Classification of Affects Using Head Movement, Skin Color Features and Physiological Signals

Hamed Monkaresi, M. Sazzad Hussain, and Rafael A. Calvo
School of Electrical and Information Engineering
University of Sydney
Sydney, Australia
{Hamed.Monkaresi, Sazzad.Hussain, Rafael.Calvo}@sydney.edu.au

Abstract— **The automated detection of emotions opens the possibility to new applications in areas such as education, mental health and entertainment. There is an increasing interest on detection techniques that combine multiple modalities. In this study, we introduce automated techniques to detect users' affective states from a fusion model of facial videos and physiological measures. The natural behavior expressed on faces and their physiological responses were recorded from subjects (N=20) while they viewed images from the International Affective Picture System (IAPS). This paper provides a direct comparison between user-dependent, gender-specific, and combined-subject models for affect classification. The analysis indicates that the accuracy of the fusion model (head movement, facial color, and physiology) was statistically higher than the best individual modality for spontaneous affect expressions.**

Keywords- Affective computing, machine learning, video analysis, multichannel physiology, multimodal fusion.

I. INTRODUCTION

The number of applications using video cameras for tracking faces is growing exponentially. Cameras are constantly capturing images of human faces on cell phones, webcams, even in automobiles— often with the goal of using the facial information as a clue to understand more about the user's current state of mind [1], [2]. In recent years, researchers in the field of affective computing have developed affective sensors, computational techniques/ tools, and applications [2]. A lot of these studies primarily focus on a single modality such as facial expressions or acoustic-prosodic features of speech. However, it is unclear whether all emotions are expressed via facial expressions and paralinguistic features of speech. For example, there is some evidence that naturalistic episodes of boredom and engagement, two affective states that are ubiquitous in almost any task, cannot be reliably detected from the face [3], [4]. Another drawback of these behavioral uni-modal (facial expressions and speech patterns) affect detection systems is that, they can be masked, where users may attempt to disguise certain negative emotions. Hence, alternate channels (e.g. physiology) need to be considered along with behavioral modalities to detect the subtle expressions associated with complex emotions. On the other hand, in naturalistic situation emotions are expressed in a multimodal fashion, which cannot be detected accurately from just single modality, therefore, having multimodal information for affect recognition is advantageous [5].

Affect detection systems that integrate information from different modalities have been widely advocated, yet they are still rarely implemented [6]. In recent years, multimodal approaches of affect detection are becoming increasingly popular in affective computing due to its advantages [7], [8]. In this study, physiological measures such as heart activity, skin response, facial muscle activity and breathing patterns are considered along with face and head related features to improve the accuracy of affect recognition. There are a number of factors that distinguish the current approach from previous ones. Firstly, behavioral responses (head movement and face color) were recorded with a video camera while subjects viewed emotional images from International Affective Picture System (IAPS) [9] designed to elicit intense emotions. In this scenario the probability that we accessed natural emotional behavior is higher than in studies that used deliberately posed faces [10], [11]. Secondly, the physiological data was collected during this controlled stimulus presentation, which was used as additional features along with the video features to improve the overall accuracy of affect detection in a multimodal fashion. Thirdly, we adopted the circular order of the circumplex model [12] of affect during the image presentation for emotion stimulation. This is suitable because it helps determine the current subtle and complex affective state of the user, which can then be used to control the course of the interaction.

In this paper, we present results for affect detection from the individual modalities (physiology and video) and their fusion model. The results presented show the difference in three types of models; *user-dependent model*, *gender specific model*, and *combined-subject model*. These models are useful for human computer interaction (HCI) applications targeting general or specific group of users. It is also interesting to compare the user-dependent, gender-specific, and combined-subject models, which may provide valuable information to social scientists studying how these differences can affect the expression of emotions.

Section two gives brief background on multimodal fusion and a brief review on multimodal approaches in affective computing. Section three explains the data collection procedure and the computational model. Section four presents the results for this study followed by the conclusion in section five.

II. BACKGROUND

One of the main goals of multimodal affect recognition systems is to achieve better accuracy by integrating information from different input modalities (e.g. video and physiology) rather than each single channel. There are three methods to fuse multichannel information each depending on when information from the multiple sensors are integrated [5].

Data fusion is performed on the raw data for each signal and can only be applied when the signals have the same temporal resolution. It can be used for integrating physiological signals coming from the same recording equipment as is commonly the case with physiological signals, but it could not be used to integrate a video signal with a text transcript. It is also not commonly used because of its sensitivity to noise produced by the malfunction or misalignment of the different sensors.

Feature fusion is performed on the set of features extracted from each signal. This approach is more commonly used in multimodal HCI and has been used in affective computing, for example, in the Augsburg Bio-signal Toolbox [13]. Features for each signal (EKG, EMG, etc.) are primarily the mean, median, standard deviation, maxima, and minima, together with some unique features from each sensor. These are individually computed for each sensor and then combined across sensors.

Decision fusion is performed by merging the output of the classifier for each signal. Hence, the affective states would first be classified from each sensor and would then be integrated to obtain a global view across the various sensors. It is also the most commonly used approach for multimodal HCI [14].

Previous work on multimodal affect recognition tried to make use of both feature level and decision level fusion, but have mostly achieved better performance for feature fusion. Busso et al. integrated facial expression and speech at both feature and decision level using support vector machine (SVM) to recognize four emotions – sadness, happiness, anger and neutral [15]. They reported classification results of 89.1% for the bimodal model using feature fusion. The result for their decision level fusion was slightly lower than the feature level. Jonghwa Kim evaluated feature level, decision level, and hybrid fusion performance integrating multichannel physiological signals and speech signal for detecting valance and arousal using linear discriminant analysis (LDA) classifier [16]. Their fusion scheme reported results for feature, decision, and hybrid fusion, where the performance for the feature fusion was highest.

On a more naturalistic situation, Kapoor and Picard developed a contextually grounded probabilistic system to infer a child's interest level on the basis of upper and lower facial feature tracking, posture patterns (current posture and level of activity), and some contextual information (difficulty level and state of the game) [17]. The combination of these modalities yielded a recognition accuracy of 86 percent, which was significantly greater than that achieved from the facial features (67 percent upper face and 53 percent lower face) and contextual information (57 percent). However, the posture features alone yielded an accuracy of 82 percent that would indicate that the other channels are redundant with posture.

More recently, D'Mello and Graesser considered a combination of facial features, gross body language, and conversational cues for detecting some of the learning-centered affective states [7]. Their affect detector was based on feature fusion where their analysis indicated that the accuracy of the multichannel model was statistically higher than the individual channels for the fixed but not spontaneous judgments. They also investigated decision fusion but the classification accuracy rates were similar to feature fusion. Hussain et. al. [8] evaluated the performance of detecting valence and arousal using multichannel physiology and their fusion at various levels (feature, decision, and hybrid). They achieved significant improvement for the decision and hybrid fusion levels over the individual channels.

III. METHOD

A. Participants and Materials

The participants were 20 undergraduate/postgraduate engineering students. The participants' age ranged from 18 to 30 years and there were 8 males and 12 females. Due to sensor failure and loss of data from two subjects, results are presented for 18 subjects. The participants also signed an informed consent prior to the experiment. The experiment took approximately one hour.

The experiments were conducted indoors with a varying amount of ambient sunlight entering through windows in combination with normal artificial fluorescent light. Participants were asked to sit in front of a computer and interact normally while their video was recorded by an ordinary webcam (Logitech Webcam Pro 9000). All videos were recorded in color (24-bit RGB with 3 channels, 8 bits/channel) at 15 frames per seconds (fps) with pixel resolution of 640×480 pixels and saved in AVI format.

The participants were also equipped with physiological sensors that monitored electrocardiogram (ECG), facial electromyogram (EMG), respiration, and galvanic skin response (GSR). The physiological signals were acquired using a BIOPAC MP150 system with *AcqKnowledge* software at 1000 samples per second for all channels. ECG was collected with two electrodes placed on the wrists. Two channels of EMG were recorded from the *zygomatic* and *corrugator* muscles respectively. A respiration band was strapped around the chest and GSR was recorded from the index and middle finger of the left hand. Fig. 1 presents the experiment setup and mentioned sensors.

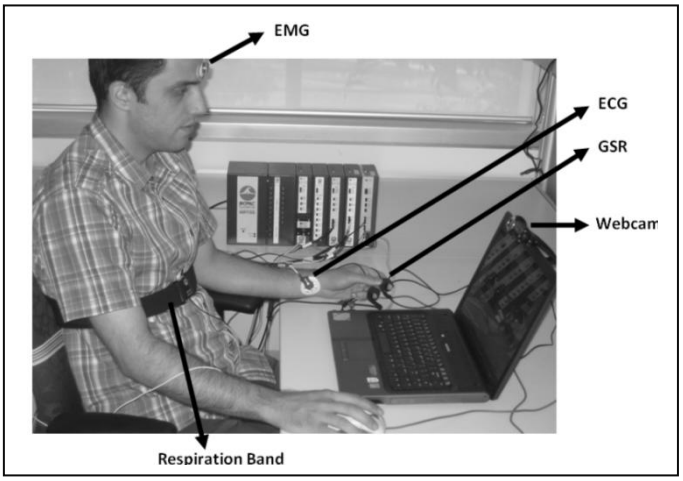

Figure 1. Experiment setup and sensors

B. Procedure

The participants viewed emotionally charged photos from the IAPS collection. A total of 90 images (three blocks of 30 images each) for 10 seconds each were presented, followed by 6 seconds pauses between the images. The images were selected so that the IAPS valence and arousal normative scores for the stimuli spanned a 3×3 valence/arousal space. Participants also self-reported their emotions by clicking radio buttons on the appropriate location of 3×3 valence/arousal grid after viewing each image.

In this paper, results are presented using the normative ratings instead of self-reports. Therefore, the computational model was trained and tested using a balanced class distribution, which could be suitable for evaluating accuracies of classification without applying any up or down sampling techniques. The IAPS normative ratings are useful because they are standardized scientifically for assessing basic and applied problems in psychology [9]. Moreover, many people do not know how to recognize, express and label/scale their own feelings, therefore self reports sometimes can be unreliable [18]. However, self-reports provide important information and should not be ignored; therefore the collected self ratings will be used as an extension of this work in future studies.

C. Feature Extraction

A total number of 279 features were extracted from video and the physiological signals. Feature vectors were calculated using 10 seconds time window corresponding to the duration of each IAPS image presentation. The feature vectors were also labeled with the normative ratings (1-3 degrees of valence/ arousal). The feature extraction process is explained briefly in the following subsections.

1) Video Features

All video recordings were analyzed offline using MATLAB and Open Computer Vision library (OpenCV). In this work, two types of image based features were explored: geometric and chromatic features. Five geometrical data of the face (x and y coordinates, width, height and area) were derived which determined the position of the head in each frame. In addition, each frame was separated into *red*, *green* and *blue* colors. The impact of different emotion on the user skin color is also explored in this study.

Statistical features were calculated based on these eight image-based features. By defining a time window, a set of statistical features can be derived. A number of statistical features such as *mean*, *median*, *standard deviation*, *minima* and *maxima* were computed. In addition, some motion features were calculated by subtracting the values of last frame of the time window from the first values. Altogether, 65 features were extracted from each video. The procedure of video feature extraction was conducted as follows:

a) Face tracking (Geometric features extraction)

First of all, for each video, we detected the face and extracted the necessary features. We utilized OpenCV for detecting and tracking the face in the video recording using an extended boosted cascade classifier (Haar classifier [19]). Most of the classifiers identify a number of false positive due to background artifacts and this rate would be raised if the movement increases in the video. In order to improve the performance of face detection module a fast and simple face tracking algorithm was developed to increase the speed of further calculation and reduce the false face positive. Accordingly, in each frame a dynamic region of interest (ROI) was selected based on the face area which was detected in previous frame (+30% of detected face area). The algorithm always looks for the face in this tracking ROI. If the face was not found in this region, the ROI was expanded to whole image.

b) Skin color (Chromatic features extraction)

In this study, two facial EMG sensors were placed, therefore, needed to be removed from the video frames before extracting the color features. This was done by tracking ROI where a rectangle of 60% width and full height of the detected face region was selected as new ROI and then was separated into the RGB channels. The average color value of all pixels in the ROI was calculated for each frame to compose raw signals for red, green and blue channels. If no face was detected then the previous values for the three channels were returned and the face detection flag for the current frame was set to zero.

2) Physiological Features

Statistical features were extracted from the different physiological channels using the Augsburg Bio-signal toolbox [13] in Matlab. Some features were common for all signals (e.g. *mean*, *median*, and *standard deviation*, *range*, *ratio*, *minimum*, and *maximum*) whereas other features were related to the characteristics of the signals (e.g. heart rate variability, respiration pulse, frequency). A total of 214 features were extracted from the five physiological channel signals (84 from ECG, 42 from EMG, 21 from GSR, and 67 for respiration).

3) Feature Fusion

Feature level fusion model was created for further analysis. In this model, all physiological features were considered as the physio modality and both head movement and skin color features were considered as the video modality. Hence, the fusion model contained all features of these two modalities. We investigated the accuracy of classification tasks through video

and physio modalities and compared them with the fusion model.

D. Classification

The Waikato Environment for Knowledge Analysis (Weka) was used for feature selection and classification. Prior to classification, in order to reduce the dimensionality of the large number of features which might decrease classification performance due to unnecessary features, correlation based feature selection (CFS) method was used for choosing the best subset of features. The correlation based feature selection technique evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them [20].

Six machine learning algorithms; k-nearest neighbor (k=1, k=3), linear support vector machine (SVM), BayesNet, J48 decision tree and linear logistic regression from the weka toolbox [21] were selected for classification. Then, a vote classifier with the *average probability* rule for combining the classifiers was applied [22]. The training and testing for both types of classes (valence and arousal) was performed separately with a 10-fold cross validation. The kappa statistic [23] was used as the overall classification performance metric because it factors out random guessing (i.e., accuracy expected due to chance). A kappa of 0.0 represents chance accuracy while a kappa of 1.0 would be indicative of perfect discrimination. ZeroR, a probabilistic classifier was used as the baseline produced kappa of 0.0. In addition, the F-measure (from precision and recall) was calculated as an indication of how well each affective state was classified. For the classification scores of precision (P) and recall (R), the F-measure (F1) is calculated by; $F1=2((P\times R)/(P+R))$.

IV. Results

The classification results for detecting 1-3 degrees (low, medium, high) of valence and arousal is divided into three main parts: user-dependent analysis, gender-specific analysis and combined-subject analysis.

A. User-dependent Analysis

The subsequent analysis focuses on developing user-dependent models. Distinct models were developed and validated for each participant. Fig. 2 presents the mean and standard deviation of kappa scores assessing the overall performance of discriminating three degrees of valence and arousal. The results indicate that the classifier was successful in discriminating between three degrees of valence and arousal

The fusion model achieves 9% and 15% improvement in kappa score for detecting valence over the video and physio channels respectively. In addition, it should be mentioned that the fusion model for 15 out of 18 subjects had kappa scores ranging from 0.40 to 0.70 in detecting three degrees of valence.

On the other hand, the classification results showed that the fusion model did not improve the accuracy of detecting degrees of arousal from video modality but showed 15% improvement over the physio modality. Maximum kappa score of 0.70 was achieved by the fusion model for detecting the degrees of valence.

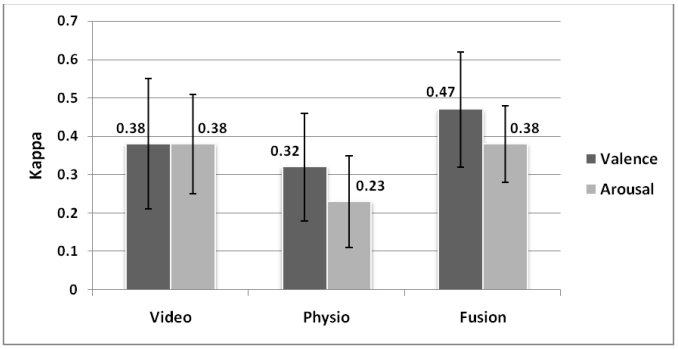

Figure 2. The mean and standard deviation of Kappa scores for valence and arousal classification (User-dependent models)

Next we investigate how well the individual degrees of valence and arousal were classified from the video, physio and fusion models. Table 1 shows the mean and standard deviation of F1 as per valence and arousal category across the 18 subjects for the vote classifier. The fusion model has the highest accuracy for detecting all three degrees of valence followed by the video modality. The performance of the physio modality was slightly lower than video for all three degrees of valence. As, for arousal, both video and the fusion model had similar F1 values for detecting *low* arousal. Video was slightly better than the fusion for detecting *medium* arousal, while it was the opposite for detecting *high* arousal. Physio is able to detect all three degrees of arousal with reasonable accuracy but lower than both video and fusion models.

TABLE I. The mean and standard deviation of F1 values for each degree of valence and arousal (User-dependent models)

		Valence			Arousal		
		Low	**Medium**	**High**	**Low**	**Medium**	**High**
Video	**Mean**	**0.64**	**0.51**	**0.60**	**0.62**	**0.49**	**0.63**
	Std	0.15	0.11	0.18	0.14	0.13	0.09
Physio	**Mean**	**0.61**	**0.44**	**0.58**	**0.52**	**0.41**	**0.51**
	Std	0.13	0.12	0.13	0.11	0.12	0.13
Fusion	**Mean**	**0.69**	**0.53**	**0.71**	**0.62**	**0.47**	**0.65**
	Std	0.12	0.15	0.14	0.11	0.12	0.09

B. Gender-specific Analysis

For this analysis, we separated our dataset into two parts, with one part containing only male subjects (n=7) and the other part only female subjects (n=11). Then, data from individual participants were standardized (converted to z-scores) to address individual variations of head behavior. We created individual classifiers for each of the datasets in order to compare their performance. The kappa score for valence and arousal using these models are shown in figures 3 and 4 respectively. From the figures, we observe that the fusion model has the highest accuracy of detecting both valence and arousal from the male dataset. However, the video modality appears to be the best detector of valence and arousal in the female dataset. The good accuracy from the video modality also suggests that video-based features are better indicators of emotions among women compared to men. Adding the physiological features with video exhibit only 3% improvement in accuracy for detecting both valence and arousal for males, whereas the accuracy of the fusion model

dropped for females. This suggests that physiological features play more important role for detecting emotional responses within males compared to females. If reproduced on other datasets, this could be a very significant result for developing affective computing systems.

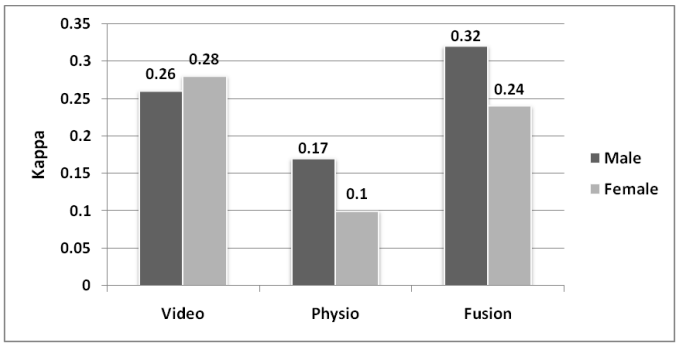

Figure 3. Kappa scores for valence classification (Gender-specific model)

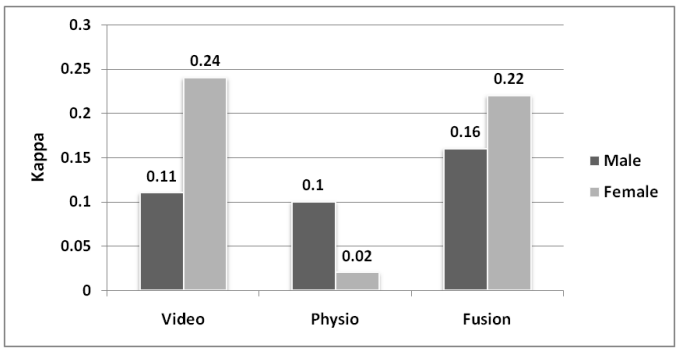

Figure 4. Kappa scores for arousal classification (Gender-specific model)

Table 2 gives F1 values for the three degrees of valence and arousal across the male and female subject classifiers. According to this table, video modality presented the best performance in detecting *low-valence* among female participants (F1=0.600) whereas the best performance among male participants was achieved by fusion model in classifying *high-valence* (F1=0.615). In contrast, the performance of all three models was not so promising in arousal classification specially through male participants (F1<0.5).

TABLE II. F1 VALUES FOR EACH DEGREE OF VALENCE AND AROUSAL (GENDER-SPECIFIC MODELS)

		Valence			Arousal		
		Low	**Medium**	**High**	**Low**	**Medium**	**High**
Video	**M**	0.494	0.448	0.588	0.406	0.415	0.413
	F	**0.600**	0.462	0.510	**0.544**	0.431	**0.517**
Physio	**M**	0.463	0.334	0.543	0.434	0.371	0.404
	F	0.377	0.398	0.425	0.333	0.335	0.389
Fusion	**M**	0.591	0.437	**0.615**	0.464	0.412	0.465
	F	0.523	0.442	0.524	0.515	0.416	**0.517**

M=Male, F=Female

C. Combined-subject Analysis

Data from individual participants were first standardized and then combined to yield one large data set with 1620 instances. The dimensionality of the data was also reduced by selecting the best features prior to the classification task. Fig. 5 shows the results for classification detecting three degrees of valence and arousal from this dataset. An interesting pattern emerges from the combined-subject analysis, where it is observed that the video modality is more suitable for detecting degrees of arousal (kappa=0.23) but physiological features are more suitable for detecting degrees of valence (kappa=0.16). The fusion of video and physio exhibits higher accuracy for valence (kappa=0.22) but slightly lower for arousal (kappa=0.19). Overall, combined-subject model showed lower performance compare to user-dependent model, since the combined-subject model considered all subjects while user-dependent models were optimized for the specific characteristic of each subject.

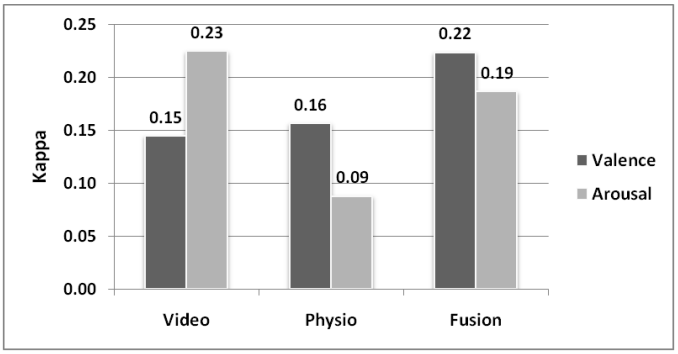

Figure 5. Kappa scores for valence and arousal classifiaction (Combined-subject model)

Table 3 shows the F1 values for detecting 1-3 degrees of normative valence and arousal. The F1 value indicates that the fusion model was best at detecting *high* and *low* valence. The video modality has a slightly higher F1 value over the fusion for detecting *medium* valence. However, all three degrees of arousal is best detecting from video.

TABLE III. F1 VALUES FOR DETECTING 1-3 DEGREES OF VALENCE AND AROUSAL (COMBINED-SUBJECT MODEL)

	Valence			Arousal		
	Low	**Medium**	**High**	**Low**	**Medium**	**High**
Video	0.438	**0.388**	0.464	**0.529**	**0.390**	**0.515**
Physio	0.445	0.366	0.500	0.381	0.381	0.415
Fusion	**0.486**	0.381	**0.536**	0.485	0.374	0.504

V. CONCLUSION

In this study we proposed a new approach for classifying naturalistic expressions of affect through head movements, skin color, heart activity, skin response and respiration. We evaluated user-dependent, gender-specific, and combined-subject models.

The present study provides evidence it is feasible to create a fusion model for detecting valence during naturalistic HCI (kappa=0.22). It also indicated that adding physiological signals did not improve the accuracy of arousal detection over the video modality. We believe, based on the research reviewed that by adding dynamic facial features the accuracy of affect detection could be improved. As future work, the proposed model will be improved by adding other facial features in order to build an applicable user-independent naturalistic emotion recognition system. Furthermore, current

results show a strong correlation between a combination of head movement, skin color features and physiological signals and level of affect across subjects. The mean kappa score of 0.47 for valence prediction and of 0.38 for arousal prediction demonstrate the feasibility of using fusion model as an affect predictor in user-dependent models.

Another notable result is that valence can be much better predicted than arousal using the fusion model. This has also been confirmed by other related work on dimensional emotion recognition [24]. Whether such conclusions hold for different context and different data remain to be evaluated. Among 1-3 degree of valence and arousal, the low and the high degrees were the best predictable classes that can be detected by proposed models.

REFERENCES

[1] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: audio, visual, and spontaneous expressions.," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 1, pp. 39-58, Jan. 2009.

[2] R. A. Calvo and S. D'Mello, "Affect Detection : An Interdisciplinary Review of Models , Methods , and Their Applications," *Affective Computing, IEEE Transaction on*, vol. 1, no. 1, pp. 18-37, 2010.

[3] S. D. Craig, S. D'Mello, A. Witherspoon, and A. Graesser, "Emote aloud during learning with AutoTutor: Applying the Facial Action Coding System to cognitive–affective states during learning," *Cognition & Emotion*, vol. 22, no. 5, pp. 777-788, Aug. 2008.

[4] B. McDaniel, S. D'Mello, B. King, P. Chipman, K. Tapp, and A. Graesser, "Facial Features for Affective State Detection in Learning Environments," in *29th Annual meeting of the cognitive science society*, 2007, pp. 467–472.

[5] R. Sharma, V. I. Pavlovic, and T. S. Huang, "Toward Multimodal Human – Computer Interface," *Proceedings of the IEEE*, vol. 86, no. 5, pp. 853-869, 1998.

[6] A. Jaimes and N. Sebe, "Multimodal human–computer interaction: A survey," *Computer Vision and Image Understanding*, vol. 108, no. 1–2, pp. 116-134, Oct. 2007.

[7] S. D'Mello and A. Graesser, "Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features," *User Modeling and User-Adapted Interaction*, vol. 20, no. 2, pp. 147-187, May 2010.

[8] S. Hussain, R. A. Calvo, and P. A. Pour, "Hybrid Fusion Approach for Detecting Affects from Multichannel Physiology," in *Affective Computing and Intelligent Interaction*, S. D'Mello, A. Graesser, B. Schuller, and J.-C. Martin, Eds. Springer, 2011, pp. 568-577.

[9] P. Lang and M. Bradley, "International affective picture system (IAPS): Technical manual and affective ratings," *Psychology*, 1997.

[10] P. Ekman, R. J. Davidson, and W. V. Friesen, "The Duchenne smile: emotional expression and brain physiology II.," *Journal of personality and social psychology*, vol. 58, no. 2, pp. 342-53, Feb. 1990.

[11] P. Ekman, *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage*. New York: W. W. Norton & Company, 2001.

[12] J. A. Russell, "A Circumplex Model of Affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6. pp. 1161-1178, 1980.

[13] J. Wagner, J. Kim, and E. Andre, "From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification," in *2005 IEEE International Conference on Multimedia and Expo*, 2005, pp. 940–943.

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

[15] C. Busso et al., "Analysis of Emotion Recognition using Facial Expressions , Speech and Multimodal Information," in *Proceedings of the 6th International Conference on Multimodal Interfaces*, 2004, pp. 205-211.

[16] J. Kim, "Bimodal emotion recognition using speech and physiological changes," *Robust Speech Recognition and Understanding*, pp. 265–280, 2007.

[17] A. Kapoor and R. W. Picard, "Multimodal affect recognition in learning environments," in *Proceedings of the 13th annual ACM international conference on Multimedia*, 2005, pp. 677–682.

[18] R. Picard, "Affective computing: challenges," *International Journal of Human-Computer Studies*, vol. 59, no. 1–2, pp. 55-64, Jul. 2003.

[19] R. Lienhart and J. Maydt, "An extended set of Haar-like features for rapid object detection," in *IEEE ICIP 2002*, 2002, vol. 1, pp. 900-903.

[20] M. A. Hall, "Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning," in *ICML '00 Proceedings of the Seventeenth International Conference on Machine Learning*, 2000, pp. 359-366.

[21] I. H. Witten and E. Frank, *Data mining: practical machine learning tools and techniques*. Morgan Kaufmann, 2005, p. 525.

[22] L. I. Kuncheva, *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons, 2004, p. 350.

[23] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37-46, Apr. 1960.

[24] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous Prediction of Spontaneous Affect from Multiple Cues and Modalities in Valence–Arousal Space," *Affective Computing, IEEE Transactions on*, vol. 2, no. 2, pp. 92-105, 2011.