A multimodal corpus for gesture expressivity analysis

G. Caridakis[1], J. Wagner[2], A. Raouzaiou[1], Z. Curto[3], E. Andre[2], K. Karpouzis[1]

[1]Image, Video and Multimedia Systems Laboratory
National Technical University of Athens
Iroon Polytexneiou 9, 15780 Zografou, Greece
[2]Multimedia Concepts and their Applications Laboratory
Augsburg University
Universitatsstr. 6a, 86159 Augsburg
[3]Humanware S.r.l.
via Garofani 1,Pisa - 56125 Italy
{gcari,araouz,kkarpou}@image.ntua.gr, {johannes.wagner,andre}@informatik.uni-augsburg.de, z.curto@hmw.it

Abstract

This work presents the design and implementation of corpus recording sessions along with some preliminary processing results. Captured modalities include speech and facial expressions but the focus is on hand gesture expressivity. Thus, this is the primary modality and is recorded using three methods: bare hands, Nintendo Wii remote controls and datagloves. Such a setup allows for multimodal affective analysis and potentially provide quantitative parameters for synthesis of systems affectively aware and able to convey affect, such as Embodied Conversational Agents. Additionally, comparative studies of gesture expressivity based on different recording techniques could be based on the introduced corpus. Cross cultural affective behavior issues are also incorporated since the experiment was performed in three countries i.e. Greece, Germany and Italy.

1. Introduction

An abundance of research within the fields of psychology and cognitive science related to the non verbal behavior and communication stress out the importance of qualitative expressive characteristics of body motion, posture, gestures and in general human action during an interaction session (Trenholm and Jensen, 2007). Although such research work studies primarily and mainly context of human to human interaction such approach can be extended to human computer interaction. Some work has incorporated gesture expressivity in HCI context but the vast majority concentrates on the expressively enhanced synthesis of gestures by virtual agents and ECAs (Pelachaud, 2009). Currently, research on the automatic analysis of gesture expressivity is still immature and this fold of human action analysis is asymmetrically studied with reference to the synthesis counterpart.

Recent psychology studies suggest that body language does constitute a significant source of emotional information. Nevertheless, it is hard to identify specific characteristics of body language that could help us assess a user's emotional state. First of all, there is no clear mapping from gestures to emotional states. Secondly, the use of gestures differs from person to person and from situation to situation. How people express bodily emotions depends on a variety of factors including the social context, people's personality and their cultural background. Corpus studies of human behavior may provide useful insights in human behavioral patterns. In the social sciences, corpus studies have a long tradition as a basis to analyze human behavior. In computer science, corpora of human behavior have been employed both for generation and analysis tasks. First, a number of attempts have been made to extract human behavioral patterns from corpora to guide the design of virtual agents. Secondly, corpora have played an important role in recognition tasks where classifiers have been trained based on labeled data of human behavior. The acquisition of corpora enables us to ground research on gesture generation and recognition in empirical data. Unfortunately, corpora with annotated communicative gestures expressing emotional content are rare. In particular, there are no corpora available that enable us to investigate culture-specific aspects of gestural emotions. As a consequence, we decided to collect our own corpus.

The objective of our work is threefold: First of all, we aim at studying the use of emotional gestures in combination with other modalities, such as facial expressions and speech. As a consequence, data need to be collected in a synchronized manner. Secondly, we investigate how bodily emotions are expressed in different cultures. As a first step, we focus on three European countries: Greece, Italy and Germany. Thirdly, we are interested in finding out how the use of interaction devices influences people's gestures and the robustness of the recognition process. In particular, we focus on: video-based gesture recognition, Wiimote-based gesture recognition and gesture recognition based on a data glove. All these three technologies come with their own advantages and disadvantages. Computer vision is the less obtrusive means to capture information about the user's body movements. However, it is rather sensitive against lighting conditions. Gesture recognition using a data glove or the Wiimote does not suffer from these problems. However, it is much more obtrusive. In particular, using a Wiimote for performing affective gestures is rather unnatural since users have to carry a device in their hands which might influence their way of gesturing.

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 a 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 are considered ideal for validating real life affective analysis systems, although such behavior is relatively rare,short lived,and filled with subtle context-based changes. On the other hand there is a large number of issues and internal processes that influence the final result involved in the affective elicitation methods. Additionally, the selection of the recorded modalities incorporates intrusion issues and/or content based, modality related issues. 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.

Besides the implications reported above the necessity for creating reusable databases consisting affectively enriched human behavior has resulted in a number of attempts for creating multimodal corpora. The importance of each corpus is determined both 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 (Douglas-Cowie et al., 2003) was constructed by the Queen's University of Belfast and mainly consists of sedentary interactions, from chat shows, religious programs and discussions between old acquaintances. It consists of 125 English speaking subjects experiencing a wide range of positive and negative emotions and of emotional intensities. The FeelTrace (Cowie et al., 2000) tool was used for labeling the corpus recording the perceived emotional state via dimensional rating. The EmoTV corpus (Abrilian et al., 2005) 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 (Kipp, 2001) 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 (Kessous et al., 2009) was also performed within the HUMAINE EU-IST project framework during its Third Summer School held in Genova 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 (Bänziger et al., 2006). Ten participants participated in the recordings representing 5 nationalities, incorporating cross cultural issues, and data on facial expressions, body movement, gestures and speech were simultaneously recorded. 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). The GEMEP (Geneva Multimodal Emotion Portrayals) corpus (Bänziger et al., 2006) constitutes a repository of portrayed emotional expressions. The researchers argue that portrayals produced by appropriately instructed actors are analogue to expressions that do occur in selected real life contexts as opposed to induced or real-life sampled emotional expressions that display expressive variability and therefore constitute excellent material for the systematic study of nonverbal communication of emotions. Ten professional French-speaking actors portrayed 15 affective states under the direction of a professional stage director, recording audio, facial expressions and head orientations, body postures and gestures from two viewpoints (perspective of an interlocutor and sideways).

This corpus construction innovation lies in its focus on gesture expressivity, the inclusion of multiple cultures and multiple human behavior capturing techniques, i.e. video, Wiimote and Datagloves. The introduced multicultural corpus allows for intercultural affective analysis, while the variety of technologies used to record human body behavior supports studies on their obtrusiveness effect. So, the motivation for this experiment is threefold. 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

3.1. Affective immersion

The adopted emotion elicitation method was inspired by the Velten mood induction technique (Velten, 1998) where people had to read aloud a number of sentences that put them in particular emotional state. First, we displayed a sentence with a clear emotional message and gave the user sufficient time to read it silently. Then the projection turned blank and the user was asked to express the according emotion through gesture and speech. The users were encouraged to use their own words as long as they helped them feel a particular emotion. The sentences were shown in three coherent blocks with first positive, then neutral and finally negative sentences in order to put the users gradually into the desired mood. We selected in total 120 sentences (40 for each target class) such as:

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

We decided to choose the order positive-neutral-negative in order not have 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 blocks is again divided into three sections, during which we equip the user with different interaction tools. During the first 20 sentences subjects are wearing a data glove by HumanWare. The next 10 sentences the glove is exchanged by two Wii remote controls, which the users hold in their hands. Finally, the remaining 10 sentences were performed with free hands.

3.2. Hardware setup

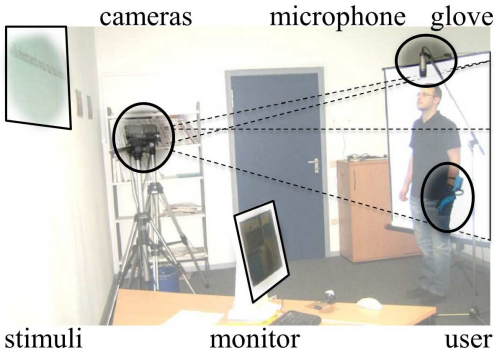

Figure 1: Picture of the experimental setting in which the capture devices are highlighted. The monitor in the front is used by the experimenter to overview the recording and run the stimuli script.

During the recordings the user stands in front of a neutral background. The stimuli, i. e. the Velten sentences, is projected 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 (720x576 pixels, 25 fps, 24 bit colour depth) are placed. The first camera is setup 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, 16kHz, 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 50Hz: 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.

3.3. Recording Software

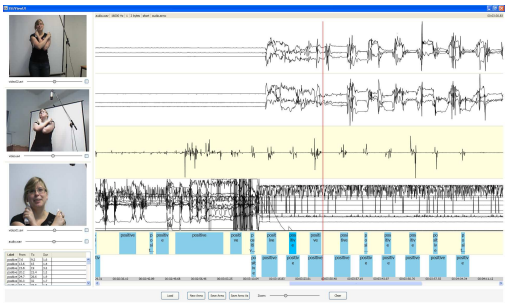

Figure 2: Reviewing the recordings using the SSI viewer tool features multiple annotation layers and simultaneous playback of modalities.

During the analysis of the corpus it is our purpose to not only look at modalities individually, but also investigate the relations between the different channels and explore ways to fuse the different kind of information. This, however, requires a proper synchronization between the modalities. To obtain synchronized recordings we use Smart Sensor Integration (SSI), a software framework for multi-modal signal processing in real-time, developed at the University of Augsburg (Wagner et al., 2009). In SSI synchronization is achieved by constantly updating the incoming signal streams to a global clock. As SSI already supports common sensor devices including webcam/camcorder, microphone and Wii remote control, integration of the modalities is straight-forward. For the data glove, which is not supported by default, we fall back to the possibility to capture a signal from a socket stream. The accomplishment of the experiment is supported by a graphical user interface, which allows the experimenter to display a sequence of HTML documents containing the according stimuli sentences.
Recordings from all devices are synchronized throughout the whole session and directly stored to disk. To avoid artefacts no compression is used on the high-quality video streams, which leads to a data transfer of $\approx$ 30 MB/s per video. For this reason the two videos are recorded on a separate pc, which is synchronized by a broadcast message sent from the main machine. This way it becomes possible to distribute recordings on any number of machines. During a recording SSI is used to detect parts of the signals with high activity, e.g. when a user is talking or performing a gesture. Based on these events preliminary annotations are generated for each signal. Signals and annotations can be reviewed and adjusted using a graphical tool shown in Figure 2. This way, SSI not only supports the recording, but also the analysis of the corpus.

3.4. Procedure

Apart from the setup of the equipment, participants were necessary for our experiment. Each partner was responsible to recruit local subjects and run the experiment at his lab. An according translation of the same set of Velten sentences

was used. 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 was varying between 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 well distributed. In case of the German experiment, subjects were given an allowance of 20 €. So far, 15h:14m:24s of interaction has been recorded in the three countries (which includes parts when users were changing devices).

Subjects training 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 possible 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. The issues covered in this form are described in Section 3.4.. During the experiment we presented to every participant what she should do and she could re-watch the video or be present while another person was following the procedure. We were at her disposal to settle any queries and two persons were constantly present during the experiment to guarantee its success and to help the participants using the Wiimote, wearing the glove etc.

Ethical issues As already mentioned, the participants should sign a consent form which ensures 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. The data is stored, so they should have in mind that, although the samples collected are anonymous, the voice or the face of a subject might be recognized. The consent form gives them the right to ask for erasure or blocking of the data that concerns her/him and to withdraw from the experiment at any time.

4. Corpus preliminary analysis

4.1. Speech

To analyse the expressivity in speech we use EmoVoice [1], a tool for real-time recognition of emotions from acoustic properties of speech (Vogt et al., 2008). The acoustic features used by EmoVoice are mainly based on short-term acoustic observations, including pitch, signal energy, Mel-frequency cepstral coefficients, spectral and voicing information, and the harmonics-to-noise ratio. Overall, a set of 1316 features is obtained from each speech segment[2]. For evaluation purpose a Naïve Bayes classifier is trained and tested using on two-fold cross-validation.

So far, the German and Italian sentences have been proceeded and some of the results are listed in Table 2. The quoted recognition rates were obtained by taking the sentences of all subjects: first separately for each country and then combined. In addition, the last two columns list the worst and best performance of a speaker-dependent classification. In both cases, a difference of more than 20% proves the high variability between users. Compared to that the drop from 53.83% (German) and 51.16% (Italian), respectively, to 48.91% (combined) appears rather small. This, for instance, suggests that differences between individuals are actually more relevant than differences between the two cultures.

nation	positive	neutral	negative	average	min	max
GER	36.36%	67.11%	58.01%	53.83%	54.17%	75.83%
IT	36.39%	60.83%	56.25%	51.16%	49.17%	74.17%
BOTH	31.81%	53.60%	61.31%	48.91%	-	-

Table 2: EmoVoice classification results

4.2. Gestures

4.2.1. Gesture expressivity features

Behavior expressiveness is an integral part of the communication process since it can provide information on the current emotional state and the personality of the interlocutor (Mehrabian, 2007). Many researchers have studied characteristics of human movement and coded them in binary categories such as slow/fast, restricted/wide, weak/strong, small/big, unpleasant/pleasant in order to properly model expressivity. We adapted six expressivity dimensions described in (Hartmann et al., 2005), as the most complete approach to expressivity modeling, since it covers the entire spectrum of expressivity parameters related to emotion and affect. Derived from the field of expressivity synthesis five parameters have been defined, consisting a subset of the six expressivity dimensions: Overall activation, Spatial extent, Temporal, Fluidity and Power.

Overall activation is considered as the quantity of movement during a dialogic discourse and, given the above definitions, is formally defined as the sum instantaneous quantities of motion of the two hands. Spatial extent is expressed with the expansion or the condensation of the used space in front of the user (gesturing space). In order to provide a strict definition of this expressivity feature spatial extent is considered as the maximum value of the instantaneous spatial extent during a gesture. The temporal expressivity parameter denotes the speed of hand movement during a gesture and dissociates fast from slow gestures. On the other hand, the energy expressivity parameter refers to the movement of the hands at during the stroke phase of the gesture. This parameter is associated 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.

4.2.2. Bare hands

Regarding the hand and head detection and tracking problem which is a required step for extracting expressivity features from a gesture several approaches have been reviewed. Amongst them only video based methods were considered since motion capture or other intrusive techniques would interfere with the person's emotional state which is a crucial issue in this kind of analysis. The major factors taken under consideration are computational cost and robustness, resulting in an accurate near real-time skin

[1] https://mm-werkstatt.informatik.uni-augsburg.de/EmoVoice.html

[2] Here: the utterance of one Velten sentence

detection and tracking module. The overall process is described in detail in (Caridakis et al., 2007) and includes creation of moving skin masks and tracking the centroid of these skin masks among the subsequent frames of the video depicting a gesture. The described algorithm is lightweight, allowing real time processing and indicative results and intermediate steps can be seen at 3.

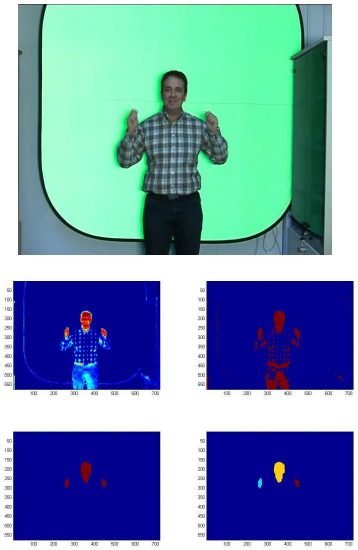

Figure 3: Image processing intermediate steps and results

Extraction of the expressivity features based on the coordinates of the hands, as detected using the image processing described above, is actually a process based on calculations over the motion vectors resulting from the coordinates. Overall activation is computed as the sum of the motion vectors' norm, Spatial extent is modeled by the maximum Euclidean distance of the position of the two hands and Fluidity is the variance of the norms of the motion vectors. Power is actually identical to the first derivative of the motion vectors norms, calculated previously.

4.2.3. Glove

Regarding finger movement analysis, our approach is based on the extraction of some features (or expressivity dimensions) very similar to the ones seen before. In particular we have extracted: Overall activation, Power, Spatial extent and Fluidity.

The dataglove measures the angles of finger joints. The angles are then normalized according to a multi-user transformation function, accessible through a calibration procedure. We can define the finger's motion energy as a sum of the fingers joint angular velocities and use this value as an overall activation feature. The power of one gesture is approximated by the sum of the root mean square of all the finger joint angular accelerations while a measure of fluidity of the gesture can be determined using motion jerk, i.e. the derivative of acceleration. Finally, the spatial extent of the movement is given by the maximum distance a finger joint has moved during the sentence.

5. Conclusions

The work presented here discusses issues related to the design and implementation of an experiment, resulting in a multimodal corpus of affective behavior, incorporating acoustic prosody, facial expressions and gesture expressivity, and briefly introduces the methods that will be used in the future to process the resultant corpus. During corpus affective analysis, that is considered ongoing and future work, significant conclusions are expected to be drawn, especially for the analysis of gesture expressivity, its correlation with other modalities and related cross cultural and interaction obtrusiveness issues. The ambition of this research work is that the constructed multimodal corpus, once synchronized and formatted, will be established as a benchmark multimodal corpora standard focused on gesture expressivity. Feature extraction, multimodal analysis and synchronization and fusion techniques from the involved research teams will be applied to the corpus and, hopefully, will provide reference point for future attempts within the affective computing community.

6. Acknowledgements

This work has been funded by the FP6 IP Callas (Conveying Affectiveness in Leading edge Living Adaptive Systems), Contract Number IST-34800.

7. References

S. Abrilian, L. Devillers, S. Buisine, and J.C. Martin. 2005. Emotv1: Annotation of real-life emotions for the specification of multimodal affective interfaces. In *HCI International*.

T. Bänziger, H. Pirker, and K.R. Scherer. 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 Tuesday 23 rd May 2006*, page 15. Citeseer.

G. Caridakis, A. Raouzaiou, E. Bevacqua, M. Mancini, K. Karpouzis, L. Malatesta, and C. Pelachaud. 2007. Virtual agent multimodal mimicry of humans. *Language Resources and Evaluation 41 (3-4), Special issue on Multimodal Corpora, pp. 367-388, Springer*.

R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder. 2000. 'FEELTRACE': An instrument for recording perceived emotion in real time. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*. Citeseer.

E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach. 2003. Emotional speech: Towards a new generation of databases. *Speech Communication*, 40(1-2):33–60.

B. Hartmann, M. Mancini, S. Buisine, and C. Pelachaud. 2005. Design and evaluation of expressive gesture synthesis for embodied conversational agents. In *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, page 1096. ACM.

L. Kessous, G. Castellano, and G. Caridakis. 2009. Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. *Journal on Multimodal User Interfaces, Springer, DOI 10.1007/s12193-009-0025-5*.

M. Kipp. 2001. Anvil-a generic annotation tool for multimodal dialogue. In *Seventh European Conference on Speech Communication and Technology*. ISCA.

A. Mehrabian. 2007. *Nonverbal communication*. Aldine.

C. Pelachaud. 2009. Studies on gesture expressivity for a virtual agent. *Speech Communication*, 51(7):630–639.

S. Trenholm and A. Jensen. 2007. *Interpersonal communication*. Oxford University Press, USA.

E. Velten. 1998. A laboratory task for induction of mood states. *Behaviour Research and Therapy*, 35:72–82.

T. Vogt, E. André, and N. Bee. 2008. Emovoice - a framework for online recognition of emotions from voice. In *Proceedings of Workshop on Perception and Interactive Technologies for Speech-Based Systems*, Kloster Irsee, Germany, June. Springer.

J. Wagner, E. André, and F. Jung. 2009. Smart sensor integration: A framework for multimodal emotion recognition in real-time. In *Affective Computing and Intelligent Interaction (ACII 2009)*.