Bimodal Modelling of Facial and Upper-Body Gesture for Affective HCI

Hatice Gunes, Massimo Picardi, Tony Jan
Faculty of Information Technology, University of Technology, Sydney (UTS)
Sydney, NSW, 2007, Australia
{haticeg, massimo, jant}@it.uts.edu.au

Abstract

Multimodal systems allow humans to interact with machines through multiple modalities such as speech, facial expression, gesture, and gaze. This paper presents a bimodal model of facial and upper-body gesture for affective HCI suitable for use in a vision-based multimodal system. What distinguishes the present study from its predecessors is that, this model combines Facial Action Units (FAUs) and Body Action Units (BAUs) to encode affective states. To our best knowledge there has been no attempt to combine face and body gesture for multimodal affect recognition yet.

Keywords

Affective HCI, perceptual interface, bimodal emotion model, action units (AUs), gesture recognition, facial expression recognition.

INTRODUCTION

Emotion research has grown significantly in the last decade in the field of psychology, neuroscience and sociology revealing that emotions have an important role in various aspects of human's life. Not only do they effect human-human communications by conveying intentions and tendencies but also they are tightly coupled with all functions we, humans, are engaged with: attention, perception, learning, reasoning, decision making, planning, action selection, memory storage and retrieval. (Damasio, 1994; Isen, 2000; Scherer *et al.*, 2001).

Such significant findings in social science triggered further research in computer science community particularly in artificial intelligence and HCI. A new area called *affective computing* has emerged and has provided inspiration to various researchers for enabling affective HCI, designing machines and interfaces that will recognize, understand and interpret human emotional states via language, speech, facial and bodily gesture (Picard, 1997).

Emotions can be communicated by various physical changes in the body: changes invisible to others (e.g., blood chemistry, brain activity and neurotransmitters) and/or physical changes that can be differentiated by humans (voice, tone, face and gesture). Hence, a broad range of modalities is available, including speech and language, gesture and head movement, body movement and posture, as well as facial expression. Expressive facial and bodily gesture is one of the main non-verbal communication channels in human-human interaction (HHI) (Mehrabian, 1968; De Meijer, 1989). According to Mehrabian (Mehrabian, 1968), 93% of our communication is nonverbal and humans display their emotions most expressively through facial expressions and body gestures. Considering the effect of the message as a whole, spoken words of a message contribute only for 7%, the vocal part contributes for 38%, while facial expression of the speaker contributes for 55% to the effect of the spoken message. Hence, understanding human emotions through these nonverbal means is one of the necessary skills both for humans and also for the computers to interact intelligently and effectively with their human counterparts. Computers can also measure affect that is clearly expressed to them, for instance, using vision and visually accessible modalities. It is possible to measure facial and bodily activities that might not be visible using electromyography (EMG) (Lisetti&Schiano, 2000). However, we are interested in the visible activities, "visual" communicative cues and "facial and bodily display" for their "signification of intent".

Automatic facial expression recognition has attracted interest of artificial intelligence and computer vision research communities for the past decade. Significant research results have been reported in recognition of basic emotions using facial expressions (e.g. Cohen *et al.*, 2002). Growing amount of research has also investigated movement and gesture as a main channel of non-verbal communication in HHI and HCI. However, existing literature on automatic emotion recognition has focused mainly on the face, the expressive information that body gestures carry has not been explored (Hudlicka, 2003). On the other hand, there exist emotion recognition studies in psychology revealing relationships between particular body movements and specific emotions (Clynes, 1975; Scherer *et al.*, 1986; Wallbott&Scherer, 1988; De Meijer, 1989; Sogon & Masutani, 1989; Nakamura *et al.*, 1990; Boone and Cunningham, 1998; Givens, 2001).

Taking into account these findings, the goal of our research is to build a computer system which will understand two of the non-verbal communication channels namely face and upper-body and extract the affective information they convey using image processing, computer vision and machine learning techniques. The rationale for this attempt of combining face and body gesture for a better understanding of human non-verbal behaviour is the recent interest and advances in multi-modal interfaces. According to Hudlicka (Hudlicka, 2003), while much progress has been achieved in affect assessment using a single measure type, reliable assessment typically requires the concurrent use of multiple methods. Multimodal systems employ information coming from several channels and combine different modalities (i.e. speech, facial expression, gesture, and gaze) that occur together to function in a more efficient and reliable way in various HCI applications. Pantic and Rothkrantz (Pantic&Rothkrantz, 2003) clearly state the importance of a multimodal affect analyser for research in emotion recognition. The modalities considered are visual and auditory, by processing facial expression and vocal cues. Examples of such bimodal systems are the works of (Chen&Huang, 2000*)* and (De Silva&Ng, 2000). The interpretation of other visual cues such as body movements and gestures is not explicitly addressed in (Pantic&Rothkrantz, 2003) due to the fact that emotion recognition via body movements and gestures has only recently started attracting the attention of computer science and HCI communities (Hudlicka, 2003). It is believed that their use is limited in traditional desktop computer settings, both in terms of feasibility and utility (Hudlicka, 2003). However, according to Hudlicka (Hudlicka, 2003) once we leave the desktop setting and enter the world of wearable sensing devices, virtual environments and synthetic avatars, these modalities become more relevant. Several papers focus on movement and gestures as means of expressing and recognizing emotions (Camurri *et al.*, 2003; Paiva *et al.*, 2003). The research in analysing expressive cues of a dancer at the University of Genoa is the only bodily expression recognition system that we are aware of (Camurri *et al.*, 2003). From this point of view, part of our work is similar to the one described in (Camurri *et al.*, 2003), however, analysing the affect from dance is a different problem than analysing affect from combination of expressive face and body display.

What distinguishes the present study from its predecessors is that, face and upper-body gestures are combined in a bimodal/multimodal manner to distinguish between various expressive cues that will help computers recognize particular emotions. This paper presents a bimodal model of facial and upper-body gesture for affective HCI suitable for use in a vision-based multimodal system.

THEORETICAL BACKGROUND

In this section we present the theoretical background for expressive face and body gesture analysis.

Expressive Face

Over the past decade, emotion recognition studies have suggested that neutral, anger, disgust, fear, happiness, sadness and surprise are universally recognized from facial expressions (Ekman&Friesen, 1968, 1978). Vision-based systems that automatically analyse the facial expressions are based on these findings and can be classified into two categories: (1) systems that recognize prototypic facial expressions; (2) systems that recognize facial actions.

(1) Systems that recognize prototypic facial expressions: There has been a significant amount of research on creating systems that recognize a small set of prototypic emotional expressions (i.e., surprise, anger) from static images or image sequences. This focus on recognizing prototypic facial expressions follows from the work of Ekman (Ekman&Friesen, 1968). Bassili suggested that motion in the image of a face would allow emotions to be identified even with minimal information about the spatial arrangement of features (Bassili, 1978). Thus, facial expression recognition from image sequences is based on categorizing prototypic facial expressions by tracking facial features and measuring the amount of facial movement. There are various approaches that have been explored. Some of those include analysis of facial motion, measurements of the shapes and facial features, their spatial arrangements and holistic spatial pattern analysis (see (Pantic&Rothkrantz, 2000) for details).

(2) Systems that recognize facial actions: Seven universal emotion categories are not sufficient to describe all facial expressions (Ekman&Friesen, 1978). Although prototypic expressions, like happy, surprise and fear, are natural, they occur infrequently in everyday life and provide an incomplete description of facial expression. To capture the subtlety of human emotion and paralinguistic communication, automated recognition of fine-grained changes in facial expression is needed (Ekman&Friesen, 1978). Ekman and Friesen (Ekman&Friesen, 1978) developed the Facial Action Coding System (FACS) for describing facial expressions by action units (AUs). The system is based on the enumeration of all "action units" of a face that cause facial movements. Of 44 FACS AUs defined, 30 AUs are anatomically related to the contractions of specific facial muscles: 12 are for upper face, and 18 are for lower face. AUs can be classified either individually or in combination. Vision-based systems that attempt to recognize action units (AUs) are motivated by FACS.

Some of the previous work to automatically recognize action units has used optical flow across the entire face or employed facial feature measurement. Donato *et al.* (Donato *et al.*, 1999) compared optical flow, principal component analysis, independent component analysis, local feature analysis, and Gabor wavelet representation to recognize eight individual AUs (6 upper face and 2 lower face AUs) and four combinations of lower face AUs (e.g. 9+25). Tian *et al.* (Tian *et al.*, 2001) developed an automatic AU analysis system using facial features to recognize 16 action units and any combination of those by describing the shape of facial features by multistate templates. (Pantic&Rothkrantz, 2003) developed a facial gesture recognition system from both frontal and profile image sequences, 32 individual AUs occurring alone or in combination are recognized using rule-based reasoning.

Expressive Body Gesture

Expressive body movement/gesture is any of several changes in the physical location, place, or position of the material parts of the human form (e.g., of the eyelids, hands, or shoulders) that conveys expressive information. Recent studies show that adults as well as children show great skill in their ability to decode emotions from body movements (Clynes, 1975; Scherer *et al.*, 1986; Wallbott&Scherer, 1988; De Meijer, 1989; Sogon & Masutani, 1989; Nakamura *et al.*, 1990; Boone and Cunningham, 1998; Givens, 2001). However, the computer science community has mostly attempted to integrate gestures that are employed in the absence of speech and emotional facial displays for command entry or interface control. Hence, there are very few papers focused on recognizing affect from body movements and gestures (Hudlicka, 2003).

Camurri *et al.* in (Camurri *et al.*, 2003) presents analysis and classification of expressive gesture in human full-body movement and particularly in dance performances by comparing system classification with spectators' ratings. They base their research on the fact that distinct emotions are often associated with distinct qualities of body movements, and develop a system for capturing and analysing human body movement and mapping it onto one of four basic emotions (anger, fear, grief and joy). For this purpose they use the non-propositional movement qualities (e.g. tempo and force of the movement), rather than specific gestures expressing particular emotions. Paiva *et al.* (Paiva *et al.*, 2003) on the other hand focuses on expression of affect with gestures and head movements using a smart doll, the SenToy. SenToy is an input device that allows the user to perform gestures or movements that the sensors inside the doll pick up. Users express a particular emotion by manipulating the doll's hands, legs and head. Different gestures express one of the following emotions: anger, fear, surprise, sadness, gloating and happiness (e.g. hands covering the eyes showing fear). The focus of this study is more on established symbols for particular emotions (e.g. the emotion "sad" is detected when the SenToy is bent forward). However, according to Hudlicka (Hudlicka, 2003), the combination of non-propositional movement qualities, such as those analysed by (Camurri *et al.*, 2003), with symbols reflecting specific emotions, used by (Paiva *et al.*, 2003), appears particularly promising. And that is particularly what we are intending to cover in our research by analysing the affective content of the upper-body gestures.

For this purpose, we first analyse the spatial extent the gesture occurs as conveyer of expressive information. We exploit *proxemics*, the study of humankind's perception and use of space (Hall, 1963). According to its founder, Hall (Hall, 1963), like facial expressions, gestures, and postures, space "speaks". Hall (Hall, 1963) identified four bodily distances-- *intimate* (0 to 18 inches or 0 to 45.72 cm), *personal-casual* (1.5 to 4 feet or 45.72cm to 1.213 m), *social-consultive* (4 to 10 feet or 1.213 m to 3.05 m), and *public* (10 feet and beyond or 3.05 m or beyond)--as key points. Laban (Laban&Lawrence, 1974) also analysed spatial extent and divided the space used by the dancer into two: personal and global. Taking into account these two studies, we carry out our analysis in a spatial extent determined as the "personal space" which covers first two bodily distances defined by Hall (0-4 feet). For instance, in case of a user in front of the computer, the space occupied by his body is considered his "personal space". When analysing expressive gestures we will consider movement of the upper-body (amount of expansion/contraction) and movement of the various body parts (e.g. hands, shoulders) in the "personal space" as shown in Fig. 1a.

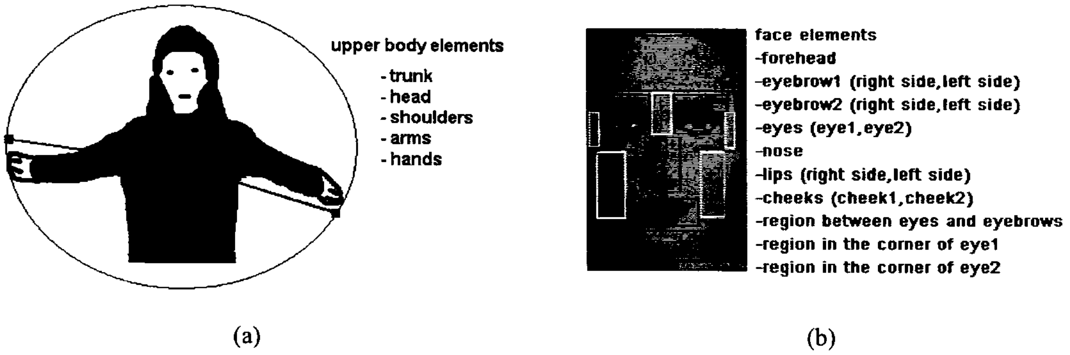

upper body elements
- trunk
- head
- shoulders
- arms
- hands
face elements
-forehead
-eyebrow1 (right side,left side)
-eyebrow2 (right side,left side)
-eyes (eye1,eye2)
-nose
-lips (right side,left side)
-cheeks (cheek1,cheek2)
-region between eyes and eyebrows
-region in the corner of eye1
-region in the corner of eye2
(a)
(b)

Figure 1. (a) Upper-body model in the personal space (b) Facial Model

In order to extract the features in the upper-body movement conveying affective information we exploit "kinesics" that studies behavioural patterns of nonverbal communication and expressive force of gestures and body movement (Givens, 2001). Boone and Cunningham (Boone&Cunningham, 1998) distinguish between propositional and non-propositional aspects of movement. Propositional expressive gestures are described as specific movements of specific local features/bodily parts or postures corresponding to stereotypical emotions (e.g., clenched fists showing anger; bowed head and dropped shoulders showing sadness). Non-propositional expressive gestures are, instead, not coded as specific movements but form the quality of movements by providing descriptions such as how direct/flexible, sudden/sustained, firm/gentle or free/bound a movement is (Boone&Cunningham, 1998).

In our attempt to analyse and recognize expressive upper-body gestures we will base our research on both types of gestures as suggested by Hudlicka (Hudlicka, 2003): propositional and non-propositional. For this purpose, we categorize the body posture and movement of bodily parts such as hands and shoulders under *propositional expressive gestures* and we analyse them using the findings in psychology. The gestures that are not in this category are defined as *non-propositional expressive gestures* and we analyse the quality of non-propositional gestures using the studies of Laban (Laban& Ullmann, 1988).

Propositional expressive gestures/actions: There exist emotion recognition studies in psychology suggesting that it is possible to obtain information about the emotional state of a person by looking at his/her non-facial cues. Wallbott and Scherer (Wallbott&Scherer, 1988) found cross-cultural agreement in self-report on certain emotion movement patterns. De Meijer (DeMeijer, 1989) found relationships between whole body movements and specific emotions. Boone and Cunningham (Boone&Cunningham, 1998) starting from previous studies by De Meijer identified six expressive cues involved in the recognition of the four basic emotions anger, fear, grief, and happiness. The six cues are "frequency of upward arm movement, the duration of time arms were kept close to the body, the amount of muscle tension, the duration of time an individual leaned forward, the number of directional changes in face and torso, and the number of tempo changes an individual made in a given action sequence". Mehrabian (Mehrabian, 1969) identified four postural categories that provide clues about emotional state of the person: forward lean (attentiveness), drawing back or turning away (negative, refusing), expansion (proud, conceited, arrogant), forward-leaning trunk, bowed head, dropped shoulders, and sunken chest (depressed, downcast, dejected). Givens (Givens, 2001) collected the expressive body cues in a non-verbal dictionary. Other studies have also revealed relationships between particular body movements and specific emotions (Clynes, 1975; Scherer *et al.*, 1986; Wallbott&Scherer, 1988; Sogon & Masutani, 1989; Nakamura *et al.*, 1990). In spite of the research conducted, there is not a common trend when analysing the expressive gestures as in the case of facial expression.

Non-propositional expressive gestures/actions: In our work we use non-propositional gestures to define the quality of the expressive body actions by extracting parameters such inwardness/outwardness of the action, contraction/expansion of the body, trajectories (i.e. paths followed during the movement) formed by bodily parts (i.e. straight/smooth), duration of the body action (i.e. long/short) etc. These parameters are based on Laban's theory. Laban (Laban&Lawrence, 1974) in his "Theory of Effort" considers movement as a communication medium and extracts parameters related to its expressive power. Effort describes the *quality of movement* and is manifested in bodily actions through space, time, weight, and flow elements. The Space dimension is considered by measuring to what extent limbs are contracted or expanded in relation to the body centre and how much movements are "direct" (a straight trajectory in direction of the movement, sensation threadlike) or "flexible" (wavy trajectory, in direction and of a movement sensation of pliant extent in space, or a feel of every-whereness) and which direction is influential considering the continuous motion (Boone&Cunningham, 1998, showed that joy is characterized by a higher amount of upward movements). The Time dimension is considered in terms of overall duration of the action and duration of action pauses. In terms of time component an action can be described as "sudden" or "sustained". Weight component is used to measure how much strength and weight (firm/gentle) is exerted in an action and has been mainly associated with the vertical component of acceleration. Flow component is considered in terms of analysis of shapes of speed, pause phases and amount of acceleration/deceleration during action phases to find out how free/bound an action is. According to Laban, bodily attitudes during movement are determined by two main action forms. One of these forms flows from the centre of the body outwards, while the other flows from the periphery of the space surrounding the body inwards to the centre (Laban & Ullmann, 1988). In our study, we use the descriptive elements from Laban's theory by trying to answer the following questions:

- Which part of the body moves? – head /torso/shoulders/hands
- In which direction of space is the movement exerted? – forward/backward/left/right/up/down
- What extent the body is contracted/expanded? –contraction/expansion in relation to the body centre
- What shape the moving parts create? –direct trajectories /smooth trajectories
- How long does the movement take? – the overall duration/ phases
- At what speed does the movement progress? – acceleration/deceleration

In this sense, this part of our work is similar to the work described by (Camurri *et al.*, 2003).

CONCEPTUAL FRAMEWORK

We present our conceptual framework by firstly describing the face and upper-body model we employed for our system. We then describe how we combine the face and upper-body actions in order to relate them to specific emotional states.

Facial Modelling

Most of the facial expression recognition systems use either 3D explicit wire frame model of the face or 2D models with characteristic points (Pantic&Rothkrantz, 2000). Designing a 3D face model to accurately represent facial geometric properties is a difficult task. Initially, the 3D face model is constructed based on the locations of facial feature points that are provided interactively by a human operator (e.g. Cohen *et al.*, 2002). Similarly, most of the facial expression recognition systems that use 2D point-based models do initialisation manually in the first frame (e.g. Tian *et al.* 2001). These initial manual interventions cause most of these systems to be semi-automatic. Most of these approaches use a *feature-based face representation* based on distances, angles, and areas between features such as eyes, nose, or chin. These parameters are extracted from full facial views or profiles based on facial feature changes, and are stored as a feature vector prior to expression classification (e.g. Pantic&Rothkrantz, 2000). There exist some other approaches that analyse motion from image sequences using *optical flow*. *Optical flow method* estimates motion vectors as an array of points over the face, which is called the *flow field*. It is usually represented by an array of arrows (or flow vectors) indicating direction and magnitude of the displacement at each image location.

We choose to define our model as a frontal-view face model that consists of feature bounding rectangles called ROIs (region of interest) as illustrated in Fig.1b. In this sense our model is similar to the model defined by Colmenarez *et al.* In their model the appearance of each facial feature is provided by the image sub-window located around its position and the model consists of four facial regions (eyebrows, eyes and nose-mouth) and nine facial features (upper eyebrows, lower eyebrows, eyes corners, tip of the nose and lip corners) (Colmenarez *et al.*, 1999). However, our facial model is more complex because of the fact that we not only include the forehead and the chin as additional features, but also combine them with salient feature regions such as the cheeks. Hence, we define a hybrid model combining feature-based and motion-based representations with eight features, eleven permanent feature regions and additionally, five salient feature regions. We first locate the eight facial features namely eyes, eyebrows, lips, nose, chin and forehead automatically in the neutral frame. We then define the bounding rectangle for each of these features shown with black rectangles in Fig. 1a. Additionally, we divide the eyebrow region to two separate regions namely upper eyebrow and lower eyebrow region. We also divide the lip region to two separate regions namely right lip and left lip regions. This is done to obtain better analysis results when attempting to recognize FAUs since the movement of the upper eyebrow is a different FAU from the movement of the lower eyebrow (AU1 vs. AU4) (Ekman & Friesen, 1978). Hence, we define eleven permanent feature rectangles: forehead, upper and lower eyebrows, eyes, nose, left and right lip region and chin. We also define five bounding rectangles for salient features located between stable features namely, the region between eyes and eyebrows, corner of right eye, corner of left eye, right cheek and left cheek. These are shown with white rectangles in Fig. 1b.

Hereby, we briefly describe how we use this face model in our system. In our work, we explore a middle ground between feature based and optical flow based approaches. We first compute locally averaged optical flow within each ROI (e.g. forehead, right eye, and left lip region) and we name this procedure *ROI based optical flow*. We also compare the texture within the ROI of the salient features in the expressive frame with respect to the neutral frame. We do that by calculating the texture ratio (*T*) for each region by dividing the texture in the expressive frame within the given region (*Te*) by the texture in the neutral frame within the given region (*Tn*) (*T=Te/Tn)*. We then compare this ratio against a threshold that is obtained with experimental results.

If (T> threshold) furrow present; else furrow absent

The changes in the movement of the permanent regions and salient regions are observable from the results obtained from optical flow and texture analysis, respectively.

Body Modelling

Currently, there are not many bodily expression recognition systems. Analysis of body movements that convey expressive information is a challenging task due to the fact that body motion contains a higher degree of flexibility than facial motion. Camurri *et al.* (Camurri *et al.*, 2003) uses full body model to analyse expressive cues of a dancer. Using a full body model is a logical step if the task is to recognize expressive cues from dance motions since dancers use all the bodily parts and perform on the stage; hence usage of space tends to be larger.

However, our task is to analyse expressive cues within HHI and HCI which mostly takes place as dialogues in sitting position, hence, the expressiveness of the lower part of the body is ignored in this paper.

We choose to define our upper-body model as a frontal-view model that consists of eight bodily features namely, face, torso, shoulders, hands and arms (Fig.1a) also by taking into account the body model defined by Laban (Laban&Ullmann, 1988). For the sake of simplicity, currently we do not consider elbow and wrist in our model, however further work is being conducted on distinguishing these parts as individual body features. In our system we use this model by first locating the face and the hands using skin colour. We then calculate the centre of these regions namely, centre of the upper-body (Bx,By), centre of the face (Fx,Fy); centre of the right hand (HRx,HRy) and centre of the left hand (HLx,HLy). Our motivation at choosing this model is that the changes in the position of the centre of each feature are directly observable and trackable due to their distinguishing colour. Moreover, this model avoids manual initialisation.

BIMODAL MODELLING OF EMOTION WITH FAUs AND BAUs

Since our aim is designing a multimodal system that will recognize emotions conveyed in expressive facial and bodily gesture, hereby we create an emotion model for combining Facial Action Units (FAUs) and Body Action Units (BAUs) in a bimodal manner.

Description of Facial Action Units (FAUs) and Body Action Units (BAUs)

We described our face and body model in the previous section. When coupled together, the two models reveal significant information about emotional expression. For instance, in Fig. 2, if we only consider the facial information the emotion revealed, it could be described as *sadness*. However, when we look at the whole body and consider the expressiveness of the face and body jointly coupled together, we could describe the affective state of the person as *puzzled*. Therefore, when coupled together, the two models do not contain redundant information about the affective state of the person. These two models are used further to encode affective states in terms of FAUs and BAUs.

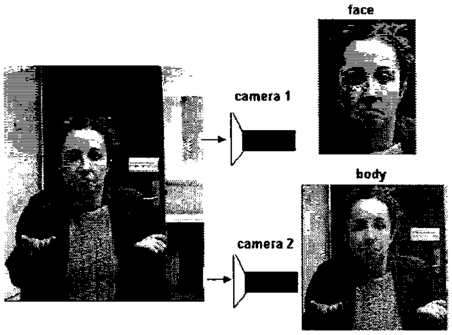


Figure 2. Our system settings

FAUs were defined exhaustively by Ekman&Friesen in FACS (Ekman&Friesen, 1978) so we will not discuss them here in detail. We will simply give an example of how each emotion is defined in terms of FAUs. For instance, surprise is defined in the FACS as a combination of four FAUs:

Surprise = {FAU 1}+ {FAU 2}+ {FAU 5}+ {FAU 26}; OR *{FAU 1}+ {FAU 2}+ {FAU 5}+ {FAU 27};*

(FAU 1: Inner Brow Raised; FAU2: Outer Brow Raised; FAU5B: Upper Lid Raised; FAU26: Jaw Dropped; FAU27: Mouth Stretched). The emotion surprise is defined to be additive of these FAUs.

For further description of each emotion in terms of FAUs see (Ekman&Friesen, 1978).

To date, BAUs that carry expressive information have not been defined with a Body Action Coding System. For this reason, in this paper, we employ the body movements that carry expressive information and call them Body Action Unit (BAUs) to create a Body Action Coding System (BACS). Table 1 provides a list of the defined BAUs in terms of features grouped under specific emotion categories taking into account the psychological framework (see the section on Theoretical Background) together with the results obtained from our experiments.

Table 1. Association between expressive movements/postures and conveyed basic emotions based on references in psychology. The table is strongly rooted on the literature in psychology together with the results obtained from our experiments.

Emotional state	BAU	Propositional BAUs	Non-propositional cues
Happiness	1 9	open/expanded body arms raised or away from the body	movements reaching out from body centre frequency of arms up duration of arms away from torso
Sadness	2 3 6 8 20b 12	contracted/closed body body shift- forward leaning trunk bowed head dropped shoulders covering the face with two hands self-touch (disbelief)/ covering the body parts/ arms around the body/shoulders	long duration of time slow movement duration of time body leaning forward
Surprise	4 20c 20d	body shift- backing self - touch two hands covering only the cheeks self - touch two hands together covering the mouth	
Fear	2 4 12 20b 20d 14	contracted/closed body body shift- backing self-touch (disbelief)/ covering the body parts/ arms around the body/shoulders self-touch (disbelief) covering the face with hands self touch-two hands together covering the mouth closed hands / clenched fist	long stops between changes movements kept close to body centre
Anger	1 14 16 17 18	open/expanded body closed hands / clenched fist palm-down gesture hands-on-hips finger point	short duration of time short stops between change movements reaching out from body centre directional changes in face and torso
Disgust	4 5	body shift- backing body shift/orientation changed/moving to the right or left	directional changes in face and torso
Uncertainty	11 7 15 20(a-g) 13	shoulder-shrug head-tilt-side palm-up gesture self-touch gestures (head/face/cheek/ mouth / forehead / ear /chin/) hands behind the head	side-to-side head shakes
Anxiety	20f 20d 19	self-touch ear with one hand self touch-two hands together covering the mouth hands pressed together in a moving sequence	
Frustration	18 14 10 20f 16	finger point closed hands / clenched fist crossing the arms self-touch ear with one hand palm-down	
Negative rejection refusing	2 10 5 13	contracted/closed body crossing the arms body shift/orientation changed/moving to the right or left hands behind the head.	directional changes in face and torso

An Experiment on Human Emotion Recognition from BAUs

We aim to relate the emotion categories to single BAU or combination of BAUs ({*Emotion→BAUs*} relationship). For this purpose, we performed an experiment with ten judges (five female, five male) from different ethnical backgrounds to find out how consistently humans relate certain emotions to specific body gestures. The judges were presented with nineteen short movies of one person performing a single BAU or combination of BAUs with either hands/arms/body or shoulder, and with no faces available (see Table 2). We recorded the test sequences simultaneously using two fixed cameras connected to two different PC's with a simple setup and uniform background (see Fig. 2). One camera was placed in front of the person capturing the head only and the second camera was placed further away from the body in order to capture upper-body movement. We choose to use two cameras due to the fact that current technology still does not provide us with frames with required quality to process detailed upper-body and facial information. We then asked each participant to map the viewed video to one of the emotional categories provided in the list with descriptions for each emotion category.

Initially, the emotions elicited during our experiments were the seven prototypic expressions (Ekman& Friesen, 1968). However, with the feedback provided by the participants we decided to add four additional emotion categories since the seven emotions were not found descriptive enough by the participants. Therefore, the emotions were changed from seven to eleven, namely: neutral, happiness, sadness, surprise, fear, anger, disgust, uncertainty, anxiety, frustration and rejection. Our experimental results (see Table 2) show that there was significant agreement about what emotion each gesture represented in rows numbered as 1-5,10,13,15,18 and 20 (Table 2). On the other hand, there was a little less agreement about what emotion each gesture represented for rows numbered as 6-9,11,12 14, 16,17 and 19. However, this result is normal and expected; the lack of consistent results from the ten participants is part of our thesis about why bimodal analysis is required. We conclude that humans are capable of mapping certain expressive body gestures, performed alone or in combination, to specific emotions; thus providing the ground truth to build a computer system that can analyse face and upper-body gesture in a bimodal manner. Collecting data to suggest that body gesture added to facial gesture analysis will improve understanding of emotion is part of our future experiment. How to measure these AUs automatically is also not in the scope of this paper but part of our future paper.

Table 2. Survey results on human mapping of Body Action Units (BAUs) to emotions. Results in the table prove that humans are capable of mapping certain expressive body gestures to specific emotions.

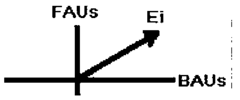
		mapped emotion →	happiness	sadness	surprise	fear	anger	disgust	uncertainty/ confusion/ puzzlement/thinking	frustration	anxiety	negative rejecting refusing	other
row#	BAUs shown	description of BAUs ↓	the number of people categorising BAUs shown as indicator of the specific emotion										
1	1+9+14	expanded body+arms raised aside the body+with clenched fists	8		1		1						
2	1+17	expanded body+hands on hips					8					2	
3	2+12	contracted body+arms around the shoulders		2		7					1		
4	6+8	bowed head + dropped shoulders		10									
5	11+15	shoulder shrug+ palms up			2				8				
6	13+20a	moving hands from body center to behind head+ touching the head			1				5			2	2
7	4	body shift- backing			5	5		5					
8	5	body shift- moving to the reight or left					1	2				5	2
9	10	crossing the arms						1	1	2		6	
10	14	clenched fists in continous sequence	1				7			2			
11	16	palms down					5				1	4	
12	18	finger point in a moving sequence					5			2		3	
13	19	hands pressed together in amoving sequence								1	9		
	20	self touch gestures(with one or two hands)											
14	20a	touching the head											
15	20b	covering the face (two hands)		2		7				1			
16	20c	covering the cheeks (two hands)		1	5	1			1	2			
17	20d	covering the mouth	1	1	7	7					1		
18	20e	touching the forehead							9	1			
19	20f	touching the ear							3	1	3		3
20	20g	touching the chin							10				

Bimodal Model

In our system, we need to measure FAUs and BAUs and relate them to emotion categories to recognize a particular emotion (*{BAUs→ Emotion}* relationship). In order to provide a model on how to map the AUs to certain emotion categories we define an *emotion space* by describing each emotion as a linear function of FAUs and BAUs.

$$Ei = F\,(FAUs,\ BAUs)\ ;$$

We define eleven emotional states obtained with experimental results, namely: neutral, happiness, sadness, surprise, fear, anger, disgust, uncertainty, anxiety, frustration, rejection (E0=Neutral; E1=Happiness; E2=Sadness; E3=Surprise; E4=Fear; E5=Anger; E6=Disgust; E7=Uncertainty/confusion/puzzlement; E8=Anxiety; E9=Frustration; E10= Negative/refusing/rejecting). Each emotion *Ei* consists of weighted *FAUs* and *BAUs* according to an estimated importance and has a unique position within the emotion space as shown in Fig. 3.



Ei=F (FAUs, BAUs)
Ei=ΣW(j)*FAU(j) + ΣW(k)*BAU(k);

Figure 3. Illustration of the Ei function defined as a combination of FAUs and BAUs.

Weight *W* can vary as having a value of either 0 or 1 when defining each emotion category. For instance, if an emotion is not displayed with FAU(1), the weight for FAU(1) can be defined as W(1)=0.

$$e.g.\ \ Ei=0*FAU(1)+1*FAU(2)+0*BAU(1)+1*BAU(5);$$

Weights can be taken into account for normalization of the AUs. The value of the *Expression Function=Ei* varies between 0 and 1 indicating the trustiness/intensity of the emotion from the FAUs and BAUs extracted from the frames: $0<=Ei<=1.$ Given a threshold, we can also define a rule-based decision-making based on the value of Ei.

If (Ei >threshold) true/ real/ intense emotion; else uncertain/ slight/ moderate emotion

In our bimodal emotion model we combine the FAUs and BAUs to recognize the emotion displayed. For instance, for the emotion *anger* we first check the body posture and propositional gestures since they can be detected independently in each frame, where as non-propositional gestures need to be tracked over time and recognized after the gesture has ended. Having determined the expressive propositional gesture, we analyse the face. Then we combine information from two sources to recognize the emotion. In this sense, we apply late fusion of the multmodal information provided from two cameras and expressive sources.

Variations from prototypical expressions are explained as reflecting the elicitation of blends of the fundamental emotions (Ekman&Friesen, 1978). For instance there exist blend of expressions corresponding to different states: sad-angry, angry-afraid, dazed-surprise, questioning-surprise etc. In this context, our bimodal model will also help to recognize blends of the fundamental emotions since we describe our model as combinations of AUs from face and/or body. Hence, it is easier to figure out which AU is absent or present prior to decision making. For instance, in a case where the face is showing *surprise* and the body is reflecting *happiness*; if only the extracted FAUs are used the emotion would be interpreted as *surprise*, however, looking at the body gesture the emotion would be interpreted as *happiness*, hence, we will be able to classify the overall emotion "happy-surprise". BAUs can be considered as the auxiliary mode in this case, when emotion is a blend of various emotions or has variations within itself.

CONCLUSION

This paper presented a bimodal model of facial and upper-body gesture for affective HCI suitable for use in a vision-based multimodal system. Our experimental results show that humans are capable of mapping certain expressive body gestures, performed alone or in combination, to specific emotions. This provides the ground truth to build a computer system that can analyse face and upper-body gesture in a bimodal manner to recognize particular emotions. The model proposed in this paper is currently being used in our vision-based bimodal system that extracts features automatically from expressive face and upper-body display. Our work is an attempt towards recognizing expressive face and upper-body gesture for affective multimodal HCI. Our future work will provide extensive results from our bimodal interface.

REFERENCES

Bassili, J. N. (1978) " Facial motion in the perception of faces and of emotional expression", *J. Experimental Psyschology,* 4, 373-379.

Boone, R.T. and Cunningham, J.G. (1998) Children's decoding of emotion in expressive body movement: The development of cue attunement, *J. Developmental Psychology*, 48, 32-44.

Camurri, A., Lagerlof, I., Volpe, G. (2003) Recognizing Emotion from Dance Movement: Comparison of Spectator Recognition and Automated Techniques, *Int. J. of Human-Computer Studies,* 59(1- 2) , 213-225.

Chen, L. S. and Huang, T. S.(2000) Emotional expressions in audiovisual human computer interaction, *Proc. IEEE ICME*, 423–426.

Clynes, M. (1975) Communication and generation of emotion through essentic form. In L.Levi(Ed.), *Emotions: Their parameters and measurement*, New York:Raven Press.

Cohen, I., Sebe, N., Garg, A. and Huang ,T.S. (2002) Facial expression recognition from video sequences. *Proc. IEEE ICME.*

Colmenarez, A., Frey, B., Huang, T.S. *(1999)* A probabilistic framework for embedded face and facial expression recognition, *Proc. of IEEE CVPR*, Vol. 1, pp. 597.

Damasio, A. (1994) Descartes' Error: Emotion, Reason, and the Human Brain. Gosset/Putnam Press, New York.

De Meijer, M. (1989) The contribution of general features of body movement on the attributions of emotions, *J. of NonVerbal Behavior,* 13, 247-268.

De Silva, L. C. and Ng, P. C. (2000) Bimodal emotion recognition, *Proc. FG*, 332–335.

Donato, G., Bartlett, M., Hager, J., Ekman, P. and Sejnowski, T. (1999) Classifying facial actions, *IEEE PAMI*, 21(10), 974-989.

Ekman, P. and Friesen, W. V. (1978), *The Facial Action Coding System: A Technique for Measurement of Facial Movement*, Consulting Psychologists Press, San Francisco, CA.

Ekman, P. and Friesen, W. V. (1968), Nonverbal Behavior in Psychotherapy Research, In John Shlien (Ed.), *Research in Psychotherapy* (Washington, D.C.: American Psychological Association) , 179-216.

Givens, David B. (2001) The Nonverbal Dictionary of Gestures, Signs & Body Language Cues, *Nonverbal Communication*, 2nd Ed., Martin Remland, Houghton Mifflin Co.

Hall, E.T. (1963) A System for the Notation of Proxemic Behavior, *American Anthropologist,* 65, 1003-1026.

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

Isen, A.M. (2000) Positive Affect and Decision Making. In: Lewis, M., Haviland, J. (Eds.), Handbook of Emotions. Guilford, New York.

Laban , R. and Lawrence, F.C. (1974) *Effort* (2nd Edition) MacDonald and Evans, London.

Laban, R. and Ullmann, L.(1988) The Mastery of Movement, Northcote House Educational Publishers.

Lisetti, C. L. and Schiano, D. (2000) Facial Expression Recognition: Where Human-Computer Interaction, Artificial Intelligence, and Cognitive Science Intersect. *Pragmatics and Cognition,* Vol. 8(1): 185-235.

Massumoto, D. and Kudoh, T (1985) Cross-cultural examination of semantic dimensions of body postures, *J. of Personality and Social Psychology,* 55, 36-42.

Mehrabian, A. (1969) Significance of Posture and Position in the Communication of Attitude and Status Relationships, *Psychological Bulletin,* 71, 359-72 .

Mehrabian, A. (1968) Communication without words, *Psychol. Today*, 2(4), 53–56.

Nakamura, M., Buck, R., Kenney, D.A. (1990) Relative contribution of expressive behavior and contextual information on the judgement of the emotional state of another, *J. of Personality and Social Psychology*, 59, 1032-1039.

Paiva, A.*et al.*(2003) SenToy: an affective sympathetic interface, *Int. J. Hum.-Comput. Stud.*, 59(1-2), 227-235.

Pantic, M. and Rothkrantz L.J.M.(2000) Automatic analysis of facial expressions: the state of the art, *IEEE PAMI,* 22 (12), 1424-1445.

Pantic, M. and Rothkrantz, L.J.M. (2004) Facial Action Recognition for Facial Expression Analysis From Static Face Images, *Proceedings of the IEEE Ttransactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 34(3), 1449-1461

Picard, R. W. (1997), Affective Computing, MIT Press, Cambridge.

Picard, R.W. (2000) Towards computers that recognize and respond to user emotion, *IBM Systems Journal,* 39 (3–4), 705–719.

Picard, R. W. (2003) Affective computing: challenges, *Int. J. Human-Computer Studies*, 59 , 55–64

Richmond, V.P. *et al.* (1991). *Nonverbal Behavior in Interpersonal Relations* (2nd Ed., Englewood Cliffs, New Jersey: Prentice Hall).

Scherer, K.R. *et al.* (1986) Emotional experience in cultural context: A comparison between Europe, Japan and United States, In K.R. Scherer(Ed.), *Facets of emotions*, Hillsdale, NJ:Erlbaum, 5-30.

Scherer, I.R. *et al.* (2001) Appraisal Processes in Emotion:Theory, Methods, Research. Oxford University Press, Oxford.

Sogon, S. and Masutani, M. Identification of emotion from body movements: A cross-cultural study of Americans and Japanese, *Psychological Reports*, 65, 35-46.

Tian,Y. *et al.* (2001) Recognizing action units for facial expression analysis, *IEEE PAMI*, 23(2)

Wallbott, H. and Scherer, K.R.(1988) How universal and specific is emotional experience? Evidence from 27 countries on five continents, *Social Science Information*, 25, 763-795.

COPYRIGHT