Affect Recognition from Face and Body Early Fusion vs. Late Fusion

Hatice Gunes
Computer Vision Research Group
Faculty of Information Technology
University of Technology, Sydney (UTS)
Australia
haticeg@it.uts.edu.au

Massimo Piccardi
Computer Vision Research Group
Faculty of Information Technology
University of Technology, Sydney (UTS)
Australia
massimo@it.uts.edu.au

Abstract - *This paper presents preliminary results of automatic visual emotion recognition from two modalities: face and body. Firstly, individual classifiers are trained from individual modalities. Secondly, we fuse facial expression and affective body gesture information first at a feature-level, in which the data from both modalities are combined before classification and later at a decision-level, in which we integrate the outputs of the mono-modal systems by the use of suitable criteria. We then evaluate these two fusion approaches, in terms of performance over mono-modal emotion recognition based on facial expression modality only. The emotion classification using the two modalities achieves a better recognition accuracy outperforming the classification using the individual facial modality. Moreover, fusion at a feature level performs better recognition than fusion at a decision level.*

Keywords: Input fusion, affect recognition, action unit recognition, facial expression, body gesture.

1 Introduction

Cassell's research [5] shows that humans are more likely to consider computers to be human-like when those computers understand and display appropriate nonverbal communicative behavior. 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 emotion 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 facial expressions (e.g. [2]). Emotion recognition via body movements and gestures has only recently started attracting the attention of computer science and HCI communities [14]. However, the interest is growing with works similar to the one presented in [1]. So far, most of the work in affective computing focuses 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) that occur together[14].

Integrating multiple modalities for emotion recognition is motivated by human-human interaction. People naturally communicate multi-modally by combining language, tone, gesture and head movement, body movement and posture and facial expression and possess a refined mechanism for data fusion. Machines, to date, are less able to emulate this ability. This issue is central to current research in affective multimodal HCI [22]. Multimodal interfaces function in a more efficient and reliable way, modalities usually complement each other and help improve the accuracy and robustness of affective and perceptual interfaces. Mathematical reasons for combining modalities is to increase certainty for decision making since combination of multiple observations (even from the same source) is statistically advantageous by increasing accuracy of measurements.

Relatively few efforts have focused on implementing emotion recognition systems using multimodal data [14]. The most common approach has been to combine facial expressions with audio information [22]. De Silva et al. [9] proposed a rule-based audio-visual emotion recognition system. The outputs from audio (prosodic features), and video (the maximum distances and velocities between six specific facial points) classifiers are fused at the decision-level [9]. Chen et al. [6] track the predefined basic motions on the face and use prosodic features including pitch, energy and rate of speech for the audio mode. They concatenated the best audio and video features to form a bimodal feature vector and found out that recognition using the bimodal approach improves tremendously. Yoshitomi et al. proposed a multimodal system that combines speech and visual information with thermal distribution acquired by an infrared camera at a decision-level with pre-determined weights [30]. Balomenos et al. [1] combined facial expressions and hand gestures for recognition of 6 emotions. They use facial points from MPEG-4 compatible animation by first defining a mapping between these points and the movement of specific feature points. Under each emotion category the facial feature movements and hand movements are defined. Eventually, they use weights to account for the reliability of the two subsystems as far as the emotional state estimation is concerned [1].

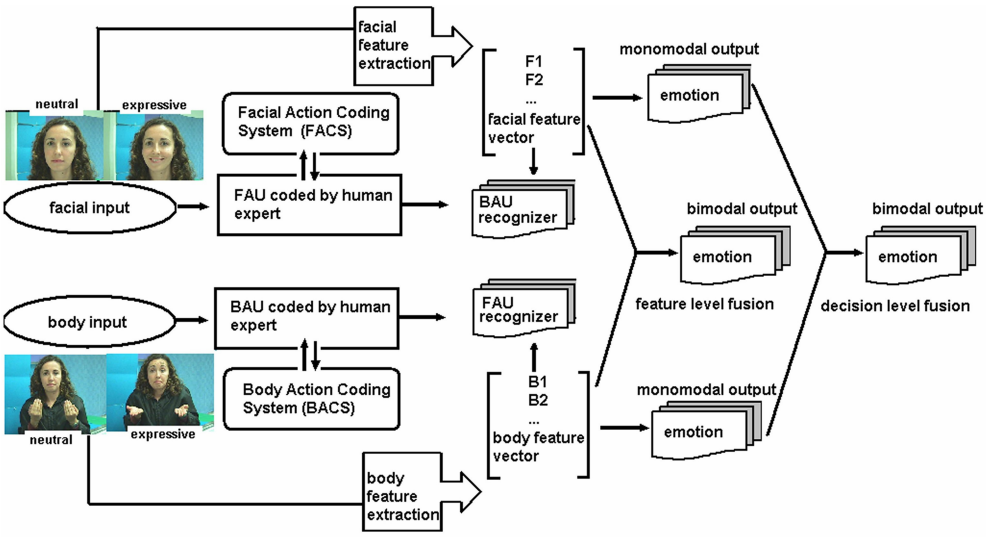

Figure 1. System framework for FAUs, BAUs and emotion recognition.

Kapoor et al. focused on machine recognition of affect using multiple modalities [15]. They look at the problem of detecting the affective states of high-interest, low-interest and "refreshing" in a child who is solving a puzzle. They combine sensory information from the face, the postures and the state of the puzzle in a probabilistic framework. The raw data from the camera, the posture sensor and the game being played is first analyzed to extract relevant features. Based on the extracted features, all of these experts predict the affective state independently. Probabilistic models of error and the critics, which predict the performance of the individual expert on the current input, are used to combine the experts' beliefs about the correct class. They demonstrated that combining multiple modalities achieves much better recognition accuracy than classification based on individual channels [15].

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. Our motivation is based on the fact that all of the studies mentioned above have improved the performance of emotion recognition systems by the use of multimodal information (e.g. [1,6,9,14,30]). Initially, we focus on FAUs and specific BAUs (i.e. shoulder shrug) and analyze the static images, namely neutral and expressive frames. After describing the feature extraction techniques for face and body briefly, classification results from three subjects are presented. Firstly, individual classifiers are trained separately with face and body features for classification into BAU and FAU categories. Secondly, the same procedure is applied for mono-modal classification into labeled emotion categories. Finally, we fuse affective face and body modalities for classification into combined emotion categories (a) at a feature-level, in which the data from both modalities are combined before classification (b) at a decision-level, in which the outputs of the mono-modal systems are integrated by the use of average and weight criteria. We observe that the emotion classification using the two modalities achieves a better recognition accuracy outperforming the classification using the individual facial modality. With these preliminary experiments, our aim is to compare which fusion approach is more suitable for our vision-based multimodal framework that uses face and body gesture for affect recognition.

2 Methodology

Initially, we analyze the two modalities, namely facial action units (FAUs) and body action units (BAUs) separately, as described in the following sections. Our task is to analyze expressive cues within HHI and HCI which mostly takes place as dialogues in sitting position, hence, the expressiveness of the lower part of the body is ignored in 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. The general system framework for both mono-modal and bi-modal emotion recognition is depicted in Figure 1.

2.1 Modality 1: Facial Action Units

The leading study of Ekman and Frisen [9,10] 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.

However, according to [9], six universal emotion categories provide an incomplete description of all facial expressions. In order to capture the subtlety of human emotion, recognition of fine-grained changes in facial expressions is needed [9]. Ekman and Friesen [9] developed the Facial

Action Coding System (FACS) for describing facial expressions by facial action units (FAUs). Of 44 FACS AUs defined, 30 AUs are anatomically related to the contractions of specific facial muscles: 12 are for upper face, and 18 are for lower face. AUs can be classified either individually or in combination. In order to show how each emotion is defined in FACS we present an example below of how the emotion "surprise" is defined as a combination of four FAUs [9]:

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 action units (AUs) [2]. Table 1 provides the list of the FAUs recognized by our system.

Table 1. List of the FAUs recognized by our system and their description.

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

2.2 Modality 2: Body Action Units

Propositional expressive gestures are described as specific movements of specific bodily parts or postures corresponding to stereotypical emotions (e.g. bowed head and dropped shoulders showing sadness) [3]. Non-propositional expressive gestures are, instead, not coded as specific movements but form the quality of movements (e.g. direct/flexible) [18]. In this paper, we analyze the propositional gestures from static frames since analysis of non-propositional gestures would require time-stamped analysis (not addressed in this work). 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 [12,18,20,21,24,27,28] together with the results obtained from our experiments, in [13]. Table 2 provides the list of the BAUs and the correlation between the BAUs and the emotion categories recognized by our system.

Table 2. List of the BAUs recognized by our system and the correlation between the BAUs and emotion labels.

BAU	BAU description	Emotion	BAU	BAU description	Emotion
0	neutral	Neutral	9	left hand on left shoulder	sad_fear
1	body extended	angry_happy	10	right hand on right shoulder	sad_fear
2	body contracted	fear_sad	11	shoulder shrug	uncertain
3	left hand moved up	angry_disgust	11+12	shoulder shrug+ palms up	uncertain_angry
4	right hand moved up	angry_disgust	12	palms up	uncertain_angry
5	left hand touching the head	sad_surprise	13	two hands up	angry_happy
6	right hand touching the head	sad_surprise	14	two hands touching the head	fear_sad
7	left hand touching the face	sad_surprise	15	two hands touching the face	fear_sad
8	right hand touching the face	sad_surprise	16	arms crossed	angry_fear

3 Feature Extraction

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. 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 still images of the face and body, independently by exploiting skin color information. Hand displacement is computed as the motion of the centroid coordinate. We employ the Camshift algorithm [1] 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 [1]. We also calculate the optical flow by comparing the displacement from the neutral face to the expressive face using the Lucas-Kanade Algorithm [4].

In the first frame, the body is in neutral position (hands held in front of the torso). In the following frames, the system can handle in-line rotation of the face and hands. The first and last frames (neutral and peak) were used for training and testing of FAUs and BAUs. All samples were initially AU coded by two human experts.

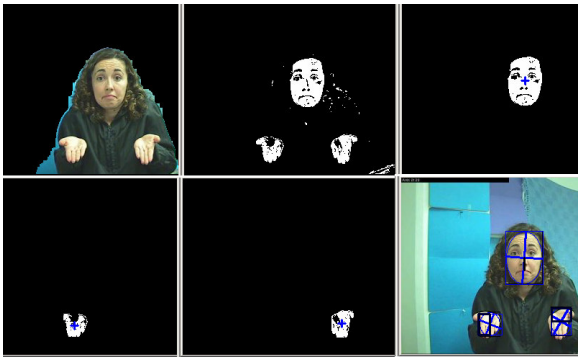

Figure 2. (first row)Expressive silhouette, body parts located, face located; (second row) left and right hand located, body parts tracked with Camshift.

4 Monomodal Emotion Recognition

For FAU and BAU recognition we used Weka-3-4, a tool for automatic classification [23]. Amongst the various classifiers provided by this tool, BayesNet provided the best classification result with 10-fold cross validation for FAUs and BAUs recognition. The results are presented in Table 3.

Table 3. FAUs and BAUs classification results for 3 subjects using BayesNet

	Instances	Attributes	Number of Classes	Correctly classified
Whole Face	313	67	65	69.329 %
Upper Face	246	67	20	73.170 %
Lower face	273	67	28	72.527 %
Body	297	140	22	76.431 %

For FAU and BAU classification, we created a separate class for each different combination of single AUs, for face and body separately. For FAU classification, we divided the instances for classification into upper and lower FAUs, separately. The classification accuracy for the upper face seems to be better than the lower face or whole face AU classification. As mentioned earlier these results are preliminary. We believe that increasing the training and testing instances 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.

We then used the same procedure to classify the data from expressive face and body into labeled emotion categories separately. Amongst the various classifiers provided by Weka [23], Decision Trees provided the best classification result with 10-fold cross validation. The results are presented in Table 4.

Table 4. Emotion recognition results for 3 subjects using 156 training samples and 50 testing instances. Emotion categories used for face are disgust, happiness, surprise, fear, anger, sadness, happy_surpise, uncertanity. Emotion categories used for upper-body are anger_disgust, anger_fear, anger_happy, fear_sad_surprise, uncertanity_fear_surprise, uncertanity_surprise.

	Attributes	Number of Classes	Classifier	Correctly classified
Body*	140	6	Bayes Net	90 %
Body	140	6	Decision Trees	80 %
Face	67	8	Bayes Net	74 %
Face*	67	8	Decision Trees	78 %

* The classification results used for late fusion.

5 Bimodal Emotion Recognition

In general, modality fusion is to integrate all incoming single modalities into a combined single representation [29]. One of the key issues in multimodal data processing is to decide when to combine the information [22]. Typically, fusion is either done at a feature level or deferred to the decision level [29]. To make the fusion issue tractable the individual modalities are usually assumed independent of each other. This simplification ignores the relationship between the modalities (i.e., using the facial expression recognition information to inform the gestural recognition processing).

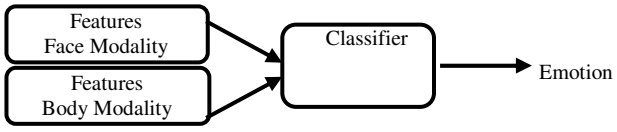

Figure 3. Feature-level fusion

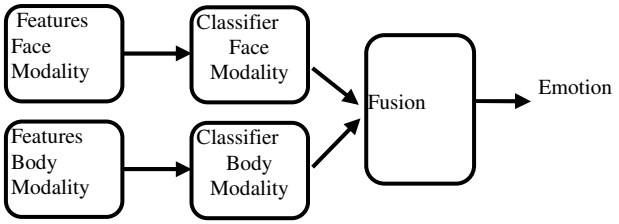

Figure 4. Decision-level fusion

In this work, our fusion strategy can be referred to as micro temporal fusion, combining information that is produced either in parallel or over overlapping time intervals [8]. We fuse the static frames of expressive face and body that carry information on the emotion displayed being in their apex, at the time of the recording. To fuse the affective facial and body information we implemented two different approaches: feature-level fusion, in which a single classifier with features of both modalities are used (Figure 1); and, decision level fusion, in which a separate classifier is used for each modality, and the outputs are combined using some criteria (Fig. 3 and Fig.4).

5.1 Feature Level Fusion

Feature-level fusion is performed by using the extracted features from each modality and concatenating these features into one large vector (See Fig.3). We transform the images into a representation that decomposes the images into features (e.g. movement of facial features, shoulders, hands etc.) and perform fusion in this domain. The resulting feature set can be quite large (in our case increases from 67 and 140 to 206). Therefore, it is possible to use a feature selection technique to find the features from both modalities that maximize the performance of the classifier. We apply attribute selection on combined input data with Best first search method in Weka 3.4 [23]. It searches the space of attribute subsets by greedy hill-climbing augmented with a backtracking facility. Setting the number of consecutive non-improving nodes allowed controls the level of backtracking done. Best first may start with the empty set of attributes and search forward, or start with the full set of attributes and search backward, or start at any point and search in both directions (by considering all possible single attribute additions and deletions at a given point). In our case, it evaluated a total number of 8259 subsets and found the subset with a merit of 83%. The number of features selected was 39 among 206 features. This selection criteria is good in terms of decreasing the dimensionality of the problem, however for better results we need to increase the number of training and testing samples.

Table 5. Emotion classification of the combined feature vector with Decision Trees and BayesNet into 8 emotion categories (happiness, sadness, fear, anger, disgust, surprise, happy_surprise, uncertanity).

	Training	Testing	Attributes	Classifier	Correctly classified
Face& Body	156	50	206	Bayes Net	88 %
Face& Body	156	50	206	Decision Trees	94%
Face& Body	206	10-fold cross	206	Bayes Net	83.49%
Face& Body	206	10-fold cross	206	Decision Trees	89.81%
Face& Body	156	50	45	Bayes Net	96 %
Face& Body	156	50	45	Decision Trees	82 %
Face& Body	206	10-fold cross	45	BayesNet	92.72%
Face& Body	206	10-fold cross	45	Decision Trees	84.46%

After the feature selection process 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 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 "angry-happy" for body; and "angry" or "happy" for face). Eventually, the fused vector inherits the face.

We experiment C4.5 with 10-fold cross validation on a data set that consists of 156 training and 50 testing instances. We firstly tested the classifiers with a set of 206 initially combined features and later with a set of 39 features. For the feature set of 206 attributes, Decision Trees provided the best classification accuracy. Once we used the feature selection criteria and obtained the reduced feature set with 39 features, recognition with BayesNet improved dramatically. The results are presented in Table 5.

5.2 Decision Level Fusion

If the modalities are asynchronous but temporally correlated, like in our case with face and body gesture, decision level (late) integration is the most common way of integrating the modalities [29]. Decision fusion (late integration) is most commonly found in speech and gesture combination [29]. It can be described as fusion of each modality that is first pre-classified independently and the final classification is based on the fusion of the outputs of the different modalities.

Usually performing late integration is chosen over performing early integration for two primary reasons [29]. First, the feature concatenation used in early integration would result in a high dimensional data space, making a large multi-modal database necessary for robust statistical model training. Second, late integration provides greater flexibility in modeling. For instance in our case, with late integration, it is possible to train the face and body classifiers on different data sources and different classifiers that provide the best accuracy for each modality separately.

In late integration, the face data and body data are analyzed by separate classifiers. Each classifier processes its own data stream, and the two sets of outputs are combined in a later stage to produce the final hypothesis. Designing optimal strategies for decision-level fusion has been of interest to researchers in the fields of pattern recognition, machine learning, neural networks and more recently in data mining, knowledge discovery and data fusion [17]. Various statistical approaches have been used: product rule, sum rule,max/min/median rule, majority vote or adaptation of weights[17].

Among the various statistical approaches present for decision level fusion we used the most appropriate techniques for our system: product, average and weight criteria. This was done due to the fact that body emotion classes are not exactly the same as the facial emotion classes. Emotion classes for body were generated taking into account how humans classify body action units into emotions without using the facial information. When

combining the two; facial emotion category takes over by being the primary mode and body being an auxiliary mode to confirm and/or to improve the emotion classification from facial expression.

We used three criteria to combine the posterior probabilities of the mono-modal systems at the decision-level: product, in which the posterior probabilities are multiplied and the maximum is selected; average, in which the posterior probabilities of each modalities are equally weighted, average is calculated and the maximum is selected; and, weight, in which different weights are applied to face and body modalities. We describe the general approach of late integration of the individual classifier outputs as follows:

$p(x_1,....x_R|\omega_k)$ represents the joint probability distribution of the measurements extracted by the classifiers. Let us assume that the representations used are conditionally statistically independent.

Pattern Z is to be assigned to one of m possible classes $(\omega_1,\ldots,\omega_m)$

There are R classifiers with feature representations $x_1, x_2, \ldots, x_R$

Model each class Wk by probability

density function $p(x_i|\omega_k)$ with a priori probability $P(\omega_k)$. Assign $Z \rightarrow \omega_j$

if $P(\omega_j|x_1,\ldots,x_R) = \max_k P(\omega_k|x_1,\ldots,x_R)$

We used WEKA 3.4 [23] for individual emotion classification on a training set of 115 and on a testing set of 35 instances. C4.5 Decision trees with 10-fold cross validation provided the best results for both individual modalities. Then we fused the a posteriori probabilities by averaging and by the use of weights as described in the following sub-sections.

Fusion by the product rule

The first criteria we used for combining the posterior probabilities of the mono-modal classifiers at the decision level is the product rule. The joint probability distribution of the measurements extracted by the classifiers are assumed to be conditionally statistically independent.

$$P(x_j,...,x_R \mid \omega_k) = \prod_{i=1}^{R} p(x_i|\omega_k)$$

Under this assumption, the way we used the Product Rule can be described as:

assign $Z \rightarrow \omega_j$

if $\prod_{i=1}^{R} p(\omega_j \mid x_i) = \max_{k=i}^{m} \prod_{i=1}^{R} p(\omega_k \mid x_i)$

The results of this fusion are presented in Table 6.

Table 6. Emotion recognition results for late fusion using the Products Rule, Average and Weight criteria on the testing set of 50 samples.

Emotion	Recognition results on the testing set		
	Product Rule	**Average Criteria**	**Weight criteria (λf= 0.70, λb=0.30)**
Overall	0.8	0.86	0.82
Anger	0.6	0.7	0.7
Disgust	0.875	1	1
Fear	1	1	1
Happiness	0.8	0.8	0.8
Sadness	0.4	0.6	0.4
Surprise	0.909	0.909	0.909
Happy_surprise	1	1	1
Uncertainty	1	1	0.857

Fusion by the average rule

Under the equal prior assumption, computing the average a posteriori probability for each class over all the classifier outputs can be defined as :

Assign $Z \rightarrow \omega_j$

if $\frac{1}{R}\sum_{i=1}^{R} P(\omega_j \mid \chi_i) = \max_{k=1}^{m} \frac{1}{R}\sum_{i=1}^{R} P(\omega_k \mid \chi_i)$

Thus, the rule assigns a pattern to that class the average a posteriori probability of which is maximum. If any of the classifiers outputs an a posteriori probability for some class which is an outlier, it will affect the average and this in turn could lead to an incorrect decision [17]. However, in our case, during the averaging procedure, the facial modality has the lead. For instance, in the case where emotion classifier based on the body outputs “anger_disgust” and the emotion classifier based on the face outputs “happiness”, then the final decision of the averaging rule will be “happiness” based on the a posteriori probability of the face emotion classifier.

The results of fusion based on averaging are presented in Table 6.

Fusion by weight criteria

This method is based on adaptation of weights, in which different weights are applied to the different mono-modal systems and results are interpreted jointly. For each class the output from individual classifier is weighted and the sum is calculated. The class which receives the biggest value is then selected as the final decision as described below:

Assign $Z \rightarrow \omega_j$

if $\sum_{i=1}^{R} \lambda_i P(\omega_j \mid \chi_i) = \max_{k=1}^{m} \sum_{i=1}^{R} \lambda_i P(\omega_k \mid \chi_i)$

$(\lambda_1,..,\lambda_R)$: pre-assigned weights for each classifier output

In our case facial modality has the lead and the body modality needs to be integrated. Therefore, weights are pre-assigned as follows: 0.70 for facial modality and 0.30 for body modality. The sum of these values is then used to calculate the final result. The emotion recognition results are shown in Table 6. When we changed the weights as 0.60 for facial modality and 0.40 for body modality, the result did not change.

6 Discussions and Conclusions

Results reveal that emotion classification using the two modalities achieves 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. Moreover, early fusion seems to achieve a better recognition accuracy compared to late fusion. In late integration averaging seems to be the right way to fuse the two modalities since decreasing the weight of the body modality causes the accuracy to decrease visibly.

When using both affective face and body information, the results are promising, although we expect the error to increase as we more thoroughly test the system. Moreover we use a small database, so the generalizability and robustness of the results should be tested with a larger data set.

Late integration allows adaptive channel weighting between the face and body modalities. Additionally, late integration allows asynchronous processing of the two streams. However, the kind of fusion strategy to choose may not only depend upon the input modalities. Multimodal fusion patterns may depend upon the particular task at hand. A comprehensive analysis of experimental data may therefore help gather insights and knowledge about the integration patterns thus leading to the choice of the best fusion approach for the application, modalities and task at hand.

In this work, our fusion strategy can be referred to as micro temporal fusion, combining information that is produced either in parallel or over overlapping time intervals [8]. However, facial expression and body gesture modalities are asynchronous and temporally correlated. The facial expression is produced and completed (onset-apex-offset) in tens or even hundreds of milliseconds before the hand and/or body gestures are actually produced. The difference between time responses of face and body gestures can be very large, therefore, a gesture recognition system needs more time to recognize a gesture than a facial expression recognition system. Eventually we need to do macro temporal fusion that takes care of either sequential inputs or time intervals that do not overlap but belong to the same temporal time window. A variety of methods for modeling visual asynchrony have been proposed in the literature (i.e. coupled-hidden Markov models, linked-hidden Markov models). Our temporal fusion needs to rely on time proximity: time-stamped features from different input channels will merged if they occur within a pre-defined time window (i.e. 5 sec). When only one modality is available (only facial expression while body is neutral) the system will have to rely on the mono-modal recognition results only. Our future work will be towards the macro-temporal fusion of the face and body modalities.

References

[1] T.Balomenos et al.,"Emotion Analysis In Man-Machine Interaction Systems", Springer-Verlag Heidelberg MLMI 2004 Lecture Notes ,Volume 3361, pp. 318, 2005.

[2] M. S. Bartlett, et al., “Machine Learning Methods for Fully Automatic Recognition of Facial Expressions and Facial Actions”, Proc. IEEE SMC, The Hague, The Netherlands, pp. 592-597, 2004.

[3] R. T. Boone, 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. Cassell, “A framework for gesture generation and interpretation”. In R. Cipolla and A. Pentland (editors), Computer vision in human-machine interaction. Cambridge University Press, 2000.

[6] L.S. Chen, H. Tao, T.S. Huang, T. Miyasato, R. Nakatsu, “Emotion recognition from audiovisual information”, IEEE Second Workshop on Multimedia Signal Processing, ,7-9 Dec., pp.83 – 88, 1998.

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

[8] A. Corradini, M. Mehta, N.O. Bernsen, J.-C. Martin, “Multimodal Input Fusion In Human Computer Interaction On The Example Of The On-Going NICE Project”, Proc. The NATO-ASI Conference On Data Fusion For Situation Monitoring, Incident Detection, Alert And Response Management, Yerevan, Armenia, August 18th-29th 2003.

[9] L. C. De Silva and P. C. Ng, "Bimodal emotion recognition," in Proc.FG, 2000, pp. 332–335.

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

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

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

[13] H. Gunes et al., "Bimodal Emotion Modelling from Facial and Upper-Body Gesture for Affective HCI", Proc. of OZCHI 2004, Wollongong, Australia.

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

[15] A. Kapoor, R. W. Picard, Y. Ivanov, "Probabilistic Combination of Multiple Modalities to Detect Interest", Proc. IEEE ICPR 2004.

[16] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas, "On Combining Classifiers", *IEEE PAMI*, Vol. 20, No. 3, March 1998.

[17] L. I. Kuncheva, "A Theoretical Study on Six Classifier Fusion Strategies", *IEEE PAMI,* Vol. 24, No. 2, February 2002.

[18] R. Laban and L. Ullmann, The Mastery of Movement, Northcote House Educational Publishers, 1988.

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

[20] A. Mehrabian, "Communication without words", *Psychol. Today*, Vol. 2(4), pp. 53–56, 1968.

[21] 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.

[22] M. Pantic, and L.J.M. Rothkrantz, "Towards an Affect-Sensitive Multimodal Human-Computer Interaction", Proc. of the IEEE, 91(9) (2003) 1370-1390.

[23] J. R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufman Publishers, San Mateo, California, 1993.

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

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

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

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

[28] H. Wallbott and K.R. Scherer, "How universal and specific is emotional experience? Evidence from 27 countries on five continents", *Social Science Information*, Vol. 25, pp. 763-795, 1988.

[29] L. Wu, S. L. Oviatt, P. R. Cohen, "Multimodal Integration-A Statistical View", *IEEE Transactions on Multimedia*, Vol. 1(4), pp. 334-341, 1999.

[30] Y. Yoshitomi et al., "Expression recognition using thermal image processing and neural network," in Proc. ROMAN, 1997, pp. 380–385.