Fusing Face and Body Gesture for Machine Recognition of Emotions

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* - Research shows that humans are more likely to consider computers to be human-like when those computers understand and display appropriate nonverbal communicative behavior. Most of the existing systems attempting to analyze the human nonverbal behavior focus only on the face; research that aims to integrate gesture as an expression mean has only recently emerged. This paper presents an approach to automatic visual recognition of expressive face and upper body action units (FAUs and BAUs) suitable for use in a vision-based affective multimodal framework. After describing the feature extraction techniques, classification results from three subjects are presented. Firstly, individual classifiers are trained separately with face and body features for classification into FAU and BAU categories. Secondly, the same procedure is applied for classification into labeled emotion categories. Finally, we fuse face and body information for classification into combined emotion categories. In our experiments, the emotion classification using the two modalities achieved a better recognition accuracy outperforming the classification using the individual face modality.**

Index Terms – face expression, body gesture, action unit recognition, emotion recognition, fusion.

I. INTRODUCTION

Emotions can be communicated by various modalities, including speech and language, gesture and head movement, body movement and posture, as well as face expression. According to Mehrabian [16], in human-human interaction (HHI) spoken words only account for 7% of what a listener comprehends; the remaining 93% consist of the speaker's nonverbal communicative behavior (i.e. body language and intonation). There exist other findings claiming that humans display their emotions most expressively through face expressions and body gestures [9, 17]. Moreover, research shows that humans are more likely to consider computers to be human-like when those computers understand and display appropriate nonverbal communicative behavior [6]. Therefore, the interaction between humans and computers will be more natural if computers are able to understand the nonverbal behavior of their human counterparts and recognize their affective state.

Automatic facial expression recognition has attracted the interest of artificial intelligence and computer vision research communities for the past decade. Significant research results have been reported in recognition of emotions using face expressions (e.g. [2]). Growing amount of research has also investigated movement and gesture as one of the main channels of nonverbal communication in human-human interaction (HHI) and human computer interaction (HCI). However, existing literature on automatic emotion recognition has focused mainly on the face; research that aims to integrate gesture as an expression mean in HCI has recently started [13].

According to Hudlicka [13], while much progress has been achieved in affect assessment using a single measure type, reliable assessment typically requires the concurrent use of multiple modalities (i.e. speech, face expression, gesture, and gaze) that occur together to function in a more efficient and reliable way. Pantic and Rothkrantz [19] clearly state the importance of a multimodal affect analyzer for research in automatic emotion recognition. The modalities considered are face expressions and audio information for bimodal emotion recognition. The interpretation of other visual cues such as body movements is not explicitly addressed in [19] 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 [13]. However, the interest is growing with works similar to the ones presented in [1] and [14].

Taking into account these findings, the aim of our research is to combine face and upper-body gestures in a bimodal manner to distinguish between various expressive cues that will help computers recognize particular emotions. In this paper, we present experimental results of automatic recognition of expressive face and upper body action units (FAUs and BAUs) and associated emotions suitable for use in a vision-based affective multimodal framework.

II. METHODOLOGY

Our task is to analyze expressive cues within HHI and HCI which mostly take place as dialogues in a sitting position; hence, we focus on the expressiveness of the upper part of the body in our work.We assume that initially the person is in frontal view; the complete upper body, two hands and the face are visible and not occluding each other. We first analyze the two modalities, namely face action units (FAUs) and body action units (BAUs) separately and then we apply fusion as described in the following sections. The general system framework for both unimodal and bimodal emotion recognition is depicted in Figure 1.

A. *Modality 1: Face Action Units*

The leading study of Ekman and Frisen [8, 9] formed the basis of visual automatic face expression recognition. Their studies suggested that anger, disgust, fear, happiness, sadness and surprise are the six basic prototypical face expressions recognized universally.

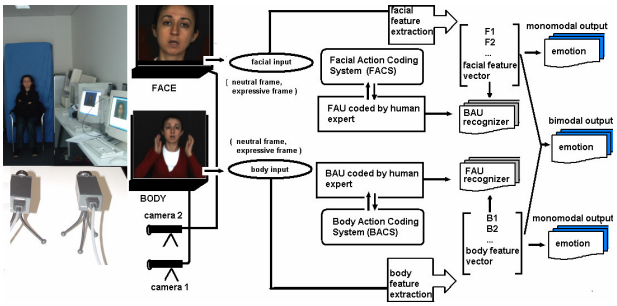

Fig. 1 System framework for FAUs, BAUs and emotion recognition.

However, six universal emotion categories are not sufficient to describe all facial expressions in detail. In order to capture the subtlety of human emotion, recognition of fine-grained changes and atomic movements of the face is needed [8]. Ekman and Friesen [8] developed the Facial Action Coding System (FACS) for describing face expressions by face action units (FAUs). 44 face action units (FAUs) are defined. 30 FAUs are anatomically related to the contractions of specific face muscles: 12 are for upper face, and 18 are for lower face. AUs can be classified either individually or in combination. In order to show how FAUs are linked to emotions in FACS we present an example below of how the emotion "surprise" is defined as a combination of four FAUs [7] (the notation "+" refers to the linear combination of FAUs occurring together):

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; FAU5: Upper Lid Raised; FAU26: Jaw Dropped; FAU27: Mouth Stretched. The emotion surprise is defined to be additive of these FAUs.)

FACS is the most commonly used coding system in vision-based systems attempting to recognize FAUs [2]. Table I provides the list of the FAUs and their description and Table II provides the correlation between the FAUs and the emotion categories recognized by our system, respectively. Tables I and II leave gaps between numbering and combinations of FAUs due to two reasons: (a) the numbering of FAUs in Table I is based on [8]; (b) our system does not attempt to recognize all listed FAUs in [8], it attempts to recognize the combinations available at hand from the subjects recorded.

B. *Modality 2: Body Action Units*

Propositional expressive gestures are described as specific movements of specific body parts or postures corresponding to stereotypical emotions (e.g. bowed head and dropped shoulders showing sadness). Non-propositional expressive gestures are not coded as specific movements but form the quality of movements (e.g. direct/flexible) [3]. In this paper, we focus on the propositional gestures only since they can be easily extracted from static frames. We employ the propositional body movements that carry expressive information and call them Body Action Unit (BAU) to create the Body Action Coding System (BACS). Since there is not a readily available BACS we defined the BAUs used in our system in terms of features grouped under specific emotion categories taking into account the psychological studies together with the results obtained from our experiments, in [12]. Table III provides the list of the BAUs and the correlation between the BAUs and the emotion categories recognized by our system.

TABLE I

LIST OF THE FAUS RECOGNIZED BY OUR SYSTEM AND THEIR DESCRIPTION (BASED ON [8]).

FAU	FAU description	FAU	FAU description	FAU	FAU description
1	inner brow raised	12	lip corner pull	26	jaw dropped
2	outer brow raised	13	cheek puff	27	mouth stretched
4	brow lowered	14	dimpler	28	lips sucked in
5	upper lid raised	15	lip corner depressed	41	lid dropped
6	cheek raised	17	chin raised	43	eyes closed
7	lower lid tight	20	lip stretched	61	eyes turned left
9	nose wrinkle	23	lip tightened	62	eyes turned right
10	upper lip raised	24	lips pressed	63	eyes turned up
		25	lips parted	64	eyes turned down

TABLE II

LIST OF THE EMOTIONS RECOGNIZED BY OUR SYSTEM AND THE CORRELATION BETWEEN THE FAUS AND EMOTION LABELS.

Emotion	FAU combination	Emotion	FAU combination	Emotion	FAU combination	Emotion	FAU combination
disgust	4+6+9+17+20 4+6+9+24 4+6+9+12+25 4+7+9+25 4+9+17+24 6+9+17+20 7+10+41 10	**happiness**	1+2+6+12 12 12+14 12+25 12+25+41 6+12+25+41 6+12 6+12+41 6+12+25 6+12+25	**surprise**	1+2 1+2+5 1+2+5+24 1+2+5+24+63 1+2+5 1+2+5+25 1+2+5+25 1+2+5+26 1+2+5+27	**anger**	4 4+7 4+7+9+17 4+7+10 4+7+23 4+7+24 4+17+24 4+24 7+17+24 12+14+17+24
happy_surprise	1+2+5+6+12 1+2+5+6+12+25 1+2+5+6+12+26 1+2+5+6+12+27	**fear**	1+4+5+12 1+6+17+20 1+17+20 17+20	**sadness**	1+2+15+17+24 1+15+17+24 1+15+17 1+15 15+17 15+17+41 17+24	**uncertainty**	25 25+41 25+43 25+61 25+64 28 41+62 43 61,62,63,64

TABLE III

LIST OF THE BAUS RECOGNIZED BY OUR SYSTEM AND THE CORRELATION BETWEEN THE BAUS AND EMOTION LABELS. THE NOTATION "+" REFERS TO BAUS OCCURING TOGETHER AS A COMBINATION. THE NOTATION "(1+)" IMPLIES THAT THE BAU OCCURS EITHER WITH "BAU 1" OR WITHOUT IT.

BAU	BAU description	BAU	BAU description	emotion	BAU combination
0	neutral	13	left hand touching the neck	**anger_happiness**	1, 20, 1+20
1	body extended	14	right hand touching the neck	**anger_disgust**	3, 4 , 1+3, 1+4
2	body contracted	15	left hand on left shoulder	**anger_fear**	25
3	left hand moved up	16	right hand on right shoulder	**fear_sadness_surprise**	2+18, 1+21, 1+22, 1+23, 1+24
4	right hand moved up	17	shoulder shrug		(1+) 5, (1+) 6, (1+) 7, (1+) 8,

5	left hand touching the head	18	shoulder drop		(1+) 9, (1+) 10,
6	right hand touching the head	19	palms up		(1+) 11, (1+) 12,
7	left hand about to touch the head	20	two hands up		(1+) 13, (1+) 14,
8	right hand about to touch the head	21	two hands touching the head		(1+) 15, (1+) 16,
9	left hand touching the face/ facial parts	22	two hands about to touch the head	**uncertainty_fear_surprise**	(1+) 17, (1+) 19,
10	right hand touching the face/facial parts	23	two hands touching the face		(1+) 17+19,
11	left hand about to touch the face	24	two hands about to touch the face		(1+) 17+25
12	right hand about to touch the face	25	arms crossed		

III. FEATURE DETECTION AND EXTRACTION

In this work, we choose to use the well-known methods proposed in face, body and hand detection approaches since such methods have proven reliable and computationally efficient. We assume that initially the person is in frontal view, the upper body, hands and face are visible and not occluding each other. In our experiments we select a whole frame sequence where an expression is formed in order to perform feature extraction and tracking. Our feature vector consists of displacement measures between two major frames; namely a frame with the neutral expression ("neutral frame") and one where the expression is at its apex ("expressive frame").

A. *Face feature extraction*

The first step in automatic FAU analysis is to locate the face in the image. Firstly, morphological operations are used to smooth the image [22]. We then apply skin color segmentation based on HSV color space [22]. We obtain the face region by choosing the largest connected component among the candidate skin areas [23]. We then employ closing (dilation and erosion) and find the contour of the face that returns the filled face region [22]. Once the face and its features are detected, for tracking the face and obtaining its orientation for the next sequence we employ the Camshift algorithm [4]. On convergence, the Camshift algorithm returns orientation, length, and width, hence enabling the estimation of face rotation [4].

We detect the key features in the neutral frame and define the bounding rectangle for each facial feature. For feature extraction we apply two basic methods. The first one is based on the gray-level information of the face region combined with edge maps and the second one is based on the min-max analysis by Sobottka and Pitas [23]. All the edge maps or edge information mentioned in this paper are obtained by using the Canny Edge Detector [5]. We first enhance the face region by histogram equalization [22]. We improve the contrast of the features by thresholding the image into binary. For example, in the case of the eyes, this is due to the color of the pupils and the sunken eye-sockets. Our method also uses min-max analysis introduced by Sobottka and Pitas [23] to detect the eyebrows, eyes, mouth and chin, by evaluating the topographic gray-level relief. After binarizing the image, face histograms are determined by the X- and Y- axis projection. We use the information of expected locations of face parts to restrict the searching area within the face region. In the following we provide the detailed steps for the various features.

1) Eyes: After estimating the bounding rectangle for the face, we use knowledge of feature locations to restrict search areas for the eyes to the upper half of the face region. For eye detection, the horizontal histogram of the skin-region is computed. The rows containing the eyes are located in correspondence of a histogram local minimum in the upper face part. Further, to obtain exact location of the eyes, we apply band pass filtering and morphological operations on the enhanced face region. Connected components are then identified as the areas of candidate eye regions. Motion of the eyes is measured by optical flow calculation using the Lucas-Kanade Algorithm [15]. We also model the state of the eyes with two states: open and closed. We first assume that the eye state in the first frame is neutral and open. After binarizing the eye pixels, we obtain the horizontal projection of the eye region. This projection is further used to determine the current state of the eyes.

2) Lip region: Once eyes have been detected, the mouth area is searched according to inter-eyes distance by finding the lip color in the lower half of the face region. Lips can be easily discriminated from skin based on their different intensity levels and color. We detect the lips based on the technique described in [11]. We then apply connected component labeling for the candidate lip regions with defined color feature to obtain the biggest connected component in the pre-defined search space [22]. We also model the atomic movement of the lips with six states: closed, tightly closed, sucked in, open, jaw dropped and stretched. We first assume that the lip state in the first frame is neutral and closed. Moreover, we use the color information when identifying whether the mouth is closed or open. For the open mouth and the tightly closed mouth, there are non-lip pixels inside the lip region. Based on the lip and non-lip colored pixels, we obtain the horizontal X-projection of the mouth region.

3) Eyebrows, nostrils and chin: We use the knowledge of the previously detected feature locations to restrict search areas for the eyebrows, nose and chin. We first apply a second derivative Gaussian filter, elongated at an aspect ratio of 3 to 1, to the face region. Interest points, detected at the local maxima in the filter response, indicate the possible locations of these features. Eyebrows are expected to be located in the upper part of the face and are the first non-skin components on the face region below the forehead. We also examine the edges around the upper part of the eyes by applying a horizontal edge enhancement to obtain the eyebrow curves. Then the upper and lower bounding rectangles are defined based on the boundary information. Motion of the eyebrow is measured by using optical flow [15]. The search space for chin is arranged according to the lip line and the horizontal lower limit of the face region. The tip of the chin is localized as the first minima starting from the horizontal lower limit of the face region.

4) Detecting face motion: After detecting the key features in the neutral frame and defining the bounding rectangles for face features, we consider the temporal information in the subsequent frames by extracting the movement in these pre-defined bounding rectangles. We calculate the optical flow by comparing the displacement from neutral to expressive face using the Lucas-Kanade Algorithm [15]. We estimate averaged optical flow within each region of interest. We then calculate the direction of the dominant motion vector.

5) Wrinkle analysis: Psychological studies proved that for face expression analysis the appearance of wrinkles in four main areas is important: forehead, in between eyebrows, outer comers of eyes and mouth corners [9]. Therefore, we analyze the wrinkle change within these regions by using edge density per unit area against some threshold. We compare the wrinkling within the bounding rectangle of the transient features in the expressive frame with respect to the neutral frame.

B. *Body feature extraction*

In each frame a segmentation process based on a background subtraction method is applied in order to obtain the silhouette of the upper body. We then apply thresholding, noise cleaning and morphological filtering [22]. After thresholding, one iteration of 3*3 dilation is applied on the binary image. Then, a binary connected component operator is used to find the foreground regions, and small regions are eliminated [22]. Since the remaining region is bigger than the original one it is restored to its original size by the erosion procedure [22]. We then generate a set of features for the detected foreground object, including its centroid, area, bounding box and expansion/contraction ratio for comparison purpose (see Fig.2).

1) Segmentation and tracking of the body parts: We first locate the face and the hands exploiting skin color information. Among the detected candidate regions, the largest connected component gives the face region; the second and third largest connected components give the hands, respectively (see Fig.2). We then calculate the centroid of these regions in order to use them as reference points for the body movement. We employ color since we need to detect the hands even if they are located within the silhouette. Hand displacement is computed as the motion of the centroid coordinate. We employ Camshift technique [4] for tracking the hands and comparison of bounding rectangles is used to predict their locations in subsequent frames.

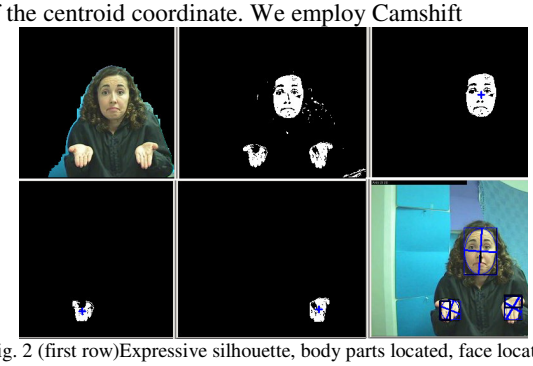

Fig. 2 (first row)Expressive silhouette, body parts located, face located; (second row) left and right hand located, body parts tracked with Camshift.

2) Locating shoulders: We locate shoulders based on the model knowledge of where they usually occur with respect to the face, upper body and hands. According to our upper-body model, in the neutral frame, shoulders are the widest point of the upper half of the silhouette. First, we compute the 1D horizontal projections of the silhouette. We then assume that most people present a narrower row in skin blob at neck level and a much wider row at the shoulder level, compared to the neck level. Thus, starting from the face centroid, we search for the widest row in the upper body blob. We also compute the 1D vertical projection of the silhouette and locate the shoulders as two minimums on left and right hand side of the bounding rectangle for the head. For recognizing "shoulder shrug", we compare the horizontal position of the shoulders with respect to the neutral frame.

Fig. 3 Camshift tracking when hands are about to merge with the face. The darker rectangles represent the position of the fingers.

3) Region merging: When two hands merge or when the hand(s) cover the face region due to their skin color, they might be segmented as one foreground region by the Camshift algorithm. Camshift applies a simple analysis of the predicted bounding boxes of the tracked objects and the bounding box of the detected foreground region (see Fig.3). When the merged region splits, the localization procedure is run again to obtain and re-initialize the current location of each region.

4) Hand pose and orientation estimation: Orientation feature helps to discriminate between different poses of the hand. On convergence, the Camshift algorithm returns orientation, length and width of the bounding rectangle for the hand, hence, enabling the estimation of hand rotation [4]. Using this information we decide if the hand is in vertical or horizontal position. After estimating the initial pose of the hand it is possible to find out the position of the fingers (see Fig.3). Edges have proven useful features for discriminating between different poses of the hand [18]. We define four categories for finger position estimation: up, down, right and left. We use this information when classifying the feature vectors into various BAUs (e.g. arms crossed, hands touching the head etc.).

IV. Experiments with Unimodal Data

We recorded the test sequences simultaneously using two fixed cameras, connected to two separate PCs with a simple setup and uniform background. We created a setting with a simple background in order to reduce implications in the background removal and processing of

the upper body features. One camera was placed specifically capturing the head only and the second camera was placed in order to capture upper-body movement from the waist above. We chose to use two cameras due to the fact that current technology still does not provide us with frames with the required quality to process detailed upper-body and face information together. We recorded three subjects performing FAUs and BAUs alone or in combination. In the first frame, the body is in neutral position. In the following frames, the system can handle in-line rotation of the face and hands. The "neutral frame" and "expressive frame" were used for training and testing of FAUs and BAUs. All samples were initially AU coded by two human experts.

TABLE IV
FAUS AND BAUS CLASSIFICATION RESULTS FOR 3 SUBJECTS.

	Instances	Attributes	Number of Classes	Classifier	Correctly classified
whole face	311	67	88	BayesNet	70.09 %
upper face	249	67	23	BayesNet	77.10 %
lower face	267	67	30	BayesNet	74.90 %
body	296	140	23	C4.5	81.41 %

Firstly, for FAU and BAU recognition we used Weka, a tool for automatic classification [20]. Amongst the various classifiers provided by this tool, BayesNet provided the best classification result with 10-fold cross validation for FAUs and C4.5 provided the best classification results for BAUs recognition. The results are presented in Table IV. For FAU and BAU classification, we created a separate class for each different combination of single AUs, for face and body separately. Moreover, for FAU classification, we divided the instances for classification into upper and lower FAUs. The classification accuracy for the upper face seems to be better than the lower face or whole face AU classification. These results are preliminary and we believe that increasing the training set will improve the classification. Yet, the accuracy achieved proves that the dimensionality of the problem is lower than the estimate provided by the product of the number of attributes by the number of classes, meaning that some of the classes are not statistically independent.

Secondly, we used Weka [20] to classify the data from expressive face and body into labeled emotion categories. We created a separate class for each emotion, for face and body separately. For face, we created eight classes: happiness, sadness, fear, anger, disgust, surprise, happy_surprise and uncertainty. The six basic emotion classes are based on [8]. If the face displays a combination of happiness and surprise then we classify it as "happy_surprise". Moreover, during our experiments the three subjects manipulated their faces in various ways, therefore for the expressions that did not match any of the seven categories mentioned above we created an extra category and named it as "uncertainty". For the emotion classification based on the body gestures we created classes that are combinations of two or three emotion categories. This is done due to the fact that the face modality is the primary mode and the body modality is an auxiliary mode in our system. We are not intending to use the body classification results alone for the final emotion classification. Emotion categories used for upper-body are anger_happiness, anger_disgust, anger_fear, fear_sadness_surprise, uncertainty_fear_surprise.

For emotion classification from face and body, C4.5 [19] with 10-fold cross validation provided the best classification result. The results are presented in Table V.

TABLE V
EMOTION RECOGNITION RESULTS FOR 3 SUBJECTS USING C4.5.

	Instances	Attributes	Number of classes	Correctly classified
whole face	265	67	8	72.83 %
body	297	140	5	92.25 %
face and body combined	206	206	8	89.80 %

V. EXPERIMENTS WITH BIMODAL DATA

The few studies that are present in the literature (e.g. [1, 7, 14]) have shown that the performance of emotion recognition systems can be improved by the use of multimodal information. This motivated us to combine affective face and body information for more efficient emotion recognition.

When it comes to integrating the multiple modalities the major issue is when and how to integrate them. Depending on how closely coupled the modalities are there are three different levels of integration: data level, feature level and decision level. Fusion at the feature level is appropriate for closely coupled and synchronized modalities (e.g. speech and lip-movements) [24]. If the modalities are asynchronous but temporally correlated, like in our case with face and body gesture, decision level integration is the most common way of integrating the modalities [24]. However, in this paper, we use the static frames of neutral and peak expressions of face and body images; face and body actions are considered to be synchronous and the temporal correlation is ignored. Therefore, we fuse the face expression and body gesture information at the feature-level. This is performed by concatenating the feature vectors from each modality and using a single classifier.

We transform the images into a representation that decomposes the images into features (e.g. movement of face features, shoulders, hands etc.) and perform fusion in this domain. We fuse face and body features only if the category for the face vector and that for the body vector are the same, or the body category includes the face category (such as "anger-happiness" for body; and "anger" or "happiness" for face). The fused vector inherits the face. For bimodal emotion recognition at the feature level, C4.5 with 10-fold cross-validation provided the best classification results. See Table V for the emotion recognition results for three subjects.

For the emotions considered, we observe that using the two modalities achieves a better recognition accuracy in general, outperforming the classification using the face modality only, suggesting that using expressive body information adds value to the emotion recognition based solely on the face.

Boosting: Boosting, a popular approach in machine learning, is based on the observation that finding many rough rules of thumb can be a lot easier than finding a single, highly accurate prediction rule [21]. To apply the

boosting approach, a method or algorithm for finding the rough rules of thumb is needed [20]. The boosting algorithm calls this "weak" or "base" learning algorithm repeatedly, each time feeding it a different subset of the training examples. Each time it is called, the base learning algorithm generates a new weak prediction rule, and after many rounds, the boosting algorithm must combine these weak rules into a single prediction rule that, hopefully, will be much more accurate than any one of the weak rules. Boosting, appears to work well for unstable classifiers, such as decision trees, in which a small perturbation in the training set may lead to a significant change in constructed classifier [20, 21]. As a consequence, we decided to test it on the C4.5 classification that we used for emotion recognition.

We experimented the Adaboost M1 method which is a class for boosting our nominal classifier, C4.5 decision tree [21]. Emotion recognition results for all three cases: whole face, body and integration of face and body improved significantly (see Table VI). Even if boosting often dramatically improves the performance, sometimes over-fitting can be a major problem [21].

TABLE VI
EMOTION RECOGNITION RESULTS FOR 3 SUBJECTS USING ADABOOST M1 WITH C4.5.

	Instances	Attributes	Number of Classes	Correctly classified
whole face	265	67	8	87.54 %
body	297	140	5	94.94 %
face and body combined	206	206	8	94.66 %

VI. CONCLUSIONS AND FUTURE WORK

This paper presented an approach to automatic visual recognition of expressive face and upper body action units (FAUs and BAUs) and associated emotions suitable for use in a vision-based affective multimodal framework. In our experiments, the emotion classification using the two modalities achieved a better recognition accuracy outperforming the classification using the individual face modality, suggesting that using expressive body information adds value to the emotion recognition based solely on the face. Moreover, using boosting for the nominal classifier C4.5 improved the emotion recognition results significantly for all three cases: whole face, body and integration of face and body.

The main issue when fusing affective information issued from face and body is to decide on which criteria to use and at what abstraction level to do this fusion. When fusing the bimodal information at the feature level, feature set can be quite large (like in our case). Therefore, it is possible to use a feature selection technique to find the features from both modalities that maximize the performance of the classifier(s). Fusion at the decision level is generally more robust because it exploits many more criteria.

Time is an important factor when integrating the two modalities. In this work, we have combined information from both the face and the body as if it co-occurred exactly at the same time. As future work, we will attempt to use the relationship between the two channels with increased number of subjects and data, with time-stamped analysis, as an added value. We also aim to experiment late fusion of the two modalities in the interpretation process to improve the recognition accuracy as well as using the face mode as principal and the body mode as auxiliary.

REFERENCES

[1] T. Balomenos et al., "Emotion Analysis in Man-Machine Interaction Systems", Proc. of Machine Learning for Multimodal Interaction, pp. 318 – 328, 2004.

[2] M.S. Bartlett, G. Littlewort, C. Lainscsek, I. Fasel and J. Movellan, "Machine learning methods for fully automatic recognition of face expressions and face actions", Proc. of IEEE SMC, pp. 592-597, 2004.

[3] R. T. Boone and J. G. Cunningham, "Children's decoding of emotion in expressive body movement: The development of cue attunement", *Developmental Psychology*, vol. 34, pp. 1007-1016, 1998.

[4] G. R. Bradski, "Computer vision face tracking for use in a perceptual user interface", *Intel Technology Journal*, 2nd Quarter, 1998.

[5] J. Canny, "A computational approach to edge detection", *Proc. of IEEE PAMI*, vol. 8, no. 6, pp. 679-698, 1986.

[6] J. Cassell, "A framework for gesture generation and interpretation", In R. Cipolla and A. Pentland (eds.), *Computer vision in human-machine interaction*, Cambridge University Press (2000).

[7] L.S. Chen and T.S.Huang, "Emotional expressions in audiovisual human computer interaction", Proc. of IEEE ICME, vol. 1, pp. 423-426, 2000.

[8] P. Ekman and W. V. Friesen, *The Face Action Coding System*, Consulting Psychologists Press, San Francisco, CA,1978.

[9] P. Ekman and W. V. Friesen, *Unmasking the face: a guide to recognizing emotions from facial clues*, Imprint Englewood Cliffs, N.J. : Prentice-Hall, 1975.

[10] P. Ellis, "Recognizing faces", *British J. of Psychology*, vol. 66, no.4, pp. 409-426, 1975.

[11] N. Eveno, A. Caplier and P.Y. Coulon, "Key points based segmentation of lips", Proc. of IEEE ICME, vol. 22, pp.125 – 128, 2002.

[12] H. Gunes, M. Piccardi and T. Jan, "Bimodal emotion modelling from face and upper-body gesture for affective HCI", Proc. of OZCHI, CD-ROM (ISBN: 1 74128 079), 10 pages, 2004.

[13] E. Hudlicka, "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, 2003.

[14] A. Kapoor, R. W. Picard and Y. Ivanov, "Probabilistic combination of multiple modalities to detect interest", Proc. of IEEE ICPR, 2004.

[15] B.D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision", Proc. of 7th Int. Joint Conference on Artificial Intelligence, pp. 674–680, 1981.

[16] A. Mehrabian, *Nonverbal Communication*, Aldine-Atherton, Chicago, Illinois, 1972.

[17] M.D. Meijer, "The contribution of general features of body movement on the attributions of emotions", *J. of Nonverbal Behavior*, vol. 13, pp. 247-268, 1989.

[18] J. MacCormick and M. Isard, "Partitioned sampling, articulated objects, and interface-quality hand tracking", Proc. of ECCV, vol.2, pp. 3–19, 2000.

[19] M. Pantic and L.J.M. Rothkrantz, "Towards an affect-sensitive multimodal human-computer interaction", *Proc. of the IEEE*, vol. 91, no. 9, pp. 1370-1390, 2003.

[20] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools with Java implementations*, Morgan Kaufmann, San Francisco, 2000.

[21] R. E. Schapire, "The boosting approach to machine learning: An overview", Proc. of MSRI Workshop on Nonlinear Estimation and Classification, 2002.

[22] L.G. Shapiro and A. Rosenfeld, *Computer Vision and Image Processing*, Boston, Academic Press, 1992.

[23] K. Sobottka and I. Pitas, "A novel method for automatic face segmentation, face feature extraction and tracking", *Image Communication*, Elsevier, 1997.

[24] L. Wu, S. L. Oviatt, and P. R. Cohen, "Multimodal Integration-A Statistical View", *IEEE Transactions on Multimedia*, vol. 1, no. 4 , pp. 334-341, 1999.