RGB-D Database for Affective Multimodal Human-Robot Interaction

C. Doblado, E. Mogena, F. Cid, L. V. Calderita and P. Núñez

***Abstract*—Affective Human-Robot Interaction is an interesting topic for social robotics. The ability to recognize human affective states is the core of most of these interaction systems. It allows robots to adapt their actions and reactions according to previously learned intentions and affective behaviors. Thus, databases containing representative samples of human affective behavior are needed. In this paper, a multimodal RGB-D database for affective Human-Robot Interaction is presented. It consists of 20 subjects with different forced and spontaneous emotional states, such as happiness, sadness, anger, fear and neutral. Body gestures and facial expressiveness have been recorded using a Microsoft sensor. The proposed database is used as a first step in a feature detection and extraction algorithm based on the body language analysis, which can be used in a multimodal emotion recognition system.**

***Index Terms*—Social Robotics, Emotion Recognition, Affective Human-Robot-Interaction.**

I. Introduction

In the last decades, affective human robot interaction (HRI) has become one of the main research areas in social robotics. The goal of current HRI systems is to achieve natural communication between humans and robots, in the same way that humans interact with each other in their daily life. In order to interact with humans, a robotic system should be able not only to understand user's behavior and intentions, but also to estimate their affective state. Determining human emotions helps robots in adapting the communication in real time, improving and enriching the interaction [1].

Natural human-human interaction takes advantage of many communication channels, which can be categorized as either verbal or nonverbal. To express emotions, humans speak, point, gesture, use facial expressions, head motion, and eye contact (see Fig. 1). Thus, several input channels or *modes*, such as speech, body language, gestures or facial expressiveness have been typically used for the current state-of-art multimodal emotion recognition systems [2]. Facial expression is the essential means of transmitting emotions. However, there are other important channels, such as body gestures, that plays an important role in affective state recognition. During a speech, for instance, body language accounts for more than half of the message that is communicated to the listeners.

Though there are many databases that can be used for emotion recognition (see surveys [12], [6], [7]), there are not that many public databases including both 3D body gestures and faces. In one hand, this paper purposes such a database, for which a Microsoft Kinect RGB-D sensor is used. Kinect is capable of providing simultaneously color and depth information at real-time. Due to its low-cost and availability, this RGB-D camera has been extremely popular in the last years. Besides, Kinect sensor is optimized to work from a close range as 0.5m to 3m, which allows to use it for typical HRI. On the other hand, the analysis of body gestures for affective HRI using 3D data is not extensive in the literature. Thus, two are the *major contributions* of this paper: i) to develop a database using both, color and depth image, for emotion recognition; and ii) to define a set of invariant features from body language for using in multimodal emotion recognition system.

C. Doblado, E. Mogena, F. Cid, L.V. Calderita and P. Núñez are members of Robotics and Artificial Vision Lab. *Robolab* Group, University of Extremadura, Spain.
E-mail: cdoblado@alumnos.unex.es; emogenac@alumnos.unex.es; fecidb@alumnos.unex.es; pnuntru@unex.es

		Social Behaviour	Technology		
		emotion	speech anlysis	computer vision	biometry
Gesture and posture	hand gestures	√		√	√
	posture	√		√	√
	walking			√	√
Face and eyes behaviour	facial expressions	√		√	√
	gaze behaviour	√		√	
	focus of attention	√		√	
Vocal behaviour	prosody	√	√		
	turn taking	√	√		
	vocal outbursts	√	√		
	silence	√	√		
Space and Environment	distance	√		√	

Fig. 1. Behavior cues associated to emotion as well as the technologies involved in their automatic detection.

The rest of this paper is organized as follows: the current state-of-art in emotion databases and body languages features for affective HRI are shown in Section II. Section III presents a description of the proposed database for emotion recognition based on body gestures. In Section IV, a description of the set of features extracted from the body gesture analysis is given. Finally, Section V describes the conclusions and future lines of this work.

II. Related Works

This paper describes initial attempts to collect a multimodal affective RGB-D database, which can be used for emotion recognition. In the field of affective HRI, different automatic emotion recognition systems have been studied in the literature. Most of these works are based on the analysis of only one mode or channel of information, such as video

sequences (facial expresiveness, for instance) or audio signals. Independent of the nature of the information source, the raw data is processed and a set of features is extracted. After that, this set of features allows to classify the input data into an affective state which are the basis of the proposed approach.

On one hand, facial expressions have been usually exploited to detect and recognize human emotions. An interesting and updated review was shown in [5]. This work shows the evolution of the research, from 2D to 3D approaches. Commonly, these frameworks use the Facial Action Coding System (FACS) proposed by Ekman et al.'s [10], which is based on facial muscle deformations. On the other hand, in the last years, body movements and gestures have been already used for multimodal emotion recognition [11]. In this work, the authors successfully introduce new features for 2D video sequence analysis.

Developing robust affective multimodal interaction systems requires databases containing representative samples of human expressive behavior. Unfortunately most of the available affective databases are only focused on emotion recognition from facial expressiveness or speech (interesting surveys are found in [12], [6] and [7], respectively). Multimodal databases for emotion recognition have been appearing in the last years. In [8], a multimodal database is described, which consists of audio-visual and gesture information of 125 subjects. A database of spontaneous emotions based on audio-visual emotion is also presented in [9].

Currently, most of the facial expressiveness or body gesture databases are based on 2D RGB frames (static or dynamic images), but very little has been made using 3D data. RGBD sensors have been recently used for several applications. In the field of HRI, a face database using Microsoft Kinect sensor is described [13]. It consists of 1581 color and depth images taken from 31 persons in 17 different poses and facial expressions. In [14], the authors present a RGBD database of 15 persons containing synchronized color-depth video streams for the task of human daily activity recognition. Contrary to these previous approaches, an affective RGBD database containing facial expresiveness and body gestures is presented in this paper. This database is explained with details in the following sections. It consists of different subjects' emotional state.

Finally, body gestures features have been used for emotion recognition. In [11], a set of descriptors is defined. However these descriptors are extracted using RGB images. Similar features are presented in [3], also using 2D information. This paper extends these descriptors by taking RGB-D information from the Microsoft Kinect sensor.

III. Multimodal Affective Database

In this paper, a RGB-D database for affective multimodal HRI is presented. Currently, the most of emotion recognition system is based on facial expressiveness. A set of 20 subjects have been selected and recorded during different sessions. From each person, four different emotions (*i.e.,* anger, fear, happiness and sadness) plus a neutral affective state have been studied in two cases: acted and spontaneous. From the RGB-D sequence, both, facial expressiveness and body gestures, are analyzed in real-time for extracting features that can be used for emotion recognition. An overview of the proposed methodology is shown in Fig. 2.

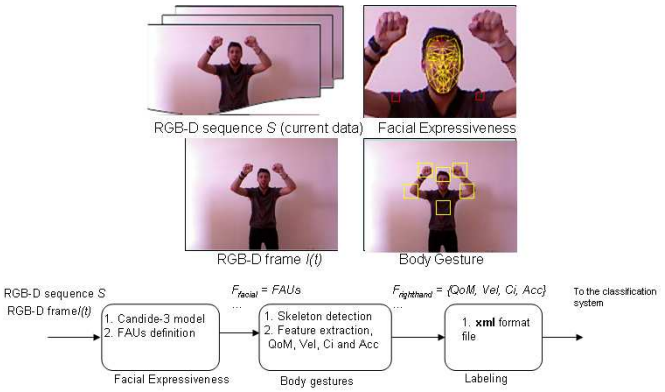

Fig. 2. Overview of the proposed methodology. Both, facial expressiveness and body gestures are recorded from each subject.

TABLE I
PrimeSense RGB-D sensor specification

Parameter	Microsoft Kinect sensor
Dimension (mm)	14cm x 3.5cm x 5cm
Field of View (Horizontal, Vertical, Diagonal)	58^o H, 45^o V, 70^o D
Depth image size	VGA (640x480)
Spatial x/y resolution (@ 2m distance from sensor)	3mm
Operation range	0.8m - 3.5m
Operation environment (every lighting condition)	Indoor

A. RGB-D sensor

The PrimeSense RGB-D camera provides two different images: a depth image and a color image. Depth image contains a pixel matrix, where each pixel contains a value representing the distance of the object covered by this specific pixel from the sensor. On the other hand, the color image is a standard output of a 2D digital camera. Both, depth and color images, are able to be combined producing a real-time 30fps stream of VGA frames. Table I describes some of the major characteristics of the sensor given by the manufacturer.

Fig. 3a illustrates the 12 Degree of Freedom (DOF) IADEX[1] Muecas robot head. This expressive robotic head is mounted on Loki, a social robot used for interacting with humans taking into account their affective states. Kinect sensor is located on top of the head. In Fig. 3b the RGB image acquired by the robot is drawn. Fig. 3c shows the depth image. In order to efficiently use the RGB-D sensor for HRI, an extrinsic calibration method based on Burrus' work [17], has been achieved. Fig. 3d shows the effect of the correction after applying the calibration method for the images drawn in Fig. 3b-c (red color represents those points in the RGB image without depth information available). .

[1] www.iadex.es

Fig. 3. a) PrimeSense sensor (kinect style) mounted on the Muecas robot head; b) RGB image acquired by the RGB-D sensor; c) depth information associated to the RGB image shown in b); and d) RGB image after callibrating.

B. Affective HRI Scenario

The working scenario is inside RoboHome, a living lab located at University of Extremadura. This research lab consists of a seventy square meters standard apartment with two rooms, a hall, a bathroom and standard furniture. Fig. 4 illustrates a three-dimensional representation of RoboHome. Subject and robotic platform's poses have been labelled in the figure. Besides, a more detailed description of the recording scenario has been shown on the right. As is drawn in Fig. 4, RGB-D sensor is located at about 2 meters from the subject. The studio is a private place which offers the intimate atmosphere that is needed for the subjects to show their emotions in a natural way.

Two different recording sessions have been achieved for each subject. In the first one, the person was subjected to a set of experiences in order to get spontaneous emotions. In the second one, the subject was trained by actors to reproduce pre-defined movements for each emotion. These learned movements are motivated by Stock et al.'s work [15]. Each emotion is recorded in separate files.

C. Subjects

The database is composed of 20 volunteers. All of them are engineering students (10 women and 10 men) in their twenties. In this age range, variation of skeletal structure and height is less pronounced, and therefore useful for evaluating the body gesture features. The height of the volunteers varies from about 1.60 to 1.80 meters.

D. Set of emotions

Based on Ekman's work [16], who described emotions as discrete, measurable and physiologically distinct, most of the emotion recognition approaches are based on a set of six basic emotions: anger, disgust, fear, happiness, sadness and surprise. Currently, some psycological theories describe the emotions as a continuum, into two dimensions known as valences (how negative or positive the experience is) and arousal (extent of reaction to stimuli). The database described in this paper contains five different basic emotions, that are a subset of Ekman's basic emotions: anger, fear, happiness, sadness and a neutral state. Of course, these basic emotions can be used in the space of valence-arousal.

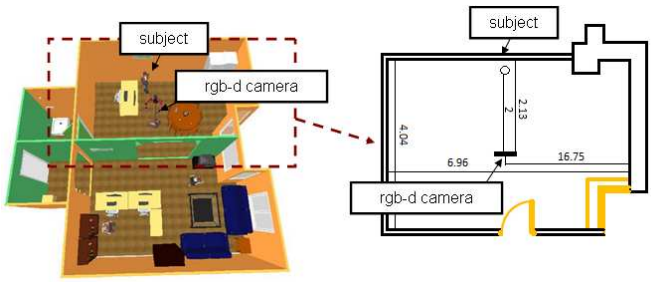

Fig. 4. RoboHome living lab layout (University of Extremadura). A more detailed description of the recording scenario is illustrated on the right. Both, subject and robot's poses are labelled.

Fig. 5 shows the basic movements of each forced emotion. As illustrated, *Sad* its acted by a small slow movement with the arms in the lower position and the shoulder's position a little bit under his typical situation. *Neutral* is a fluid movement at medium-low high. On the other hand, *Fear* is a quick action of covering the own head as protecting from something external. *Angry* is an energetic medium-high movement, sometimes with a forward development of the hands. Finally, *Happy* is an action of raising the arms beyond the head and move them lively.

E. Database organization

Files and folders follows the structure drawn in Fig. 6. As it was afore-mentioned, these five basic emotions are captured in two ways: forced and spontaneous. Thus, the final database consists of 20 folders, one for each subject. These folders contain other two subfolders, for both, forced and spontaneous recording sessions, respectively. Finally, emotions are saved in five *.oni* files[2].

In this same folder, a *.xml* file is included. This file has a description of the emotion (*i.e.,* category, start [ms] and end [ms]), and also the body gestures and facial features associated to each emotion (*i.e.,* QoM, Vel_{left}, Vel_{right}, $Prox$, H_{left} and H_{right}). A set of elements and attributes has been written into the file in order to annotate these features and other information of interest for future researchers. This annotation is detailed in Fig. 7. Finally, RGB-D images are acquired at 33 frames per second, with the file size of about 270Mb per minute (640x480, uncompressed).

[2]The Kinect OpenNI library uses .oni as a custom video file format to store videos that contain RGB-D information.

Fig. 5. Body gestures associated to each basic emotion: sad, neutral, fear, angry and happy, respectively.

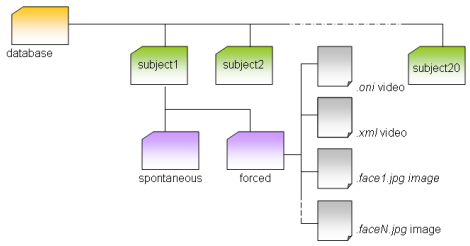

Fig. 6. Files and folder structure for the database proposed in this work. For more information, visit *www.robolab.unex.es*.

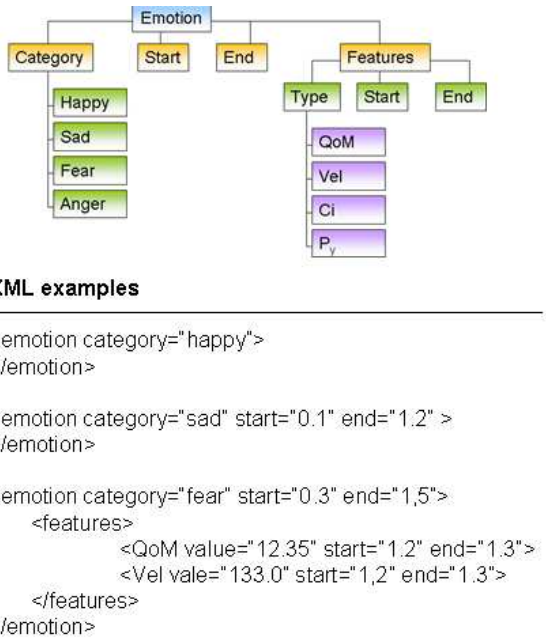

Fig. 7. Description of the *.xml* file included in the folder. This file allows to annotate the most important features of the recorded emotional state.

IV. Feature Detection and Extraction for Automatic Emotion Recognition

This section introduces the methods for detecting and extracting features from the subjects' movements. In order to choose the set of features has been taken into account its use to recognize emotions. Both, facial expressiveness and body gestures have been analysed as is described in the following subsections.

A. Facial Extraction for Emotion Recognition

The proposed methodology consists of a robust feature extraction algorithm, which uses the Candide-3 reconstruction model described in [18]. Facial features are extracted from the deformation of the mesh model and are directly related to the Action Units (AUs) described in FACS [10]. Fig. 8b illustrates the mesh model over the face drawn in Fig. 8a. An example of the features extraction is shown in Fig. 8c. Finally, in Fig. 8d, the set of AUs is illustrated.

B. Body Feature Extraction for Emotion Recognition

In this section, the body feature extraction algorithm is described. The proposed method consists of tracking the human skeleton in real-time. The 3D body mapping and

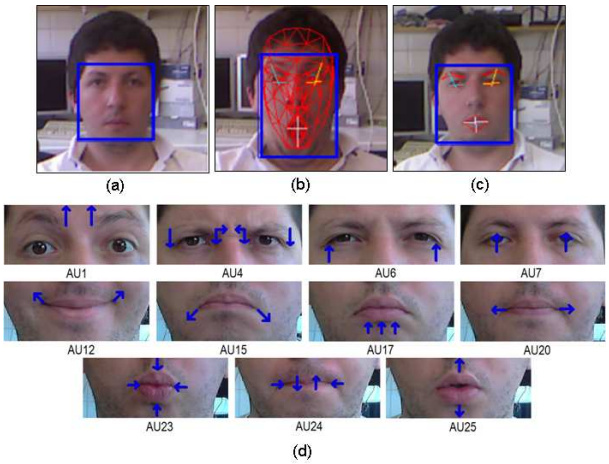

Fig. 8. a) Image of the face to be processed; b) Candide-3 reconstruction model over the face shown in a); c) An example of features extracted from b); and d) The set of AUs extracted from the face image.

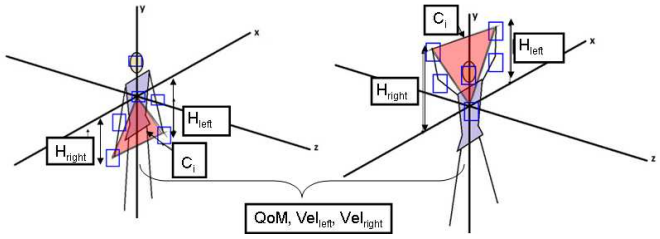

Fig. 9. Graphical representation of the body gestures features proposed in this paper.

skeletal tracking is conducted using the Kinect sensor and the algorithm included in OpenNI library [4]. Given the human skeleton, the 3D positions of hands, elbows, chest and head are extracted at consecutive instants of time. Three seven expressive motion cues are extracted in the Cartesian coordinate system: velocities and positions of the left and right hands, Quantity of Motion, Proximity and Contraction Index. A graphical representation of these features is illustrated in Fig. 9. Let N be the number of consecutive frames, these features are extracted as follows:

- Velocity (Vel) value is related to the trajectory followed by the human body. To extract the velocity, the current and the N previous positions are considered to give a more accurate data. Let x_i^{lh} be the 3D positions of the left hand at an instant of time i. Then, the velocity at this instant of time i is given by:

$$Vel_i^{lh} = \frac{1}{N} \cdot \sum_{k=0}^{N-1} \frac{(x_k^{lh} - x_{k-1}^{lh})}{t} \tag{1}$$

  where t is the data acquisition time of the RGBD sensor. This same equation is used to calculate the vel_i^{rh} value using the 3D positions of the right hand x_i^{rh}.
- Normalized height of the hands is also related to the 3D positions of the left and right hands, x_i^{lh} and x_i^{rh} respectively. These values, H_{lh} and H_{rh} are first calculated only using the y component of the position vector. These values are normalized to the y component of the chest position.
- Contraction index (C_i) is a value for the contraction degree of the body focusing in the relationship of the chest and the hands position. This relationship has been carried out by the area of the triangle defined by the x coordinate of these three points. First, the semiperimeter, s of the triangle is calculated:

$$s = \frac{u + v + w}{2} \tag{2}$$

  where u, v and w are the sides of the triangle. The Contraction Index, using the Heron's Method, remains as follows:

$$C_i = \sqrt{s \cdot (s-u) \cdot (s-v) \cdot (s-w)} \tag{3}$$

- Proximity (P_z) refers to the direction of the chest motion. This value is normalized to the z coordinate of the chest position so when P_z is positive the chest is approaching to the camera.
- Quantity of Motion (QoM) is related to the amount of detected motion. In this paper, QoM is calculated by using the 3D position of the body upper joints detected from the skeleton (*i.e.,* left and right hands and elbows, chest and head). Let x_i^A be the 3D position of joint at an instant of time i, Then, QoM^A is defined as:

$$QoM_i^A = \frac{1}{N} \cdot \sum_{k=0}^{N} x_i^A - x_{i-1}^A \tag{4}$$

  where $A \in$ (*left hand, right hand, left elbow, right elbow, chest, head)*. Finally, the total QoM is evaluated as the average value of QoM^A.

In the Fig. 10 and 11, is illustrated the evolution of the features descriptor vector during the four emotional states used in this work. In Fig. 10 happiness (red) and sadness (blue) are illustrated. As shown in the figure, some features (such as the velocities or heights of the hands) are distinct, which can be used for emotion reconition system from body gesture in order to improve the results. Similar results are shown in Fig. 11, where features associated to the fear (green) and anger (orange) emotional estates are illustrated.

V. Conclusion

In the Human Robot Interaction literature, the most recent works in emotion recognition are based on multimodal approaches. Multimodal emotion recognition systems use more than one input channel (*i.e., mode*) to detect emotions. These approaches combine usually facial expressiveness, speech and body gestures. On one hand, this paper describes a multimodal RGB-D database for affective Human-Robot Interaction. Both, the facial expressiveness and the body gestures associated to different emotions are included in the database. Based on Ekman's work [16], a set of basic emotions has been recorded for 20 different subjects (happiness, sadness, anger, fear and neutral).

On the other hand, a feature description vector is calculated in order to characterize body gestures and facial expressiveness associated to the emotions. The proposed database can be used

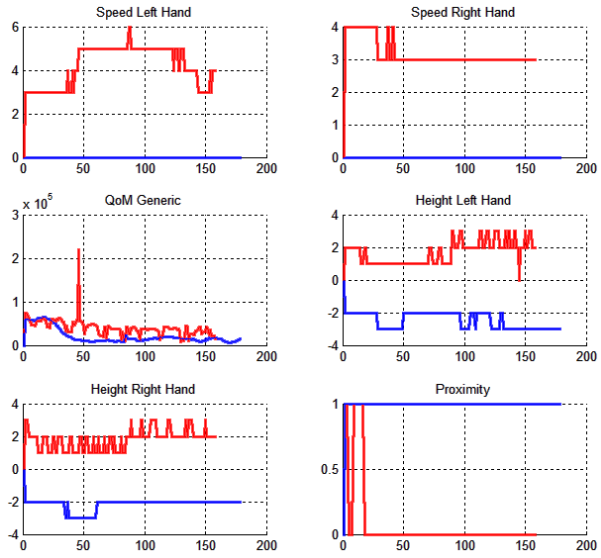

Fig. 10. Evolution of the features descriptor vector during two emotional state: happiness (red) and sadness (blue).

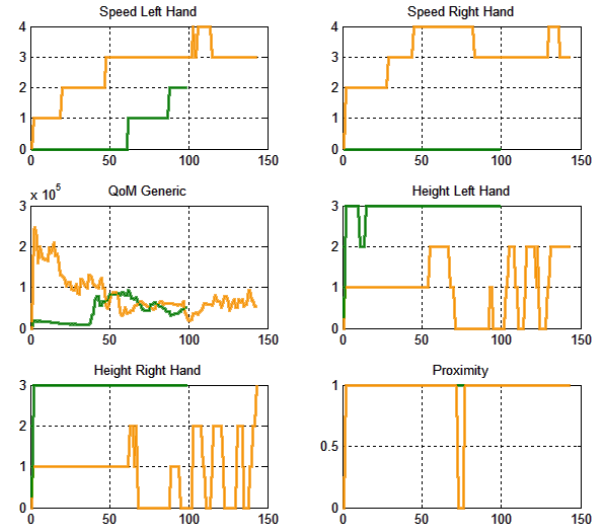

Fig. 11. Evolution of the features descriptor vector during two emotional state: fear (green) and anger (orange).

as a first step in a feature detection and extraction algorithm based on the facial expressiveness and the body language analysis. In the future, these feature description vectors will be used in a multi-modal emotion recognition system.

ACKNOWLEDGMENT

This work has been partially supported by the MICINN Project TIN2012-38079-C03-01.

REFERENCES

[1] R. W. Picard, "Affective Computing". MIT Press, pp. 88-91, 2000.
[2] N. Sebe, I. Cohen, T. Gevers, T. S. Huang, "Multimodal Approaches for Emotion Recognition: A Survey", In *Internet Imaging VI, SPIE'05*, USA, 2005.
[3] M. Mancas, D. Glowinski, G. Volpe, P. Coletta and A. Camurri. "Gesture Saliency: a Context-aware Analysis". In *Gesture in Embodied Communication and Human-Computer Interaction*, Lecture Notes in Computer Science, Vol. 5934, pp. 146-157, 2010.
[4] OPENNI - In URL: *http://www.openni.org/*.
[5] V. Bettadapura, "Face Expression Recognition and Analysis: The State of the Art", Tech Report, College of Computing, Georgia Institute of Technology, 2012.
[6] D. Ververidis and C. Kotropoulos. "A Review of Emotional Speech Databases", In *Proceedings 9th Panhellenic Conference on informatics (PCI)*, pp. 560-574, 2003.
[7] S. Ramakrishnan. "Recognition of Emotion from Speech: A Review", In Speech Enhancement, Modeling and Recognition- Algorithms and Applications, InTech 2012.
[8] E. Douglas-cowie, R. Cowie, I. Sneddon, C. Cox , O. Lowry, M. Mcrorie , J-C. Martin , L. Devillers, S. Abrilian, A. Batliner, N. Amir and K. Karpouzis. "The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data", In *Proceedings ACII '07 Proceedings of the 2nd international conference on Affective Computing and Intelligent Interaction*, pp. 488-500, 2010.
[9] R. Gajsek, V. Struc, B. Vesnicer, A. Podlesek, L. Komidar, F. Mihelic. "Analysis and Assessment of AvID: Multi-Modal Emotional Database", In Text, Speech and Dialogue, Lecture Notes in Computer Science Volume 5729, pp. 266-273, 2009.
[10] P. Ekman, WV. Friesen, JC. Hager. "Facial Action Coding System FACS", The manual, 2002.
[11] L. Kessous, G. Castellano, G. Caridakis. "Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis", In Mar. *Journal on Multimodal User Interfaces*, Vol. 3, No. 1., pp. 33-48, 2010.
[12] H. Gunes and M. Piccardi, "Creating and Annotating Affect Databases from Face and Body Display: A Contemporary Survey", In textitIEEE International Conference on Systems, Man and Cybernetics, Vol. 3, pp. 2426-2431, 2006.
[13] R.I. Hg, P. Jasek, C. Rofidal, K. Nasrollahi, T.B. Moeslund, and G. Tranchet. "An RGB-D Database Using Microsofts Kinect for Windows for Face Detection", In *Eighth International Conference on Signal Image Technology and Internet Based Systems*, pp. 42-46, 2012.
[14] B. Ni, G. Wang and P. Moulin. "RGBD-HuDaAct: A Color-Depth Video Database For Human Daily Activity Recognition", In *IEEE International Conference onComputer Vision Workshops*, pp. 1147-1153, 2011.
[15] J. V. Stock, R. Righart and B Gelder. "Body Expressions Influence Recognition of Emotions in the Face and Voice". Emotion, Vol. 7, No. 3, pp.487-494, 2007.
[16] P. Ekman, "Basic Emotions". Handbook of Cognition and Emotion. John Wiley & Sons Ltd, Sussex, UK, 1999.
[17] N. Burrus. "Kinect calibration". In URL *http://nicolas.burrus.name/index.php/Research/KinectCalibration*. 2011.
[18] J. Ahlberg, "CANDIDE-3 - an updated parameterized face". Report No. LiTH-ISY-R-2326, Dept. of Electrical Engineering, Linkping University, Sweden, 2001.