiness", and "Positive". The corpus will be made available for research purposes. MAHNOB-HCI [30] is another interesting multimodal corpus. The corpus consists of synchronized

recording of face videos, audio signals, eye gaze data, and peripheral/central nervous system physiological signals. 27 participants from both genders and different cultural backgrounds participated in two experiments. The recorded videos and bodily responses were segmented and stored in a database, available to the academic community.

The innovation of the corpus construction in our paper lies in its focus on gesture expressivity, the inclusion of multiple cultures and multiple human behavior capturing techniques,

Table 1 Example of the used Velten sentences per emotion category

The hike was fantastic! You won't believe it! But we made it to the top!	Positive
The names on the mailing list are alphabetically ordered	Neutral
Sometimes I wonder whether my effort is all that worthwhile	Negative

i.e. video, Wiimote and Datagloves. The stimuli used for emotion elicitation included only Velten sentences [17] and no other visual or auditory stimuli. The multicultural corpus introduced, allows for intercultural affective analysis, while the variety of technologies used to record human body behavior supports studies on their obtrusiveness effect. Aspects of the presented research work include: the data is multicultural, from different modalities and captured with different techniques. German, Greeks and Italians, while speaking, use their hands in a different way. The described experiment is providing us with the means to compare the expressiveness not only between the different cultures, but also between the different capturing techniques and the different emotional characterizations. Furthermore, the data is synchronized, so analyzing the affective behavior of the user allows us to extract conclusions for the correlation of gesture expressivity with acoustic prosody and facial expressions.

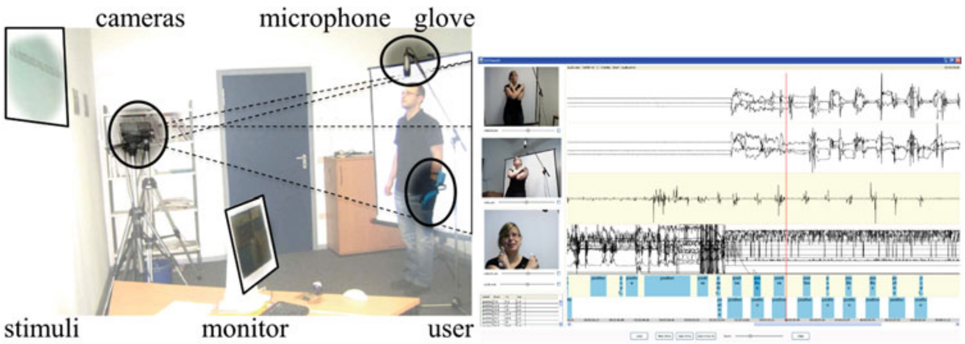
3 Corpus construction

The overall corpus construction process involved technical setup in terms of devices and software used for recordings, subjects recruiting, training, affective immersion and privacy protection. The corpus construction is described in detail in [10].

Affective immersion and procedure The adopted emotion elicitation method was inspired by the Velten mood induction technique [31] where people had to read aloud a number of sentences that put them in particular emotional state. The sentence, containing a clear emotional message, was displayed and the user was asked to express the corresponding emotion through gesture and speech. We selected in total 120 sentences (40 for each target class) such as the ones illustrated in Table 1. We decided to choose the sequence positive-neutral-negative in order not to switch directly between the two emotional extremes. Furthermore, users usually feel less motivated towards the end of the experiment and it would be harder to put them into a positive emotional state. Each of the three emotional sessions is again divided into three sections, during which we equip the user with different interaction tools (data glove, Wii and free gesturing). In every country, the experiment was conducted, the institute responsible for the recordings, recruited local subjects and performed the experiment. The same set of Velten sentences were accordingly translated and adapted by native speakers and this translated set was used in the experiments conducted in the respective country. In Greece, 11 subjects (6 male and 5 female) between 23 and 40 years old took part in the experiment, while in Germany 21 subjects (11 male, 10 female) were following our scenario. Their age varied between 20 and 28 years old, while in Italy 19 (11 males and 8 females) took part in the experiment, between 24 and 48 years old.

Subjects training and ethical issues Before the experiment we recorded a video with the whole procedure. A trained person was executing the gestures with Wii-mote, glove or bare hands. The candidate participants were offered the opportunity to watch the video and/or read the Velten sentences and to pose any questions they want regarding the experiment. Once they agreed to participate, they were given a consent form to sign, ensuring that they are informed about the scope of the experiment, their involvement and that they can assess the risks that might occur from the processing of data.

Hardware setup and recording software During the recordings the user stands in front of a neutral background (Fig. 1). The stimuli, i.e. the Velten sentences, is projected



(a) Hardware **(b)** SSI recording software

Fig. 1 Experimental setup

on a screen in front of him. The projection is adjusted in a way that the user can read the displayed text without the need to turn his head. Below the projection, in a distance of two meters and approximately at the height of the user's face, two high-quality cameras (720 × 576 pixels, 25 fps, 24 bit colour depth) are placed. The first camera is set-up to capture the user's complete body including arm gestures, while the second camera aims at the user's face and captures a close-up of shoulder and head. In addition the whole scene is captured at a lower resolution with a webcam, primarily for annotation and monitoring purpose. Audio is recorded with an USB microphone (Samson C01U, 16 kHz, mono, 16 bit). To avoid occlusions in the videos a stand is used to locate the microphone on top of the user's head. Each recording is divided in three parts characterized by different interaction modes. During the first mode the user is wearing a data-glove on one hand. The data-glove is provided by HumanWare and is used primarily to record finger movements during the experiment, to verify whether (and how much) users gesticulate with their hands and fingers. The dataglove records 26 signals at a sampling rate of 50 Hz: 15 signals for flexions of all fingers on one hand, 2 signals for flexion and ad/abduction of the wrist and since it embeds an IMU (inertial measurement unit) in the forearm it also records a 3-axial magnetic field, 3-axial acceleration and 3 angular velocities. The data transfer is done over a wireless Bluetooth connection. During the second interaction mode the user holds Nintendo's Wii remote control in each hand, which measures 3D acceleration. The last interaction mode is freehand.

During the corpus affective analysis we aim to, additionally to unimodal analysis, also investigate the relations between modalities and explore ways to fuse different affective cues. This, however, requires a proper synchronization between the modalities. To obtain synchronized recordings we use Smart Sensor Integration (SSI), a software framework for multimodal signal processing in real-time, developed at the University of Augsburg [33].

4 Affective annotation

Our intention was to record affective behavior, which reflects the variability in real-life. Thus, we decided not to perform the experiment with professional actors. In the past this has been often practised (e.g. Database of Facial Expressions (DaFEx) [5] or Berlin Database of Emotional Speech (Emo-DB) [6]) and led to very homogeneous databases with clearly differentiable classes. This is because actors are particularly skilled to control their body movements and expressions, which allows them to repeat certain behavior in a similar way over and over again. Besides, they are used to adopt prototypical behaviors that is easily recognized by the audience. Hence, recruiting professionals helps to obtain homogeneous and well distinguishable samples, but at the same time also limits the variance in the data and leads to exaggerated and less natural observations. Having this in mind, we decided in favor of participants without acting background and advised them to use whatever body language and vocal expressivity they felt appropriate. Following such an approach we were able to collect samples covering a large variety of behavioral patterns, inline with our research goals. As a side effect, however, we often observed a discrepancy between what was the intended emotional class (given by the Velten sentence) and the behavior expressed by the participant. Consequently it became indispensable to rework the pre-annotations, we had automatically generated from the stimuli scripts.

4.1 Scheme and motivation

In order to apply post recording annotations we had to decide whether to keep the original emotion classes (*positive*, *neutral* and *negative*), or go for a different annotation scheme. When reviewing some of the samples it became obvious that even though the emotion inducing sentences used are more categorized along the valence axis, participants were expressing their emotions at different arousal levels. This applied not only between classes, but even within the same emotional class. Especially when looking at negative sentences, some samples tend to a depressed and sad mood (low arousal), while others are expressed in an aroused and angry way (high arousal). One way to include arousal would be by using an annotation based on dimensional axes, e.g. activation and evaluation. Measuring emotions in an activation-evaluation space has a long tradition [27,28] and the advantage of allowing intermediate and continuous ratings. This is especially useful to describe shaded emotions and emotional changes within the same episode. An appropriate labeling tool for this task would be for instance FEELTRACE [12] developed at Queen's University Belfast. However, as a result of the experimental design we do not expect many blended or masked emotions to occur, nor do we encounter sudden emotional changes within a sentence (which in our case defines a clearly closed emotional episode). Hence, we decided to stick to the categorical approach, but exchange the original categories with a new set of four classes based on the activation-evaluation space, namely *positive-low*, *positive-high*, *negative-low* and *negative-high*. We decided not to include neutral as a fifth class, as nearly all neutral and part of the positive observations share a calm and optimistic subtone, in our new scheme represented by *positive-low*.

A second decision concerns the modalities to serve as source throughout the labeling process. In the corpus at hand we can choose among three available modalities, the audio, face and gesture channel. Of course, any combination of the three is possible, too. In fact, presenting all available modalities to the raters would provide the most comprehensive information. As long as emotions are consistently expressed in all modalities this would be the most obvious option. However, there is evidence that when emotions are expressed in everyday interaction, different channels are as likely to conflict as to complement. Douglas-Cowie et al. [16] showed this by means of two multimodal databases, the EmoTV corpus [1] and the Belfast naturalistic database [14]. Since participants in our corpus had no acting background and were not particularly instructed to express emotions uniformly across modalities, we may reckon a similar effect. Therefore, we decided to generate two annotations: based on the audio and video channel, which would allow us to investigate a possible divergence at least between the two channels.[1] Both runs were completed independently of each other with a temporal distance of several weeks (same annotators, though).

4.2 Statistical analysis

The analyzed German sub-corpus consists of 2520 samples (21 participants * 3 classes * 40 sentences) almost equally distributed across gender. From these, 7 samples were excluded, either because subjects refused to perform or due to technical problems. All remaining 2513 segments were labeled by three raters in two independent loops. The raters' native language and nationality were identical to the ones of the subjects of the recordings they were annotating and the same procedure will be applied to the other two sub-corpora in the near future. To achieve a high immersion, headphones were used and raters could individually adjust volume during playback. Segments were replayed in chronological order and annotators could loop an utterance as often as needed and even jump forth and back in order to repeat older segments and re-assign labels. Final combination of differing annotations is done via majority decision, as three assessments are given to each orientation of valence or arousal respectively, decisions are definite. For instance, if the 1st rater assigns label *low* and *positive*, the 2nd *low* and *negative*, and the 3rd *high* and *negative*, the segment is finally labeled as *low and negative*. A couple of weeks later, the procedure was repeated, but this time solely video recordings of the participant's face were shown. Again, raters could go forth and back, and repeat a sequence if wished.

Table 2 Agreement between raters under both conditions

	Fleiss' Kappa value		
	Valence-arousal	Valence	Arousal
Audio	0.52	0.84	0.38
Video	0.52	0.71	0.48

To report inter-rater reliability we calculate the kappa value according to Fleiss et al. [21]. Fleiss' Kappa value is a common way to measure the agreement over multiple raters. It is expressed as a number between 0 and 1, where 1 indicates a perfect agreement. Table 2 gives Kappa values for both conditions, i.e. when labeled on audio only versus labeled on video only). In both cases the kappa value for all four classes amounts to 0.52, which expresses a moderate agreement. If we look at the kappa value for valence and arousal independently, there are two facts to notice. For both conditions the agreement for valence is much higher (0.84 and 0.71, which implies high agreement) as it is for arousal (0.38 and 0.48, which implies only a fair agreement). This could derive from the fact that expressed sentences were selected to be either negative, neutral or positive and that people are better in judging valence than arousal [3]. Thus, it should be easier for the raters to agree on the valence of an utterance than its level of arousal. Apart from this, we observe a higher agreement for valence if annotation is based on audio recordings, while the agreement for arousal is higher if based on video recordings. This is another sign that emotions in the studied corpus are not always expressed in a coherent way across modalities. In Fig. 2 class distributions are visualized for each of the three raters, as well as, for the combined case. Indeed, the relative amount of samples assigned to the four classes differs not only between raters, but also for the final combination. For instance, in the video annotation we find one third less samples within the class *positive-high*, whereas in the audio annotation the number of samples falling into class *positive-low* is about 12 % smaller.

5 Affective analysis

With the corpus presented in this article we hope to contribute to the field of affective computing in two ways: first, offering recordings repeated under the same conditions in three European countries in order to investigate cultural differences in the expression of emotion. Second, offering a multimodal emotion corpus that allows researchers to test fusion algorithms on semi-spontaneous and inhomogeneous data rather than purely acted data that is perhaps too uniform to reflect human behavior in natural situations. While there is still a large burden of work to accomplish in terms of annotation

[1] Initially, we had expected that emotions are more or less homogeneously expressed across the three modalities. But first analysis, yet based exclusively on annotations obtained for the audio channel, showed surprisingly low improvements when adding information of other modalities to the audio channel [26]. This was taken as a first hint for a possible discrepancy between the modalities.

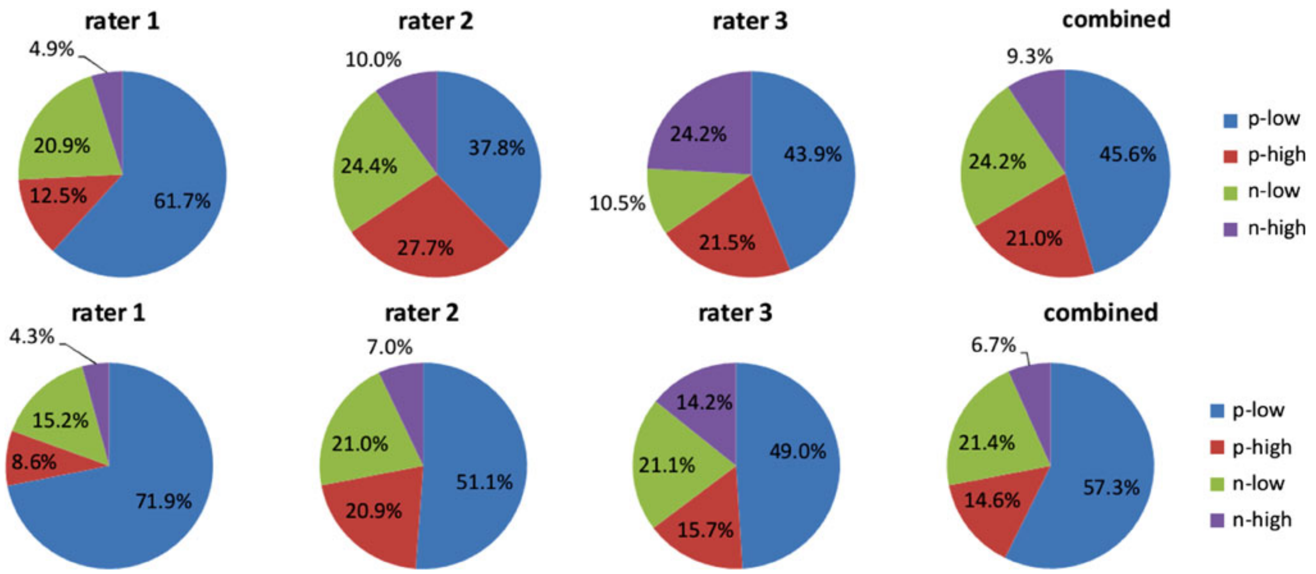


Fig. 2 Class distribution for individual raters and in combination on basis of the audio (*top*) and video channel (*bottom*)

and analysis, we can already present first results with respect to both approaches.

5.1 Gestural analysis

5.1.1 Image processing for hand detection and tracking

Regarding the hand and head detection and tracking required step for extracting expressivity features from a gesture, we adopted a video based, non obtrusive approach which focuses on low computational cost and robustness. The overall process, described in detail in [8], includes creation of moving skin masks and tracking the centroid of these skin masks among the subsequent frames of the video depicting a gesture. Real time color model of the human skin is constructed by sampling the upper area of the box containing the head which corresponds to the forehead of the user, thus tackling illumination issues which often impede natural interaction processing. Object correspondence between two frames is performed by a heuristic algorithm and the fusion of color and motion information eliminates any background noise or artifacts, thus reinforcing robustness. The overall process is depicted in Fig. 3.

5.1.2 Gesture expressivity modeling

Gesture expressivity is modeled using six parameters, namely *overall activation* (OA), *speed* (TE), *power* (PO), *fluidity* (FL), *spatial extent* (SEmax and SEmean), each capturing a certain aspect of natural hand interaction. Thereby it is not relevant what the actual gesture looks like, but rather *how* it is performed.

Overall activation is considered as the quantity of movement during a dialogic discourse and is formally defined as the sum instantaneous quantities of motion. Spatial extent is expressed with the expansion or the condensation of the used space in front of the user (gesturing space). The temporal expressivity parameter denotes the speed of hand movement during a gesture and dissociates fast from slow gestures. The Power expressivity parameter refers to the movement of the hands at during the stroke phase of the gesture. Detecting the stroke phase of the gesture is far from trivial and thus we opted to associate this parameter qualitatively with the acceleration of hands during a gesture. Fluidity differentiates smooth/elegant from the sudden/abrupt gestures. This concept attempts to denote the continuity between hand movements and is suitable for modeling modifications in the acceleration of the upper limbs. Under this prism, we formally define as the gesture's fluidity the variation of power. According to the latter formalization, fluidity expressivity parameter corresponds a quantity that is reversely proportional to the notion of fluidity.

5.1.3 Formalization for bare hands

Defining a gesture G as a sequence, of (x^G_{li}, y^G_{li}) and (x^G_{li}, y^G_{li}), relative to the head coordinates of the left and right hand respectively, for a duration of T frames thus $i \in [1, T]$. *Overall activation* is formally defined as the sum instantaneous quantities of motion: $OA_G = \sum_{i=1}^{T-1} D^G_{li} + D^G_{ri}$, $D_i = \left|\overrightarrow{(x_i, y_i)(x_{i+1}, y_{i+1})}\right|$. In order to provide a strict definition of the *Spatial Extent* expressivity feature spatial extent is considered as the maximum

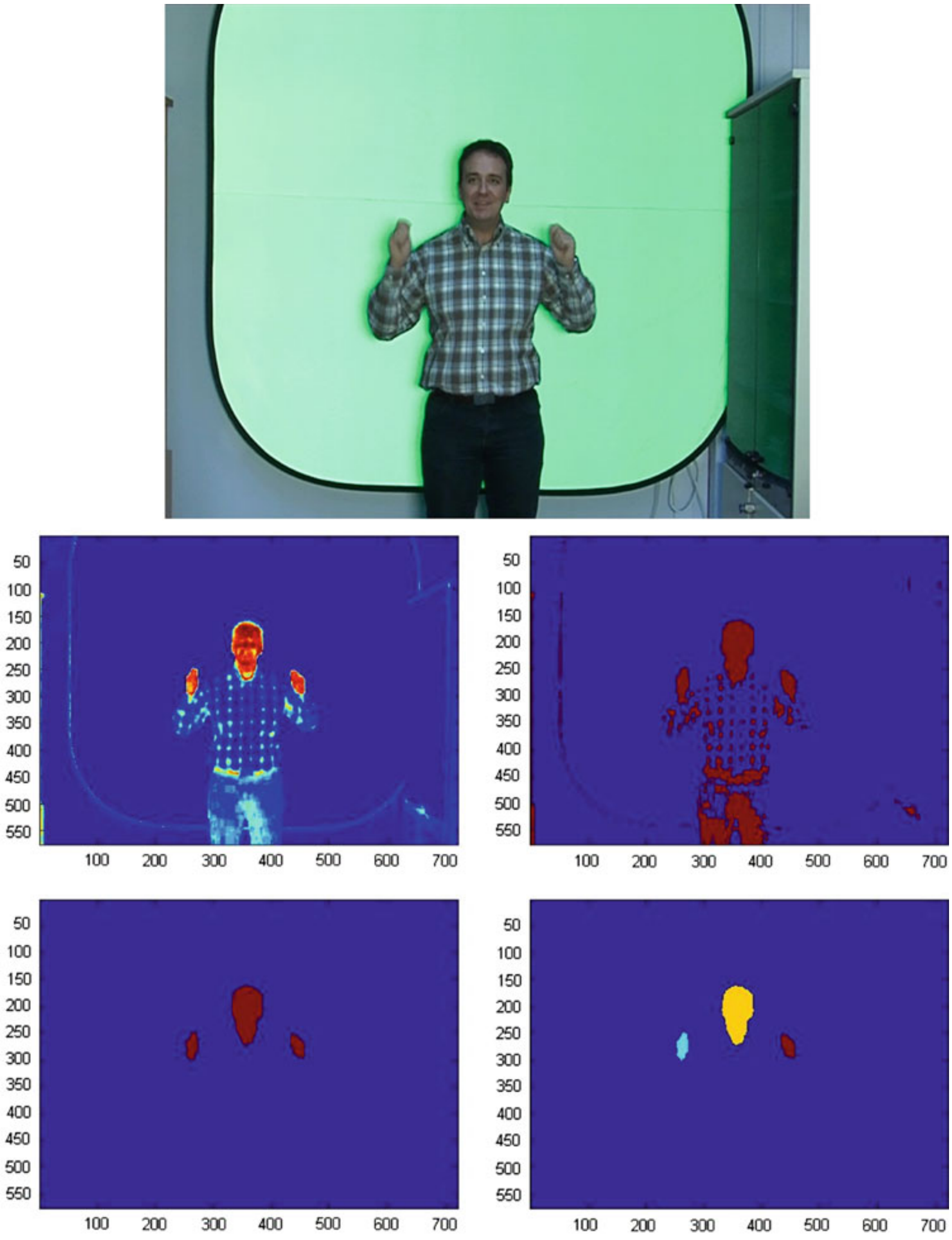


Fig. 3 Image processing intermediate steps and final result for hand detection

value of the instantaneous spatial extent during a gesture: $SE_G = \max e_i, i \in [1, T], e_i = \left|\overrightarrow{(x_{ri}, y_{ri})(x_{li}, y_{li})}\right|$. The *Temporal* expressivity parameter is defined as the as the arithmetic mean of this quantity and since OA_G, as defined earlier and corresponds to its discrete integral $TE_G = \frac{OA_G}{T}$. The *Energy* expressivity parameter is associates qualitatively with the first derivative of the norm of D which refers to the acceleration of hands during a gesture $PO_G = |D|'$. *Fluidity* is formally defined as the variance of the energy expressivity parameter as described in the previous paragraph $FL_G = var(PO_G)$. A detailed description of the bare hand modeling approach can be found in [9].

5.1.4 Formalization using Wii

Wii is equipped with built-in accelerometers and thus provides readings of the acceleration along three axes, eliminating the need for a feature extraction module. Only post-processing is required in order to reduce the influence of gravity. In accordance to bare hand formalization, the expressivity features can be calculated as if each hand's motion vector **mv** is calculated by the respective acceleration measurements $|\mathbf{mv}| = \sqrt{(\int\int acc_x)^2 + (\int\int acc_y)^2 + (\int\int acc_z)^2}$. As a result, the *Overall activation*, is modeled as $OA = \sum_0^n |\mathbf{mv_l}| + |\mathbf{mv_r}|$ for a gesture of n sampling points. The

Table 3 Overview of pre-processing steps and feature extraction methods applied in our experiments

Modality	Channels	Pre-processing	Short-term features	Long-term feature	Total
Voice	Mono audio, 16 kHz	Pre-emphasis filter	Pitch, energy, MFCCs, spectral, voice quality	Mean, median, maximum, minimum, variance, median, lower/upper quartile, absolute/quartile range	1316
Face	RGB video, 720 × 576, 25 fps	Conversion to gray image	Bounding box of face, position of eyes, mouth and nose, opening of mouth, facial expression happy/angry/sad surprised	Mean, energy, standard deviation, minimum, maximum, range, position minimum/maximum, number crossings/peaks, length	264

rest of the expressivity features are formalised accordingly (e.g. $TE = \int_{t=0}^{n}(acc_x + acc_y + acc_z)dt$).

5.2 Multimodal analysis

In addition to the gesture modality, cross-cultural expressivity analysis of which is presented in the previous section, in the following we will focus on the vocal and facial modality within the subcorpus described in Sect. 4.2.

5.2.1 Feature extraction

Description of the raw signal is done by features extracted from the recordings. Different pre-processing steps were used per modality in order to suppress unwanted aspects of the signals. Afterwards the features are calculated and depending on whether the features are extracted on a small running window of fixed size or on longer chunks of variable length, we denote them as short- or long-term feature. Table 3 offers a summary of applied processing methods and calculated feature types for each modality. For vocal feature extraction we use features calculated by the EmoVoice component [32]. Video processing is provided by *SHORE*, a library for facial emotion detection developed by Fraunhofer IIS[2] [24]. For a detailed description of the feature extraction procedure please refer to [34].

5.2.2 Evaluation techniques

For evaluation, we adopted the realistic and user independent Leave-One-Speaker-Out approach, where we consecutively draw samples belonging to one single subject out of the set. Remaining samples are used for training of the classification model which are finally tested against the isolated samples. As classification scheme we chose the simple but efficient Naive Bayes approach, which employs the Bayes Theorem:

$$P(E_i|f_1,\ldots,f_n) = \frac{P(E_i)\prod_{j=1}^{n} P(f_j|E_i)}{P(f_1,\ldots,f_n)}$$

Probability of the emotion E_i, given an observed feature vector $(f_1,\ldots,f_n)$ of dimension n, depends on the *a-priori* probability $P(E_i)$ of the emotion, multiplied by the product of the probability of each feature f_i given the emotion, divided by the *a-priori* probability of the feature vector. As classification result, the emotion E_i from a set of N emotions $E_1,\ldots,E_N$ that maximises the equation is chosen. Parameters for the probability distributions $P(E_i)$ and $P(f_j|E_i)$ are gained from the annotated training data.

5.2.3 Multimodal fusion

Combining facial and vocal information has been reported to improve recognition accuracy in emotion recognition tasks [36]. Fusion of modalities can be achieved at feature-level or decision-level. In case of feature-level fusion, features extracted from the different channels are simply merged into a single and high dimensional feature set. Afterwards a single classifier is trained for the task of classification. For decision-level fusion several methods have been proposed. They all have in common that one classifier is trained for each modality and probabilities are combined to build a final decision. In [26] we have carried out a systematic comparison of a total of 16 fusion methods. Among the tested decision-level algorithms *Product Rule* performed stable across different datasets. As the name implies, the *Product Rule* combines probabilities by multiplying outputs of the classifiers per class. Final decision is found by keeping the class with the highest arithmetic product. In the following we will apply feature-level fusion, as well as, *Product Rule* to include both fusion strategies.

[2] http://www.iis.fraunhofer.de/en/bf/bv/ks/gpe/demo/.

Table 4 Results for single and fused modalities based on both, the audio and the video annotation (separated by |)

	Recognition results in %				
	Positive-low	Positive-high	Negative-low	Negative-high	Average
Voice	58 | 53	47 | 43	49 | 58	47 | 44	50 | 49
Face	42 | 41	72 | 71	32 | 63	53 | 39	50 | 53
Feat. fus.	54 | 54	64 | 59	40 | 58	50 | 41	52 | 53
Dec. fus.	59 | 54	60 | 61	45 | 57	49 | 43	53 | 54
	Arousal				
	High	Low	Average		
Voice	64 | 60	71 | 74	68 | 67		
Face	66 | 70	70 | 79	68 | 75		
Feat. fus.	68 | 71	73 | 77	71 | 74		
Dec. fus.	66 | 70	74 | 77	70 | 74		
	Valence				
	Positive	Negative	Average		
Voice	65 | 63	71 | 74	68 | 68		
Face	63 | 48	66 | 75	65 | 62		
Feat. fus.	68 | 64	70 | 74	69 | 69		
Dec. fus.	69 | 64	71 | 73	70 | 69		

5.2.4 Results

Table 4 shows classification results for the audio and video annotations. In case of audio based annotation, unweighted average recall (class-wise average) for four classes is identical for both modalities, with a notable high score in *positive-high* achieved by facial observations. Voice and face modality perform nearly equal for arousal classification, *low* arousal is better detected than *high*. The vocal modality seems slightly better suited for valence recognition, overall are *negative* emotions better recognized than *positive* ones.

Table 4 also gives recognition rates gained with the video annotation. With this annotation, the facial modality outperforms the vocal one for four classes. The high score in *positive-high* is still achieved. Arousal classification is now clearly dominated by classification of the face, *low* arousal is still better detected than *high*. *Negative* emotions Negative emotions are again better recognized when trying to detect along the valence axis. Surprisingly, the facial modality performs even worse with the video annotation.

When comparing both annotations, best results on four classes are achieved with the facial analysis on video annotation. The facial modality recognizes the *positive-high* emotion very well with both annotations—presumably because it is well suited for detection of smiles and movements of the face associated with laughter. This leads to the highest score for detection of high arousal (*70 %*), gained with the video annotation. Recognition rates for the vocal modality are stable across both annotations. It is overall better suited for valence classification.

Accuracies for single modalities, along with results for the two fusion approaches are listed. For runs based on the audio annotation we see an improvement of about 2–3 %. For video based annotation, however, where recognition rates vary between the channels this positive effect is no longer observed. The fusion approach always gains a result similar to the outcome of the stronger modality, though. We believe that this moderate impact follows from the inhomogeneous expression of the emotions across modalities as was already discussed in Sect. 4.2. We have published a detailed discussion on this problem elsewhere [26].

5.3 Cultural analysis

Current section presents results of analysis carried out focusing from a cultural point of view. Section 5.3.1 discusses, the focal point of the research work presented here, cross-cultural gestural analysis. Additional within-cultural (Sect. 5.3.2) and cross-cultural classification (Sect. 5.3.3) are presented in order to compare classification rates and explore differences between both cultures as well as the possibilities of a generalized, universal recognition framework (Sect. 5.3.4). This cultural analysis section aims to investigate the following questions and the findings are discussed in Sect. 5.3.5:

Table 5 The table summarizes the result of a multiple comparison test for the six expressivity features extracted for the three countries

Further explanation can be found in the text

	OA	TE	PO ($\cdot 10^{-4}$)	FL	SEmax	SEmean
Mean values						
de	5.97	0.05	2.53	0.02	1.37	0.97
it	11.19	0.07	3.08	0.05	1.67	1.24
gr	11.30	0.05	0.36	0.02	2.32	1.45
Pairwise comparison ($p = 0.01$)						
de-it	↓	↓	–	–	↓	↓
de-gr	↑	–	–	–	↓	↓
it-gr	–	↑	–	–	↓	↓

- Which are the gestural expressivity characteristics for each of the examined cultures and how are they related?
- Do cultural differences between European origins influence the expression of emotion and how does this affect automatic emotion recognition?
- Are within-cultural emotion recognition frameworks more potent than a universal one?
- Which modality is more suitable for cross-cultural emotion recognition (facial, vocal or a combined approach)?

5.3.1 Cross-cultural gestural analysis

As period of interest we took again the turns of the Velten sentences. Features were extracted for each subject and turn, and collected in a single feature matrix describing the gestural repertoire of that country. This is repeated for all three countries. Next, we perform a multiple comparison test to look for significant differences between cultural groups. Therefore, we apply one-way analysis of variance and compare the means of the groups to test whether the hypothesis that they are all the same holds. If the so called p-value falls under the significance level (in our case 0.01), this suggests that the means of the two groups are significantly different with respect to that feature.

Results of the multiple comparison test are summarized in Table 5. The upper part of the table shows the mean values for each expressivity feature calculated from all samples belonging to one of the following three cultural groups: German (de), Italian (it) and Greek (gr). The values already suggest differences for some of the features, e.g. the overall activity in the Italian and Greek sub-corpus is almost twice as high as in the German sub-corpus. The lower part of the table assigns to each feature and combination x–y of two countries one of three symbols: if no significant difference was found, ↓ to express that the particular feature is significantly lower for country x compared to country y, and ↑ if it is the other way round. Hence, we can derive that the gestures in the German sub-corpus are in average performed with less activity and at a lower speed. Same applies for the spatial extent, which takes less space compared to gestures recorded from Italian and Greek subjects. No significant differences was found in terms of power and fluidity, though. Spatial extent is generally highest for Greek gestures, but executed with less speed compared to Italian gestures.

Table 6 Results of experiments done on the German sub-corpus

	Neutral	Positive	Negative	Average
Voice	66.13	50.92	60.30	59.12
Face	65.46	72.90	33.27	57.21
Feat. fus.	65.40	70.02	49.91	61.78
Dec. fus.	68.87	67.15	47.45	61.16

Table 7 Results of experiments done on the Italian sub-corpus

	Neutral	Positive	Negative	Average
Voice	58.16	50.79	56.57	55.17
Face	71.51	69.29	28.10	56.30
Feat. fus.	57.96	70.08	46.35	58.13
Dec. fus.	72.50	68.11	28.47	56.36

5.3.2 Within-culture

As basis for our estimations about cross-cultural emotion recognition and a universal framework we have to take a look at classification accuracies within cultures. Therefore subjects from Germany and Italy are separated for training and testing. Table 6 shows results for experiments done with only German participants, Table 7 with only Italian subjects respectively. In both cases vocal and facial classification perform on a nearly equivalent level. The German framework slightly favors the spoken modality over facial expressions, on the Italian corpus it is the other way round. Both show a significant lack of accuracy for detecting negative emotions with the facial modality. This trend migrates into feature level fusion as well as decision level fusion. Overall classification within the two cultures seems to perform on a highly comparable level.

Table 8 Off-corpus experiment: train with German sub-corpus and test with Italian sub-corpus

	Neutral	Positive	Negative	Average
Voice	50.15	12.99	43.43	35.52
Face	60.34	68.90	7.30	45.51
Feat. fus.	54.01	66.93	27.01	49.32
Dec. fus.	57.07	66.93	24.45	49.48

Table 9 Off-corpus experiment: train with Italian sub-corpus and test with German sub-corpus

	Neutral	Positive	Negative	Average
Voice	34.00	69.40	18.71	40.70
Face	62.59	45.38	34.03	47.33
Feat. fus.	55.91	67.97	29.49	51.12
Dec. fus	66.33	58.73	29.11	51.39

5.3.3 Cross-culture

The consequential next step in evaluating cultural differences for automatic classification of emotions between Germans and Italians is to train classification models with users belonging to one specific culture and to test the generated classifiers against subjects from the other country. Table 8 shows results of training data taken from Germany and test-samples of Italian origin. Table 9 describes the reverse approach.

Cross-cultural classification results in a nutshell: They completely fall off compared to within-culture categorization. Note that we took care of standardizing all recorded signals, so this phenomenon can really be tracked down to differences in emotional expressions between recorded subjects. German and Italian emotions are better recognized by the facial cues, but to a much lower degree than seen in Tables 6 and 7. Nearly every modality completely lacks in terms of at least one emotion category. Feature fusion and decision fusion do even out described problems and lift overall classification accuracy.

5.3.4 Universal

The final experiment carried out in the course of this work is the approach of a universal framework for cross-cultural emotion recognition. All recorded users are mixed up into one group for training and testing the mentioned modalities and fusion schemes (Table 10). This way we can estimate the usefulness of a collection of training data recorded from several cultures for identifying emotional categories of single subjects.

Table 10 Universal approach carried out on the whole corpus including Italian and German subjects

	Neutral	Positive	Negative	Average
Voice	70.53	47.50	41.10	53.04
Face	65.31	69.91	24.91	53.38
Feat. fus.	64.83	68.02	27.52	53.46
Dec. fus.	66.67	65.05	28.77	53.50

5.3.5 Discussion of cultural analysis

Presented results show that after incorporating both observed cultures into the training data, we regain classification accuracy slightly worse but nevertheless resembling within-culture emotion recognition. The vocal and facial modality as well as feature fusion and decision fusion perform on an equal level whilst having the same troubles with the negative emotion category as stated in Tables 6 and 7. After systematically investigating within-cultural, cross-cultural and universal emotion-classification of German and Italian expressivity-corpora we can now try to answer the major research-questions of this study.

The major cultural gestural expressivity characteristics were identified as Overall Activation, Spatial Extent and Speed. Additionally, the comparison of culture specific expressivity revealed that, as intuitively expected, Greeks and Italians gesture more actively than Germans while occupying more gesturing space. Gesture's performed by Italians are also quicker and jerkier than Greeks and Germans. Findings also justifying the widely known type of gesture that illustrates cultural specificity, the Italianate gesture. The issue of generalizing these findings on more cultures or culture groups (e.g. Northern Europe, Mediterranean) remains but hopefully the corpus release, discussed in Sect. 6, will contribute towards this by enabling more cultures to be included.

Do cultural differences between European origins influence the expression of emotion?—When doing within-cultural classifications, characteristics of results resemble each other. A good example would be the bad performance of the facial modality for identification of negativity in both cultures. The differences and their impact on automatic emotion recognition really start to come into effect when doing cross-cultural categorization. Results become unpredictable and recognition accuracy drops by a vast amount.

Are within-cultural emotion recognition frameworks more potent than a universal one?—Yes, in our case they are. But the universal one is not drastically outperformed. When incorporating all observed cultures into the training efforts, generalization seems to be possible. We can conclude this generic possibility from resembling overall performance and the problems of facial classification of negative emotions that can be observed in within-cultural as well as universal classification. Interestingly the answers to these first two questions confirm recent multi-corpora studies like [29].

The last question to be answered is whether the facial or vocal modality or a combined approach is best suited for emotion recognition across different cultures. First let us take a look at the vocal and facial modalities. In within-cultural experiments Germans favor vocal cues, Italians the facial ones—in both cases they are close. The universal framework performs pretty even on both modalities. But negativity is not well recognized by the facial classifiers across all mentioned experiments. On the other hand the facial modality turns out to be the most overall reliable one when one should really consider cross-cultural classification as shown in Tables 8 and 9. All in all fusion approaches nearly always outperform single modalities.

6 Corpus release

Aiming to contribute to the Affective Computing research community we plan to release the described corpus to the public domain. By releasing the corpus, researchers will have the opportunity to apply, test and benchmark their methodologies to the multimodal content of the corpus enhanced with the accompanying annotation and reference analysis results. One of the major issues in work related to affective analysis of emotionally enhanced human behavior is the usage of heterogeneous corpora, as described in Sect. 2 throughout different research work and by releasing current corpus we aim to tackle this issue. Understandably, it is a strong statement to argue that this corpus will satisfy all the requirements set by research approaches. On the other hand, by providing the corpus construction and annotation procedure and useful guidelines and best practices will enable researchers to reproduce such corpora or even enhance them in terms of affective aspects such as emotional categories or other emotion representation approaches, modalities, annotation schemas or simply recordings from other cultures.

The corpus will be public domain when the research teams (NTUA and UoA) involved in the construction, annotation and analysis have completed their research and published the results. In preparation of such a release a number of issues have to be resolved. Audiovisual digital content, especially visual content captured at a high resolution from two inputs, when uncompressed or compressed with a lossless compression algorithm requires extremely large storage capabilities (often many TB). Additionally, when this content should be shared among many interested parties, in different locations, the network bandwidth requirements are as challenging as the corresponding storage requirements. This large volume of data are currently hosted by NTUA storage equipment and accessed through File Transfer Protocol, but in view of the upcoming corpus release and the expected rise of network access perhaps this is a temporary solution. The need for a more appropriate storage and access facility will perhaps become more imperative once researchers start contributing new corpora. An appropriate repository, based on Cloud computing, would be Amazon's Web Services (AWS) Public Data Sets [2]. Hosting public data sets at no charge for the community, and like all AWS services, users are charged only for the compute and storage they use for their own applications. Providing access to data sets from their Amazon Elastic Compute Cloud (Amazon EC^2) instances will provide researchers a unified platform for corpus sharing. Such a solution could prove useful for dealing with the storage and distribution problem of the corpus but further investigation is certainly needed towards the suitability, adequacy and sustainability of this approach.

Additionally to hosting and access to the corpus data, copyrighting and licensing for using, modifying and redistributing the corpus should be catered for. We are oriented towards a custom license agreement based on Creative Commons Attribution-NonCommercial-ShareAlike 3.0 (CC BY-NC-SA 3.0) [13] making the corpus freely available for research and educational use. The latter CC license allows the user to copy, adapt, distribute and transmit the work under the condition the users attribute the work in the manner specified within the license. Altering, extending and transforming the corpus is allowed under the term the resulting work is distributed only under the same or similar license to the one used in the original release. Commercial usage is not allowed but this restriction, as all the conditions of the license, can be waived given the permission of the copyright holders or negotiated with the interested party. The corpus will be protected by copyright via an agreement signed by the researchers involved while forming a legal entity representing the interested parties is under consideration. Finally, privacy and personal data protection issues of the corpus participants have been already dealt with the terms and conditions of the consent form the participants signed during their involvement in the corpus construction process, as was described in Sect. 3. The participants signed a disclaimer to their right over the content of the corpus while they retain their right to be excluded from the corpus and receive information on the processing and distribution status. The multimodal content of the corpus will be accompanied with the respective annotation. The annotation procedure and results are described in Sect. 4. The annotation will be aligned to the content and statistical analysis will be provided.

Although the release of the corpus as public domain is still considered future work we have already agreed on a workplan and procedure. The interested researchers will have the opportunity to receive information regarding the corpus on a web page either hosted by an institution's web site or in a dedicated web site with a characteristic domain name. Instructions for accessing the corpus and information on licensing and terms and conditions of usage will also be available on this web site. Corpus instances containing audio-visual content with limited (appropriate for web publishing) duration and analysis will be available in order for the interested researcher to obtain a clearer understanding of the corpus. Furthermore, the visitor will have the option to download a restricted part of the corpus in order to roughly test his affective analysis architecture before requesting access to the full extent of the corpus. This restricted part of the corpus will be available with a more relaxed license (perhaps even without one) since its restricted size will not prove useful for an extensive research.

7 Discussion and conclusions

The work presented here discusses issues related to the design and implementation of an experiment, resulting in a multimodal cross-cultural corpus of affective behavior, incorporating acoustic prosody, facial expressions and gesture expressivity as well as three cultures, i.e. German, Greek and Italian. Due to the definition of a common protocol, the corpus enables us to identify characteristic patterns of emotional expression in different cultures. This was exemplified for the analysis of video-based hand gestures for which we found culture-specific differences in terms of expressivity features. For example, the German gestures were executed with less activity and lower speed than the Italian or Greek gestures. Gestural behavior, which was the focus in this article, is only one aspect of emotional behavior. Other aspects include the vocal and facial modality. Since the focus of the research work presented here is the actual cross-cultural multimodal affective corpus, analysis results are not presented exhaustively. Partial results are presented in Sect. 5 to confirm the validity of the corpus construction and annotation approaches presented in Sects. 3 and 4 respectively. During the corpus release to the public domain, according to the plan described in Sect. 6, it will be accompanied with analysis results that will serve as reference point for future research work using the corpus.

The ambition of this research work is that the constructed corpus will be established as a benchmark multimodal cross-cultural affective corpora standard and, hopefully, will provide reference point for future attempts within the affective computing community. At the same time we must bear in mind the rapid development of better and cheaper sensing technology in the near future, which will lead to new standards for the community. We try to take this into account by offering a portable and adjustable experimental setting, which can be repeated in order to collect data not only within a different cultural context, but also using additional sensor devices. This will also be depicted in the guidelines, provisions and best practices accompanying the released corpus. Future studies can build on the presented research work and extend it leading to new insights.

Acknowledgments This work was partially funded by the European Commission under grant agreement eCute (FP7-ICT-2009-5), ILHAIRE (FP7-ICT-2009.8.0) and CEEDS (FP7-ICT-2009-5).

References

1. Abrilian S, Devillers L, Buisine S, Martin JC (2005) EmoTV1: annotation of real-life emotions for the specification of multimodal affective interfaces. In: International proceedings of HCI
2. Amazon Web Services: Public Data Sets (2012) http://aws.amazon.com/publicdatasets/. Accessed 31 Jan 2012
3. Amir N, Weiss A, Hadad R (2009) Is there a dominant channel in perception of emotions? In: 3rd International conference on affective computing and intelligent interaction and workshops, 2009 (ACII 2009). IEEE, Amsterdam, pp 1–6
4. Bänziger T, Pirker H, Scherer K (2006) GEMEP-GEneva multimodal emotion portrayals: a corpus for the study of multimodal emotional expressions. In: The workshop programme corpora for research on emotion and affect, 23 May 2006. Citeseer, p 15
5. Battocchi A, Pianesi F, Goren-Bar D (2005) A first evaluation study of a database of kinetic facial expressions (dafex). In: Proceedings of the 7th international conference on multimodal interfaces (ICMI '05). ACM, New York, pp 214–221
6. Burkhardt F, Paeschke A, Rolfes M, Sendlmeier W, Weiss B (2005) A database of German emotional speech. In: Proceedings of interspeech, Lissabon, pp 1517–1520
7. Busso C, Bulut M, Lee C, Kazemzadeh A, Mower E, Kim S, Chang J, Lee S, Narayanan S (2008) Iemocap: interactive emotional dyadic motion capture database. Lang Resour Eval 42(4):335–359
8. Caridakis G, Raouzaiou A, Bevacqua E, Mancini M, Karpouzis K, Malatesta L, Pelachaud C (2007) Virtual agent multimodal mimicry of humans. Special issue on multimodal corpora. Lang Resour Eval 41(3–4): pp 367–388. Springer, Berlin. http://www.image.ece.ntua.gr/publications.php
9. Caridakis G, Raouzaiou A, Karpouzis K, Kollias S (2006) Synthesizing gesture expressivity based on real sequences. Workshop on multimodal corpora: from multimodal behaviour theories to usable models. In: LREC 2006 conference, Genoa, Italy, 24–26 May 2006. http://www.image.ece.ntua.gr/publications.php
10. Caridakis G, Wagner J, Raouzaiou A, Curto Z, Andre E, Karpouzis K (2010) A multimodal corpus for gesture expressivity analysis. In: Multimodal corpora: advances in capturing, coding and analyzing multimodality, LREC, Malta, 17–23 May 2010
11. Castellano G, Leite I, Pereira A, Martinho C, Paiva A, McOwan P (2010) Inter-act: an affective and contextually rich multimodal video corpus for studying interaction with robots. In: Proceedings of the international conference on multimedia. ACM, New York, pp 1031–1034
12. Cowie R, Douglas-Cowie E, Savvidou S, McMahon E, Sawey M, Schröder M (2000) 'FEELTRACE': an instrument for recording perceived emotion in real time. In: ISCA tutorial and research workshop (ITRW) on speech and emotion, Citeseer
13. Creative Commons: BY-NC-SA 3.0 (2012) http://creativecommons.org/licenses/by-nc-sa/3.0/. Accessed 2 Feb 2012
14. Douglas-Cowie E, Campbell N, Cowie R, Roach P (2003) Emotional speech: towards a new generation of databases. Speech Commun 40(1–2):33–60
15. Douglas-Cowie E, Cowie R, Sneddon I, Cox C, Lowry O, Mcrorie M, Martin J, Devillers L, Abrilian S, Batliner A et al (2007) The

humaine database: addressing the collection and annotation of naturalistic and induced emotional data. In: Affective computing and intelligent interaction, pp 488–500

16. Douglas-Cowie E, Devillers L, Martin JC, Cowie R, Savvidou S, Abrilian S, Cox C (2005) Multimodal databases of everyday emotion: facing up to complexity. In: INTERSPEECH 2005, pp 813–816
17. Velten E (1968) A laboratory task for induction of mood states. Behav Res Ther 6:473–482
18. Ekman P et al (1971) Universals and cultural differences in facial expressions of emotion. University of Nebraska Press, Lincoln
19. Elfenbein H, Beaupré M, Lévesque M, Hess U (2007) Toward a dialect theory: cultural differences in the expression and recognition of posed facial expressions. Emotion 7(1):131
20. Fanelli G, Gall J, Romsdorfer H, Weise T, Van Gool L (2010) A 3-d audio-visual corpus of affective communication. IEEE Trans Multimedia 12(6):591–598
21. Fleiss J, Levin B, Paik M (2003) Statistical methods for rates and proportions. Wiley series in probability and mathematical statistics. Probability and mathematical statistics. Wiley, New York
22. Kessous L, Castellano G, Caridakis G (2009) Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. J Multimodal User Interfaces. doi:10.1007/s12193-009-0025-5. http://www.image.ece.ntua.gr/publications.php
23. Kipp M (2001) Anvil-a generic annotation tool for multimodal dialogue. In: Seventh European conference on speech communication and technology. ISCA
24. Küblbeck C, Ernst A (2006) Face detection and tracking in video sequences using the modified census transformation. Image Vis Comput 24:564–572
25. Leite I, Pereira A, Martinho C, Paiva A (2008) Are emotional robots more fun to play with? In: 17th IEEE international symposium on robot and human interactive communication, 2008, RO-MAN 2008. IEEE, pp 77–82
26. Lingenfelser F, Wagner J, André E (2011) A systematic discussion of fusion techniques for multi-modal affect recognition tasks. In: ICMI, pp 19–26
27. Plutchik R (1994) The psychology and biology of emotion. HarperCollins College Publishers
28. Russell JA (1980) A circumplex model of affect. J Pers Soc Psychol 39:1161–1178. doi:10.1037/h0077714
29. Shami M, Verhelst W (2007) Automatic classification of expressiveness in speech: a multi-corpus study. Speaker classification II, pp 43–56
30. Soleymani M, Lichtenauer J, Pun T, Pantic M (2011) A multi-modal affective database for affect recognition and implicit tagging. IEEE Transactions on Affective Computing, vol 99, p 1
31. Velten E (1998) A laboratory task for induction of mood states. Behav Res Ther 35:72–82
32. Vogt T, André E (2009) Exploring the benefits of discretization of acoustic features for speech emotion recognition. In: Proceedings of 10th conference of the international speech communication association (INTERSPEECH). ISCA, Brighton, UK, pp 328–331
33. Wagner J, Lingenfelser F, André E (2011) The social signal interpretation framework (SSI) for real time signal processing and recognition. In: Proceedings of Interspeech 2011
34. Wagner J, Lingenfelser F, André E, Kim J (2011) Exploring fusion methods for multimodal emotion recognition with missing data. IEEE Transactions on Affective Computing 99(PrePrints)
35. Zara A, Maffiolo V, Martin J, Devillers L (2007) Collection and annotation of a corpus of human-human multimodal interactions: emotion and others anthropomorphic characteristics. In: Affective computing and intelligent interaction, pp 464–475
36. Zeng Z, Pantic M, Roisman G, Huang T (2009) A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans Pattern Anal Mach Intell 31(1):39–58