2006 IEEE Conference on
Systems, Man, and Cybernetics
October 8-11, 2006, Taipei, Taiwan

Creating and Annotating Affect Databases from Face and Body Display: A Contemporary Survey

Hatice Gunes and Massimo Piccardi

Abstract— **Databases containing representative samples of human multi-modal expressive behavior are needed for the development of affect recognition systems. However, at present publicly-available databases exist mainly for single expressive modalities such as facial expressions, static and dynamic hand postures, and dynamic hand gestures. Only recently, a first bimodal affect database consisting of expressive face and upper-body display has been released. To foster development of affect recognition systems, this paper presents a comprehensive survey of the current state-of-the art in affect database creation from face and body display and elicits the requirements of an ideal multi-modal affect database.**

I. INTRODUCTION

Human-computer interfaces have improved significantly over the last decade, but mainly in one direction: from the computer to the human. Computers are now capable of visualizing complex, dynamic graphics and synthesizing expressive audio, thus substantially enriching the communication towards the human side. However, computers have not developed a corresponding capability of understanding the human emotional display. Input from humans to computers is still operated through traditional devices such as keyboards and pointers. Even more sophisticated input modalities such as speech recognizers and gaze detectors lack the capability of sensing the affective state of the human user. Such a capability could be exploited to make the interaction with computers less unnatural (or frustrating) and more effective to a desirable extent similar to that of human-human interaction (HHI). Strong evidence from psychology, neuroscience and sociology supports the feasibility of quantitative affect analysis from perceivable features of human beings starting from the fundamental work in [6]. Humans exploit such features to sense each others affective state in daily interaction. Moreover, the findings from the aforementioned sciences have recently triggered research in the computer science community for the automation of the affect analysis. A new area called *affective computing* has emerged and provided inspiration to various researchers for designing interfaces that will sense, recognize, understand and interpret human emotional states via language, speech, face and body gesture [25].

One major present limitation of affective computing is that most of the past research has focused on emotion recognition from one single sensorial source, or modality: the face display [23]. While it is true that the face is the main display of a human's affective state, other sources can improve the recognition accuracy. However, relatively few works have focused on implementing emotion recognition systems using affective multi-modal data [23]. Developing robust affective multi-modal systems requires databases containing representative samples of human multi-modal expressive behavior. There have been some attempts to create comprehensive test-beds for comparative studies of facial expression analysis, gesture recognition and multi-modal affect analysis. Emotion recognition via body movements and gestures has only recently started attracting the attention of computer science and human-computer interaction (HCI) communities [14]. The interest is growing with works similar to these presented in [2],[3], [10] and [16]. Moreover, a fundamental study by Ambady and Rosenthal suggests that the most significant channels for judging behavioral cues of humans appear to be the visual channels of facial expressions and body gestures [1]. Therefore, a reliable automatic affect recognizer should certainly attempt to combine facial expressions and body gestures.

Accordingly, this paper surveys current efforts in affective face and body gesture database creation and discusses the issues and challenges for creating and annotating an ideal multi-modal affect database. We provide a comprehensive review (although not exhaustively) of the various databases by grouping them into three categories: (a) facial expression databases; (b) body gesture databases and (c) multi-modal affect databases.

II. FACIAL EXPRESSION DATABASES

There have been some attempts to create comprehensive test-beds for comparative studies of facial expression and facial action unit (FAU) analysis following the work of Ekman and Friesen [6], [7]. As in the case of automated facial expression analysis systems, available databases can be grouped in two categories, based on the type of facial data they contain: (1) prototypical facial expressions or (2) FAU activations. The first group follows from [6] and contains facial display of the six basic emotions happiness, sadness, fear, disgust, surprise and anger from either single images or image sequences. The second group of databases contain more subtle changes in facial features (i.e. FAUs) and are coded using the Facial Action Coding System (FACS) [7]. A set of translation rules are used to link the FAU coding into basic emotions. For instance, the presence of four FAUs can be interpreted as the emotion "surprise" [7]: *FAU1+ FAU2+ FAU5+ FAU26=Surprise* (FAU1: Inner Brow Raised; FAU2: outer brow raised; FAU5: upper lid raised; FAU26: jaw

Authors are with the Faculty of Information Technology, University of Technology, Sydney (UTS) P.O. Box 123, Broadway 2007, NSW, Australia. (haticeg,massimo)@it.uts.edu.au

TABLE I

A SELECTION OF REPRESENTATIVE FACIAL EXPRESSION DATABASES AND A COMPARISON OF THEIR FEATURES

database	JAFFE	Maryland	Cohn-Kanade	MMI	FABO
posed(p)/spontaneous(s)?	p	p	p	p	p
head movement?	no	yes	no	no	yes
uniform background?	yes	yes	yes	mostly	yes
controlled light?	yes	yes	yes	yes	yes
occluded features?	no	no	no	no	some
# of data	213	70	2105	1500	430x2 mono-modal,607x2 bimodal
static images(s)/videos(v)?	s	v	v	740 s; 848 v	v
resolution	256x256	560x240	640x480	720x576	1024x768
length	NA	9 sec	10 sec	40-520 frames	10 sec
gray scale(g)/color(c)?	g	g	mostly g, few c	c	c
more than one view at a time?	no	no	no	yes,frontal and profile	no
# of subjects	10	40	100	19	23
ethnically diverse?	no	yes	yes	yes	yes
both genders?	female only	yes	yes	yes	yes
subject age group	young		18-30	19-62	18-42
subjects with beards/glasses?	no	no	no	yes	yes
single FAUs?	no	no	no	yes	yes
combination of FAUs?	no	no	yes	yes	yes
basic expressions?	yes	yes	yes	yes	yes
starts and ends with neutral?	NA	unknown	no	yes	yes
videos contain more than one expressive display?	NA	yes	no	few	yes
labelers	60 Japanese subjects	experimenters	FACS coders	2 FACS coders	experimenters& independent observers
AU coded?	no	no	last frames only	yes	no
expression coded?	yes	yes	no	yes	yes
temporal annotation provided?	no	no	no	partially	in progress
used for validating automated systems?	yes, many	yes	yes, many	yes	yes

dropped). We present the details about these databases in the following sub-sections and provide a comparison in Table 1.

A. Japanese Female Facial Expression (JAFFE) Database

JAFFE contains 213 images of 7 facial expressions (6 basic + 1 neutral) posed by 10 Japanese female models [14]. The images in the database are of 256x256 resolution. Each image has been rated on 6 emotion adjectives by 60 Japanese subjects. The rating results are distributed along with the images. The JAFFE database contains only gray scale static images and is publicly available for use in non-commercial research. The database contains only fully formed expressions and does not contain FAUs occurring alone or in combination. Representative images are shown in Fig.1(a).

(a) (b)

Fig. 1. Example images from (a) the JAFFE database and (b) the University of Maryland database.

B. The University of Maryland Database

The University of Maryland database contains 70 image sequences of 40 subjects who displayed emotions of their own choice and were free to move their heads while avoiding profile views [30]. The subjects varied in race and gender and performed a total of 145 expressions. Each sequence contains one to three expressions and is approximately 9 seconds long. Images in the database are of 560x240 resolution. The obtained sequences were manually *expression annotated* after the recordings by the experimenters. The database contains only fully formed expressions and does not contain FAUs occurring alone or in combination. Representative images are shown in Fig.1(b).

(a) (b)

Fig. 2. Example images from (a) the Cohn-Kanade AU-Coded Facial Expression database and (b) the MMI Facial Expression database.

C. Cohn-Kanade AU-Coded Facial Expression Database

Subjects' facial behavior was recorded in a laboratory setting. Image sequences with in-plane and limited out-of-plane motion were included [15]. However, only limited image data from the frontal camera are available for distribution. The released portion of the database contains 100 university students ranging in age from 18 to 30 years. The image sequences began with a neutral face and ended in full formed facial expression display. The sequences were digitized into 640x480 pixel arrays with mainly 8-bit gray-scale values.

The available portion mostly contains facial expressions and combination of facial action units; single action units are not present. FACS coding for each subject and expression performed are provided together with the database. The codes refer only to the final frame in the image sequence. The database does not contain the FAU codes for every frame and the videos are not temporally annotated (i.e. segmented into neutral-onset-apex-offset-neutral phases). Representative images are shown in Fig 2(a).

D. MMI Face Database

This database contains approximately 1500 samples of both static images and image sequences from 19 male and female subjects in frontal and in profile view [24]. The database was obtained by instructing participants on how to pose in laboratory settings with a uniform background; although sequences with non-uniform background are also provided. The database contains single FAUs, FAU combinations and prototypical emotion displays. The sequences are 24- bit true color and 720x576 pixel/frame. Dual-view images combine frontal and profile view of the face and were recorded using a mirror. The sequences last between 40 and 520 frames, and follow the pattern of neutral-expressive-neutral facial display. The images were FAU annotated by two FACS coders (certified experts who "dissect" an observed expression, decomposing it into the specific FAUs). 169 samples at the time of distribution are provided with their temporal annotation. Representative images are shown in Fig 2(b).

III. BODY GESTURE DATABASES

To our best knowledge, unlike the facial expression databases, there is not a publicly available general purpose benchmark database for gesture recognition. Gesture databases exist for static hand postures and dynamic hand gestures, mostly for command entry purposes [29]. Existing databases mainly consist of non-affective one hand gestures only, and do not take into consideration the relationship between body parts (i.e. between hands; hands and the face; hands, face and shoulders etc.). The details of such databases are explained in the following subsections and a comparison is provided in Table 2.

A. The Massey Hand Gesture Database

This database contains about 1500 images of different hand postures, in different lighting conditions [5]. The data were collected from a hand gesture in front of a dark background, and in different lighting environments, including normal light and dark room with artificial light. The clipped images vary in size with maximum resolution of 640x480 with 24 bit RGB color. So far, the database contains material gathered from 5 different individuals. Images representing the database are provided in Fig. 3(a).

B. Sebastien Marcel's Static and Dynamic Hand Posture/Gesture Database

The static image database contains images in .pnm format [26]. Images were acquired for 6 hand postures (a, b, c, point, five, v) from about 10 different persons. The dynamic hand posture database contains image sequences for 4 hand gestures (Click, Rotate, Stop-Grasp-Ok, No) from 10 different people. The dynamic hand gesture database contains 2D hand trajectories in a normalized body-face space, 4 hand gestures, from 10 different people repeated many times. Example images are provided in Fig. 3(b).

(a) (b)

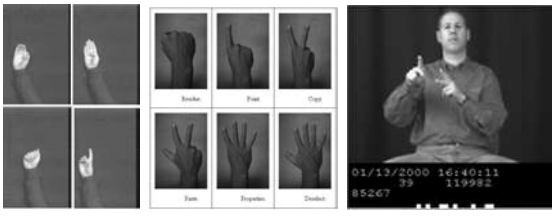
Fig. 3. Example images from (a) the Massey Hand Gesture database [5], and (b) Sebastien Marcel's Static Hand Posture database [26]

C. Thomas Moeslund Gesture Recognition Database

The database consists of 2060 images of a hand performing the different static signs used in the international sign language alphabet [18]. The images in the database are gray-scale images in .tif format with a resolution of 248x256 pixels. Each of the gestures/signs are performed in front of a dark background and the user's arm is covered with a similar black piece of cloth. Each gesture is performed at various scales, translations, and rotations in the plane parallel to the image-plane. Example images are shown in Fig. 4(a).

D. Holte and Stoerring Pointing and Command Gesture Database

The database contains pointing and command gestures under mixed illumination conditions [12]. The gesture vocabulary consists of 13 gestures; 9 gestures are static and 4 are dynamic. The video sequences are recorded in PAL resolution, 768 x 576 pixels (see Fig. 4(b)). The scenario script is created to make participants believe that they interact with the object placed on the table. The gestures occur under a special light setup on a table containing various objects. Annotation is provided in the form of "start of a gesture, start of a stroke, end of a stroke, end of a gesture".



(a) (b) (c)

Fig. 4. Example images from (a) Thomas Moeslund's Gesture Recognition database [18], (b) Holte and Stoerring Pointing and Command Gesture database [12] and (c) the CARE ASL database [19].

TABLE II

A SELECTION OF REPRESENTATIVE BODY GESTURE DATABASES AND A COMPARISON OF THEIR FEATURES

database	Massey	Sebastien M.	Thomas M.	Holte&Stoe.	Boston Univ.	Max Planck	FABO
purpose	various	command entry	sign language	pointing& command entry	sign language	speech related gesture	emotion recognition
emotion intent?	no	no	no	no	no	no	yes
posed(p)/spontaneous(s)?	p	p	p	p	p	s	p
controlled light?	yes	yes	yes	yes	yes	unknown	yes
both hands recorded?	no	no	no	no	yes	yes	yes
upper-body visible?	no	no	no	no	yes	unknown	yes
more than one body part activated at a time?	no	no	no	no	yes	yes	yes
face/facial features visible?	NA	NA	NA	NA	yes	unknown	yes
uniform background?	yes	yes	yes	yes	yes	unknown	yes
occluded features?	no	no	no	no	yes	unknown	yes
# of data	1500	repeated gestures	2060	repeated gestures	unknown	unknown	607x2 bimodal
# of postures/gestures	unknown	6 postures, 4 gestures	alphabet length	9 static, 4 dynamic	various	unknown	various
dynamic posture/gesture?	unknown	yes	no		yes	yes	yes
static images(s)/video sequences(v)?	v	s	s	s	v	v	v
resolution	640x480	unknown	248x256	768x576	648x484	unknown	1024x768
length	unknown	unknown	1 frame	unknown	various	various	60 - 350
gray scale(g)/color(c)?	c	g	g	g	unknown	unknown	c
more than one view at a time?	no	no	no	no	yes, 4	no	yes
# of subjects	5	10	1	unknown	unknown	unknown	23
ethnically diverse?	unknown	unknown	no	yes	unknown	yes	yes
both genders?	unknown	unknown	no	yes	unknown	yes	yes
subject age group	unknown	unknown	unknown	unknown	unknown	various	18-45
starts and ends with neutral?	unknown	unknown	no	unknown	unknown	unknown	yes
videos contain more than one expressive display?	no	no	no	no	no	unknown	yes
labeled data?	no	yes	yes	yes	yes	unknown	yes
labelers	NA	experimenters	experimenters	experimenters	experimenters	unknown	experimenters& independent observers
expression coded?	no	NA	NA	NA	NA	no	yes
temporal annotation provided?	no	no	NA	yes	no	unknown	in progress
publicly available?	yes	yes	yes	yes	yes	no	yes

E. Video Sequences of American Sign Language (ASL)

This database is obtained by the National Center for Sign Language and Gesture Resources at Boston University and contains annotated Quicktime movies of American Sign Language (ASL) sentences [19]. The signing was captured simultaneously from four different cameras, at a frame rate of 60 frames per second and at an image resolution of 648x484 pixels. Therefore, the database contains samples of upper-body image sequences. However, the focus is entirely on sign language communication and not affective body expression (see Fig. 4(c)).

F. The Gesture Database from the Max Planck Institute

The Gesture Database from the Max Planck Institute consists of the video recordings of speech and gestures that spontaneously accompany speech, and the annotations regarding gesture and speech in the recording [8]. The recordings were made in different countries, including Italy, the USA, Japan, Turkey, Ghana etc. Speech events are recorded eliciting spontaneous gestures, such as narration of traditional stories and autobiographical stories, description of the local environment, and route direction. However, the database is not publicly available yet.

IV. VISUAL MULTI-MODAL AFFECT DATABASES

The detailed comparison of the databases included in this section is provided in Table 3.

A. The Database Collected at the University of Amsterdam

This database was recorded by creating a video kiosk with a hidden camera which would display segments from recent movie trailers. Modalities are face and audio. The database contains recordings from 28 students [27]. After each subject had seen the video trailers, they were interviewed to find out their emotional state corresponding to the hidden camera video footage. As the experiment itself was conducted in a spontaneous way only expressions corresponding to the naturally occurring emotions could be captured (neutral, joy, surprise, or disgust). The database is not publicly available yet.

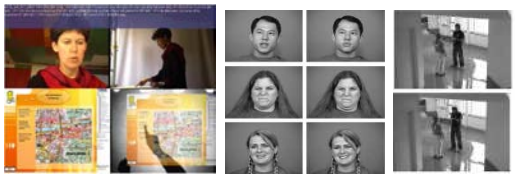
TABLE III

A SELECTION OF REPRESENTATIVE BI-MODAL DATABASES AND A COMPARISON OF THEIR FEATURES

database	Univ. of Amsterdam	SmartKom	Univ. of Texas	FABO
purpose	HCI	HCI	HHI	affect analysis and recognition
emotion intent?	yes	no	for face videos, yes	yes
bi-modal components	face and audio	upper-body and speech	face and speech/body and speech	face and upper-body
posed(p)/spontaneous(s)?	s	s	partially p and s	p
controlled light?	yes	yes	yes	yes
both hands visible?	no	yes	yes	yes
upper-body visible?	no	yes	yes	yes
more than one body part activated at a time?	unknown	yes	yes	yes
face visible?	yes	yes	no	yes
uniform background?	yes	unknown	yes	yes
occluded features?	unknown	unknown	yes, for body	yes
# of data	unknown	many	many	430x2 mono-modal, 1644x2 bimodal
# of postures/gestures	unknown	unknown	unknown	various(over 30)
dynamic posture/gesture?	unknown	yes	yes	yes
static images(s)/videos(v)?	unknown	v	v	v
resolution	unknown	not specified	720x480	1024x768
length	unknown	4,5 min	10 s	60 - 350 frames
gray scale(g)/color(c)?	unknown	c	c	c
more than one view at a time?	no	yes	no	yes
# of subjects	28	224	284	23
ethnically diverse?	yes	yes	yes	yes
both genders?	yes	yes	yes	yes
subject's age group	young	not specified	18-25	18-42
starts/ends with neutral?	no	no	no	yes
more than one expressive display in videos?	yes	yes	no	yes
labeled data?	yes	yes	yes	yes
labelers	subjects themselves	experimenters	experimenters	experimenters& independent observers
expression coded?	yes	no	yes	yes
temporal annotation provided?	no	unknown	no	in progress
publicly available?	not yet	yes	yes	yes
used for validating automated systems?	yes	yes	unknown	yes

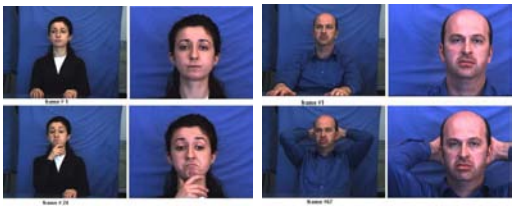
B. The SmartKom corpora (SKP)

The SmartKom multi-modal database was produced (1999 - 2003) at the Bavarian Archive for Speech Signals (BAS), University of Munich [28]. The primary aim of the database was the empirical study of HCI in a number of different tasks and scenarios. The database consists of multi-modal recordings of 224 persons in a Wizard-of-Oz setting, a common methodology in HCI practice. This methodology makes the user believe that he is interacting with a computer program when in fact he is interacting with a human. The hand gestures were recorded via an infrared camera (to capture the 2D gestures) as input to the Siemens Virtual Touchscreen (SIVIT) device. Various gestures were observed such as pointing with/without touching the display, reading/moving hand, searching, counting etc. Eventually, the multi-modal recordings were obtained from four different views (front camera, side camera, graphical system output, infrared camera merged with the system output) and an audio signal (see Fig. 5(a)). Although the database contains multiple communicative modalities the main focus is on the audio modality. Emotion labeling is obtained only for the video from the frontal camera together with the audio signal.



(a) (b) (c)

Fig. 5. Example images from the databases collected at (a) BAS, illustration of the recordings obtained from four sensors [28]; and the University of Texas [21], (b) videos from the facial expression recordings and (c) the conversational setting.



(a) (b)

Fig. 6. Example sequences from FABO obtained from body (left columns) and face (right columns) cameras.

C. The University of Texas Database

This database contains both close-up videos of faces (dynamic information about the face across a range of viewpoints and emotional expressions) and longer-range video clips that contain whole body shots [21]. 76 of the subjects were male and 208 female. They were mostly Caucasians, between the ages of 18 and 25. Multiple experimenters coded different data sets, though only one experimenter coded each individual video clip. For the dynamic facial expression videos the subject watched a 10-minute video, which contained scenes from various movies and television programs intended to elicit different emotions. According to the judgment of the experimenter coding the data the videos contain instances of: happiness, sadness, fear, disgust, anger, puzzlement, laughter, surprise, boredom, or disbelief (see Fig. 5(b)). Head and eye movements often accompany the expressions. Some clips also contain more than one expression (e.g., a puzzled expression, which turns to surprise or disbelief, and ultimately laughter). The conversation videos recorded conversation between the subject and a laboratory staff member by placing the camera 8 m away and 3.5 m high (see Fig. 5(c)). To capture some natural gesturing in these videos, the subject was asked to give directions to a building on campus. These videos last 10 seconds. Although the database contains bimodal data in the conversation videos (audio-visual data), its drawback is the fact that simultaneous recordings were not obtained for face (using close up shooting) and body (shooting at a moderate distance) using multiple cameras at the same time. The moderate distance defined is too far (8 m) to enable analysis of subtle movements of the facial features and body gesture at the same time. Moreover, the subjects were not asked to take part in a more emotion-oriented context.

D. FABO

We created a bimodal database that consists of recordings of facial expressions alone and combined face and body expressions [11]. We recorded the sequences simultaneously using two fixed cameras with a simple setup and uniform background. One camera was placed specifically capturing the head only and the second camera was placed in order to capture upper-body movement from the waist above. Prior to recordings subjects were instructed to take a neutral position, facing the camera with hands visible and placed on the table. The subjects were first asked to perform Facial Action Units (e.g. raise the upper eyebrows) either occurring alone or in combination (e.g. AU8 can not be displayed without AU25). In the second stage the subjects were asked to perform facial expressions, namely: neutral, happiness, neutral surprise, positive surprise, negative surprise, fear, anger, sadness and disgust. In the last stage, the subjects were asked to perform face and body gestures simultaneously by looking at the facial camera constantly. During the recordings we used a *scenario approach* where subjects are provided with short scenarios describing an emotion eliciting situation. They were told, for instance, what they would do when "it was just announced that they won the biggest prize in lottery" etc. In some cases the subjects came up with a variety of combinations of face and body gestures. As a result of the feedback and suggestions obtained from the subjects, the number and combination of face and body gestures performed by each subject varies slightly (see [11] for details). Fig. 6 shows images obtained simultaneously by the body and face cameras. The annotation of the data contained in the FABO database consist of (a) experimenters' labeling as experts (b) the subjects' own labeling of their own performance and (c) the annotation of the visual data (each face and body video separately) by six independent human observers. We developed a survey for face and body videos separately by using the labeling schemes for emotion content (e.g. happiness) and signs (e.g. how contracted the body is). The FABO database contains more than 900 mono-modal and more than 1000 bimodal videos. It has been already used for the validation of the approach proposed in [10] which could not have been possible with any existing databases due to their lack of combined affective face and body displays.

V. Discussion and Open Issues

After providing representative examples of visual affect databases, in this section we list the limitations of each category and summarize the open issues for future research.

A. Facial Expression Databases

All of the publicly available facial expression databases collected data by instructing the subjects on how to perform the desired actions with limited out of plane head motions. The audio-visual database collected at the University of Texas is the only facial expression database containing a wide range of spontaneous facial expressions. However, even then the background and lights were strictly controlled and people were asked to wear tops of the same gray color. There is no single facial expression database of images that is used commonly by all different facial expression research communities. In general, each research community has created and used their own facial expression database. To date, the Cohn-Kanade AU-Coded Facial Expression Database is the most commonly used database in research on automated FAU/facial expression analysis.

The development of a common annotation scheme of facial muscle movement, namely the Facial Action Coding System (FACS) for describing facial expressions by action units (FAUs), has helped significantly in creating extensive facial expression databases and validating the automatic recognizers.

Consequently, limitations of the current facial expression databases can be listed as follows:

- None of the existing databases contain spontaneous facial expressions with freedom of out of plane head movement of any degree.
- All of the databases were created with controlled lights and background.
- The image sequence databases contain one or more neutral-expressive-neutral facial displays; none of them contain variable-expressive-variable behavior [24].

- None of the facial expression databases contain occluded faces or facial features.
- Although one database (i.e. MMI) contains profile-view none of them attempt to record the expressive facial display using multiple cameras simultaneously.
- Although one database (i.e. MMI) contains partial annotation of the temporal segments of FAUs (neutral-onset-apex-offset-neutral), none of them attempt to generate temporal annotation for either facial expression or FAU data or both.

B. Body Gesture Databases

The body and hand gestures are much more varied than facial gestures. There is an unlimited vocabulary of body postures and gestures with combinations of movements of various body parts [4], [9]. Despite the effort of Laban in analyzing and annotating body movement [17], unlike the FAUs, body action units (BAUs) that carry expressive information have not been defined with a Body Action Coding System (BACS). Therefore, it is even harder to create a common benchmark database for affective gesture recognition. Moreover, communication of emotions by body gestures is still an unresolved area in psychology. Unlike the facial action units (FAUs), there is not one common annotation scheme that can be adopted by all the research groups. The most common annotation has been *command-purpose annotation*, for instance calling the gesture as "rotate" or "click" gesture. Another type of annotation has been based on the *gesture phase*, e.g. "start of gesture stroke-peak of gesture stroke-end of gesture stroke". However, a more detailed annotation scheme, similar to that of FACS is needed. A general body gesture annotation scheme, possibly named as Body Action Unit Coding System (BACS), should include information and description as follows: *body part* (e.g. left hand), *direction* (e.g. up/down), *speed* (e.g. fast/slow), *shape* (hands made into fists), *space* (flexible/direct), *weight* (light/strong), *time* (sustained/quick), and *flow* (fluent/controlled) as defined by Laban and Ullman [17]. Additionally, temporal segments (neutral-start of gesture stroke-peak of gesture stroke-end of gesture stroke-neutral) of the gestures should be included as part of the annotation scheme.

Limitations of the present body gesture databases can be summarized as follows:

- Except for the sign language ones, all the databases contain only gestures for command entry and/or navigation purposes (e.g. Point and move the book, close the hand, paste and close the hand again)
- All of the databases contain only one hand-gestures; two-hand gestures are not included.
- In general, only slow movements are allowed during the recording process.
- None of the databases aim to contain representative samples of human affective body display.
- Except for the sign language ones, whole upper-body, both hands and head are not included in the recordings.

In general, the aforementioned databases lack expressiveness of the body and ignore the relationship between various body parts (e.g. head and shoulders, hands and head etc.) and therefore cannot be used for analysis of human nonverbal communicative behaviors.

C. Multi-modal Databases

After providing the strong and weak points of currently existing face and body gesture databases, in this subsection, we pose a question on "how an ideal multi-modal affect database" should be and try to provide answers to it. Firstly, the extent and the nature of affective data collection depends on the following factors [25]:

- *Posed/spontaneous:* Is the emotion elicited by the subject upon request or is there an actual reason or situation creating the affective activation?
- *Expression/emotion:* Is the actual target on expression (how people externalize) or emotion (what people feel internally)?
- *Laboratory setting/real life:* Is the recording obtained in a laboratory with controlled background/lights/noise or in real life with unconstrained conditions?
- *Open recording/hidden recording:* Does the subject know that s(he) is being recorded?
- *Emotion-purpose/other-purpose:* Does the subject know that s(he) is expected to create emotional response?

To foster development of natural human-computer interfaces, an ideal multi-modal affect database should contain data obtained in a natural setup. In other words, data that is spontaneous and obtained in real life situation with non-emotion purpose. Taking into account the aforementioned factors, an ideal multi-modal affect database thus should have the following features:

- The subject is present in his/her *natural environment* (i.e. office or house).
- The subject is in a particular affective state due to some real-life event or trigger of events (i.e. stressed at work).
- The subject *does not try to hide* what s(he) feels, on the contrary, displays what s(he) feels using multiple communicative channels (i.e. facial expression, head movement, body gestures, voice etc.).
- The subject *is not aware of the recording*, hence will not restrain himself/herself unlike the case when s(he) is part of an experiment.
- There are occurrences of *occlusions* (i.e. hands occluding each other or hand occluding the face) and noise (i.e. in audio recordings).
- There are *multiple sensing devices* (i.e. multiple cameras, multiple microphones, haptic sensors etc.).
- *Viewing and lighting conditions are not uniform.*
- The *sessions are long*, expanding between one day and possibly a couple of weeks, capturing *all variations* of expressive1-expressive2-expressive3-neutral behavior in every possible order or combination [23].
- The subjects are of *diverse age, gender and ethnic background.*

Once the multi-modal data have been acquired, they need to be annotated and analyzed to form the ground truth for

machine understanding of the human affective multi-modal behavior. Annotation of the data in a bi-modal/multi-modal database is a very tiresome procedure overall as it requires extra effort and time to view and label the sequences with a consistent level of alertness and interest (e.g. it takes more than *one hour* to FAU code *one minute* of facial video). Hence, obtaining the emotion- and quality-coding for all the visual data contained in bi-modal databases is very difficult to achieve. Moreover, we believe that for the annotation purposes it is almost impossible to use emotion words that are agreed upon by everybody. The problem of what different emotion words are used to refer to the same emotion display is not, of course, a problem that is unique to this; it is by itself a topic of research for emotion theorists and psychologists. It is a problem deriving from the vagueness of language, especially with respect to terms that refer to psychological states [20]. As a rule of thumb, at least two main labeling schemes, in line with the psychological literature on descriptors of emotions, should be used: verbal categorical labeling (perceptually determined, i.e. happiness) and broad dimensional labeling: arousal(arousal–sleep) and valence (activated-deactivated). This labeling is in accordance with emotion theories in psychology: (a) Ekman's theory of emotion universality [6] and (b) Russell's theory of arousal and valence [22]. Taking into account these facts an ideal multi-modal affect database should be annotated as follows. (1) Experimenters, preferably a group consisting of an expert in the affective computing field or an emotion researcher, should view and label the multi-modal data. (2) Subjects' own evaluation should be obtained by asking the subjects after the recordings, to view and fill in a survey about their expressions. This feedback will form the subject's own evaluation of his affective state. (3) The multi-modal data should additionally be annotated by independent human observers with different ethnic and/or cultural background in order to obtain independent interpretations. Moreover, it could be further analyzed whether being exposed to the expressions (hearing/seeing etc.) from one sensor (face camera only) or another (body camera only), or from multiple sensors simultaneously (cameras and headphones) affects the observer's interpretations.

VI. Conclusion

Current bi-modal/multi-modal databases are yet to improve their features, content and annotation schemes to achieve the level of specifications listed above. Creating a spontaneous multi-modal affect database is a challenging task involving ethical and privacy concerns together with technical difficulties (high resolution, illumination, multiple sensors, consistency, repeatability etc.). Given these restrictions, a database of directed emotional display has been the only alternative possible to date. Another challenging issue is that of creating a database that contains samples of both staged and spontaneous data in order to study the differences between these and how this procedure can be automated. However, the research field of multi-modal affect recognition is relatively new and future efforts have to follow.

References

[1] N. Ambady and R. Rosenthal, "Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis," *Psychological Bulletin*, vol. 11, pp. 256-274, 1992.

[2] Balomenos, T., et al. "Emotion analysis in man-machine interaction systems," *Proc. of MLMI 2004*, LNCS 3361, pp. 318 - 328.

[3] Burgoon, J. K., et al., "Augmenting human identification of emotional states in video," *Proc. of Int. Conference on Intelligent Data Analysis*, 2005.

[4] Coulson, M., "Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence," *J. of Nonverbal Behavior*, vol. 28, pp. 117-139, 2004.

[5] F. Dadgostar, A. L. C. Barczak, and A. Sarrafzadeh, "A color hand gesture database for evaluating and improving algorithms on hand gesture and posture recognition," *Research Letters in the Information and Mathematical Sciences*, vol. 7, pp. 127-134, 2005.

[6] P. Ekman and W. V. Friesen, *Unmasking the face: a guide to recognizing emotions from facial clues*,Prentice-Hall: NJ; 1975.

[7] P. Ekman and W. V. Friesen, *The Facial Action Coding System: A technique for measurement of facial movement*, Consulting Psychologists Press, San Francisco:CA; 1978.

[8] Gesture Database from the Max Planck Institute: http://www.mpi.nl/ISLE/overview/Overview_GDB.html

[9] D. B. Givens, *The Nonverbal dictionary of gestures, signs and body language cues*, Center for Nonverbal Studies Press: Washington; 2005.

[10] H. Gunes and M. Piccardi, "Affect recognition from face and body: Early Fusion vs. Late Fusion," *Proc. IEEE SMC*, 2005, pp. 3437-3443.

[11] H. Gunes and M. Piccardi, "A Bimodal Face and Body Gesture Database for Automatic Analysis of Human Nonverbal Affective Behavior," *Proc. ICPR 2006* (in press).

[12] Holte and Stoerring Pointing and Command Gesture Database:http://www.cvmt.dk/ fgnet/pcgdb/

[13] E. Hudlicka, "To feel or not to feel: The role of affect in human-computer interaction", *Int. J. Hum.-Comput. Stud.*, Vol. 59(1-2), pp. 1-32, 2003.

[14] Japanese Female Facial Expression (JAFFE): Database:http://www.mis.atr.co.jp/mlyons/jaffe.html

[15] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," *Proc. IEEE FGR*, 2000, pp. 46-53.

[16] A. Kapoor and R. W. Picard, "Multimodal affect recognition in learning environments," *Proc. ACM Multimedia*, 2005, pp. 677-682

[17] R. Laban and L. Ullmann,*The Mastery of movement*, 4th Rev. and Ed.: Princeton Book Company Publishers, 1988.

[18] T. Moeslund Gesture Recognition Database: http://www.vision.auc.dk/%7Etbm/Gestures/database.html.

[19] National Center video sequences of ASL: http://csr.bu.edu/asl/html/sequences.html

[20] A. Ortony and T. J. Turner, "What's basic about basic emotions?," *Psychological Review*, vol. 97, pp. 315-331, 1990.

[21] A. J. O'Toole et al., "A video database of moving faces and people," *Tran. IEEE PAMI*, vol. 27, No. 5, 2005, pp. 812-816.

[22] J. A. Russell, "Facial expression of emotion: What lies beyond minimal universality?," *Psychological Bulletin*, vol. 118, pp. 379-391, 1995.

[23] M. Pantic et al., "Affective multimodal human-computer interaction," *Proc. ACM Multimedia*, 2005 pp. 669-676.

[24] M. Pantic et al., "Web-Based database for facial expression analysis," *Proc. IEEE ICME*, 2005, pp. 317-321.

[25] R. W. Picard, *"Affective computing"*, MIT Press, Cambridge; 1997.

[26] Sebastien Marcel's Static and Dynamic Hand Posture/Gesture Database http://www.idiap.ch/ marcel/Databases/gestures/main.php

[27] N. Sebe et al., "Authentic facial expression analysis," *Proc. IEEE FGR*, 2004, pp. 517-522.

[28] SmartKom Corpora: http://www.phonetik.uni-muenchen.de/Bas/BasProjectseng.html

[29] M. Turk, "Gesture recognition," in Handbook of Virtual Environments : Design, Implementation, and Applications, K. Stanney, Ed.: Lawrence Erlbaum Associates, Inc.(Draft version), Dec. 2001.

[30] University of Maryland Database: http://www.umiacs.umd.edu/users/yaser/DATA/index.html