Available online at www.sciencedirect.com

ScienceDirect

Journal of Network and
Computer Applications 30 (2007) 1334–1345

Journal of NETWORK and COMPUTER APPLICATIONS

www.elsevier.com/locate/jnca

Bi-modal emotion recognition from expressive face and body gestures

Hatice Gunes*, Massimo Piccardi

Computer Vision Research Group, Faculty of Information Technology, University of Technology, Sydney (UTS), PO Box 123, Broadway, NSW, 2007, Australia

Received 20 September 2006; accepted 20 September 2006

Abstract

Psychological research findings suggest that humans rely on the combined visual channels of face and body more than any other channel when they make judgments about human communicative behavior. However, most of the existing systems attempting to analyze the human nonverbal behavior are mono-modal and focus only on the face. Research that aims to integrate gestures as an expression mean has only recently emerged. Accordingly, this paper presents an approach to automatic visual recognition of expressive face and upper-body gestures from video sequences suitable for use in a vision-based affective multi-modal framework. Face and body movements are captured simultaneously using two separate cameras. For each video sequence single expressive frames both from face and body are selected manually for analysis and recognition of emotions. Firstly, individual classifiers are trained from individual modalities. 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 classification using the individual facial or bodily modality alone.

Keywords: Bi-modal emotion recognition; Facial expression; Expressive body gestures; Feature-level fusion; Decision-level fusion

*Corresponding author. Tel.: +61 2 95144526; fax: +61 2 95144535.
E-mail address: haticeg@it.uts.edu.au (H. Gunes).

1084-8045/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jnca.2006.09.007

1. Introduction and methodology

Automated emotion recognition plays an important role in affective computing, a new paradigm of human–computer interaction (HCI) conceptualising computers as affective entities. However, automated emotion recognition proves in itself a very challenging task. To date, the most promising results have been achieved in recognition of emotions from facial expressions. Other modalities such as body movements and gestures have only recently started attracting the attention of the HCI community (e.g. Hudlicka, 2003). Moreover, despite the common use of multiple modalities in human–human interaction (HHI), relatively few works have focused on implementing emotion recognition systems using affective multimodal data. The most common approach has been to combine facial expression with audio information Pantic et al. (2005). Kapoor et al. (2004) addressed the problem of detecting the affective states of high-interest, low-interest and "refreshing" in a child who is solving a puzzle. To this aim, they combined sensory information from the face video, the posture sensor (a chair sensor) and the game being played in a probabilistic framework. Balomenos et al. (2004) combined facial expressions and hand gestures for the recognition of six prototypical emotions.

In our work, we aim to extend the affective channels used for emotion recognition to include spontaneous body gestures from the whole upper body. Our motivation stems from a study by Ambady and Rosenthal (1992) suggesting that the most significant channel for judging behavioral cues of humans appears to be the visual channel for the facial expressions and body gestures. Accordingly, 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. We focus on facial expressions and body gestures (i.e. shoulder shrug) separately and analyze the individual frames, namely neutral and expressive frames. After describing the feature extraction techniques for face and body briefly, classification results from four subjects are presented. Firstly, individual classifiers are trained separately with face and body features for mono-modal classification into labeled emotion categories. Then, we fuse affective face and body modalities for classification into combined emotion categories at (a) at the feature-level; and (b) at the decision-level. The system framework illustrating these steps is shown in Fig. 1.

1.1. *Modality 1: facial expression*

The leading study of Ekman and Friesen (1975) 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. Brave and Nass (2002) provide details of the facial cues for the displayed emotions. 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 recognized by our system based on the visual changes occurring on the face.

1.2. *Modality 2: expressive upper-body gestures*

Human recognition of emotions from body movements and postures is still an unresolved area of research in psychology and non-verbal communication. Ambady and Rosenthal (1992) found out that humans rely on the combined visual channels of face and

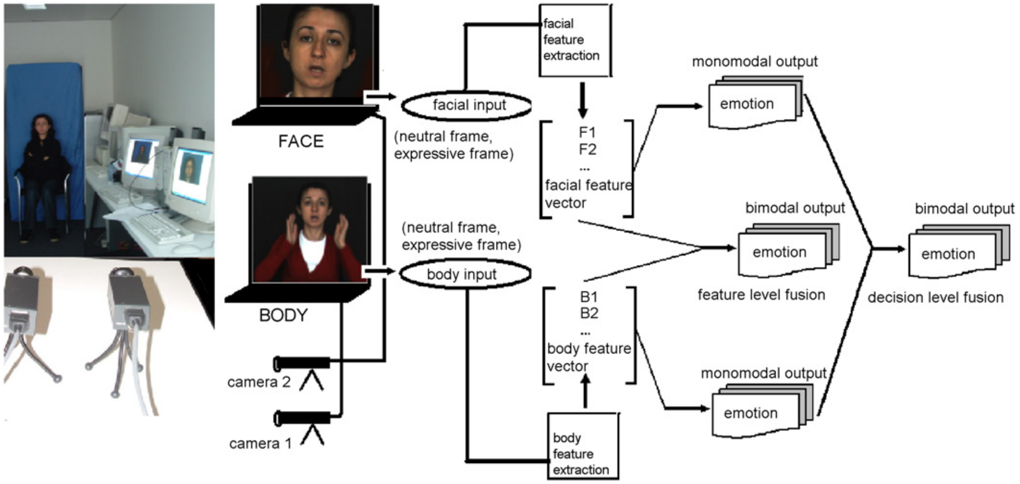

Fig. 1. System framework for mono-modal and bi-modal emotion recognition.

Table 1
List of the facial emotions recognized by our system and the changes that occur on the face when they are displayed

Anxiety
Lip bite; stretching of the mouth; eyes turn up/down/left/right; lip wipe

Anger
Brows lowered and drawn together; lines appear between brows; lower lid is tensed and may or may not be raised; upper lid is tense and may or may not be lowered due to brows action; lips are either pressed firmly together with corners straight or down or open

Disgust
Upper lip is raised; lower lip is raised and pushed up to upper lip or it is lowered; nose is wrinkled; cheeks are raised; brows are lowered; tongue out

Fear
Brows raised and drawn together; forehead wrinkles drawn to the center; upper eyelid is raised and lower eyelid is drawn up; mouth is open; lips are slightly tense or stretched and drawn back

Happiness
Corners of lips are drawn back and up; mouth may or may not be parted with teeth exposed or not; cheeks are raised; lower eyelid shows wrinkles below it, and may be raised but not tense; wrinkles around the outer corners of the eyes

Uncertainty
Lid drop; inner brow raised; outer brow raised; chin raised; jaw sideways; corners of the lips are drawn downwards

body more than any other channel when they make judgments about human communicative behavior. In his paper Coulson (1992) presented experimental results on attribution of six emotions (anger, disgust, fear, happiness, sadness and surprise) to static body postures by using computer-generated figures. 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. (2005) clearly discuss the issue of emotion recognition from bodily cues and provide useful references in a recent publication in the context of national security. We provide a table based on the cues described by Coulson (1992) and Burgoon et al. (2005) with the list of the expressive body gestures and the correlation between the gestures and the emotion categories used in the recordings of our database (see Table 2).

1.3. Data collection

There have been some attempts to create comprehensive test-bed for comparative studies of facial expression analysis (see Pantic et al., 2005 for a detailed study). However, existing databases lack expressive quality of the body and do not take into consideration the relationship between the bodily parts (i.e. between hands; hands and the face; hands, face and shoulders, etc.). Therefore, they cannot be used for the extensive analysis of human nonverbal communicative behavior. Moreover, none of the aforementioned works Balomenos et al. (2004); Kapoor et al. (2004) created an extensive affective gesture database for common research use (please see Gunes and Piccardi, 2006b for a comparative study on existing affect databases and their limitations). To cope with the existing limitations, we created a bimodal database (FABO) that consists of recordings of facial expressions alone and combined face and body expressions (Gunes and Piccardi, 2006a). We recorded the sequences simultaneously using two fixed SONY XCD-X710CR cameras, connected to two different PCs with a simple setup and uniform background. One camera was placed specifically capturing the head only and the second camera was placed in order to capture upper-body movement from the waist above. We choose to use two cameras due to the fact that current off-the-shelf technology still does not provide us with frames with the required quality to process detailed upper-body and face information together. 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. Examples of

Table 2
List of the bodily emotions recognized by our system and the changes that occur on the body when they are displayed

Anxiety
Hands close to the table surface; fingers moving; fingers tapping on the table

Anger
Body extended; hands on the waist; hands made into fists and kept low, close to the table surface

Disgust
Body backing; left/right hand touching the neck or face

Fear
Body contracted; body backing; hands high up, trying to cover bodily parts

Happiness
Body extended; hands kept high; hands made into fists and kept high

Uncertainty
Shoulder shrug; palms up

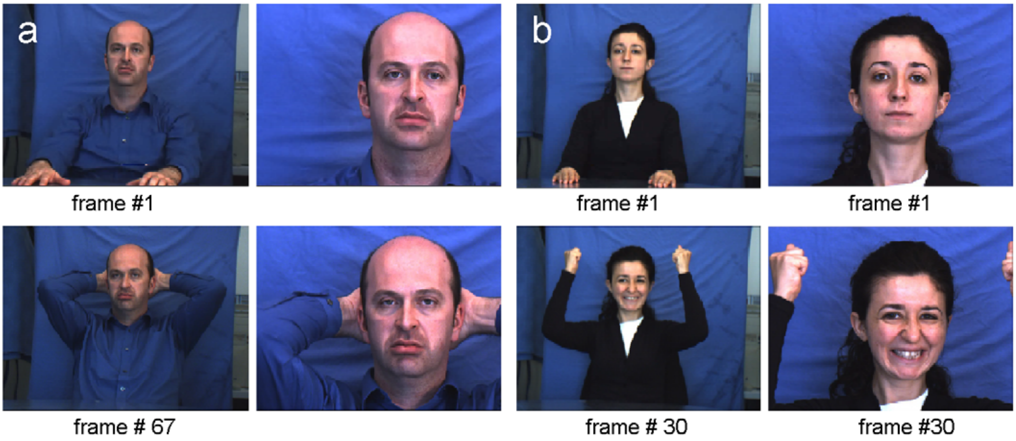

Fig. 2. Example sequences from FABO obtained from body (left columns) and face (right columns) cameras.

data sequences recorded by camera 1 (for body) and camera 2 (for face) can be seen in Fig. 2.

2. Feature extraction

A vast literature covers techniques for facial feature extraction (i.e. Yang et al., 2002). 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. 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’’).

2.1. Face feature extraction

Firstly, morphological operations are used to smooth the image. We then apply skin color segmentation based on HSV color space. We obtain the face region by choosing the largest connected component among the candidate skin areas. We then employ closing (dilation and erosion) and find the contour of the face that returns the filled face region. 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 (1998). We first enhance the face region by histogram equalization. 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 (1998) 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. We detect and locate eyes, eyebrows, nostrils, and chin. 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 computing the optical flow in such bounding rectangles. Furthermore, we analyze the wrinkle changes by using edge density per unit area against a threshold.

2.2. *Body feature extraction*

Our body model is a combination of a silhouette based and color based body models to determine the location of the upper-body parts while the person is in a sitting posture. It is used to predict the locations of the body parts (head, torso, shoulders and hands). 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. 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. Since the remaining region is bigger than the original one it is restored to its original size by the erosion procedure. 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.

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. We then calculate the centroid of these regions in order to use them as reference points for the body movement. We employ the Camshift technique Bradski, 1998 for tracking the hands and comparison of bounding rectangles is used to predict their locations in subsequent frames (see Fig. 3).

Hand pose and orientation estimation: Orientation 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 Bradski, 1998. Using this information we decide if the hand is in a vertical or horizontal position. After estimating the initial pose of the hand it is possible to determine the position of the fingers. We define four categories for finger position estimation: up, down, right and left. We use this information when classifying the feature vectors into various body movements (e.g. arms crossed, hands touching the head, etc.).

3. Emotion recognition and experimental results

In our experiments we select a whole frame sequence where an expression is formed in order to perform emotion recognition. We processed 54 sequences in total, 27 for face and 27 for body from four subjects. We processed about 1500 frames for the face and 1500 for the body. 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. The ground truth in this experiment is based on the fact that (a) subjects were asked to perform expressive face and body gestures corresponding to particular emotions and (b) the subjects believed that they were performing accordingly. After obtaining the feature vector for face and body separately we performed emotion recognition using Weka, a publicly available toolbox for

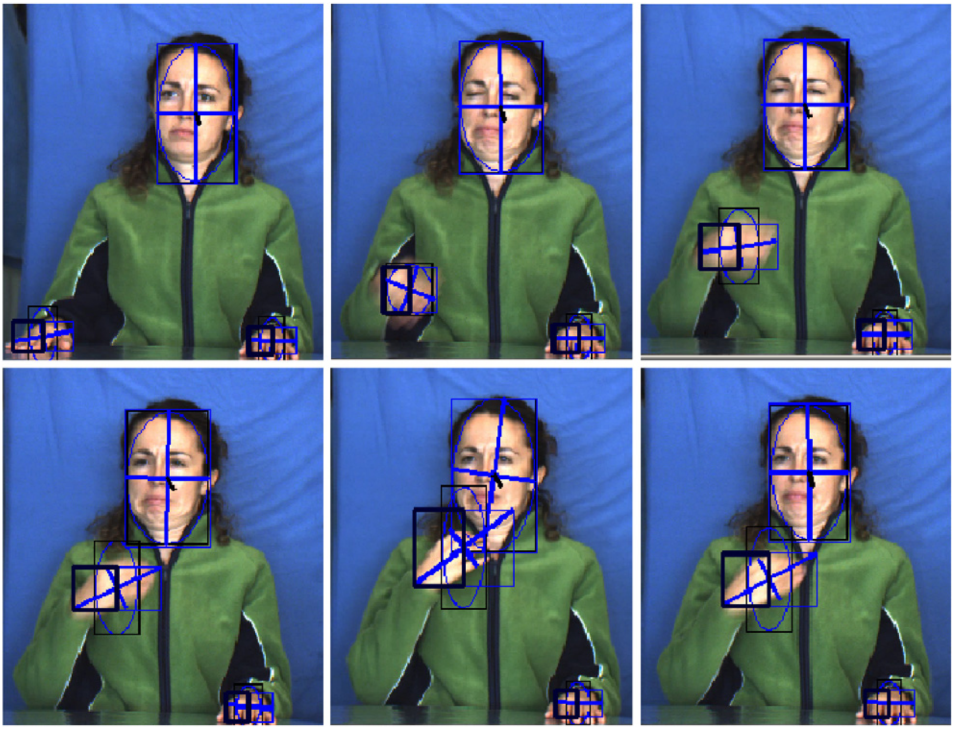

Fig. 3. Camshift tracking when one hand merges with the face.

automatic classification Witten and Frank, 1999. We performed emotion recognition in two stages: mono-modal and bi-modal emotion recognition. The details of these procedures are explained in the following sections.

3.1. Mono-modal emotion recognition

The face feature vector and the body feature vector consist of 148 and 140 features, respectively. Such values are relatively high and may pose a challenge to the training of the classification algorithms. A basic rule for deciding the minimum number of samples in a training set is for that to be proportional to the number of features used. Since our number of training samples is limited, a classifier with a reduced feature set could be better trained than a classifier using the whole feature set. Therefore, we explored a feature selection technique to find the feature subset maximizing the performance of the classifiers. We applied attribute selection in Weka by a best-first search method in forward direction. This selection method searches the space of attribute subsets by greedy hill-climbing augmented with a backtracking facility. Setting the number of consecutive non-improving nodes allowed us control of the level of backtracking done.

Feature selection: The best-first search method in Weka evaluated a total number of 4478 subsets and found the best subset with a merit of 74.6% for the facial input data. The number of features selected was 29 among 148 features. The same method evaluated a total number of 2111 subsets and found the best subset with a merit of 93% for the body input data. The number of features selected was 11 among 140 features.

Mono-modal recognition based on the face: After the feature selection procedure, we fed the reduced face feature vector to the classifiers for mono-modal emotion recognition. The best recognition results in Weka were obtained with the BayesNet classification algorithm; results are presented in Table 3 (first row). Additionally, Table 4 presents the full "confusion matrix" for the reduced facial feature vector (29 features) for the four subjects. From the left column of Table 4 it can be seen that a number of "anger" samples were classified as "disgust". This might be due to a certain self-similarity between these two classes. Moreover, some of the "uncertainty" samples were classified as "happiness". This can be explained with the fact that all of the happiness expressions in the experiments are performed with open mouth, hence lower cheek regions are pulled down; similar movement is done when performing uncertainty by pulling down the lower cheek regions.

Mono-modal recognition based on the body: After the feature selection procedure, we similarly fed the reduced body feature vectors to the classifiers for mono-modal emotion recognition. The best recognition results in Weka were again obtained with the BayesNet classification algorithm; results are presented in Table 3 (second row). Additionally, the right column of Table 4 presents the full "confusion matrix" for the reduced body feature vector (11 features) for the four subjects. From the right column of Table 4, it can be seen that a number of "anger" samples are classified as "anxiety". This can be explained with the fact that "anxiety" was performed by tapping fingers on the table and part of "anger" was performed by making the hands into fists. For both of these gestures the amount of hand movement is limited if compared to other classes such as disgust, happiness and fear.

Table 3
Emotion recognition results for the reduced face and body feature vector

Modality	Classifier	Training	Test	Attributes	Number of classes	Correctly classified
Face	BayesNet	414	386	29	6	76.40%
Body	BayesNet	424	386	11	6	89.90%

Table 4
Confusion matrices for the reduced face and body feature vectors

Confusion matrix for the reduced facial feature vector (29 attributes) for four subjects using BayesNet. Rows: true class; columns: actual classification	Confusion matrix for the reduced body feature vector (11 attributes) for four subjects using BayesNet. Rows: true class; columns: actual classification
a b c d e f	a b c d e f
41 0 0 1 0 0 |a = disgust	42 0 0 0 0 0 |a = disgust
0 83 0 1 0 0 |b = happiness	0 80 4 0 0 0 |b = happiness
0 1 14 3 0 0 |c = fear	0 0 18 0 0 0 c = fear
22 4 0 81 0 0 |d = anger	0 0 0 71 0 35 |d = anger
1 15 0 1 28 1 |e = uncertainty	0 0 0 0 47 0 |e = uncertainty
13 0 0 0 3 73 |f = anxiety	0 0 0 0 0 89 |f = anxiety

Overall, from Table 3 we can conclude that body movements are more distinguishable between themselves than facial movements. This is due to the fact that facial movements are small movements and even high resolution might not be able to provide absolute recognition accuracy.

3.2. *Bi-modal emotion recognition*

In general, modality fusion is about integrating all single modalities into a combined representation (Wu et al., 1999). One of the key issues in multimodal data processing is to decide when to combine the information. Typically, fusion is either done at the feature-level or deferred to the decision-level (Wu et al., 1999). To make the fusion issue tractable, the individual modalities are usually assumed independent of each other.

Feature-level fusion: Feature-level fusion is performed by using the extracted features from each modality and concatenating these features into one larger vector. The resulting vector is input to a single classifier which uses the combined information to assign the test samples into appropriate classes. We fuse face and body features of the corresponding expressive frames from the videos obtained from face and body cameras. However, we obtain a large feature vector consisting of 288 features. Similarly to the mono-modal emotion recognition, we decided to use a feature selection method prior to classification. Best-first search method was used with ten-fold cross validation to obtain a decisive reduction in the features' number (14 after selection). We experimented various classifiers on a data set consisting of 412 training and 386 testing instances. For the feature set with 14 attributes, BayesNet provided, again, the best classification accuracy. The recognition results and the confusion matrix obtained are presented in Table 5. Table 5 shows that a number of "anger" samples were classified as "anxiety". This case is similar to the mono-modal emotion recognition results based on body feature set alone (compare with Table 4, right column), however the number of misclassifications is much reduced. In general, for the emotions considered, we observe that using the two modalities achieves better recognition accuracy, outperforming the classification using the face or body modality alone, suggesting that using expressive face and body information adds accuracy to the emotion recognition based solely either on the face or the body. To correctly interpret these results, it is important to recall that our experiment tests unseen instances from the

Table 5
Emotion classification for the combined feature vector with BayesNet into 6 emotion categories (disgust, happiness, fear, anger, uncertainty and anxiety)

	Correctly classified (%)	Confusion matrix
Overall	94.02	a b c d e f < — classified as
Anger	100	42 0 0 0 0 0 |a = disgust
Disgust	100	0 80 4 0 0 0 |b = happiness
Fear	81.8	0 0 18 0 0 0 |c = fear
Happiness	100	0 0 0 87 0 19 |d = anger
Uncertainty	100	0 0 0 0 46 0 |e = uncertainty
Anxiety	82.4	0 0 0 0 0 89 |f = anxiety

same subjects used for the training phase. Accuracy might be lower for totally unseen subjects.

Decision-level fusion: Decision-level fusion enables each modality to be first pre-classified independently and the final classification is based on the fusion of the outputs the different modalities. Designing optimal strategies for decision-level fusion is still an open research issue, depending also on the framework chosen for optimality. Various approaches have been proposed including the sum rule, product rule, using weights, maximum/minimum/median rule, majority vote, etc. Kuncheva, 2002. We analyzed the first three techniques mentioned above for our system, namely the 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. Under a Maximum-a posteriori (MAP) approach, X must be assigned to that of M possible classes, $(w_1, ..., w_k, ..., w_M)$, having maximum posterior probability $p(w_k|X)$. An early integration approach would compute such a probability explicitly. In late integration, instead, two separate classifiers provide the posterior probabilities $p(w_k|x_f)$ and $p(w_k|x_b)$ for face and body, respectively, to be combined into a single posterior probability $p(w_k|X)$ with one of the fusion methods described in the following. Moreover, in the infrequent case in which the combined $p(w_k|X)$ has exactly the same value for two or more classes, we resort to the classification provided by the face classifier since we believe this is the "major" mode in our bi-modal approach. If the same happens for $p(w_k|x_f)$, we arbitrarily retain the first class in appearance order. The description of the three criteria we compared is given in Table 6. In our case, the face modality is assumed to be the main modality. Thus, we assigned arbitrary weights as follows: $\sigma_f = 0.7$ for the face modality and $\sigma_b = 0.3$ for the body modality. The late fusion results for sum, product and weight criteria are all presented in Table 7. According to our experimental results, with 91.1% recognition accuracy sum rule provides better fusion results than product or weight criteria.

Table 6
Description of the three late-fusion criteria used: sum, product and weight

Assign $X \rightarrow w_k$	Sum rule	$k = argmax_{k=1}^{M} p(w_k|x_f) + p(w_k|x_b)$
	Product rule	$k = argmax_{k=1}^{M} p(w_k|x_f) * p(w_k|x_b)$
	Weight criterion	$k = argmax_{k=1}^{M} \sigma_f p(w_k|x_f) + \sigma_b p(w_k|x_b)$

Table 7
Emotion recognition results (%) for late fusion using sum, product and weight criteria on the test set of 386 samples

	Sum rule	Product rule	Weight criterion ($\sigma_f = 0.70$, $\sigma_f = 0.30$)
Overall	91.1	87.3	79.7
Anger	77.5	67.2	71.9
Disgust	100	100	76.1
Fear	100	100	88
Happiness	97.6	97.6	97.6
Uncertainty	82.6	80.4	60.8
Anxiety	100	96.6	82

4. Conclusions

This paper presented an approach to automatic visual analysis of expressive face and upper-body gestures and associated emotions suitable for use in a vision-based affective multimodal framework. In our work, we focused on facial expressions and body gestures separately and analyzed individual frames, namely neutral and expressive frames. Firstly, two classifiers were trained separately with face and body features for mono-modal classification into labeled emotion categories. We then fused affective face and body modalities for classification into combined emotion categories (a) at the feature-level, in which the data from both modalities are combined before classification and (b) at the decision-level, in which the outputs of the mono-modal systems are integrated by the use of product, sum and weight criteria. Our experimental results show that: the emotion classification using the two modalities combined achieves better recognition accuracy in general, outperforming the classification using the face modality or body modality alone; by comparing Tables 5 and 7, early fusion seems to achieve better recognition accuracy compared to late fusion; and that amongst the three late fusion approaches, the sum rule proved the best way to fuse such two modalities. Future extensions of this work will verify the consistency of these findings on full-length expressive video sequences.

References

Ambady N, Rosenthal R. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. Psychological Bulletin 1992;111(2):256–74.

Balomenos T. Raouzaiou A, Ioannou S, Drosopoulos A, Karpouzis K, Kollias SD. Emotion analysis in man–machine interaction systems. In: Lecture notes in computer science, vol. 3361. Berlin: Springer; 2004. p. 318–28.

Bradski GR. Computer vision face tracking for use in a perceptual user interface. Intel Technology Journal Second Quarter, 1998.

Brave S, Nass C. Emotion in HCI. In: Jacko J, Sears A, editors. The human–computer interaction handbook: fundamentals, evolving technologies and emerging applications. Hillsdale, NJ: Lawrence Erlbaum Associates; 2002.

Burgoon JK, Jensen ML, Meservy TO, Kruse J, Nunamaker JF. Augmenting human identification of emotional states in video. In: Proceedings of the international conference on intelligent data analysis, 2005, https://analysis.mitre.org/proceedings/Final_Papers_Files/344_Camera_Ready_Paper.pdf.

Coulson M. Attributing emotion to static body postures: recognition accuracy, confusions, and viewpoint dependence. Journal of Nonverbal Behavior 1992;28(2):117–39.

Ekman P, Friesen WV. Unmasking the face: a guide to recognizing emotions from facial clues. Englewood Cliffs, NJ: Prentice-Hall; 1975.

Gunes H, Piccardi M. A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In: Proceedings of IEEE international conference on pattern recognition, 2006a. p. 1148–53.

Gunes H, Piccardi M. Creating and annotating affect databases from face and body display: A contemporary survey. In: Proceedings of IEEE international conference on systems, man and cybernetics, 2006b. p. 2426–33.

Hudlicka E. To feel or not to feel: the role of affect in human–computer interaction. International Journal of Human–Computer Studies 2003;59(1–2):1–32.

Kapoor A, Picard RW. Ivanov Y. Probabilistic combination of multiple modalities to detect interest. In: Proceedings of IEEE international conference on pattern recognition, 2004. p. 969–72.

Kuncheva LI. A theoretical study on six classifier fusion strategies. IEEE Transactions on Pattern Analysis and Machine Intelligence 2002;24(2):281–6.

Pantic M, Sebe N, Cohn J, Huang T. Affective multimodal human–computer interaction. In: Proceedings of ACM international conference on multimedia, 2005. p. 669–76.

Sobottka K, Pitas I. A novel method for automatic face segmentation facial feature extraction and tracking. Journal of Signal Processing and Image Communication 1998;12(3):263–81.

Witten H, Frank E. Data mining: practical machine learning tools and techniques with java implementations. San Francisco: Morgan Kaufmann; 1999.

Wu L, Oviatt SL, Cohen PR. Multimodal integration-a statistical view. IEEE Transactions on Multimedia 1999;1(4):334–41.

Yang M, Kriegman D, Ahuja N. Detecting faces in images: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 2002;24(1):34–58.