Semi-feature Level Fusion for Bimodal Affect Regression Based on Facial and Bodily Expressions

Yang Zhang
Computational Intelligence Research Group
Department of Computer Science and Digital Technologies
Faculty of Engineering and Environment
Northumbria University
Newcastle NE1 8ST, UK
yang4.zhang@northumbria.ac.uk

Li Zhang
Computational Intelligence Research Group
Department of Computer Science and Digital Technologies
Faculty of Engineering and Environment
Northumbria University
Newcastle NE1 8ST, UK
li.zhang@northumbria.ac.uk

ABSTRACT
Automatic emotion recognition has been widely studied and applied to various computer vision tasks (e.g. health monitoring, driver state surveillance, personalized learning, and security monitoring). As revealed by recent psychological and behavioral research, facial expressions are good in communicating categorical emotions (e.g. happy, sad, surprise, etc.), while bodily expressions could contribute more to the perception of dimensional emotional states (e.g. arousal and valence). In this paper, we propose a semi-feature level fusion framework that incorporates affective information of both the facial and bodily modalities to draw a more reliable interpretation of users' emotional states in a valence–arousal space. The Genetic Algorithm is also applied to conduct automatic feature optimization. We subsequently propose an ensemble regression model to robustly predict users' continuous affective dimensions in the valence–arousal space. The empirical findings indicate that by combining the optimal discriminative bodily features and the derived Action Unit intensities as inputs, the proposed system with adaptive ensemble regressors achieves the best performance for the regression of both the arousal and valence dimensions.

Categories and Subject Descriptors
I.2 [ARTIFICIAL INTELLIGENCE]: Miscellaneous;

General Terms
Algorithms, Performance, Experimentation, Human Factors.

Keywords
Affective computing; multimodal affect sensing; adaptive ensemble models; feature selection; optimization.

1. Introduction
Automatic emotion recognition is a well-established and fast growing field, and there is an extensive literature available on emotion recognition from different modalities or their combinations (e.g. [1]-[6]). It has been widely acknowledged that the use of multimodal information allows for a more complete emotional description and enables more accurate recognition results.

Moreover, recent multimodal emotion recognition research has mostly focused on the recognition of facial and vocal expressions in terms of a small number of discrete emotion categories (e.g. [7]-[10]). However, people tend to exhibit non-basic, subtle and rather complex emotional states in real-life interactions, which may pose a great challenge to the aforementioned categorical recognition systems. In this research, we propose a semi-feature level fusion framework that incorporates affective information of both the facial and bodily modalities to draw a more reliable interpretation of users' emotional states in a valence–arousal space. The Genetic Algorithm (GA) based feature optimization and an ensemble regression model are also proposed to respectively identify the most optimal bodily and facial muscular features and robustly predict continuous affective dimensions.

The rest of the paper is organized as follows. Section 2 reviews the state-of-the-art developments in bimodal/multimodal emotion recognition. In Section 3, we present the detailed methodology of the proposed semi-feature level fusion. Section 4 discusses the GA based automatic feature optimization. Section 5 presents the proposed adaptive ensemble model for affective dimension regression, together with the other two benchmark single regression methods. Experiment, evaluation and discussion are presented in Section 6. Finally, we draw conclusions and identify future work in Section 7.

2. Related Work
In this section, we firstly introduce the discrete and dimensional conceptualization of emotions. We then summarize the state-of-the-art automatic multimodal dimensional emotion recognition systems and developments.

2.1 Different Emotion Theories
In the field of psychology, the modelling of emotions has been well studied. In literature, there are a number of widely acknowledged theories (e.g. OCC model [11] and Scherer theory [12]). In this research, we focus on two representative ones for emotion modelling: (1) categorical and (2) dimensional approaches.

Appears in: *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2015)*, Bordini, Elkind, Weiss, Yolum (eds.), May, 4-8, 2015, Istanbul, Turkey.

Table 1 Summary of multimodal and dimensional affect recognition systems (SAL: SAL database [41], BLSTM-NN: Bidirectional Long Short-Term Memory Neural Network, LSTM: Long Short-Term Memory Neural Network, BPNN: Backpropagation Neural Network, LDA: Linear Discriminant Analysis, SVM: Support Vector Machine, SVR: Support Vector Regression, GMM: Gaussian Mixture Model)

System	Modality/Feature type	Database/Number of sample	Learning/Classification model	Fusion strategy	Results
Karpouzis et al. [19]	Various visual & acoustic features	SAL, 4 subjects, 76 passages	Recurrent Network with 4 class-outputs	not reported	Negative/positive/active/passive (discrete), 67% recognition accuracy with vision, 73% with prosody, 82% after fusion
Kim [21]	Speech & physiological signals	Private database, 3 subjects, 343 samples	Modality-specific LDA-based classification	Integration of feature and model-level fusion	4 Arousal-Valence quadrants (discrete), 55% for feature fusion, 52% for decision fusion, 54% for hybrid fusion
Nicolaou et al. [17]	Facial expression, shoulder gesture, audio cues	SAL, 4 subjects, 30,000 visual and 60,000 audio samples	HMM and likelihood space via SVM	Model-level fusion, likelihood space fusion	Negative vs. positive valence (discrete), 91.76% by facial expressions, 94% by modal fusion
Nicolaou et al. [22]	Facial expression, shoulder gesture, audio cues	SAL, 4 subjects, 30,000 visual and 60,000 audio samples	SVR and BLSTM-NN	Feature/model-level, output-associative fusion	Valence and arousal (continuous), best results: RMSE=0.15 and CORR=0.796 for valence; RMSE=0.21 and CORR=0.642 for arousal
Metallinou et al. [18]	Body language and speech cues	Private database, 16 subjects, 100 recordings	LSTM and GMM-based prediction	Feature-level fusion	Valence, arousal and dominance (continuous), CORR=0.584, 0.056, 0.337, respectively
This work	Facial and whole-body expressions	Private database, 11 subjects, 40,000 samples (frames)	BPNN, SVR, and the proposed ensemble regression model	Semi-feature level fusion	Valence and arousal (continuous), MSE= 0.077 and CORR= 0.886 for valence; MSE= 0.056 and CORR= 0.907 for arousal

The former advocates that affective state is able to be represented by a small number of prototypical emotions or their mixtures, which are basic, hard-wired in our brain, and recognized universally, such as the six basic emotions (i.e. happiness, surprise, fear, anger, sadness, and disgust) identified by Ekman and his colleagues [13]-[15].

The latter argues that affective state could be described by certain continuous attributes. A representative model proposed by Posner et al. [16] suggested that the majority of affect variability is able to be covered by two orthogonal dimensions, i.e. arousal and valence. The arousal dimension refers to the intensity of the emotional experience, and it ranges from apathetic sleepiness to frantic excitement. The valence dimension describes the level of pleasure of an emotion, and it ranges from negative unpleasant feelings to positive pleasant feelings.

The dimensional model could be a more natural, flexible and effective way to interpret emotions [17, 18]. Thus, we employ the dimensions of arousal and valence in a continuous scale for the automatic interpretation of users' emotional states in this research.

2.2 Review of State-of-the-Art Developments

Recently, a growing body of research has focused on dimensional affect recognition based on various combinations of modalities. For example, Karpouzis et al. [19] employed a Recurrent Neural Network which lends itself well to modeling dynamic events in both users' facial expressions and speech for the recognition of emotion in naturalistic video sequences. In their work, a quantized dimensional representation of users' emotional states (i.e. activation and valence) was applied, instead of detecting discrete emotion categories. Kanluan et al. [20] employed late fusion of facial expression and audio channels by using weighted linear combinations of their outputs respectively obtained by Support Vector Machines for regression to estimate the valence, activation, and dominance dimensions (on a 5-point scale, for each dimension).

Most recently, a few attempts have been proposed for actual continuous affective dimension regression (without quantization). For example, Nicolaou et al. [17] employed three modalities including facial expression, shoulder gesture and vocal cues for continuous tracking of the valence and arousal affective dimensions using Support Vector Regression (SVR) and Long-

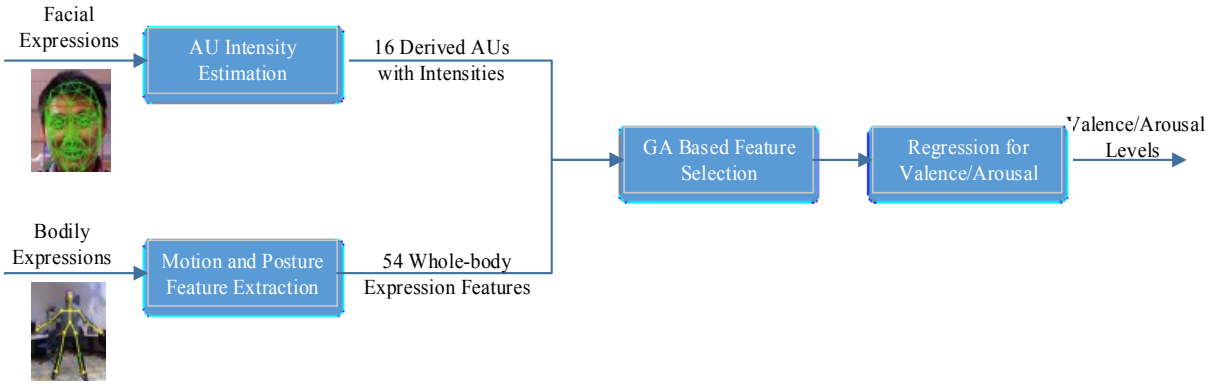

Figure 1 The proposed semi-feature level fusion framework

Short Term memory (LSTM) regression. Metallinou et al. [18] proposed a Gaussian Mixture Model-based approach to continuously predict levels of participants' activation, valence and dominance during the course of affective dynamic interactions using body language and speech features. They also produced a statistical analysis of each single bodily feature in order to select a subset of de-correlated informative features for each affective dimension. Promising results were obtained for the tracking of the arousal and dominance dimensions. For a more clear comparison, in Table 1, we briefly summarize some state-of-the-art applications that employ multiple modalities to model and recognize affect in terms of affective dimensional space, together with our work presented in this paper. Although some earlier applications listed in Table 1 ([19], [21], [22]) applied a discretized classification scheme rather than a continuous dimensional space, we still include them as they are relevant to this study.

In comparison to the existing work listed in Table 1, our research presents the first semi-feature level fusion framework in the literature that effectively combines users' whole-body features and facial Action Unit intensities to improve regression performance for affective dimensions. By employing the GA based feature optimization and the proposed adaptive ensemble regression models, our system achieves the best performance in terms of both Mean Squared Error (MSE) and Pearson correlation coefficient (CORR) measurements. The overall system is developed based on a Microsoft Kinect platform. The detailed semi-feature level fusion methodology is presented in the following.

3. Fusion Strategies for Facial and Bodily Modalities

In this section, we firstly describe the facial and bodily expression features that have been extracted and employed. We subsequently detail the proposed semi-feature level fusion framework, followed by the automatic feature selection based on the GA optimization.

3.1 Facial Expression Features

In this research, Microsoft Kinect has been used to extract initial raw facial features. Then we employ facial Action Unit [14] intensity estimators proposed in Zhang et al. [23] to measure the intensities of 16 diagnostic facial AUs, which are then used as input facial features in this research rather than using raw features (e.g. geometric or textural facial features). This is because AU intensity features are more compact and less redundant than raw facial features and can well reflect users' emotional states [23].

These AU intensity estimators automatically select 16 motion-based facial feature sets using minimal-redundancy-maximal-relevance criterion based optimization and robustly estimate the intensities of 16 diagnostic AUs for each frame using SVRs. The 16 derived AUs are AU1 (Inner Brow Raiser), AU2 (Outer Brow Raiser), AU4 (Brow Lowerer), AU5 (Upper Lid Raiser), AU6 (Cheek Raiser), AU10 (Upper Lip Raiser), AU12 (Lip Corner Puller), AU13 (Cheek Puffer), AU15 (Lip Corner Depressor), AU17 (Chin Raiser), AU18 (Lip Puckerer), AU20 (Lip Stretcher), AU23 (Lip Tightner), AU24 (Lip Pressor), AU26 (Jaw Drop) and AU27 (Mouth Stretch).

3.2 Bodily Expression Features

Moreover, we also extract a total of 54 whole-body expression features, including both static posture (e.g. distances and angles) and dynamic motion (e.g. velocity, amplitude and acceleration) features. These features are calculated based on 20 key skeletal joints tracked by Microsoft Kinect and its Natural User Interface SDK [24] in a 3D geometric manner for each frame. These features range from lower-level features, such as the joint angles of elbow and knee, to more interpretable higher-level features, such as the lean angle of spine and the degree of body contraction/expansion. The bodily expression features extracted include the following types:

- **Body Expansion Index** measures the degree of contraction and expansion of the body, in frontal, lateral and vertical directions, respectively. Figuratively speaking, it computes a 3D bounding region, i.e., the minimum cuboid surrounding the entire body.
- **Euclidean Distance** is the distance between two given skeletal joints.
- **Lean Angle** indicates the geometric angle of spine leaning forward/backward.

- **Instantaneous Velocity** can be calculated by dividing the displacement of a given joint between the current and last frames by the time interval of the two frames. It is related to the kinetic energy of a motion.
- **Average Velocity** states the averaged value of speed, and can be calculated by dividing the total motion trajectory length of a joint by the corresponding time interval.
- **Amplitude** indicates the maximum Euclidean Distance among the positions of a given joint within a predetermined time interval.
- **Acceleration** is the rate of change of velocity between the current and last frames. It is caused by the force applied to move the body part, and can be used to distinguish between smooth and sudden motions.

3.3 Semi-Feature Level Fusion

As illustrated in Figure 1, the proposed semi-feature level fusion is realized by concatenating the derived AU intensities and the extracted bodily features into a new feature vector which is subsequently employed as inputs to affective dimensional regressors for both arousal and valence. A feature normalization procedure is also performed, in which each attribute is linearly scaled to the range of [0, +1]. We subsequently conduct a GA-based automatic feature selection to identify the most optimal discriminative feature subset for each affective dimension. Finally, we employ the adaptive ensemble regression models, together with two other benchmark single Backpropagation Neural Networks (BPNNs) and SVRs for the prediction of users' continuous affective dimensions.

Our motivation is threefold. Firstly, there is strong psychological evidence (e.g. [13], [25]) indicating that the bodily expressions could be a better indicator of the arousal dimension, whereas some facial actions convey rich information of the valence dimension (e.g. the occurrence of AU15 (Lip Corner Depressor) usually indicates a 'sad' emotion, whereas AU12 (Lip Corner Puller) normally occurs with 'happiness'). Thus, their combination is able to contribute more complementary information for dimensional affect prediction.

Secondly, we focus on dimensional interpretation of affect, because in such an approach, even complex/blended emotional expressions and subtle emotion transitions can be captured and represented properly using a continuous scale of different dimensions, which could be too difficult to deal with through the categorical approach.

Most importantly, although it remains largely unclear about how humans achieve effective fusion of multimodal affective signals for a final decision, recent literature ([10], [26]) was more supportive of an early stage fusion (e.g. feature-level fusion) rather than a late stage fusion (e.g. decision-level fusion), because the feature-level fusion is able to catch more information and relations of different modalities to inform affect interpretation. However, it is difficult to directly combine features from different modalities with various metrics, dimensionalities and temporal structures. Thus, we propose the semi-feature level fusion that appropriately integrates the derived AU intensities with bodily features for dimensional affective interpretation.

4. The GA-based Feature Selection

Although great effort spent on feature extraction process, the 70 bimodal affective features (16 derived facial AU intensities + 54 bodily features) are not necessarily of equal importance or quality. Some redundant or irrelevant features could result in inaccurate conclusion whereas a compact and optimal subset of features could benefit subsequent regression models by improving their generalization and interpretability. Thus, the GA-based automatic feature optimization is performed to identify the most optimal feature subsets for each affective dimension out of the entire set of 70 features.

Algorithm 1 GA for Feature Selection

```
1: initialize population P;
2: repeat {
3:         select two parents p1 and p2 from P;
4:         offspring = crossover (p1; p2);
5:         mutation(offspring);
6:         replace(P, offspring);
7:        }
  until (stopping condition);
9: }
```

The GA is a biologically inspired optimization search method that mimics natural evolution. It is a promising alternative to conventional feature selection methods (e.g. [27], [28]). The most distinctive aspect of this algorithm is that it maintains a set of solutions (called individuals or chromosomes) in a population and employs a mechanism of selecting fitter chromosomes at each generation through genetic crossover and mutation operations based on the Darwinian principle of 'survival of the fittest'. The GA stops when the number of iterations reaches a preset threshold or acceptable results are obtained. Algorithm 1 presents the pseudocode of the employed GA optimization. Figure 2 illustrates a cycle of the GA evolutionary process.

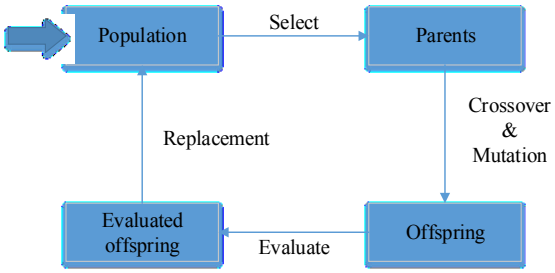

Figure 2 The evolutionary cycle of the GA

For the feature selection problem, solutions (i.e. selected features) are represented in a string with n binary digits, with each binary digit representing each feature, and values 1 and 0 meaning selected and removed features respectively. For example, chromosome '10001001' indicates the first, fifth and eighth features are selected. The GA starts with an initial population consisting of a number of d randomly generated solutions. In this research, the population size d is set to 30 according to original feature dimensions and computational complexity. We apply the following parameter setting to achieve a balance between the regression accuracy and the computational complexity:

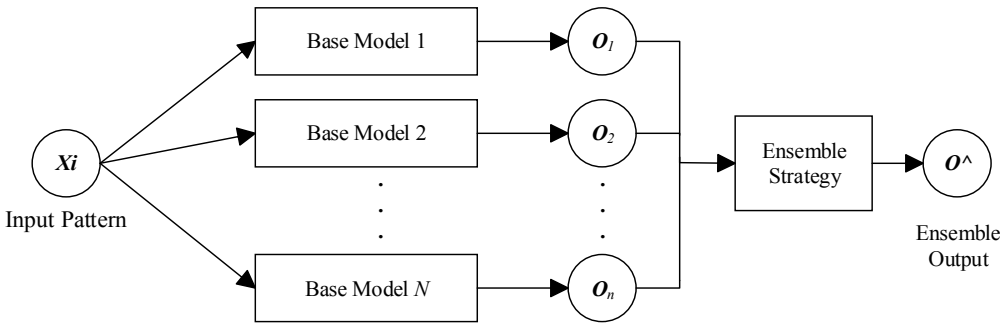

Figure 3 An example of an ensemble learning mode

control procedure: steady-state;

population size = 30;

crossover probability = 1.0;

mutation probability = 0.03;

maximum generations = 1000;

These parameters are originated by the default setting of the GA algorithm with slight adjustment to fit into our application domain which has a comparatively small feature set (i.e. 70 features).

5. Dimensional Affect Interpretation

In this research, we employ three distinctive machine learning algorithms, i.e. the proposed adaptive ensemble regression models, single BPNNs and SVRs, for the regression of affective dimensions. The latter two are commonly used for continuous affect regression problems in the existing applications (e.g. [17], [29]-[32]), and their experimental results will be used as the benchmark for comparison.

5.1 The Proposed Adaptive Ensemble Regression Model

In this research, we propose an adaptive ensemble model for the regression of valence and arousal dimensions. As illustrated in Figure 3, ensemble learning refers to approaches that generate several base models that complement each other to make a prediction. Compared to traditional single model-based methods, ensembles have the advantages of improved robustness and increased prediction accuracy [33]. We employ two instantiations of the proposed adaptive ensemble regression model to effectively handle continuous affective dimension prediction tasks, with each ensemble model dedicated to each affective dimension (i.e. either valence or arousal). The proposed ensemble regression model is developed and modified based on an ensemble classifier for novel class detection proposed by [23]. For this proposed ensemble regression model, we employ SVRs as the base regressors and use a series of adaptive ensemble mechanisms for the model generation, so that it is able to deal with regression problems efficiently. For an exhaustive review of ensemble approaches, readers may refer to [34, 35].

The ensemble model generation starts with the weight initialization for the training dataset based on a multiple linear regression analysis against the ground truth. Then a subset of training clips with higher weights is selected from the original training set to train a base model. Although a variety of algorithms, such as Decision Trees and Neural Networks, could be used as the base regressor, in this research, we employ SVRs as the base regressor. The detailed introduction of the SVR is provided in Section 5.3. Subsequently, we calculate and assign a weight to the current base model based on its regression performance for the original training dataset. We also update the weights of the training clips with the aim of increasing the weights of those clips which have higher error rates and are more difficult to predict. Overall, the above procedures iterate three times, thus three weighted base regressors are generated for the building of the ensemble model (considering a balance between performance and computational complexity). The final ensemble regression result can be therefore obtained by calculating the weighted average of the outputs of the three base models.

We have also employed two other single regression models, i.e. BPNN and SVR, for the prediction of affective dimensions, whose results are used as benchmark for comparison with those obtained by the ensemble regression models. The single regressors, BPNN and SVR, are introduced respectively in the following sections.

5.2 Feedforward Neural Network

As mentioned earlier, we employ single-hidden layer feedforward Neural Networks respectively for the regression of arousal and valence. In this research, we employ two BPNNs for the regression of the two affective dimensions respectively with each BPNN consisting of an input layer, a hidden layer, and an output layer, as shown in Figure 4. Each layer of the BPNN contains a number of nodes, which are interconnected with adjacent layers. Also, each node is a simple processing element that responds to the weighted inputs received from the preceding layer.

The feedforward Neural Networks are trained by Backpropagation algorithm [36], which iteratively adjusts the weights between the nodes in response to the errors until some targeted minimal error is achieved between the actual and target output values. We apply the following parameter setting, so that it is able to best achieve a balance between accuracy, speed and generalization performance.

learning rate = 0.2;

momentum value = 0.7;

termination error = 0.01;

number of the nodes in hidden layer = 10-50;

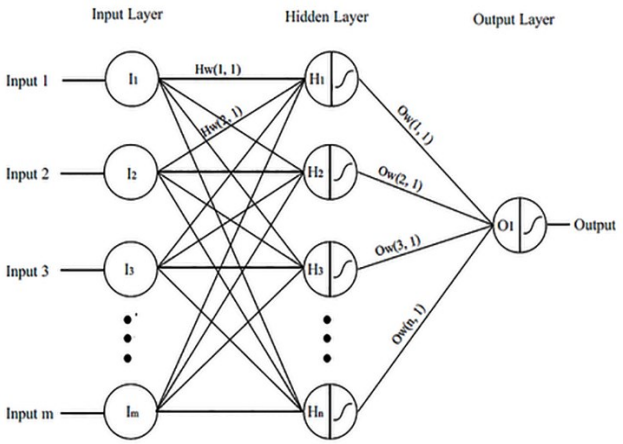

Figure 4 A sample topology of a single-hidden layer feedforward neural network

5.3 Support Vector Regression

We also employ single models such as SVRs for the regression of both valance and arousal dimensions in this research. As one of the most dominant kernel methods, Support Vector Machine (SVM) employs the convex optimization function which guarantees that the optimal solution will be found. The basic idea of SVR is to compute a linear regression function in a higher dimensional feature space where the lower dimensional inputs are mapped using a kernel function [37].

In this research, we apply non-linear radial basis function kernel (RBF) SVRs, because the RBF is able to effectively deal with nonlinear cases and has simpler hyperparameters compared to other nonlinear kernels (e.g. polynomial kernel). Please note that when the dimensions of features are very high (e.g. thousands), the RBF kernel may become unsuitable in comparison to a linear kernel. However, it is not the case in this application.

We employ the established LibSVM Library [38] for the SVR implementation. A typical "grid search" procedure with cross-validation is conducted to determine the optimal combination of the cost (C), gamma (g) and epsilon (ε) parameters [39]. More specifically, various combinations of parameter values (i.e. exponentially growing values: $C = 2^{-10}, 2^{-9}, ..., 2^{15}$; $g = 2^{-15}, 2^{-14}, ..., 2^{10}$; $\varepsilon = 2^{-10}, 2^{-9}, ..., 2^{-1}$) are conducted and the one with the lowest MSE is selected. The MSE evaluates the prediction results by taking into account the squared error of the predicted value from the ground truth.

6. Experiments, Evaluations and Discussions

In this section, we firstly present the data prepared for system evaluation. The experimental results are discussed subsequently.

6.1 Data Collection and Annotation

In this research, eleven participants, five female and six male, ranging from 25 to 40 years old, were recruited for our affective facial and bodily expression data collection. All of them were asked to take a brief training, which allowed them to get more familiar and comfortable with the Kinect sensor and laboratory conditions to enable more natural performance. In order to avoid stereotypical and strongly acted expressions, we employed more diverse and interactive methods to arouse emotional responses of participants, such as viewing tragic/comedic movie clips, telling jokes, and making improvised performances with each other, instead of directly guiding them to perform specific emotional bodily expressions. A total of 85 clips containing various emotional expressions was recorded (including both skeletal tracking data from the depth sensor and color video data from the RGB camera). The time length of each clip varies between 10 and 20 seconds (i.e. between approximate 300 and 600 frames per clip). Each clip starts from a neutral state and includes one or a few emotional expressions with bodily and facial displays.

In order to establish reliable ground truth for each affective dimension for system evaluation, we recruited five annotators to perform frame-by-frame affective dimension annotation for each clip, most of whom had essential experience in affective annotation tasks. The range of valence/arousal ratings is from -1 (the most negative/inactive) to +1 (the most positive/active). We apply the following three steps to establish the ground truth:

- We calculate the CORR for each pair of annotations, and then filter out the pair(s) with the CORR lower than a cutoff threshold;
- We calculate the mean value of each annotation, and then filter out the pair(s) with the difference of the mean values greater than a cutoff threshold;
- The rest of the annotations are selected to compute the ground truth for the corresponding clip by taking the average of them. If there is no annotation left (i.e. all the five annotations are filtered out), that clip will be excluded from our corpora, as lacking of essential inter-annotator agreement to establish the ground truth.

The cutoff thresholds for the CORR and the mean value difference are respectively set to 0.4 (a standard for moderate correlations in statistics) and 0.5, empirically. In this way, we select 58 and 60 valid emotion clips with acceptable inter-annotator agreement and well-founded ground truth for valence and arousal, respectively. The system evaluation is presented in the following.

6.2 Experimental Results

All experiments are conducted following a leave-one-subject-out cross-validation scheme, i.e. the data of ten subjects are used for training and the remaining one for testing, and each subject is tested in turn. The final result is an average over these rounds. As mentioned earlier, the merged feature vector consists of the derived AU intensities and the extracted bodily features.

In Table 2, we present the experimental results of applying the single BPNNs, SVRs, and ensemble regression models with SVRs as the base regressors for the regression of arousal and valence dimensions using the merged features automatically selected based on the GA optimization. To evaluate the effectiveness of the proposed semi-feature level fusion framework for continuous affect regression, we also conduct experiment with solely bodily features.

Table 2 Experimental results of the proposed semi-feature level fusion using single BPNNs and SVRs, and the proposed ensemble model

	Modality	Number of selected features	BPNNs		SVRs		Ensemble (SVRs)	
			CORR	MSE	CORR	MSE	CORR	MSE
Arousal	Bodily	37	0.808	0.072	0.867	0.066	0.903	0.057
	Bimodal	39 (27 bodily + 12 facial)	**0.882**	**0.068**	**0.89**	**0.059**	**0.907**	**0.056**
Valence	Bodily	25	0.723	0.121	0.791	0.103	0.815	0.093
	Bimodal	31 (24 bodily + 7 facial)	**0.865**	**0.091**	**0.883**	**0.089**	**0.886**	**0.077**

First of all, as shown in Table 2, the fusion of facial and bodily modalities provides obvious performance enhancement for both arousal and valence dimensions. Especially for valence, integrating facial AU intensity information with bodily features appears to perform much better than solely using bodily features in terms of both MSE and CORR metrics. These results are also theoretically consistent with psychological research (e.g. [13], [25]) which hypothesizes that facial expressions communicate rich and explicit affective information of the valence dimension (e.g. happiness and sadness). Moreover, by using adaptive ensemble regression models, we achieve the best prediction performance for both of the arousal (CORR = 0.907, MSE = 0.056) and valence (CORR = 0.886, MSE = 0.077) dimensions. These results demonstrate that the proposed semi-feature level fusion framework provides an effective solution for facial and bodily modality fusion, and the system achieves very impressive performance improvements. Moreover, Table 3 lists the most optimal combinations of features determined by the GA that generate the best results for both arousal and valence using the ensemble models.

7. Conclusions

There is recently a shift of focus from discrete and unimodal emotion recognition to continuous and multimodal recognition, as the latter is more flexible and reliable for the interpretation of spontaneous emotions in real-life scenarios. In this research, we proposed a semi-feature level fusion framework that effectively combines affective information from both the facial and bodily modalities to boost the performance of the dimensional affect recognition. The semi-feature level fusion is realized by concatenating the derived AU intensities and the discriminative bodily features into a merged feature vector which is subsequently optimized by the GA and then employed as inputs of the proposed ensemble model for the regression of both arousal and valence. To the best of our knowledge, this is the first attempt to combine AU intensities and whole-body features for automatic affect recognition, which overcomes the inherent shortcomings of conventional feature and decision-level fusion.

Finally, we identify the following several potential directions for future work. First of all, although we have collected sufficient data for system evaluation, these data are all recorded under laboratory conditions. As pointed out by Kleinsmith & Bianchi-Berthouze [40], a more naturalistic and extensive corpus with diverse subjects and challenging spontaneous affective expressions could better reflect the system performance in real-life scenarios. Besides, using an extensive database annotated in a richer affective space with a variety of affective dimensions, the proposed arousal-valence dimensional emotion recognition framework can be easily extended to include other additional dimensions, such as dominance and expectation. We can also

Table 3 The most optimal feature combinations for ensembles selected by the GA optimization

	Modality	Features
Arousal	Bodily (27)	Body Expansion Index (in X, Y, Z axes), Head Lean Angle, Body Lean Angle, Left/Right Elbows Joint Angle, Left/Right Knees Joint Angle, Distance between Left/Right Hands and Left/Right Shoulders (in X, Y, Z axes), Instantaneous Velocity of Hands, Instantaneous Velocity of Elbows, Amplitude of Hands, Acceleration of Hands, Acceleration of Elbows
	Facial (12)	AU1, AU2, AU4, AU5, AU6, AU12, AU13, AU15, AU20, AU23, AU26, and AU27
Valence	Bodily (24)	Body Expansion Index (in X, Y, Z axes), Head Lean Angle, Body Lean Angle, Left/Right Elbows Joint Angle, Left/Right Knees Joint Angle, Distance between Left/Right Hands and Left/Right Elbows (in X, Y, Z axes), Distance between Left/Right Hands and Left/Right Shoulders (in X, Y, Z axes), Instantaneous Velocity of Hands,
	Facial (7)	AU1, AU2, AU4, AU6, AU12, AU15, and AU23

further explore the correlations between those different affective dimensions for affect interpretation. Furthermore, literature indicates that, in some cases, the performance of ensembles could be potentially boosted by combining different types of base learning algorithms within one ensemble [34]. Thus, it shows potential to further improve the adaptive ensemble models by exploring such combinations of diverse base models.

8. REFERENCES

[1] Soyel, H. & Demirel, H. (2010). Optimal feature selection for 3D facial expression recognition using coarse-to-fine classification. *Turkish Journal of Electrical Engineering and Computer Sciences*, 18 (6), 1031–1040.

[2] Kleinsmith, A., Bianchi-Berthouze, N., & Steed, A. (2011). Automatic Recognition of Non-Acted Affective Postures. *IEEE Trans. on Systems, Man, and Cybernetics*. Part B, 41 (4), 1027–1038.

[3] Gunes, H. & Pantic, M. (2010). Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners. In *Proc. of Int. Conf. on Intelligent Virtual Agents*, 371–377.

[4] Bernhardt, D., & Robinson, P. (2007). Detecting affect from non-stylized body motions. LNCS: *Proceedings of 2nd Int. Conference on Affective Computing and Intelligent Interaction*, 59–70.

[5] Bartlett, M.S., Littlewort, G.C., Frank, M.G., Lainscsek, C., Fasel, I.R., & Movellan, J.R. (2006). Automatic recognition of facial actions in spontaneous expressions. *Journal of Multimedia*, 1 (6), 22–35.

[6] Tsalakanidou, F., & Malassiotis, S. (2010). Real-time 2D+3d Facial Action and Expression Recognition. *Pattern Recognition*, 43 (5), 1763–1775.

[7] Gunes, H., Piccardi, M., & Pantic, M. (2008). From the lab to the real world: Affect recognition using multiple cues and modalities. In Jimmy Or (Ed.), *Affective computing, focus on emotion expression, synthesis and recognition* (pp. 185–218). Vienna, Austria: I-Tech Education and Publishing.

[8] Gunes, H. & Pantic, M. (2009). Automatic temporal segment detection and affect recognition from face and body display. *IEEE Trans on Systems, Man, and Cybernetics*, Part B, 39 (1), 64–84.

[9] Cohn, J., Kreuz, T., Yang, Y., Nguyen, M., Padilla, M., Zhou, F., & Fernando, D. (2009). Detecting depression from facial actions and vocal prosody. In *Proceeding of International Conference on Affective Computing and Intelligent Interaction (ACII2009)*.

[10] Zeng, Z., Pantic, M., Roisman, G.I., & Huang, T.H. (2009). A survey of affect recognition methods: audio, visual and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31 (1), 39–58.

[11] Adam, C., Herzig, A. & Longin, D. (2009). A logical formalization of the OCC theory of emotions, *Synthese*, 168:201–248.

[12] Scherer, K. R. (2000). Emotions as episodes of subsystem synchronization driven by nonlinear appraisal processes. In M. D. Lewis & I. Granic (Eds.) *Emotion, development, and self-organization: Dynamic systems approaches to emotional development* (pp. 70-99). New York: Cambridge University Press.

[13] Ekman, P., & Friesen, W.V. (1983). *Emfacs-7: Emotional Facial Action Coding System*. University of California at San Francisco.

[14] Ekman, P., Friesen, W.V., & Hager, J.C. (2002). *Facial Action Coding System, the Manual*. Published by Research Nexus division of Network Information Research Corporation, USA.

[15] Ekman, P., Friesen, W.V., & Hager, J.C. (2002). *Facial Action Coding System Investigator's Guide*. Consulting Psychologist Press, Palo Alto, CA.

[16] Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, 17 (3), 715–734.

[17] Nicolaou, M., Gunes, H., & Pantic, M. (2011). Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Trans. on Affective Computing*, 2 (2), 92–105.

[18] Metallinou, A., Katsamanis, A., & Narayanan, S. (2013). Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information. *Image and Vision Computing*, 31 (2), 137–152.

[19] Karpouzis, K., Caridakis, G., Kessous, L., Amir, N., Raouzaiou, A., Malatesta, L., & Kollias, S. (2007). Modeling naturalistic affective states via facial, vocal, and bodily expressions recognition. *In: Lecture notes in artificial intelligence*, vol. 4451, Springer, Berlin, 91–112.

[20] Kanluan, I., Grimm, M., & Kroschel, K. (2008). Audio-visual emotion recognition using an emotion recognition space concept. *Proc. of European Signal Processing Conference*.

[21] Kim, J. (2007). Robust Speech Recognition and Understanding. *Bimodal Emotion Recognition using Speech and Physiological Changes*, I-Tech Education and Publishing, 265–280.

[22] Nicolaou, M., Gunes, H., & Pantic, M. (2010). Audio-visual classification and fusion of spontaneous affective data in likelihood space. In *Proc. of IEEE Int. Conf. on Pattern Recognition*, 3695–3699.

[23] Zhang, Y., Zhang, L. and Hossain, A. (2015). Adaptive 3D Facial Action Intensity Estimation and Emotion Recognition. *Expert Systems with Applications*, 42 (3), 1446–1464.

[24] Microsoft Corporation. (2013). *Kinect for windows SDK programming guide*, version 1.8.

[25] Ekman, P., & Friesen, W.V. (1967). Head and body cues in the judgment of emotion: A reformulation. *Perceptual and Motor Skills*, vol. 24, 711–724.

[26] Stein, B. and Meredith, M.A. (1993). *The Merging of Senses*. USA, MIT Press.

[27] Huang, C.L. & Wang, C.J. (2006). A GA-based feature selection and parameters optimization for support vector machine. *Expert Systems with Applications*, 31 (2), 231–240.

[28] Oh, I.S., Lee, J.S., & Moon, B.R. (2004). Hybrid genetic algorithms for feature selection. *IEEE Transactions on*

Pattern Analysis and Machine Intelligence, 26 (11), 1424–1437.

[29] Wollmer, M., Eyben, F., Reiter, S., Schuller, B., Cox, C., Douglas-Cowie, E., & Cowie, R. (2008). Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies. In *Proc. of 9th Inter, speech Conf.*, 597–600.

[30] Zhang, L., Jiang, M., Farid, D. & Hossain, A.M. (2013). Intelligent Facial Emotion Recognition and Semantic-based Topic Detection for a Humanoid Robot. *Expert Systems with Applications*, 40 (2013), pp. 5160-5168.

[31] Zhang, L. & Barnden, J. (2012). Affect Sensing Using Linguistic, Semantic and Cognitive Cues in Multi-threaded Improvisational Dialogue. *Cognitive Computation*. Volume 4. Issue 4. 436-459.

[32] Zhang, L. (2013). Contextual and Active Learning-based Affect-sensing from Virtual Drama Improvisation. *ACM Transactions on Speech and Language Processing (TSLP)*, Vol 9, Issue 4, Article No. 8.

[33] Garcia-Pedrajas, N., Herv´as-Mart´ınez, C., & Ortiz-Boyer, D. (2005). Cooperative Coevolution of Artificial Neural Network Ensembles for Pattern Classification. *IEEE Transactions on Evolutionary Computation*, 9(3), 271–302.

[34] Mendes-Moreira, J.A., Soares, C., Jorge, A.M., & Sousa, J.F.D. (2012). Ensemble approaches for regression: A survey. *ACM Computing Surveys*, 45(1), 1-40.

[35] Farid, D., Zhang, L., Hossain, A.M., Rahman, C.M., Strachan, R., Sexton, G. and Dahal, K. (2013). An Adaptive Ensemble Classifier for Mining Concept-Drifting Data Streams. *Expert Systems with Applications*, Vol 40, Issue 15. 5895-5906.

[36] Hecht-Nielsen, R. (1989). Theory of the Backpropagation neural network. *Neural Networks, IJCNN, International Joint Conference on*, San Diego, CA, USA.

[37] Basak, D., Pal, S., & Patranabis, D. C. (2007). Support vector regression. *Neural Information Processing* - Letters and Review, 11(10).

[38] Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27: 1-27:27, 2011.

[39] Hsu, C., Chang, C. & Lin, C. 2010. *A practical guide to support vector classification*. Department of Computer Science National, Taiwan University.

[40] Kleinsmith, A., & Bianchi-Berthouze, N. (2013). Affective *Body Expression Perception and Recognition: A Survey. Affective Computing, IEEE Transactions on*, 4 (1), 15–33.

[41] Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, O., McRorie, M. (2007). The HUMAINE Database: addressing the needs of the affective computing community. In *Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction*, Lisbon, Portugal, 488–500.