Combining Classifiers in Multimodal Affect Detection

M. S. Hussain, Hamed Monkaresi and Rafael A. Calvo
School of Electrical and Information Engineering, The University of Sydney, NSW 2006, Australia
{Sazzad.Hussain, Hamed.Monkaresi, Rafael.Calvo}@sydney.edu.au

Abstract

Affect detection where users' mental states are automatically recognized from facial expressions, speech, physiology and other modalities, requires accurate machine learning and classification techniques. This paper investigates how combined classifiers, and their base classifiers, can be used in affect detection using features from facial video and multichannel physiology. The base classifiers evaluated include function, lazy and decision trees; and the combined where implemented as vote classifiers. Results indicate that the accuracy of affect detection can be improved using the combined classifiers especially by fusing the multimodal features. The base classifiers that are more useful for certain modalities have been identified. Vote classifiers also performed best for most of the individuals compared to the base classifiers.

Keywords: Classifiers, machine learning, affective computing, data fusion.

1 Introduction

Affective computing, mostly useful in the area of human computer interaction (HCI), and particularly affect detection, heavily depends on efficient machine learning techniques (Calvo and D'Mello, 2010). Various modalities such as behavioural signatures and physiological patterns can be indicators of affect, thus pattern recognition techniques applied to a single, but mostly a combination of modalities could lead to affect detection.

A number of techniques have been developed for affect detection and studies tend to use features from audio-visual, speech-text, dialog-posture, face-body-speech, and speech-physiology, face-physiology, and multi-channel physiology (for detailed review see (Calvo and D'Mello, 2010)). Most of these studies have applied single classifiers, such as support vector machines (SVM), k-nearest neighbours (KNN), linear/quadratic discriminant analysis (LDA/QDA), decision trees, Bayesian network etc. with single and multiple modalities (mostly as feature fusion). However, finding a single classifier that works well for all modalities and individuals is difficult. Even though decision level fusion approaches have been proposed for integrating multimodal information in affect detection, in most studies it could not exceed the performance of feature level fusion (Sebe et al., 2005).

Combining classifiers is thought to provide more accurate and efficient classification results. Instead of just one classifier, a subset of classifiers (aka base classifiers) can be considered along with the best subset of features for the best combination (Kuncheva, 2004). Moreover, a certain base classifier may do well on a certain modality, but it is challenging to generalize one classifier for multiple channels or modalities. There are two important reasons for considering combined classifiers (Utthara et al., 2010): (1) A single classifier can not perform well when the nature of features are different. Using combination of classifiers with a subset of features may provide a better performance. (2) To improve generalization, where a classifier may not perform well for new data beyond that in the training set –generally very small in affective computing applications.

Two main strategies are applied for combining classifiers: fusion and selection. This study investigates the classifier fusion approach. In classifier fusion, each classifier is provided with complete information about the feature space. Combiners such as the average and majority vote are then applied for fusion. Combined classifiers may not necessarily always out-perform one single classifier, but the accuracy will be on an average better than all the base classifiers (Utthara et al., 2010, Kuncheva, 2004).

Combining classifier can be suitable in emotion studies for affect detection or classification where features contribute from multiple modalities (reason 1) and individuals (reason 2) in varying environmental setups. Omar AlZoubi et al. (2011) proposed a classifier ensemble approach using a Winnow algorithm to address the problem of day-variation in physiological signals for affect detection. However, the study used only one type of classifier (four SVM classifiers) for the ensemble. Combined classifiers have been considered in some of our previous studies related to affect detection from multimodal features (Hussain et al., 2012, Hussain et al., 2011b, Hussain et al., 2011a), however the improvements over the base classifiers have not been justified.

In this study we have applied vote classifiers to detect affects, in this context detecting how positive or negative their valence (e.g. happy vs. unhappy) and its intensity (aka arousal or activation) using features from multichannel physiology and facial video. Three types of base classifiers (function, lazy, decision trees) are considered and results are evaluated for the individual base classifiers and the vote classifiers. The study provides empirical justification of using the vote classifiers for affect detection using multimodal features collected from a variety of subjects, during controlled stimulus presentation. The vote classifiers are also briefly

evaluated for affect detection using a separate dataset (Hussain et al., 2012), collected during naturalistic interactions with an Intelligent Tutoring System (ITS).

Section two gives a brief description of the data collection procedure and section three gives the computations model. Section four gives the results with discussions followed by conclusion in section five.

2 Data Collection: Participants, Sensors, and Procedures

The data used for detecting affect (i.e. the arousal and valence dimensions) in this paper was collected in a study where participants viewed emotionally stimulating photos. The purpose of this experiment was to collect physiological signals and facial video in response to emotional stimulus (3-degrees of arousal and valence). Data was collected from 20 students (8 males and 12 females, age ranged from 18 to 30) from University of Sydney. Each session took approximately 15 to 20 minutes for preparation (consent forms, sensor setup and the explanation of experiment protocol) prior to the experiments. Physiological signals and facial video were recorded during the entire session. The participants' electrocardiogram (ECG), skin conductivity (SC), and respiration (Resp) were measured using a BIOPAC MP150 system with AcqKnowledge software. Video was recorded using Logitech Webcam Pro 9000. All videos were recorded in colour at 15 frames per seconds (fps) with pixel resolution of 640×480 pixels.

The experiment was conducted under systematic setup and considered images from the International Affective Picture System (IAPS) (Lang et al., 1997) as part of the emotion stimulation process. The IAPS collection contains set of colour photographs with normative ratings of emotion (valence, arousal, dominance) providing a set of emotional stimuli frequently used in experimental investigations of emotion and attention. The normative ratings in the IAPS collection are the result of many studies with large number of subjects.

Each session lasted approximately 40 minutes where participants viewed the photos from the IAPS collection. A total of 90 images were presented; each image was presented for 10 seconds, followed by a 6 seconds pause showing a blank screen between the images. The experiment was interrupted by short breaks after presenting every 30 images. Images were categorized based on IAPS normative ratings so that the valence and arousal scores for the stimulus spanned a 3×3 space and presented based on the affective circumplex model (Russell, 1980).

3 Computational Model

The computational model for feature extraction, feature selection and classification were implemented in Matlab with the support of in-house and third party toolboxes.

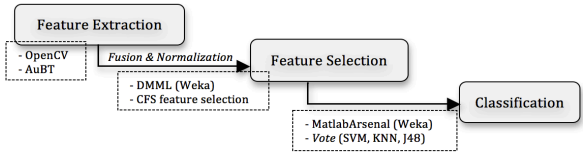

Figure 1: Overview of computational model

Figure 1 gives the overview of the computational model. The following subsections provide detailed description of the three main computational modules followed by the classifier training and testing procedure.

3.1 Feature Extraction, Normalization and Feature Fusion

A total of 287 features were extracted from the facial video and the physiological signals. Feature vectors were calculated from the time window corresponding to the duration of each stimulus presentation (10 seconds). The feature vectors were also labelled with the normative ratings (1-3 degrees of valence/arousal). The feature extraction and normalization process is explained briefly as followed.

Videos were analysed offline using MATLAB and Open Computer Vision library (OpenCV)[1]. Two types of image-based features were explored: geometric and chromatic features. Five geometrical data (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 colours in different conditions, due to movement or changing illumination sources. A total of 115 features were extracted from the videos (59 from geometric and 56 from chromatic).

Statistical features were extracted from the different physiological channels using the Augsburg Biosignal toolbox (AuBT) (Wagner et al., 2005) 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 172 features were extracted from the five physiological signals (84 from ECG, 21 from SC, and 67 for respiration).

All features were merged to achieve the fusion model (*fusion*) for further analysis. All physiological features were considered as the *physio* modality and both geometric and chromatic features were considered as the *face* modality. Hence, *fusion* contained all features of these two modalities. All features were normalized using *z-scores* before classification.

3.2 Feature Selection

The feature selection was implemented in Matlab using the DMML[2] wrapper for Weka (Hall et al., 2009). Feature selection techniques are used for discarding redundant, noisy features. This study investigates correlation based feature selection (CFS) as a way of choosing the best subset of features. The feature selection was performed separately for all individual modalities, and their fusion. The CFS technique evaluates the worth of a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them (Hall, 1999). Equation (1) gives the merit of feature subset S consisting of k features.

[1] OpenCV: opencv.willowgarage.com/wiki/

[2] DMML: featureselection.asu.edu/software.php

$$Merit_{S_k} = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \quad (1)$$

Where, $\overline{r_{cf}}$ is the average value of all feature-classification correlations, and $\overline{r_{ff}}$ is the average value of all feature-feature correlations. The subset with the highest merit, as measured by Equation (1) found during the search, is used to reduce the dimensionality of both the original training data and the testing data. The CFS is defined by Equation (2). The $\overline{r_{cf_i}}$ and $\overline{r_{if_j}}$ variables are refereed to as correlations.

$$CFS = \max_{S_k}\left[\frac{r_{cf_1} + r_{cf_2} + \ldots + r_{cf_k}}{\sqrt{k + 2(r_{f_1f_2} + \ldots + r_{if_j} + \ldots + r_{f_kf_1})}}\right] \quad (2)$$

3.3 Classification

The classification was performed in Matlab using MatlabArsenal[3], a wrapper for the classifiers in Weka (Hall et al., 2009). Three types of base classifiers: lazy, function, and tree are considered.

Firstly, the three types of classifiers are evaluated: decision trees (J48), k-nearest neighbor (KNN), and support vector machine (SVM). In particular, SVM, KNN and decision trees are popular based on their compatibility and performance in many applications (Nguyen et al., 2005). These popular supervised learning algorithms that are simple to implement, span a variety of machine learning theories and techniques (e.g. function, lazy, tree), making them suitable in combined classifiers for addressing the diversity of features and subject variability. The *CVParameterSelection*, a meta-classifier in Weka that performs parameter selection by cross-validation was used to evaluate and determine parameter values for the classifiers with our dataset. The *K* value of one was selected for KNN classification. The exponent value of 1.0 (linear kernel), complexity factor of 1.0 was set for SVM. The C4.5 decision tree was used with confidence factor set to 0.25, and considering the subtree operation when pruning.

Secondly, two types of vote classifiers (as followed) are evaluated for combining classification results from the base classifiers to achieve the final classification decisions.

Average Vote Classifier (AVC): This vote classifier is a meta-classifier that combines the probability distribution of base classifier using the average probability rule. This is categorized as combining probabilistic (soft) outputs (Utthara et al., 2010, Kuncheva, 2004). This Vote classifier determines the class probability distribution computing the mean probability distribution of the base *N* arbitrary classifiers as followed (Seewald, 2003):

$$\overline{pred} = \sum_{i=1}^{N} \frac{\overline{P_i}}{N} \quad (3)$$

Where, $\overline{P_i}$ refers to the probability given by classifier *i*. The Voting prediction for *j* classes are mapped using $\overline{P_i'}$ instead of $\overline{P_i}$ in Equation (3). $\overline{P_i'}$ is the vector of $p.$, for all *j*, where

$$P'_{i,j} = \begin{cases} 1 \;\; if\; j = argmax_j(P_{i,j}), \;\; for\; given\; i \\ 0 \;\; otherwise \end{cases} \quad (4)$$

Weighed Majority Vote (WMV): In this vote classifier more competent base classifiers are given greater power to make the final decision based on the weighted majority vote algorithm (a meta-learning algorithm). This classifier is categorized as combining class labels (crisp outputs) (Utthara et al., 2010, Kuncheva, 2004). The class labels are available from the classifier outputs. The decision by classifier *i* (from *N* arbitrary classifiers) for class *j* is defined as $d_{i,j}$. If the classifier chooses class ω_j then $d_{i,j}$=1, and 0 otherwise. The classifiers whose decisions are combined through weighted majority voting will choose class ω_k if

$$\sum_{i=1}^{N} b_i\, d_{i,k} = \max_j \sum_{i=1}^{N} b_i\, d_{i,j} \quad (5)$$

Where, b_i is the weight coefficient for classifier *i*.

3.4 Training, Testing and Evaluation

All datasets were initially shuffled and randomized. Then the training and testing was performed separately with 10-fold cross validation. In 10-fold (*k-fold*) cross-validation, each dataset or sample was randomly partitioned into 10 subsamples. Of the 10 subsamples, a single subsample was retained as the validation data for testing the model, and the remaining 9 (*k-1*) subsamples were used as training data. The cross-validation process was then repeated 10 times (the folds), with each of the 10 subsamples used exactly once as the validation data. The 10 results from the folds were then averaged to produce a single estimation.

The *ZeroR* classifier is used for determining the baseline accuracy. The accuracy score is used for reporting the overall classification performance and precision score is used for reporting performance of individual classes.

4 Results and Discussion

In this section we provide results for detecting 3-degrees of valence (negative, neutral, positive) and arousal (low, medium, high) from *physio*, *face* and *fusion* using the vote classifiers (AVC and WMV) and the base classifiers (J48, KNN, SVM). Figures 2 and 3 give the average classification accuracy and the standard deviation (error bars) over all subjects[4]. The baseline classification accuracy is 33% for both valence and arousal.

Firstly, evaluating the overall performance of detecting degrees of valence and arousal from individual modalities (*physio* and *face*) and *fusion* shows that in almost all cases (except J48 and KNN in arousal) *fusion* has higher accuracy and lower standard deviation. Secondly, evaluating the classifiers show that both AVC and WMV in general exhibits similar (compared to individual modalities) or higher (compared to *fusion*) accuracy compared to the base classifiers.

[3]MatlabArsenal: cs.siu.edu/~qcheng/featureselection/index.html

[4] Results for 19 subjects due to SC sensor failure in one subject.

Among the base classifiers, KNN exhibits the highest accuracy for *physio* (50%), *face* (60%) and *fusion* (62%) in valence (Figure 2). J48 exhibits the highest accuracy for both *physio* (49%) and *fusion (56%)* with slightly higher accuracy with KNN for *face* (57%) in arousal (Figure 3). SVM shows comparatively low accuracy in *face* for both valence and arousal. For this dataset, the vote classifiers are unable to improve the accuracy of the individual modalities over the base classifiers, except in *physio* for valence (showing 2% and 1% improvement in AVC and WMV respectively). The *fusion* exhibits 2% improvement (both AVC and WMV) in valence and 4% improvement (only AVC) in arousal compared to the accuracy of the best base classifiers. However, the improvements by AVC were statistically significant[5] only over SVM for *face* in both valence and arousal. The improvement by WMV was also significant over SVM for *face* but only for arousal.

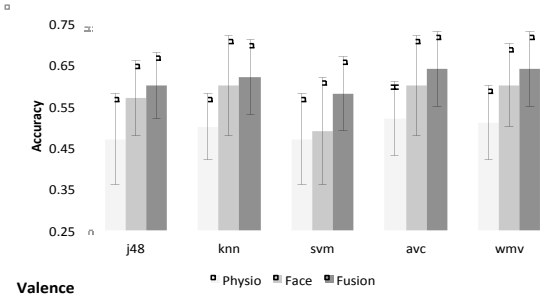

Figure 2: Accuracy (Mean, SD) of classifying 3-degrees of valence from *physio*, *face* and *fusion* using vote classfiers (AVC, WMV) and their base classfiers

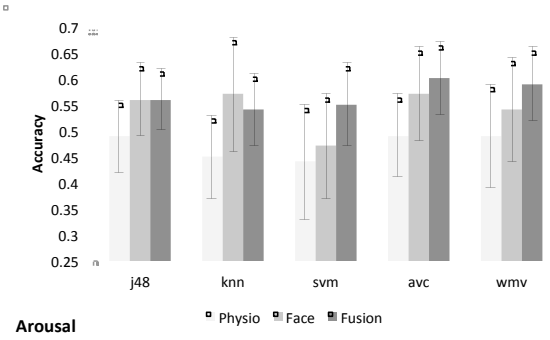

Figure 3: Accuracy (Mean, SD) of classifying 3-degrees of arousal from *physio*, *face* and *fusion* using vote classfiers (AVC, WMV) and their base classfiers

For this dataset, the base classfiers performed better for certain subjects and the vote classfiers for others. This reflects that for some subjects, where the performance of the base classfiers were poor, the vote classifiers acheived improvement. Figures 4 and 5 give the proportion of subjects representing the classifiers that performed with highest accuracy for valence and arousal respectivly. The vote classfiers (specially AVC) peformed better in most subjects for *fusion* compared to the individual modalities in both valence and arousal. This reflects that the vote classfies perform best using features from multiple modalities. For *face*, KNN perfomed better for most subjects in valence and J48 in arousal. SMV in general was less useful in most subjects except for *physio* in arousal. KNN was also less useful for *physio* and *fusion* in arousal.

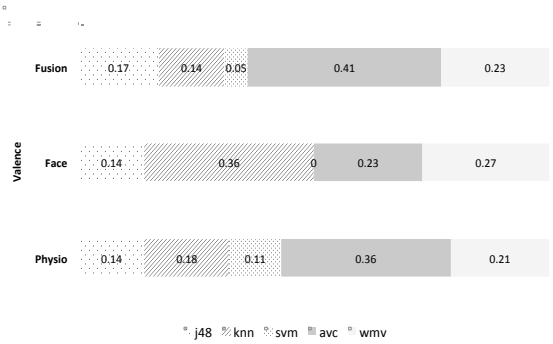

Figure 4: Proportion of subjects representing classifiers that performed with highest accuracy (val.)

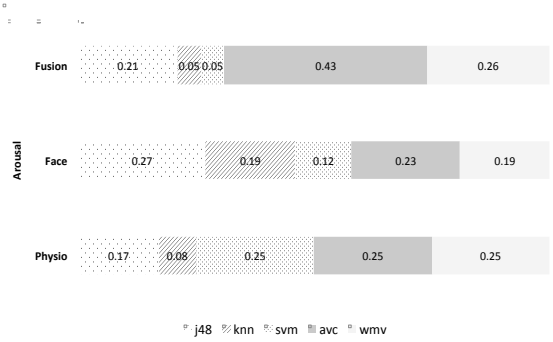

Figure 5: Proportion of subjects representing classifiers that performed with highest accuracy (ar.)

The classification was performed using balanced class distribution; therefore the precision score is used to report the classification accuracies of the individual classes, in this case individual degrees of valence and arousal. Table 1 gives the precision scores (mean and standard deviation) for classifying the individual degrees of valence and arousal from *fusion*. According to table 1, the vote classifiers exhibit higher precision compared to the base classifiers where both AVC and WMV show similar performance. For this dataset, AVC is slightly better at detecting *positive* valence and *medium* arousal whereas; WMV is best at detecting *neutral*, *negative* valence and *high* arousal. This reflects that vote classifiers have improved the accuracy of the individual affective states compared to the base classifiers.

	Valence			***Arousal***		
	Pos.	*Neu.*	*Neg.*	*High*	*Med*	*Low*
j48	0.62 (.13)	0.53 (.10)	0.64 (.12)	0.62 (.09)	0.47 (.14)	0.58 (.12)
knn	0.66 (.13)	0.53 (.14)	0.66 (.10)	0.57 (.10)	0.45 (.14)	0.61 (.12)
svm	0.66 (.17)	0.46 (.15)	0.61 (.19)	0.60 (.11)	0.39 (.18)	0.65 (.15)
avc	***0.72*** (.14)	0.53 (.13)	0.67 (.13)	0.64 (.10)	***0.50*** (.16)	***0.66*** (.11)
wmv	0.70 (.16)	***0.55*** (.13)	***0.69*** (.13)	***0.65*** (.09)	0.46 (.15)	***0.66*** (.14)

Table 1: Precision scores (Mean, SD) for detecting individual degrees of valence and arousal from *fusion*

[5] One-way ANOVA and post-hoc test with *bonferroni*

The evaluation of these classifiers with the same computational model can also be presented using another dataset, which consists of similar features (physiological and facial video), collected from participants during naturalistic interactions with an ITS. Hussain et al. (2012) collected this dataset and reported the accuracy of detecting degrees of valence and arousal with AVC from physiological and facial features (see paper for more details about the experiment). Participants had self-reported their affect (3-degrees of valence and arousal) judgment which were synchronized with the physiological and facial video features and used as labels for classification. However, the study by Hussain et al. (2012) did not address detection accuracies of the base classifiers, thus did not quantify if AVC achieved any improvements.

The classifiers selected (base classifiers, AVC, and WMV) in our study in this paper can be evaluated with this dataset that represent affects self-reported from naturalistic interactions compared to normative ratings from a controlled stimulus presentation. Following the study by Hussain et al. (2012), in Figures 6 and 7, we present the overall classification accuracies for detecting degrees of valance and arousal respectively from the base classifiers and the vote classifiers.

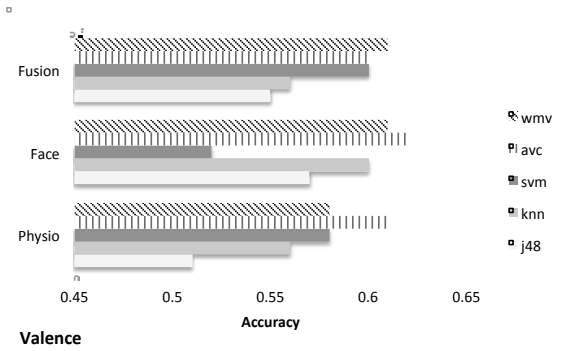

Figure 6: Detecting 3-degrees of valence (ITS dataset) from *physio*, *face* and *fusion* using vote classfiers (AVC, WMV) and their base classfiers

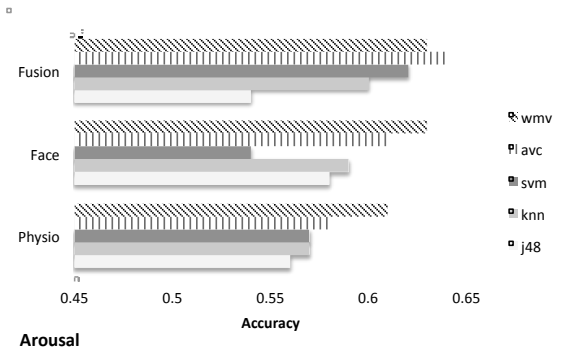

Figure 7: Detecting 3-degrees of arousal (ITS dataset) from *physio*, *face* and *fusion* using vote classfiers (AVC, WMV) and their base classfiers

The vote classifiers are able to improve the classification accuracy in valence and arousal for *face*, *physio*, and *fusion* compared to the base classifiers. In both valence and arousal, J48 was least useful for *physio* and *fusion*, whereas SVM was least useful for *face*. Similar to the IAPS dataset, KNN proved to be most useful using this dataset for *face* in both valence and arousal. Comparing the vote classifiers, WMV exhibits 1% improvement over AVC for *fusion* in valence and vice-versa in arousal. AVC shows 1% and 3% improvements for *face* and *physio* respectively over WMV in valence. However, WMV shows 2% and 3% improvements for *face* and *physio* respectively over AVC in arousal. The highest accuracy for detecting the degrees of valence is from face with 62% accuracy using AVC (similar trend as in (Hussain et al., 2012)). However, *fusion* has the highest accuracy for detecting the degrees of arousal also using AVC with 64% accuracy.

5 Conclusion

In this study we have evaluated combined classifiers and compared their performances with the base classifiers for detecting degrees of valence and arousal from multimodal features. The vote classifiers considered in this study have showed improvement over the base classifiers (J48, KNN, SVM) using our dataset, especially by fusing the multimodal features. The classifiers that are more important for certain modality have been identified, for example KNN showed to be more useful and SMV least useful for the face modality in both valence and arousal. Even though the improvements of the vote classifiers are not extremely higher than the base classifiers, they are still useful for multimodal features and subject variability in behavioural studies.

As for future work, more base classifiers can be explored to replace less useful ones (for modalities and individuals) to be used for combined classifiers. The classifier selection methods (Kuncheva, 2002) can be applied on these datasets, where every classifier can be an expert in a specific domain (modality) of the feature space for the combined classifier to improve the detection accuracy of affects.

6 References

Alzoubi, O., Hussain, M. S., D'mello, S. & Calvo, R. Affective modeling from multichannel physiology: analysis of day differences. International Conference on Affective Computing and Intelligent Interaction (ACII2011), 2011 Memphis, USA. Springer LNCS, 4-13.

Calvo, R. A. & D'mello, S. 2010. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing,* 1**,** 18-37.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I. H. 2009. The WEKA Data Mining Software: An Update. *ACM SIGKDD Explorations Newsletter,* 11**,** 10-18.

Hall, M. A. 1999. *Correlation-based feature selection for machine learning.* The University of Waikato.

Hussain, M. S., Alzoubi, O., Calvo, R. A. & D'mello, S. Affect detection from multichannel physiology during learning sessions with AutoTutor. The 15th International Conference on Artificial Intelligence in Education (AIED), 28 June - 02 July 2011a Auckland, New Zealand. Springer LNAI, 131-138.

Hussain, M. S., Calvo, R. A. & Aghaei Pour, P. Hybrid fusion approach for detecting affects from multichannel physiology. International Conference on Affective Computing and Intelligent Interaction (ACII2011), October 2011b Memphis, Tennessee, USA. Springer LNCS, 568-577.

Hussain, M. S., Hamed, M. & A., C. R. Categorical vs. dimensional representations in multimodal affect detecting during learning. 11th International Conference on Intelligent Tutoring Systems, 2012 Chania, Greece. Springer LNCS, 78-83.

Kuncheva, L. I. 2002. Switching between Selection and Fusion in Combining Classifiers: An Experiment. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on,* 32**,** 146-156.

Kuncheva, L. I. 2004. *Combining pattern classifiers: Methods and algorithms*, Wiley-Interscience.

Lang, P. J., Bradley, M. M. & Cuthbert, B. N. 1997. International affective picture system (IAPS): Technical manual and affective ratings. *Gainesville, FL: The Center for Research in Psychophysiology, University of Florida*.

Nguyen, T., Li, M., Bass, I. & Sethi, I. K. Investigation of Combining SVM and Decision Tree for Emotion Classification. Seventh IEEE International Symposium on Multimedia, 2005 Irvine, California, USA. 540-544.

Russell, J. A. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology,* 39**,** 1161–1178.

Sebe, N., Cohen, I., Gevers, T. & Huang, T. S. 2005. Multimodal Approaches for Emotion Recognition: A Survey. *Proc. SPIE,* 5670**,** 56–67.

Seewald, A. K. Towards a theoretical framework for ensemble classification. Proceedings of the 18th Int. Joint Conference on Artificial Intelligence (IJCAI-03), 2003. Morgan Kaufmann, 1443-1444.

Utthara, M., Suranjana, S., Sukhendu, D. & Pinaki, C. 2010. A Survey of Decision Fusion and Feature Fusion Strategies for Pattern Classification. *IETE Technical Review,* 27**,** 293-307.

Wagner, J., Kim, J. & Andre, E. From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification. IEEE International Conference on Multimedia and Expo 2005, 6 July 2005 Amsterdam, The Netherlands. 940-943.