# Fusing Face and Body Display for Bi-modal Emotion Recognition: Single Frame Analysis and Multi-Frame Post Integration

Hatice Gunes and Massimo Piccardi

Faculty of Information Technology, University of Technology, Sydney (UTS)
PO Box 123, Broadway, NSW, 2007, Australia
{haticeg, massimo}@it.uts.edu.au

**Abstract.** This paper presents an approach to automatic visual emotion recognition from two modalities: expressive face and body gesture. Face and body movements are captured simultaneously using two separate cameras. For each face and body image sequence single "expressive" frames are selected manually for analysis and recognition of emotions. Firstly, individual classifiers are trained from individual modalities for mono-modal emotion recognition. Secondly, we fuse facial expression and affective body gesture information at the feature and at the decision-level. In the experiments performed, the emotion classification using the two modalities achieved a better recognition accuracy outperforming the classification using the individual facial modality. We further extend the affect analysis into a whole image sequence by a multi-frame post integration approach over the single frame recognition results. In our experiments, the post integration based on the fusion of face and body has shown to be more accurate than the post integration based on the facial modality only.

## 1 Introduction

The case for affective computing is supported by the observation that humans display a rich set of emotional cues while communicating with other humans. In such human-human communications, a subject playing the role of the source-end uses a variety of cues such as gestures, tone of the voice, facial expressions that will be interpreted by the subject at the receiver-end. Such a rich set of emotional cues will be wasted in human-computer interaction until computers as the receiver-end will be capable of recognizing them to a human-like extent.

Significant research results have been reported in recognition of emotional cues from facial expressions (e.g. [2]). The level of acknowledgement for emotion recognition via body movements and gestures is lower since it has only recently started attracting the attention of computer science and HCI communities [13]. However, the interest is growing with works similar to those presented in [1], [4] and [14]. So far, most of the work in affective computing has focused on only a single channel of information (e.g. facial expression). However, reliable assessment typically requires the concurrent use of multiple modalities (i.e. speech, facial expression, gesture, and gaze)

occurring together [13]. Multimodal interfaces operate in a more efficient way, modalities usually complement each other and help improve the accuracy and robustness of affective and perceptual interfaces.

Relatively few papers have focused on implementing emotion recognition systems using affective multimodal data [17]. There exist several works in the literature (e.g. [7, 9]) that combined facial video and audio information either at a feature-level [7] or at a decision-level [9]. More recently, Kapoor et al. [14] combined sensory information from the face video (manually extracted features), the posture sensor (a chair sensor) and the game being played in a probabilistic framework to detect children's affective states. However, they do not focus on gestures of the hands and other bodily parts. Balomenos et al. [1] combined facial expressions and hand gestures for recognition of six prototypical emotions by using facial points from MPEG-4 compatible animation and defining certain hand movements under each emotion category. They fused the results from the two subsystems at a decision level using pre-defined weights.

Similarly, the aim of our research is to combine face and upper-body gestures in a bi-modal manner to recognize emotions automatically. Compared to the work in [1], we use a higher number of hand gestures and postures combined with the displacement of the other bodily parts (e.g. shoulders). Moreover, we compare the experimental results from feature-level and decision-level fusion of the face and body modalities to determine which fusion approach is more suitable for our work. Our motivation for multimodality is based on the fact that all of the aforementioned studies have improved the performance of their emotion recognition systems by the use of multimodal information [1, 7, 9, 14].

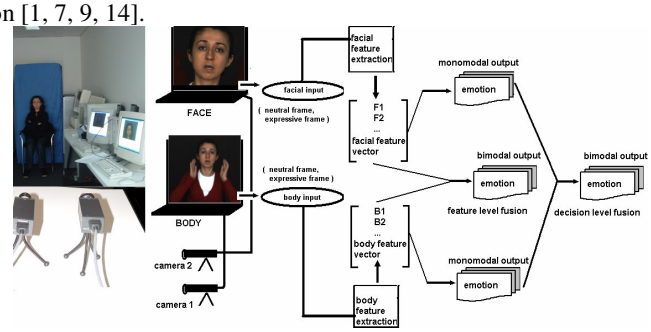


**Fig.1.** Our system framework for mono-modal and bi-modal emotion recognition.

## 2 Methodology

Initially, we present the two modalities, namely facial expressions and expressive body gestures, as described in the following sections. Our task is to analyze expressive cues within HCI which mostly take place as dialogues from a sitting position; hence, we focus on the expressiveness of the upper part of the body in our work. Since we were not able to find a publicly available database with bi-modal expressive face and

body gesture we created our own bi-modal database (FABO) by capturing face and body simultaneously from 23 people using two cameras as shown in Fig.1. Our database consists of recordings of the participants performing required face and body expressions. Moreover, after the recordings we conducted a survey asking the participants to evaluate their own performance. Based on their statement we considered a number of recorded sequences as outliers and did not include them in this work.

**Table 1.** List of the facial emotions recognized by our system and the changes that occur on the face when they are displayed (based on [6]).

| | |
|---|---|
| **anxiety**<br>▪ lip bite<br>▪ stretching of the mouth<br>▪ eyes turn up/down/left/right<br>▪ lip wipe | **fear**<br>▪ brows raised and drawn together<br>▪ forehead wrinkles drawn to the center<br>▪ upper eyelid is raised and lower eyelid is drawn up<br>▪ mouth is open<br>▪ lips are slightly tense or stretched and drawn back |
| **anger**<br>▪ brows lowered and drawn together<br>▪ lines appear between brows<br>▪ lower lid is tensed and may or may not be raised<br>▪ upper lid is tense and may or may not be lowered due to brows' action<br>▪ lips are either pressed firmly together with corners straight or down or open | **happiness**<br>▪ corners of lips are drawn back and up<br>▪ mouth may or may not be parted with teeth exposed or not<br>▪ cheeks are raised<br>▪ lower eyelid shows wrinkles below it, and may be raised but not tense<br>▪ wrinkles around the outer corners of the eyes |
| **disgust**<br>▪ upper lip is raised<br>▪ lower lip is raised and pushed up to upper lip or it is lowered<br>▪ nose is wrinkled<br>▪ cheeks are raised<br>▪ brows are lowered<br>▪ tongue out | **uncertainty**<br>▪ lid drop<br>▪ inner brow raised<br>▪ outer brow raised<br>▪ chin raised<br>▪ jaw sideways<br>▪ corners of the lips are drawn downwards |

**Table 2.** List of the body emotions recognized by our system and the changes that occur in the upper-body when they are displayed (based on [4, 8, 11]).

| | | |
|---|---|---|
| **anxiety**<br>▪ hands close to the table surface<br>▪ fingers moving<br>▪ fingers tapping on the table | **anger**<br>▪ body extended<br>▪ hands on the waist<br>▪ hands made into fists and kept low, close to the table surface | **disgust**<br>▪ body backing<br>▪ left/right hand touching the neck |
| **fear**<br>▪ body contracted<br>▪ body backing<br>▪ hands high up, trying to cover bodily parts | **happiness**<br>▪ body extended<br>▪ hands kept high<br>▪ hands made into fists and kept high | **uncertainty**<br>▪ shoulder shrug<br>▪ palms up |

### 2.1 Modality 1: Facial Expression

The leading study of Ekman and Frisen [10] formed the basis of visual automatic facial expression recognition. Their studies suggested that anger, disgust, fear, happiness, sadness and surprise are the six basic prototypical facial expressions recognized universally. Brave and Nass also provide details of the facial cues for displayed emotions in [6]. We base our facial feature extraction module on distinguishing these cues from the neutral face and from each other. Table 1 provides the list of the facial emotion categories (anger, disgust, fear, happiness, uncertainty and anxiety) recognized by our system based on the visual changes occurring on the face.

### 2.2 Modality 2: Expressive Body Gesture

Human recognition of emotions from body movements and postures is still an unresolved area of research in psychology and non-verbal communication. There are numerous works suggesting various opinions on this area.

In his paper [8], Coulson presented experimental results on attribution of 6 emotions (anger, disgust, fear, happiness, sadness and surprise) to static body postures by using computer-generated figures. He found out that in general, anger, happiness, and sadness are being attributed to certain postures, with some identified by 90%. For instance, arms open and raised above shoulder level constituted the posture receiving the highest concordance for the emotion "happiness". From his experiments he concluded that human recognition of emotion from posture is comparable to recognition from the voice, and some postures are recognized as well as facial expressions. Burgoon et al. clearly discuss the issue of emotion recognition from bodily cues and provide useful references in a recent publication in the context of national security [4]. They claim that emotional states are conveyed by a set of cues: "The natural expression, we may suppose, is a total made up of a certain facial expression, certain gestures, and a bodily posture" [4]. They focus on the identification of emotional states such as positivity, anger and tension in video from body and kinesics cues. In their paper, Boone and Cunningham [3] suggest that propositional expressive gestures are described as specific movements of certain bodily parts or postures corresponding to stereotypical emotions (e.g. bowed head and dropped shoulders showing sadness). Non-propositional expressive gestures are, instead, not coded as specific movements but form the quality of movements (e.g. direct/flexible).

In this paper we focus on the propositional gestures only as they can be easily extracted from individual frames. Table 2 is based on the cues described by Burgoon et al. [4], Coulson [8], Givens [11]; and provides the list of the body gestures and the correlation between the gestures and the emotion categories currently recognized by our system.

## 3 Feature Extraction

In our experiments we select a whole frame sequence where an expression is formed in order to perform feature extraction and tracking. For feature extraction we processed all available sequences and we classify only apex frames where an expression is fully formed. For each apex frame, we use a manually selected neutral frame and a set of previous frames for feature extraction and tracking. We assume that initially the person is in frontal view, the upper body, hands and face are visible and not occluding each other.

We apply a segmentation process based on a background subtraction method in each frame in order to obtain the silhouette of the upper body. We then apply thresholding, noise cleaning, morphological filtering and connected component labeling [18]. We generate a set of features for the detected foreground object, including its centroid, area, bounding box and expansion/contraction ratio as reference for body

movement. We extract the face and the hands automatically from individual frames of the face and body independently, by exploiting skin color information. The hand position consists of the position of the centroid and in-plane rotation. We employ the Camshift algorithm [5] for tracking the hands and predicting their locations in the subsequent frames (see Fig. 2). Orientation feature helps to discriminate between different poses of the hand together with the edge density information. For the face, we detect the key features in the neutral frame and define the bounding boxes for each facial feature (forehead, eyes, eyebrows, nose, lips and chin). Once the face and its features are detected, for tracking the face and obtaining its orientation for the next sequence we use again the Camshift algorithm. We also calculate the optical flow by comparing the displacement from the neutral face to the expressive face using the Lucas-Kanade algorithm [16]. Further details of our approach are explained in [12].

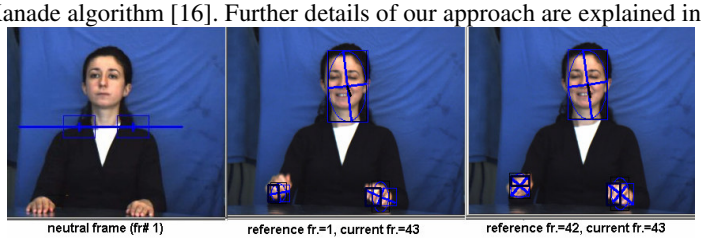


**Fig. 2.** Camshift tracking for face and two hands.

## 4 Single Frame Analysis

In this experiment, we processed 58 sequences in total, 29 for face and 29 for body from 4 subjects. We processed about 1750 frames for the face and 1750 for the body overall. However, we used only the "expressive" or "apex" frames for training and testing and we omitted the frames with intermediate movements. We used nearly half of these for training and the other half for testing purposes. For training we used the following sequences (one version for face and one for body): fear (1 sequence), happiness (3 sequences), anger (5 sequences), anxiety (2 sequences), uncertainty (2 sequences), and disgust (2 sequences). Similarly, the following sequences were used for testing: fear (1 sequence), happiness (3 sequences), anger (4 sequences), anxiety (2 sequences), uncertainty (2 sequences), and disgust (2 sequences). After obtaining the feature vector for face and body separately, we performed emotion recognition using Weka, a tool for automatic classification [19].

### 4.1 Mono-modal emotion recognition

Before the automatic recognition procedure all frames were initially labeled by two human experts. The ground truth in this work was established based on participants' own evaluation of their performance and the authors as human experts labeling all the frames. Further validation from a large pool of human experts will be done as future

work. We created a separate class for each emotion, for face and body separately. The face feature vector consists of 148 attributes and the body vector consists of 140 attributes. We then fed these feature vectors into separate classifiers for mono-modal emotion recognition.

**Table 3.** Mono-modal emotion recognition results for 4 subjects using BayesNet.

| Modality | Training | Testing | Attributes | Number of classes | Correctly classified |
|---|---|---|---|---|---|
| Face | 414 | 386 | 148 | 6 | 82.9 % |
| Body | 424 | 386 | 140 | 6 | 100 % |

For the face, we used 414 frames for training and 386 for testing. For the body, we used 424 frames for training and 386 for testing. BayesNet [19] provided the best classification results both for face and body emotion recognition (results are presented in Table 3).

### 4.2 Bi-modal emotion recognition

In general, modality fusion is to integrate all incoming single modalities into a combined single representation [20]. Typically, fusion is either done at a feature-level or deferred to the decision-level [20]. To make the fusion issue tractable the individual modalities are usually assumed independent of each other. In this work, to fuse the affective facial and body information we implemented both approaches: feature-level and decision-level fusion.

**Feature-level fusion.** Feature-level fusion is performed by using the extracted features from each modality and concatenating these features into one large vector. The resulting vector is input to a single classifier which uses the combined information to assign the testing samples into appropriate classes. We fuse face and body features of the corresponding expressive frames from the corresponding videos obtained from face and body cameras. We experimented various classifiers on a dataset that consists of 412 training and 386 testing instances. For the feature set with 288 attributes, BayesNet provided the best classification accuracy again (100% in this test). For the emotions considered, we observe that using the two modalities achieves better recognition accuracy in general, outperforming the classification using the face modality alone. To correctly interpret these results, it is important to recall that our experiments test unseen instances from the same subjects used for the training phase. Accuracy might be significantly lower for totally unseen subjects.

**Decision-level fusion.** Decision-level (late) integration is the most common way of integrating asynchronous but temporally correlated modalities [20]. Each modality is first pre-classified independently and the final classification is based on the fusion of the outputs from the different modalities. Designing optimal strategies for decision-level fusion is still an open research issue. Various approaches have been proposed: sum rule, product rule, using weights, max/min/median rule, majority vote etc. [15]. We used the first three techniques mentioned above for our system: sum, product and weight criteria. We describe the general approach of late integration of the individual

classifier outputs as follows: $X = (x_f, x_b)$ represents the overall feature vector consisting of the face feature vector, $x_f$, and the body feature vector, $x_b$. $X$ must be assigned to one of $M$ possible classes, ($\omega_1$, … $\omega_{k,,....}$ $\omega_M$) having maximum posterior probability $p(\omega_k|X)$. An early integration approach would compute such a probability explicitly. In late integration, instead, two separate classifiers provide the posterior probabilities $p(\omega_k|x_f)$ and $p(\omega_k|x_b)$ for face and body, respectively, to be combined into a single posterior probability $p(\omega_k|X)$ with one of the fusion methods described in the following. Moreover, in the infrequent case in which the combined $p(\omega_k|X)$ "fires" the same value for two or more classes, we resort to the classification provided by the face classifier since this is the major mode in our bi-modal approach. If the same happens for $p(\omega_k|x_f)$, we arbitrarily retain the first class in appearance order. The description of the three criteria we used (sum, product and weight) is given in Table 4.

**Table 4.** Description of the three late-fusion criteria used: sum, product and weight.

| | | |
|---|---|---|
| assign $X \rightarrow \omega_k$ | Sum rule | $k = \arg\max_{k=1..M} \left( \left( p(\omega_k \mid x_f) + p(\omega_k \mid x_b) \right) \right)$ |
| | Product rule | $k = \arg\max_{k=1..M} \left( p(\omega_k \mid x_f) p(\omega_k \mid x_b) \right)$ |
| | Weight criterion | $k = \arg\max_{k=1..M} \left( \lambda_f \, p(\omega_k \mid x_f) + \lambda_b \, p(\omega_k \mid x_b) \right)$ |

In our case, the face modality has the lead and the body modality needs to be integrated. Thus, when using the weight criteria we assigned arbitrary weights as follows: $\lambda_f = 0.7$ for the face modality and $\lambda_b = 0.3$ for the body modality. The late fusion results for sum, product and weight criteria are all presented in Table 5.

**Table 5.** Bi-modal emotion recognition results for 4 subjects using BayesNet.

| Emotion | Recognition rates on the testing set (%) | | |
|---|---|---|---|
| | Sum Rule | Product Rule | Weight criterion ($\lambda_a$= 0.70, $\lambda_g$=0.30) |
| Overall | 91 | 88 | 83 |
| Anger | 80 | 76 | 75 |
| Disgust | 100 | 100 | 97 |
| Fear | 94 | 83 | 77 |
| Happiness | 100 | 100 | 98 |
| Uncertainty | 78 | 76 | 63 |
| Anxiety | 98 | 93 | 83 |

## 5 Multi-Frame Post Integration

In order to provide a generalized affect analysis for a whole video we apply a multi-frame post integration approach on the single frame recognition results. The post integration combines single frame recognition results first by calculating the total number of recognized frames for each emotion category and then choosing the emotion with the maximum value as the "assigned emotion" or final decision for a whole video. We further analyze how the mono-modal and bi-modal multi-frame post integration approaches can differently prove useful for affect analysis of a whole video with two experiments:

*Experiment 1.* We used video #2 as an illustrative test case. The test results for "video #2" are shown in Table 6.
*Experiment 2.* We applied the multi-frame post integration analysis to all the testing videos. The test results are shown in Table 7.
The results obtained from both tables are analyzed in Section 6.

**Table 6.** Results of single frame recognition and multi-frame post integration for video #2.

| | | face video | body video | early fusion | late fusion | | |
|---|---|---|---|---|---|---|---|
| frame index # | actual emotion | single frame recognition result | single frame recognition result | | sum | product | weights |
| 59 | uncertainty | disgust | uncertainty | uncertainty | disgust | disgust | disgust |
| 60 | uncertainty | anxiety | uncertainty | uncertainty | uncertainty | uncertainty | anxiety |
| 61 | uncertainty | anger | uncertainty | uncertainty | uncertainty | uncertainty | uncertainty |
| 62 | uncertainty | happy | uncertainty | uncertainty | happy | disgust | happy |
| 63 | uncertainty | happy | uncertainty | uncertainty | happy | disgust | happy |
| 65 | uncertainty | happy | uncertainty | uncertainty | happy | disgust | happy |
| 66 | uncertainty | happy | uncertainty | uncertainty | happy | disgust | happy |
| 67 | uncertainty | happy | uncertainty | uncertainty | uncertainty | uncertainty | happy |
| 68 | uncertainty | happy | uncertainty | uncertainty | uncertainty | uncertainty | happy |
| 69 | uncertainty | uncertainty | uncertainty | uncertainty | uncertainty | uncertainty | uncertainty |
| 70 | uncertainty | happy | uncertainty | uncertainty | uncertainty | uncertainty | happy |
| 71 | uncertainty | happy | uncertainty | uncertainty | happy | disgust | happy |
| 72 | uncertainty | happy | uncertainty | uncertainty | happy | disgust | happy |
| 73 | uncertainty | happy | uncertainty | uncertainty | happy | disgust | happy |
| 74 | uncertainty | happy | uncertainty | uncertainty | happy | disgust | happy |
| 75 | uncertainty | happy | uncertainty | uncertainty | happy | disgust | happy |
| 76 | uncertainty | uncertainty | uncertainty | uncertainty | uncertainty | uncertainty | uncertainty |
| 77 | uncertainty | happy | uncertainty | uncertainty | uncertainty | disgust | happy |
| 78 | uncertainty | happy | uncertainty | uncertainty | uncertainty | uncertainty | happy |
| 79 | uncertainty | happy | uncertainty | uncertainty | uncertainty | uncertainty | happy |
| post integration | | disgust 1<br>happiness 15<br>fear 0<br>anger 1<br>uncertainty 2<br>anxiety 1 | disgust 0<br>happiness 0<br>fear 0<br>anger 0<br>uncertainty 20<br>anxiety 0 | disgust 0<br>happiness 0<br>fear 0<br>anger 0<br>uncertainty 20<br>anxiety 0 | disgust 1<br>happiness 9<br>fear 0<br>anger 0<br>uncertainty 10<br>anxiety 0 | disgust 11<br>happiness 0<br>fear 0<br>anger 0<br>uncertainty 9<br>anxiety 0 | disgust 1<br>happiness 15<br>fear 0<br>anger 0<br>uncertainty 3<br>anxiety 1 |
| **final decision** | | **happiness** | **uncertainty** | **uncertainty** | **uncertainty** | **disgust** | **happiness** |

**Table 7.** Results of multi-frame post integration for all testing videos.

| | **emotion recognized** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **actual emotion** | **face only** | | **early fusion** | | **late fusion** | | | |
| | | **# of frames** | | **# of frames** | **sum** | **# of frames** | **product** | **# of frames** |
| **anger** | **disgust** | **22** | **anger** | **26** | **disgust** | **21** | **disgust** | **25** |
| 26 frames | happiness | 4 | others | 0 | anger | 5 | anger | 1 |
| | others | 0 | | | others | 0 | others | 0 |
| **uncertainty** | **happiness** | **15** | **uncertainty** | **20** | **uncertainty** | **10** | **disgust** | **11** |
| 20 frames | uncertainty | 2 | others | 0 | happiness | 9 | uncertainty | 9 |
| | others | 3 | | | others | 1 | others | 0 |
| **anxiety** | **anxiety** | **26** | **anxiety** | **41** | **anxiety** | **37** | **anxiety** | **35** |
| 41 frames | disgust | 13 | others | 0 | disgust | 3 | disgust | 6 |
| | others | 2 | | | others | 1 | others | 0 |
| **disgust** | **disgust** | **14** | **disgust** | **15** | **disgust** | **15** | **disgust** | **15** |
| 15 frames | anger | 1 | others | 0 | others | 0 | others | 0 |
| | others | 0 | | | | | | |
| **disgust** | **disgust** | **27** | **disgust** | **27** | **disgust** | **27** | **disgust** | **27** |
| 27 frames | others | 0 | others | 0 | others | 0 | others | 0 |
| **fear** | **fear** | **14** | **fear** | **18** | **fear** | **17** | **fear** | **15** |
| 18 frames | anger | 3 | others | 0 | happiness | 1 | disgust | 3 |
| | others | 1 | | | others | 0 | others | 0 |
| **happiness** | **happiness** | **31** | **happiness** | **31** | **happiness** | **31** | **happiness** | **31** |
| 31 frames | others | 0 | others | 0 | others | 0 | others | 0 |
| **happiness** | **happiness** | **24** | **happiness** | **25** | **happiness** | **25** | **happiness** | **25** |
| 25 frames | others | 1 | others | 0 | others | 0 | others | 0 |
| **happiness** | **happiness** | **28** | **happiness** | **28** | **happiness** | **28** | **happiness** | **28** |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 28 frames | others | 0 | others | 0 | others | 0 | others | 0 |
| **anger** | **anger** | **35** | **anger** | **35** | **anger** | **35** | **anger** | **35** |
| 35 frames | others | 0 | others | 0 | others | 0 | others | 0 |
| **uncertainty** | **uncertainty** | **26** | **uncertainty** | **26** | **uncertainty** | **26** | **uncertainty** | **26** |
| 26 frames | others | 0 | others | 0 | others | 0 | others | 0 |
| **anger** | **anger** | **46** | **anger** | **46** | **anger** | **46** | **anger** | **46** |
| 46 frames | others | 0 | others | 0 | others | 0 | others | 0 |
| **anxiety** | **anxiety** | **47** | **anxiety** | **48** | **anxiety** | **48** | **anxiety** | **48** |
| 48 frames | uncertainty | 1 | others | 0 | others | 0 | others | 0 |
| | others | 0 | | | | | | |

## 6 Analysis and Conclusions

This paper presented an approach to automatic visual analysis of expressive face and upper body gesture and associated emotions suitable for use in a vision-based affective multimodal framework.

Initially, we focused on facial expressions and body gestures separately and analyzed the individual frames, namely neutral and expressive frames. We presented experimental results from four subjects. Firstly, individual classifiers were trained separately with face and body features for mono-modal classification into labeled emotion categories. We fused affective face and body modalities for classification into combined emotion categories (a) at the feature-level ("early" fusion), in which the data from both modalities are combined before classification and (b) at the decision-level ("late" fusion). Our experimental results show that: (a) the emotion classification using the two modalities achieved a better recognition accuracy in general, outperforming the classification using the face modality only; (b) by comparing the experimental results, early fusion seems to achieve a better recognition accuracy compared to late fusion; (c) Table 5 shows that the sum rule proved the best way to fuse the two modalities.

We further extended affect analysis into a whole video sequence by a multi-frame post integration approach over the single frame recognition results in order to output a decision for the whole video sequence with two experiments. Table 6 shows that: (a) single frame recognition accuracy from the face is not high; (b) the mono-modal multi-frame post integration from the face results in 10% accuracy for affect analysis of the whole video and wrongly labels it as "happiness"; (c) the bi-modal multi-frame post integration based on early fusion results in 100% accuracy for affect analysis of the whole video and correctly labels it as "uncertainty"; (d) the bi-modal multi-frame post integration based on late fusion with sum criterion results in 50% accuracy for affect analysis of the whole video; if the maximum is chosen (10 out of 20), then this criteria correctly labels the video as "uncertainty". Similarly, Table 7 shows that: (a) both mono-modal and bi-modal multi-frame post integration approaches prove to be useful for affect analysis of a whole video; (b) mono-modal post integration most often provides accurate results when maximum is chosen (11 out of 13 videos); (c) the bi-modal post integration based on early fusion provides 100% accuracy for affect analysis of all testing videos by correctly labeling 13 out of 13 videos; (d) the bi-modal post integration based on late fusion provides 92% accuracy by correctly labeling 12 out of 13 videos; (e) the bi-modal multi-frame post integration based on early fusion provides more accurate results than late fusion; (f) the bi-modal post integration

based on late fusion with sum criterion provides better results than the other late fusion criteria.

From our experiments we can conclude that using expressive body information adds substantial accuracy to the emotion recognition based solely on the face. Furthermore, the use of body cues helps disambiguate the recognition of emotions for those cases where emotions appear very similar in terms of facial features alone. A logical explanation for that is that body gestures can be more reliably recognized than small-scale facial actions by means of image analysis techniques in many real cases.

## References


1. Balomenos, T., et al.: Emotion Analysis in Man-Machine Interaction Systems, Springer MLMI 2004 Lecture Notes, 3361 (2004), 318-328.
2. Bartlett, M.S., et al.: Machine learning methods for fully automatic recognition of face expressions and face actions, Proc. of IEEE SMC (2004) 592-597.
3. Boone, R. T. & Cunningham, J. G.: Children's decoding of emotion in expressive body movement: The development of cue attunement, Developmental Psyc. 34 (1998)1007-1016.
4. Burgoon, J. K., et al.: Augmenting Human Identification of Emotional States in Video, Proc. of Int. Conference on Intelligent Data Analysis (2005) (in press).
5. Bradski, G. R.: Computer vision face tracking for use in a perceptual user interface, Intel Techn. J. 2nd Quarter (1998).
6. Brave, S. & Nass, C.: Emotion in HCI. In J. Jacko & A. Sears (Eds.), The HCI Handbook, Hillsdale, NJ: Lawrence Erlbaum Associates (2002).
7. Chen, L.S. & Huang, T.S.: Emotional expressions in audiovisual human computer interaction, Proc. of IEEE ICME, 1 (2000) 423-426.
8. Coulson, M.: Attributing Emotion to Static Body Postures: Recognition Accuracy, Confusions, and Viewpoint Dependence, J. of Nonverbal Behavior, 28 (2) (2004).
9. De Silva, L. C. & Ng, P. C.: Bi-modal emotion recognition, Proc. FG (2000) 332–335.
10. Ekman, P. & Friesen, W. V.: Unmasking the face: a guide to recognizing emotions from facial clues, Imprint Englewood Cliffs, N.J.: Prentice-Hall (1975).
11. Givens, D. B. , The Nonverbal Dictionary of Gestures, Signs & Body Language Cues, Washington, Center for Nonverbal Studies Press, (2005).
12. Gunes, H. & Piccardi, M.: Fusing Face and Body Gesture for Machine Recognition of Emotions, Proc. IEEE RO-MAN 2005, Nashville, USA, (2005) (in press).
13. Hudlicka, E.: To feel or not to feel: The role of affect in human-computer interaction, Int. J. Hum.-Comput. Stud., 59 (1-2) (2003) 1-32.
14. Kapoor, A., et al.: Probabilistic Combination of Multiple Modalities to Detect Interest, Proc. IEEE ICPR (2004).
15. Kuncheva, L. I.: A Theoretical Study on Six Classifier Fusion Strategies, IEEE Trans. on PAMI, 24(2) ( 2002).
16. Lucas, B.D. & Kanade, T.: An iterative image registration technique with an application to stereo vision, Proc. of 7th Int. Jnt. Conf. on Artificial Intelligence (1981) 674–680.
17. Pantic, M. & Rothkrantz, L.J.M.: Towards an affect-sensitive multimodal human-computer interaction, Proc. of the IEEE, 91(9) (2003) 1370-1390.
18. Shapiro, L.G. & Rosenfeld, A.: Computer Vision and Image Processing, Boston, Academic Press (1992).
19. Witten, H. & Frank, E.: Data Mining: Practical machine learning tools with Java implementations, Morgan Kaufmann, San Francisco (2000).
20. Wu, L., Oviatt, S. L. & Cohen, P. R. Multimodal Integration-A Statistical View, IEEE Trans. on Multimedia, 1(4) (1999)334-341.