Observer Annotation of Affective Display and Evaluation of Expressivity: Face vs. Face-and-Body

Hatice Gunes and Massimo Piccardi

Faculty of Information Technology, University of Technology, Sydney (UTS)
P.O. Box 123, Broadway 2007, NSW, Australia

{haticeg, massimo} @ it.uts.edu.au

Abstract

A first step in developing and testing a robust affective multimodal system is to obtain or access data representing human multimodal expressive behaviour. Collected affect data has to be further annotated in order to become usable for the automated systems. Most of the existing studies of emotion or affect annotation are monomodal. Instead, in this paper, we explore how independent human observers annotate affect display from monomodal face data compared to bimodal face-and-body data. To this aim we collected visual affect data by recording the face and face-and-body simultaneously. We then conducted a survey by asking human observers to view and label the face and face-and-body recordings separately. The results obtained show that in general, viewing face-and-body simultaneously helps with resolving the ambiguity in annotating emotional behaviours.

Keywords: Affective face-and-body display, bimodal affect annotation, expressivity evaluation.

1 Introduction

Affective computing aims to equip computing devices with the means to interpret and understand human emotions, moods, and possibly intentions without the user's conscious or intentional input of information—similar to the way that humans rely on their senses to assess each other's state of mind. Building systems that detect, understand, and respond to human emotions could make user experiences more efficient and amiable, customize experiences and optimize computer-learning applications.

Over the past 15 years, computer scientists have explored various methodologies to automate the process of emotion/affective state recognition. 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 (Pantic et al., 2005). Relatively few works have focused on implementing emotion recognition systems using affective multimodal data (i.e. affective data from multiple channels/sensors/modalities). While it is true that the face is the main display of a human's affective state, other sources can improve the recognition accuracy. Emotion recognition via body movements and gestures has recently started attracting the attention of computer science and human-computer interaction (HCI) communities (Hudlicka, 2003). The interest is growing with works similar to these presented in (Balomenos et al., 2003), (Burgoon et al., 2005), (Gunes and Piccardi, 2005), (Kapoor and Picard, 2005) and (Martin et al., 2005).

A first step in developing and testing a robust affective multimodal system is to obtain or access data representing human multimodal expressive behaviour. The creation or collection of such data requires a major effort in the definition of representative behaviours, the choice of expressive modalities, and the labelling of large amount of data. At present publicly-available databases exist mainly for single expressive modalities such as facial expressions, static and dynamic hand postures, and dynamic hand gestures (Gunes and Piccardi, 2006b). Only recently, a first bimodal affect database consisting of expressive face and face-and-body display has been released (Gunes and Piccardi, 2006a).

Besides acquisition, another equally challenging procedure is their annotation. Multimodal data have to be annotated in order to become usable for the automated systems.

Most of the experimental research that studied emotional behaviours or affective data collection focused only on single modalities, either facial expression or body movement. In other words, the amount of information separate channels carry for recognition of emotions has been researched separately (explained in detail in Related Work section). There also exist several studies that involve multimodal annotation specific to emotions. However, none of the studies dealing with multimodal annotation specific to emotion compared how independent human observers' annotation is affected when they are exposed to a single modality versus multiple modalities occurring together. Therefore, in this paper, we conduct a study on whether seeing emotional displays from the face camera alone or from the face-and-body camera affects the independent observers' annotations of emotion. Our investigation

focuses on evaluating monomodal versus bimodal posed affective data. Our aim is to use the annotations and results obtained from this study to train an automated system to support unassisted recognition of emotional states. However, creating, training and testing and affective multimodal system is not the focus of this paper.

2 Related Work

2.1 Emotion Research

In general, when annotating affect data two major studies from emotion research are used: Ekman's theory of emotion universality (Ekman, 2003) and Russell's theory of arousal and valence (Russell, 1980).

Ekman conducted various experiments on human judgement on still photographs of posed facial behaviour and concluded that seven basic emotions can be recognized universally, namely, neutral, happiness, sadness, surprise, fear, anger and disgust (Ekman, 2003). Several other emotions and many combinations of emotions have been studied but it remains unconfirmed whether they are universally distinguishable.

Other emotion researchers took the dimensional approach and viewed affective states not independent of one another; rather, related to one another in a systematic manner (Russell, 1980). Russell argued that emotion is best characterized in terms of a small number of latent dimensions, rather than in a small number of discrete emotion categories. Russell proposed that each of the basic emotions is a bipolar entity as part of the same emotional continuum. The proposed polarities are arousal (relaxed vs. aroused) and valence (pleasant vs. unpleasant). The model is illustrated in Figure 1.

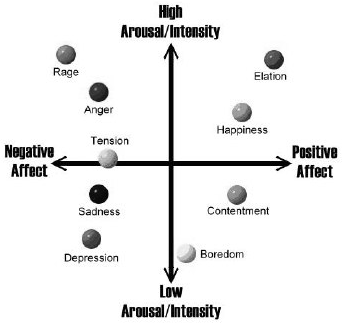

Figure 1. Illustration of Russell's circumflex model.

2.2 Affective multimodal data collection

All of the publicly available facial expression or body gesture databases collected data by instructing the subjects on how to perform the desired actions (please see (Gunes and Piccardi, 2006b) for an extensive review of publicly available visual affect databases).

2.3 Affective multimodal data annotation

Hereby, we review studies that deal with human annotation of non-verbal emotional behaviour. This review is intended to be illustrative rather than exhaustive. We do not review studies on human labelling and recognition of emotions from face expressions, as they have been extensively reviewed by Ekman (Ekman, 1982; Ekman, 2003).

In (DeMeijer, 1991), the authors studied the attribution of aggression and grief to body movements. Three parameters in particular were investigated: sex of the mover, sex of the perceiver, and expressiveness of the movement. Videos of 96 different body movements from students of expressive dance were shown to 42 adults. The results showed that the observers used seven dimensions for describing movements: trunk movement (stretching, bowing), arm movement (opening, closing), vertical direction (upward, downward), sagittal direction (forward, backward), force (strong-light), velocity (fast-slow), directness (moving straight towards the end-position versus following a lingering, s-shaped pathway). The results of this study revealed that form and motion are relevant factors when decoding emotions from body movement.

In another study on bodily expression of emotion, Wallbott recorded acted body movements for basic emotions (Wallbott, 1998). Twelve drama students were then asked to code body movement and posture performed by actors. The results revealed that the following factors appeared to be significant in the coding procedure: position of face-and-body, position of shoulders, position of head, position of arms, position of hands, movement quality (movement activity, spatial expansion, movement dynamics, energy, and power); body movements (jerky and active), body posture.

In (Montepare et al., 1999), the authors conducted an experiment on the use of body movements and gestures as cues to emotions in younger and older adults. They first recorded actors doing various body movements. In order to draw the attention of the human observers to the expression of emotions via body cues, the authors electronically blurred the faces and did not record sound. In the first part of the experiments, the observers were asked to identify the emotions displayed by young adult actors. In the second part of the experiment, the observers were asked to rate the actors' displays using characteristics of movement quality (form, tempo, force, and direction) rated on a 7-point scale and verbal descriptors (smooth / jerky, stiff / loose, soft / hard, slow / fast, expanded / contracted, and no action / a lot of action). Overall, observers evaluated age, gender and race; hand position; gait; variations in movement form, tempo, and direction; and movement quality from actors' body movements. The ratings of both younger and older groups had high agreement when linking particular body cues to emotions.

Coulson presented experimental results on attribution of six emotions (anger, disgust, fear, happiness, sadness and surprise) to static body postures by using computer-generated figures (Coulson, 2004). He found out that in general, human recognition of emotion from posture is comparable to recognition from the voice, and some postures are recognized as well as facial expressions.

All of the aforementioned studies focused on individual modalities such as facial expression or body movement; therefore, in this paper we focus on bimodal data.

There exist several studies that involve multimodal annotation specific to emotions. The Belfast naturalistic

database contains emotional interviews annotated with continuous dimensions (Douglas-Cowie et al., 2003). In (Allwood et al., 2004) authors designed a coding scheme for the annotation of 3 videos of TV interviews. Facial displays, gestures, and speech were coded using the following parameters: form of the expression and of its semantic-pragmatic function (e.g. turn managing) and the relation between modalities: repetition, addition, substitution, contradiction. (Martin et al., 2005) also designed a coding scheme for annotating multimodal behaviours during real life mixed emotions. They first collected emotionally rich TV interviews. Then they focused on the annotation of emotion specific behaviours in speech, head and torso movements, facial expressions, gaze, and hand gestures. They grounded their coding scheme on the following parameters: the expressivity of movements, the number of annotations in each modality, their temporal features (duration, alternation, repetition, and structural descriptions of gestures), the directions of movements and the functional description of relevant gestures.

The materials collected by (Martin et al., 2005) are useful multimodal data for research in affective multimodal HCI. However, as annotation in itself is challenging and ambiguous, we believe that the annotation should be done more systematically than just one observer. Moreover, the annotation in (Martin et al., 2005) focused more on actions and expressions rather than emotions.

Therefore, in this paper, we explore whether seeing emotional displays from the face camera alone or from the face-and-body camera affects the independent observers' annotations of emotion.

3 Study

In this study we are seeking answers to the following research questions.

- How do humans perceive emotions from face modality alone compared to the combination of face-and-body modalities that occur simultaneously?
- Does being exposed to the expressions from one sensor (face camera only) or from multiple sensors simultaneously (viewing face-and-body combined) affect the observers' interpretations and therefore, labelling differ (monomodal vs. bimodal)?
- Does the use of multiple modalities help simplify the human affect recognition or on the contrary makes it more complicated? Does it help with resolving ambiguity or the addition of another modality increases ambiguity?

3.1 The data set

The data set we used for this study consists of recordings of combined face and body expressions. According to five factors that were defined by Picard in (Picard et al., 2001) as influencing the affective data collection, the data we collected are: posed, obtained in laboratory settings, with an emphasis on expression rather than feelings, openly recorded and obtained with an emotion purpose. This is consistent with the characteristics of most of the available face and body gesture databases (Gunes and Piccardi, 2006b).

We recorded the video sequences simultaneously using two fixed cameras with a simple setup and uniform background. One camera was placed to specifically capture the face only and the second camera was placed in order to capture face-and-body movement from the waist above. Prior to recordings subjects were instructed to take a neutral position, facing the camera and looking straight to it with hands visible and placed on the table. The subjects were asked to perform face and body gestures simultaneously by looking at the facial camera constantly. The recordings were obtained by using a scenario approach that was also used in previous emotion research (e.g. Wallbott and Scherer, 1986). In this approach, subjects are provided with situation vignettes or short scenarios describing an emotion eliciting situation. They are instructed to imagine these situations and act out as if they were in such a situation. In our case the subjects were asked what they would do when "it was just announced that they won the biggest prize in lottery" or "the lecture is the most boring one and they can't listen to it anymore" 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 (Gunes and Piccardi, 2006a) for details). Fig. 2 shows representative images obtained simultaneously by the body and face cameras.

3.2 The annotation method

Once the multimodal data are acquired, they need to be annotated and analysed to form the ground truth for machine understanding of the human affective multi modal behaviour. Annotation of the data in a bimodal/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. Hence, obtaining the emotion- and quality-coding for all the visual data contained in bimodal databases is extremely tedious and very difficult to achieve.

We obtained the annotation of our visual multimodal data (each face and body video separately) by asking human observers to view and label the videos. The purpose of this annotation was to obtain independent interpretations of the displayed face and body expressions; evaluate the performance (i.e. how well the subjects were displaying the affect they intended to communicate using their face and bodily gesture) by few human observers from different ethnic and/or cultural background.

To this aim, we developed a survey for face and body videos separately, using the labelling schemes for emotion content and signs. We used two main labelling schemes in line with the psychological literature on descriptors of emotion: (a) verbal categorical labelling (perceptually determined, i.e. happiness) in accordance with Ekman's theory of emotion universality (Ekman, 2003) and (b) broad dimensional labelling: arousal/activation (arousal–sleep/ activated -deactivated) in accordance with Russell's theory of arousal and valence (Russell, 1980). The participants were first shown the whole set of facial videos and only after

finishing with the face they were shown the corresponding body videos. For each video they were asked to choose one label only, from the list provided: sadness, puzzlement/thinking, uncertainty/"I don't know", boredom, neutral surprise, positive surprise, negative surprise, anxiety, anger, disgust, fear, happiness.

For the face videos the procedure was as follows. We asked each participant to select labels for the numbered videos they were shown. When they had difficulty choosing a label they were encouraged to guess. Secondly, we asked each participant to choose a number between 1 and 10 as to how well the emotion is displayed (1 indicating "low" and 10 indicating "high" quality in the expressiveness).

For the body videos the procedure was as follows. We asked each participant to select labels for the numbered videos they were shown. When they had difficulty choosing a label again they were encouraged to guess. Secondly, we asked each participant to choose a number between 1 and 10 as to (a) how fast or slow the motion occurs in the display (i.e. movement speed): 1 indicating "very slow" and 10 indicating "very fast"; (b) how the movement causes the body's occupation of space in the display (i.e. movement in space): 1 indicating "very contracted/very less space coverage" and 10 indicating "very expanded/a lot of space coverage" during the movement; and (c) how powerful/energetic the movement displayed is (i.e. movement dynamics): 1 indicating "almost no action" and 10 indicating "a lot of action" in the movement.

360 face and 360 face-and-body videos were annotated in this way and results analysed.

3.3 Participants

We chose videos from 15 subjects and divided them based on the subjects into three sub-sets to make the annotation procedure easier. Eventually, the first sub-set contained 124 face and 124 body videos from five subjects and was viewed and annotated by six observers: Bulgaria (1), Turkey (1), Mexico (1), Pakistan (1), Czech Republic (1), and Australia (1). The second sub-set contained 120 face and 120 body videos from other five subjects and was viewed and annotated by six observers. Observers were from the following countries: Bulgaria (1), Turkey (1), Czech Republic (2), Slovakia (1), and China (1). The third sub-set contained 116 face and 116 body videos from other five subjects and was viewed and annotated by six observers: Bulgaria (1), Turkey (1), Czech Republic (2), Brazil (1), and China (1).

3.4 Results

For each video, all labelling provided by the six observers was analysed and the emotion category that received the highest vote as unique was used to finalize the true label of that particular video, thus, the ground truth. The display from certain subjects can be classified to a particular emotion category almost unambiguously (i.e. all six observers agree that the display is of one particular emotion category), which implies that these actors produced rather stereotyped movements irrespective of the emotion to be encoded. The classification results for other actors are observed to be more ambiguous (i.e. not all six observers agree that the display is of one particular emotion category). For face videos, "quality of expressiveness" was obtained by averaging the six quality votes provided by the observers. For body videos, results for "movement speed", "movement in space", and "movement dynamics", were similarly obtained by averaging the six votes provided by the observers.

According to the results obtained from both face and face-and-body video annotation:

- 295 out of 360 videos were labelled using the same emotion label both for the face videos and for the face-and-body videos. 65 videos were labelled differently.
- 140 out of 360 videos have more agreement for the face-and-body video than the face video alone.
- 125 out of 360 videos have same level of agreement for the face-and-body video and the face video alone.
- 95 out of 360 videos have more agreement for the face video only than the face-and-body video.

3.4.1 Results for Face Videos

The details of the independent observer agreement for face videos are presented in Table 1.

criterion	# of videos for face	# of videos for face-and-body
Higher than 3 votes	292	311
Higher than 4 votes	200	234
Higher than 5 votes	114	139
Equal to 6 votes	84	118

Table 1. The details of the independent observer agreement for face and face-and-body videos: number of videos complying with the criterion.

The emotion categories that caused more cross-confusion when labelling the face data are *puzzlement* and *anxiety*. This can be due to the fact that both emotions were expressed with similar facial displays (e.g. lip bite). Viewing face-and-body display together almost immediately helped the observers resolve their ambiguity. This in turn suggests that if physical displays for certain emotions are similar, and no specific, discriminative movement indicators exist, in independent observer labelling, these emotion displays are commonly found to be confused with one another.

When expressivity of the face videos was analysed it was found that the videos that did not have high agreement in terms of emotion labelling not necessarily were rated low in terms of expressivity. In other words, an observer rated the expressivity of the face display assuming that the person was expressing the emotion s(he) thought was the true emotion displayed.

3.4.2 Results for the Combined Face-and-Body Videos

The details of the independent observer agreement for face-and-body videos are presented in Table 1. Further results from the face-and-body video annotation are presented in Tables 2-4.

According to the results presented in Table 1 we conclude that full agreement is achieved more frequently when face-and-body are viewed together (118 compared

to 84). The results provided in Table 2 and Table 3 suggest that the emotion with lowest movement speed, least movement in space and least movement dynamics in space are *sadness*, followed by *puzzlement*, *anxiety*, and *uncertainty*.

Emotion	Average movement speed	Average movement in space	Average movement dynamics
sadness	3.89	4.63	5.11
puzzlement	4.23	4.55	4.96
uncertainty	4.55	4.53	4.72
boredom	4.56	4.80	5.25
surprise	4.83	5.01	5.29
anxiety	4.98	3.73	4.83
anger	5.32	4.76	5.46
disgust	5.41	4.62	5.61
fear	6.05	4.85	6.21
happiness	6.13	5.54	6.32

Table 2. The details of the face-and-body survey: the average movement speed, average movement in space and average movement dynamics criteria for each emotion category.

Order	average movement speed	average movement in space	average movement dynamics
1	happiness	happiness	happiness
2	fear	surprise	fear
3	disgust	fear	disgust
4	anger	boredom	anger
5	anxiety	anger	surprise
6	surprise	sadness	boredom
7	boredom	disgust	sadness
8	uncertainty	puzzlement	puzzlement
9	puzzlement	uncertainty	anxiety
10	sadness	anxiety	uncertainty

Table 3. The details of the face-and-body survey: Ranking of the emotion categories (in descending order) based on the average movement speed, average movement in space and average movement dynamics criteria.

These emotion categories fall in the "low intensity/arousal" category in Russell's circumflex model. The emotion with highest movement speed, largest movement in space and highest movement dynamics in space are *happiness*, followed by *fear*, *surprise* and *disgust*. These emotion categories fall in the "high intensity/arousal" category in Russell's circumflex model (see Fig. 1) (Russell, 1980).

According to the results compiled in Table 4, we can state that bimodal data helps with resolving ambiguity in most of the cases. The usefulness of the bimodal data for observer annotation is two-fold: (a) resolving ambiguity and (b) re-labelling of the videos.

(a) Resolving ambiguity that is present in affect annotation of the face data

Of the 65 videos that were labelled differently, ambiguity was resolved for 27 videos using the face-and-body data. This fact can be illustrated with the following examples:

- A face video that obtained divided votes by the observers (3 boredom and 3 puzzlement) was later labelled as *boredom* with much more certainty (5 votes) (video # S001-012).
- A face video that obtained divided votes by the observers (3 puzzlement and 3 anxiety) was later labelled as *anxiety* by all 6 observers (video # S001-040, see Fig. 2, left hand side).
- A face video that obtained divided votes by the observers (1 anger, 1 puzzlement 2 sadness, 1 ambiguity, 1 boredom), when viewed with face and body together, was labelled as *anxiety* by 5 observers (video # S002-010).
- A face video that obtained divided votes by the observers (3 boredom and 3 sadness), when viewed with face and body together, was labelled as *boredom* by 4 observers (video # S002-011).
- A face video that obtained divided votes by the observers (3 boredom and 3 sadness), when viewed with face and body together, was labelled as *anxiety* by all 6 observers (video # S010-039, see Fig. 2, right hand side).

(b) Changing the label of the displayed emotion obtained from face data to another label

Of the 65 videos that were labelled differently, 19 videos were re-labelled with *almost always higher agreement* when face-and-body data was viewed. This fact can be illustrated with the following examples:

- A face video that was labelled as *negative surprise*, when viewed as face and body together, was labelled as *positive surprise* (video # S001-007).
- A face video that was labelled as *puzzlement* by the observers (4 votes out of 6), when viewed as face and body together, was labelled as *anxiety* by the all 6 observers (video # S001-043).

In one case (video # S013-018), the face-and-body data helps with decreasing the level of ambiguity, but is not sufficient to resolve it. However, in 18 cases (see Table 3) the body adds ambiguity to the annotation. According to the results presented in Table 4, the emotion categories that caused more confusion in the bimodal data are *happiness* and *positive surprise* (7 out of 18 cases). Happiness was expressed as *extreme joy* and some observers labelled this display as *positive surprise*, which in fact is not wrong.

4 Discussion and Conclusions

Our investigation focused on evaluating monomodal and bimodal posed affective data for the purpose of aiding multimodal affect recognition systems that are dependent on human affective state as their input for interaction. According to the results obtained we conclude that in general, bimodal face-and-body data helps with resolving ambiguity carried by the face data alone. This in turn suggests that an automatic multimodal affect recognizer should attempt to combine facial expression and body gestures for improved recognition results.

video #	label for face video	votes	label for the combined face-and-body video	votes	changes & interpretation
s001-07	**negative surprise**	**4**	**positive surprise**	5	label changed and higher level of agreement between observers
s001-12	**boredom- puzzlement**	**3—3**	**boredom**	5	ambiguity resolved, label changed and higher level of agreement between observers
s001-23	**fear- negative surprise**	**3—3**	**fear**	5	ambiguity resolved, label changed and higher level of agreement between observers
s001-40	**puzzlement- anxiety**	**3—3**	**anxiety**	6	ambiguity resolved, label changed and full agreement between observers
s001-42	**puzzlement- anxiety**	**3—3**	**anxiety**	6	ambiguity resolved, label changed and full agreement between observers
s001-43	**puzzlement**	**4**	**anxiety**	6	label changed and full agreement between observers
s001-44	**puzzlement**	**3**	**anxiety**	6	label changed and full agreement between observers
s002-01	**happiness- positive surprise**	**3—3**	**happiness**	4	ambiguity resolved, label changed and higher level of agreement between observers
s002-10	**sadness**	**2**	**anxiety**	5	label changed and higher level of agreement between observers
s002-11	**boredom-sadness**	**3—3**	**boredom**	4	ambiguity resolved, label changed and higher level of agreement between observers
s003-01	**negative surprise**	**2**	**happiness**	3	label changed and higher level of agreement between observers
s003-05	**puzzlement**	**4**	**uncertainty-puzzlement**	2—2	bimodal data causes ambiguity
s003-08	**boredom-puzzlement**	**2—2**	**boredom**	4	ambiguity resolved, label changed and higher level of agreement between observers
s003-11	**boredom**	**3**	**boredom-anxiety**	3—3	bimodal data causes ambiguity
s004-19	**puzzlement**	**3**	**anxiety**	3	label changed
s005-07	**uncertainty**	**3**	**anger**	3	label changed
s005-14	**puzzlement- anxiety-uncertainty**	**2—2—2**	**puzzlement**	4	ambiguity resolved, label changed and higher level of agreement between observers
s005-22	**boredom-puzzlement**	**3—3**	**boredom**	5	ambiguity resolved, label changed and higher level of agreement between observers
s005-24	**disgust**	**4**	**disgust-fear**	3—3	bimodal data causes ambiguity
s005-32	**negative surprise**	**5**	**negative surprise-neutral surprise**	3—3	bimodal data causes ambiguity
s006-04	**negative surprise-fear**	**3—3**	**negative surprise**	3	label changed
s006-27	**uncertainty-puzzlement**	**3—3**	**puzzlement**	4	ambiguity resolved, label changed and higher level of agreement between observers
s006-29	**sadness-boredom**	**2—2**	**sadness**	3	ambiguity resolved, label changed and higher level of agreement between observers
s006-32	**puzzlement**	3	**uncertainty**	3	label changed
s008-02	**neutral surprise**	**3**	**negative surprise-neutral surprise**	3—3	bimodal data causes ambiguity
s008-05	**happiness**	4	**happiness-positive surprise**	3—3	bimodal data causes ambiguity
s008-07	**puzzlement**	3	**boredom**	6	label changed and full agreement between observers
s009-03	**puzzlement**	**5**	**uncertainty**	3	label changed
s009-12	**puzzlement-sadness**	**3—3**	**puzzlement**	4	ambiguity resolved, label changed and higher level of agreement between observers
s009-14	**puzzled-anxiety**	**3—3**	**anxiety**	4	ambiguity resolved, label changed and higher level of agreement between observers
s010-02	**happiness**	**5**	**happiness-positive surprise**	3—3	bimodal data causes ambiguity
s010-21	**puzzlement-anxiety**	**3—3**	**puzzlement**	5	ambiguity resolved, label changed and higher level of agreement between observers
s010-39	**sadness-boredom**	**3—3**	**anxiety**	6	ambiguity resolved, label changed and full agreement between observers
s010-42	**negative surprise**	**5**	**negative surprise-fear**	2—2	bimodal data causes ambiguity
s011-01	**happiness**	6	**happiness-positive surprise**	3—3	bimodal data causes ambiguity
s011-02	**happiness**	5	**happiness-positive surprise**	3—3	bimodal data causes ambiguity
s011-03	**anger**	3	**negative surprise-anger**	3—3	bimodal data causes ambiguity
s011-13	**puzzlement**	4	**uncertainty-puzzlement**	2—2	bimodal data causes ambiguity
s011-15	**boredom**	4	**puzzlement**	4	label changed
s011-24	**puzzlement**	2	**anxiety-boredom-puzzlement**	2—2—2	bimodal data causes ambiguity
s012-01	**happiness**	6	**positive surprise-happiness**	3—3	bimodal data causes ambiguity
s012-05	**neutral surprise**	3	**anger**	4	label changed and higher agreement between observers
s012-14	**fear-negative surprise**	**3—3**	**fear**	4	ambiguity resolved, label changed and higher level of agreement between observers
s012-20	**anxiety-puzzlement**	**2—2**	**puzzlement**	3	ambiguity resolved, label changed and higher level of agreement between observers
s013-01	**happiness**	5	**positive surprise-happiness**	3—3	bimodal data causes ambiguity
s013-03	**sadness**	4	**anger**	5	label changed and higher agreement between observers
s013-12	**happiness-positive surprise**	**3—3**	**positive surprise**	4	ambiguity resolved, label changed and higher level of agreement between observers
s013-15	**fear-disgust**	**2—2**	**fear**	5	ambiguity resolved, label changed and higher level of agreement between observers
s013-18	**anxiety-boredom-uncertainty**	**2—2—2**	**anxiety-boredom**	3—3	higher level of agreement, however ambiguity between two labels still exists
s014-01	**happiness**	5	**happiness-positive surprise**	3—3	bimodal data causes ambiguity
s014-02	**puzzlement-uncertainty**	**3—3**	**uncertainty**	4	ambiguity resolved, label changed and higher level of agreement between observers
s014-03	**anger-negative surprise**	**2—2**	**anger**	4	ambiguity resolved, label changed and higher level of agreement between observers
s014-06	**sadness-negative surprise**	**3—3**	**negative surprise**	3	ambiguity resolved and label changed
s014-07	**sadness-anxiety**	**2—2**	**anxiety**	3	ambiguity resolved, label changed and higher level of agreement between observers
s015-09	**anger-disgust**	**2—2**	**anger**	3	ambiguity resolved, label changed and higher level of agreement between observers
s015-12	**uncertainty**	3	**puzzlement**	3	label changed
s015-14	**puzzlement**	3	**boredom-uncertainty-puzzlement**	2—2—2	bimodal data causes ambiguity
s015-17	**sadness-boredom**	**2—2**	**puzzlement**	3	label changed and higher agreement between observers
s015-19	**puzzlement**	3	**boredom**	3	label changed
s015-22	**happiness**	4	**positive surprise**	5	label changed and higher agreement between observers

s015-28	**sadness-boredom**	**3—3**	**boredom**	5	ambiguity resolved, label changed and higher level of agreement between observers
s016-03	**neutral surprise**	3	**positive surprise**	4	label changed and higher agreement between observers
s016-04	**neutral surprise**	2	**positive surprise**	3	label changed and higher agreement between observers
s016-05	**anger**	3	**anger-positive surprise**	2—2	bimodal data causes ambiguity
s016-11	**puzzlement**	3	**boredom**	3	label changed

Table 4. List of the videos that were labelled differently for face and face-and-body modalities and the details of the labelling results.

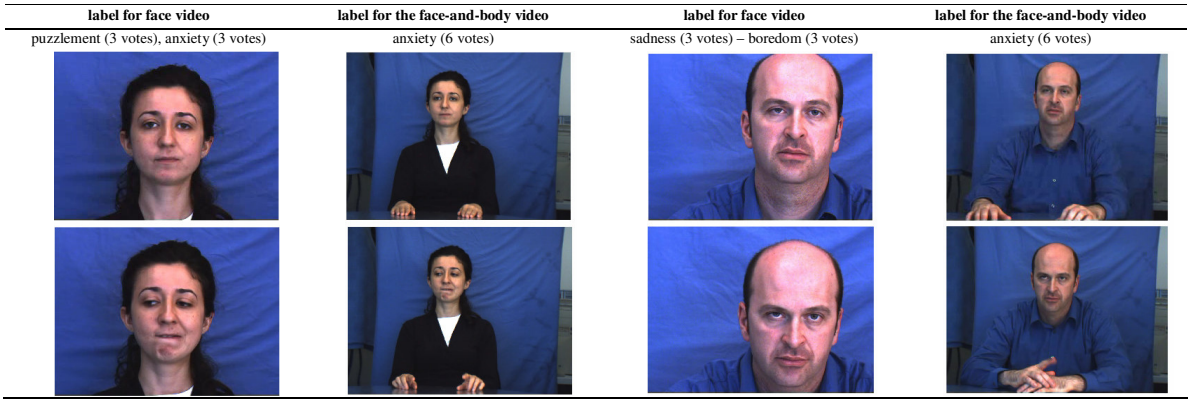

Figure 2. Example videos that were annotated differently for face and face-and-body and labels obtained from the survey: video # s001-40 (left hand side) and video # s010-039 (right hand side); **neutral** frames (first rows), expressive frames (second rows).

Overall, from the results obtained we can state that during annotation only in seldom cases do six observers fully agree on the emotion labelling. However, in general there is substantial agreement between the observers.

Affective state annotation in itself faces three main challenges (a) the type of emotion encoded, (b) the specific ability of the encoder, and (c) specific, discriminative movement indicators for certain emotions versus indicators of the general intensity of the emotional experience (Wallbott, 1998). Moreover, 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 (Ortony and Turner, 1990).

Furthermore, it is arguable that there may be differences in interpretation of the annotation scheme used to scale the expressivity of face and body. According to the results obtained we conclude that in general independent human observers tend to give average marks (i.e. 4 – 6 over a scale of 10) when rating speed, space usage and movement dynamics of the affective body movement. These results might be explained by the fact that there are some inherent difficulties in marking schemes in general (Blumhof and Stallibrass, 1994). These difficulties include:

- tendency to mark the more immediate concepts;
- tendency to mark towards the middle;
- exposing the subjectivity of marking schemes by trying to decide on, and weight, criteria. For instance, a mark of seven might represent a high mark for one observer, whereas the same mark for another observer might represent a concept of just above average.

One major finding of this study is the fact that bimodal data helps with resolving ambiguity in most of the cases (46 out of 65). However, in 18 cases (see Table 4) the body adds ambiguity to the recognition. The strategy to follow in such cases could be to ask an additional group of observers to view and label the data.

Our analysis suggests that affective information carried by the bimodal data is valuable and will aid an automatic multimodal affect recognizer achieve improved recognition results.

The relative weight given to facial expression, speech, and body cues depend both on the judgment task (i.e. what is rated and labelled) and the conditions in which the behaviour occurred (i.e. how the subjects were simulated to produce the expression) (Ekman, 1982). Despite many findings in emotion behaviour research, there is no evidence in the actual human-to-human interaction on how people attend to the various communicative channels (speech, face, body etc.). Assuming that people judge these channels separately or the information conveyed by these channels is simply additive, is misleading (Sebe et al., 2005). As future work, a study exploring these factors can be conducted.

In this study we recorded face and upper body using separate cameras, obtaining higher resolution for the face images and lower resolution for the upper body images. We did not analyse whether or not resolution poses a challenge for visual affect data interpretation and annotation. It is possible to further compare whether being exposed to face display with low resolution, face display with high resolution, and finally combined face-and-body display affects the human attention and perception of affective video data.

The experiment presented in this paper can further be extended by data obtained in natural and realistic settings.

As confirmed by many researchers in the field, directed affective face and body action tasks differ in appearance and timing from spontaneously occurring behaviour (Cohn et al., 2004). Deliberate face or body behaviour is mediated by separate motor pathways and differences between spontaneous and deliberate actions may be significant. However, collecting spontaneous multimodal affect data is a very challenging task involving ethical and privacy concerns together with technical difficulties (high resolution, illumination, multiple sensors, consistency, repeatability etc.). The research field of multimodal affective HCI is relatively new and future efforts have to follow (Pantic et al., 2005).

5 Acknowledgments

We would like to thank Aysel Gunes (Sydney Central College) for her help with the FABO recordings and the annotation procedure. We would also like to thank Michelle Murch (Production Support Coordinator, Film and Video Media Centre, Faculty of Humanities and Social Sciences, UTS) for her support regarding the technical issues for the video recordings, the anonymous participants for taking part in the recordings, and the anonymous observers for viewing and labelling the videos.

6 References

Allwood, J. et al. (2004), `The MUMIN multimodal coding scheme'. *in Proc. Workshop on Multimodal Corpora and Annotation*, Stockholm.

Balomenos, T., et al. (2004), `Emotion analysis in man-machine interaction systems', *in Proc. MLMI*, LNCS 3361, pp. 318–328.

Blumhof, J., Stallibrass, C. (1994), `Peer Assessment', Hatfield: University of Herefordshire.

Burgoon, J. K., et al. (2005), `Augmenting human identification of emotional states in video', *in Proc. Int. Conf. on Intelligent Data Analysis*.

Cohn, J.F. et al. (2004), `Multimodal coordination of facial action, head rotation, and eye motion during spontaneous smiles', *in Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pp. 129–135.

Coulson, M. (2004), `Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence', *J. of Nonverbal Behaviour*, Vol. 28, pp. 117–139.

DeMeijer, M. (1991), `The attribution of aggression and grief to body movements: the effect of sex-stereotypes'. *European Journal of Social Psychology*, Vol. 21.

Douglas-Cowie, E. et al. (2003), `Emotional speech: Towards a new generation of databases'. *Speech Communication*, Vol. 40.

Ekman, P. (1982): *Emotions in the human faces*, 2 ed., Studies in Emotion and Social Interaction, Cambridge University Press.

Ekman, P. (2003): *Emotions revealed.* Weidenfeld & Nicolson.

Gunes, H. and Piccardi, M. (2005), `Fusing Face and Body Display for Bi-Modal Emotion Recognition: Single Frame Analysis and Multi-Frame Post Integration', *in Proc. ACII*, LNCS 3784, pp. 102–111.

Gunes, H. and Piccardi, M. (2006a), `A Bimodal Face and Body Gesture Database for Automatic Analysis of Human Nonverbal Affective Behaviour', *in Proc. ICPR,* Vol. 1, pp. 1148–1153.

Gunes, H. and Piccardi, M. (2006b), `Creating and Annotating Affect Databases from Face and Body Display: A Contemporary Survey', *in Proc. IEEE SMC* (in press).

Hudlicka, E. (2003), `To feel or not to feel: The role of affect in human computer interaction', *Int. J. Hum.-Comput. Stud.,* Vol. 59, No. (1–2), pp. 1–32.

Kapoor, A. and Picard, R. W. (2005), `Multimodal affect recognition in learning environments', *in Proc. ACM Multimedia*, pp. 677–682.

Martin, J.C., Abrilian, S. and Devillers, L. (2005), `Annotating Multimodal Behaviours Occurring During Non Basic Emotions', *in Proc. ACII*, LNCS 3784, pp. 550–557.

Montepare, J. et al. (1999), `The use of body movements and gestures as cues to emotions in younger and older adults', *Journal of Nonverbal Behaviour*, Vol. 23, No. 2.

Ortony, A. and Turner, T. J. (1990), `What's basic about basic emotions?', *Psychological Review*, Vol. 97, pp. 315–331.

Pantic, M. et al. (2005), `Affective multimodal human-computer interaction', *in Proc. ACM Multimedia*, pp. 669–676.

Picard, R. W., Vyzas, E. and Healey, J. (2001), `Toward Machine Emotional Intelligence: Analysis of Affective Physiological State', *IEEE Tran. PAMI*, Vol. 23, No. 10, pp. 1175–1191.

Russell, J. A. (1980), `A circumflex model of affect', *Journal of Personality and Social Psychology*, Vol. 39, pp. 1161–1178.

Sebe, N., Cohen, I. and Huang, T.S. (2005), `Multimodal emotion recognition', Handbook of Pattern Recognition and Computer Vision, World Scientific.

Wallbott, H. G. and Scherer, K. R. (1986), `Cues and channels in emotion recognition', *Journal of Personality and Social Psychology*, Vol. 51, pp. 690–699.

Wallbott, H. G. (1998), `Bodily expression of emotion', *European Journal of Social Psychology*, Vol. 2.