Copyright ⓒ 2014 KSII

Multi-classifier Fusion Based Facial Expression Recognition Approach

Xibin Jia[1], Yanhua Zhang[1], David Powers[1, 2] and Humayra Binte Ali[2]
[1] Beijing Municipal Key Laboratory of Multimedia and Intelligent Software Technology, Beijing University of Technology, Beijing, 100124, China
[e-mail: jiaxibin@bjut.edu.cn, yh_zhang2011@emails.bjut.edu.cn]
[2] School of Computer Science, Engineering and Mathematics, Flinders University of South Australia, Adelaide, Australia
[e-mail: {david.powers, humayra.ali}@flinders.edu.au]
*Corresponding author: Xibin Jia

Received October 30, 2013; accepted Decemebr 23, 2014; published January 29, 2014

Abstract

Facial expression recognition is an important part in emotional interaction between human and machine. This paper proposes a facial expression recognition approach based on multi-classifier fusion with stacking algorithm. The kappa-error diagram is employed in base-level classifiers selection, which gains insights about which individual classifier has the better recognition performance and how diverse among them to help improve the recognition accuracy rate by fusing the complementary functions. In order to avoid the influence of the chance factor caused by guessing in algorithm evaluation and get more reliable awareness of algorithm performance, kappa and informedness besides accuracy are utilized as measure criteria in the comparison experiments. To verify the effectiveness of our approach, two public databases are used in the experiments. The experiment results show that compared with individual classifier and two other typical ensemble methods, our proposed stacked ensemble system does recognize facial expression more accurately with less standard deviation. It overcomes the individual classifier's bias and achieves more reliable recognition results.

Keywords: Multi-classifier fusion, stacking, facial expression recognition, kappa-error diagram

This work is partially supported by the Chinese Natural Science Foundation under Grants Nos.61070117, 61171169, and 61175115, the Beijing Natural Science Foundation under Grant Nos.4122004, 4132013 and 4102013, Specialized Research Fund for the Doctoral Program of Higher Education (20121103110031), and the Importation and Development of High-Caliber Talents Project of Beijing Municipal Institutions.

http://dx.doi.org/10.3837/tiis.2014.01.012

Copyright ⓒ 2014 KSII

1. Introduction

Facial expression is a primary means of conveying social information between humans, and is putatively independent of race, gender and age [1]. The good facial expression recognition is to achieve similar levels of effectiveness for Human-Computer Interaction (HCI), which is most effective when it's face-to-face between natural human beings. So, facial expression plays an important role in interpersonal communication and is explored using techniques from pattern recognition, computer vision, Psychology and Linguistics.

In 1971, Ekman and Friesen had proposed 6 basic facial expressions [2], being anger, disgust, fear, happiness, sadness and surprise which can be viewed as a K-class classification problem with K=6 (or 7 if Neutral is included). Most researchers classify facial expression based on the above K-class. Since then a lot of effort has been made to build more reliable facial expression recognition. Ekman et al. proposed FACS in 1978 and revised it in 2002 [3]. In 1997, Lanitis et al. used the active appearance models (AAM) to interpret the face images [4].

Zhang et al. used dynamic Bayesian network with the FACS (Facial Action Coding System) and realized real-time recognition facial expression substantially [5]. Shan et al. used SVM (Support Vector Machine) with Boosted-LBP (Local Binary Patterns) feature about 7-class facial expression recognition, and obtained the highest accuracy 97.5% for happiness and disgust respectively and the lowest accuracy 74.7% for sadness [6]. Peng Yang et al. divided face image into local patches according to AUs (Action Unites) and extract appearance feature from each patch, they experiment on Cohn-Kanade database by using Adaboost and obtained accuracy 92.3% on the testing set and 80.0% on the extended testing set [7].

Recently, many researchers have applied ensemble techniques that fuse the results of multiple classifiers instead of using just a single classifier. Bartlett et al. used Adaboost and SVM to get 89.1% on Cohn-Kanade database in Exp. II [8]. Sander Koelstra proposed a dynamic texture-based approach to the recognition of facial Action Units and their temporal models by using GentleBoost ensemble algorithm with Hidden Markov Model. This work tested on Cohn-Kanade database and MMI database, and obtained the highest accuracy 95.80% for AU27 and the lowest accuracy 71.33% for AU7 on the Cohn-Kanade database [9]. Thiago et al. used ensemble classifiers to recognize facial expression with Gabor and LBP, and got 88.9% on Cohn-Kanade database in the Exp. II [10]. These ensemble approaches aim to improve the classification results by integrating the several classification results obtaining on the partial selected datasets by a certain strategy. The multi-classifiers are normally same types and integrated with boosting integrating strategy, so its classification results are still determined largely by the performance of the kernel classifier. But the classifiers with the different mechanics display the discriminatory performances on the facial expression recognition under different cases, such as datasets adopted, features used. Therefore, integrating the contribution of several classifiers to improve overall classification results is one possible solution. In this paper, we aim to explore the possible solution of the multi-classifier fusion to improve overall classification results.

Considering stacking [11, 12] is an advanced form of ensemble classifier, which seeks to learn the best way of fusing several classifiers to optimize its classification performance, we propose a new emotion recognition system based on stacking in this paper. Whilst, we propose and introduce an approach of base classifiers selection referring to the achievement of

Kuncheva on algorithm evaluation with trading off the recognition error and the algorithm diversity[13]. Comprehensive comparison experiments are done in this paper to test the performance of our proposed stacking ensemble facial expression recognition system.

The rest of this paper is organized as following. Section 2 introduces the principle of our stacking ensemble emotion method. Section 3 discusses the selection way of the base classifiers based on kappa-error diagram. The tests on the two public databases JAFFE and Cohn-Kanade are demonstrated in section 4, with a detail analysis of the results and comprehensive comparison to existing methods. The summary of our present work and discussion of the future work are given in the final section.

2. Principle of Stacking Ensemble Expression Recognition Approach

Stacking is a technique to fuse multiple classifiers applied to a specific classification problem [14], and aims to improve the results of individual classifier. It outperforms the other methods to fuse multi-classifiers by simply voting or linear combination, which integrates the function of individual classifiers in expression recognition through the sample training. Although the ensemble techniques using a fixed rule such as with a simple majority voting rule is unnecessary to train with additional training data, the one using a trained rule which characterizes stacking, is potentially able to obtain a better classification result [15]. Therefore, we propose to employ stacked ensemble in face expression recognition to take full advantage of each individual classifier and obtain the better understanding of emotion by face.

The stacked fusion system is illustrated in **Fig. 1**. The lower level in **Fig. 1** is called base-level which processes the input respectively with several base classifiers. The upper level in **Fig. 1** is called meta-level which stacking relearns the results of base classifiers with using this additional level of classification, the so-called meta-classifier. The detail procedure of stacking is illustrated as follows.

Supposing there are *n* base classifiers marked as $F_1, F_2, ..., F_n$, one meta-classifier marked as M, and *m* classes marked as $C_1, C_2, ..., C_m$. For each sample S, it will be processed by following procedure.

Stacking Procedure	
Step1	i=1.
Step2	If i<=n, go to Step3. Otherwise, go to Step4.
Step3	Do classification with the base classifier F_i. For each sample S, the probability vector $\boldsymbol{P_i} = \{P_{i1}, ..., P_{ij}, ..., P_{im}\}$ under the base classifier F_i is derived, where P_{ij} indicating the probability that sample S is assigned to class C_j (j=1,2, …, m). Then go to Step2.
Step4	The classification results of all base classifiers are obtained, here marked as $\boldsymbol{P} = \{\boldsymbol{P}_1, \ldots, \boldsymbol{P}_n\}$. Then go to Step5.
Step5	The meta-classifier M processes the input data *viz.* the matrix $\boldsymbol{P}$ from base classifiers and outputs the ultimate recognition result.

Copyright ⓒ 2014 KSII

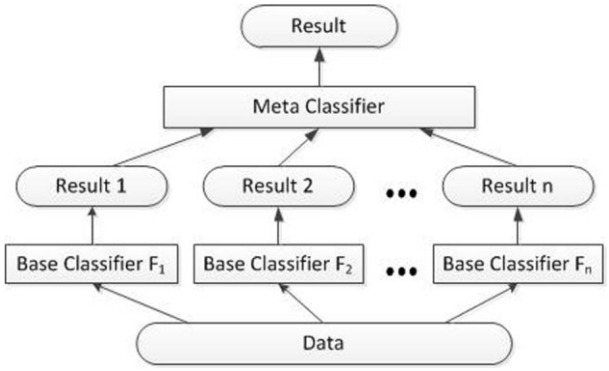

Fig. 1. Fusion system based on stacking

Note that the meta-classifier sees only the probabilities estimates for each classifier and class, across the set of fusion samples. And separate data partitions should be used for training the base classifiers, validating the base classifiers to train the meta-classifier, and testing the combined classifier. Usually this is done using cross-validation given the increased data requirements implied by the additional data partitions.

3. Selection of Classifiers According to Kappa-error Diagram

In the field of facial expression recognition, the ability of recognition system depends strongly on the classifiers selected as well as the features used. C.Shan et al. used SVM and Boosted-LBP in Ref [6]. Koutlas et al. applied ANN (Artificial Neural Network) and Gabor filters in Ref [16]. Zhang et al. adopted Dynamic Bayesian network to track run-time emotion [5]. Xu et al. employed KNN (K Nearest Neighbor) in Ref [17]. According to the achievements of the present research, the typical classification techniques: KNN, SVM, ANN and Bayesian, all have achieved a fairly good outcome under a certain context. However, these algorithms realize the classification based on very different principles. For example, KNN is based on minimizing risk, the realization of ANN depends on associative memory, the principle of SVM is maximum interval, and Bayesian is based on posterior probability. So they perform variously under the different cases. Fusing their complimentary functions to improve the effectiveness and robustness of recognition system is one of effective solution. So we explore the feasibility of integration from the above typical classification algorithm.

To make the fusion effective, the diversity between the classifiers is one of key points. Kuncheva pointed out that there are two main factors for successful fusion - individual accuracy and pairwise diversity. She proposes the bound indicating possible ensemble with trading two factors off through mathematic proof and large experiments by analyzing classifier ensemble performance of every two technique pair [18]. Referring to those public conclusions, this paper adopts the kappa-error diagram and bound conclusion in our base-level classifier selection for stacking ensemble face recognition approach.

Table 1. Contingency Table of Two Classifiers

	F_1 Correct	F_1 Wrong
F_2 Correct	*a*	*b*
F_2 Wrong	*c*	*d*

3.1 Theory of Classifier Ensemble Prune Using Kappa-error Diagram

Kappa-error diagram is a popular tool for analyze ensemble methods proposed by Margineantu and Dietterich [13]. Kappa-error diagrams visualize individual accuracy and diversity in a 2D plot, and have been used to decide which ensemble members can be pruned without much harm to the overall performance [13, 18]. The common kappa and error of two classifiers underestimated are computed as following way. Suppose F_1 and F_2 are a pair of classifiers underestimated. On a dataset, each classifier is applied to do the classification respectively. The corresponding pairwise contingency table is counted shown in **Table 1**. In the table, parameter '*a*' represents the number of samples both classifiers doing the right classification, '*b*' and '*c*' are the number of samples one classifier right and another one is wrong and '*d*' points the number of samples both classifiers wrong Then error '*e*' and kappa '*kappa*' values are computed as Eq. (1) and Eq. (2)[18].

$$e = \frac{1}{2}\left(\frac{c+d}{N} + \frac{b+d}{N}\right) \tag{1}$$

$$kappa = \frac{OA - AC}{1 - AC} \tag{2}$$

Where N is the number of samples in the dataset, that is $N = a+b+c+d$. OA and AC are computed as following Eq. (3) and Eq. (4).

$$OA = \frac{a+d}{N} \tag{3}$$

$$AC = \frac{(a+b)(a+c)+(b+d)(c+d)}{N^2} \tag{4}$$

Here 'OA' represents the average sample number that both classifiers having same classification results. So it predicts the coherence extent of two classifiers. With higher coherence, the functions of two classifiers are close and suitable to be pruned without much harm of ensemble. On the contrary, it indicates the functions are various and proper to be kept to improve the whole ensemble results with complementary contribution. Kappa is actually the derived parameter, which removes the chance factor by minus 'AC', the average of numbers of both right and both wrong. Kappa provides more objective criteria than accuracy directly. So we could conclude that the lower kappa in this definition indicates the higher diversity.

Error in Eq. (1) is the average of number of samples that each classifier doing the wrong classification. It is easy to understand that classifier with higher error will reduce the entire function of the ensemble.

Absolutely, high diversity and high accuracy are what we want in determining the base classifiers for stacking ensemble system.

Further, Kuncheva examines the bound on the region for the dichotomous case where feasible kappa-error tradeoffs are found. The paper derives bounds $k_{\min}$ on kappa in terms of the error '*e*', as in Eq. (5). The pairwise closes the bound has good performance benefit for fusion [18].

$$k_{\min} = 1 - \frac{1}{1-e}, 0 < e < 0.5 \tag{5}$$

3. 2 Base Classifiers Determination

To determine the final base classifiers from the above four typical classifiers, viz KNN, SVM, ANN and Bayesian, we made full analysis fusion performance of the each classifier pairwise. Criteria of fusion performance proposed by Kuncheva, which bases on the kappa-error

Copyright ⓒ 2014 KSII

diagram, are employed in the paper as the evaluation rule. To make the results more generally, we made the analysis on the two public databases and utilized several different features.

3.2.1 Databases

In this paper, we use two public databases: the JAFFE and the Cohn-Kanade. The JAFFE database contains 183 images from 10 different Japanese women. The Cohn-Kanade database contains 355 samples from 97 subjects. Each sample includes sequences of frames from movies of the subjects in making various expressions, we use it to test the fusion performance with input data representing in dynamic features.

3.2.2 Solution of Kappa-error Diagram Based Base Classifier Determination

The facial expression images in two public databases are preprocessed being represented in different features respectively. Here we used the Gabor feature, static geometric feature and dynamic feature. The four classifiers are used separately in expression recognition. Because the base-level classifiers should be as simple as possible, the typical algorithms: 1-NN (1-nearest neighboring), SMO (Sequential Minimal Optimization) [19], MLP (Multilayer Perceptron) and NB (Naïve Bayes) are chosen for the four classifiers mentioned above. The 6 pairs of classifiers between each other in 4 classifiers are counted with the contingency table in **Table 1**. After counting the recognition results of each pair of classifier, error and kappa about each classifier pairwise are computed according to Eq. (1) and Eq. (2). The values of pairwise kappa and pairwise error among 1-NN, SMO, NB and MLP are shown on **Table 2**. The corresponding kappa-error diagram is shown in **Fig. 2**.

To determine the base classifiers for stacking ensemble expression recognition, we do the analysis from two angels according to ensemble prune theory. Referring to Kuncheva's experiment conclusion, we first evaluate the accuracy of classifier in facial expression recognition which is the leading factor for the fusion success. The bad performance of individual classifier will cause the catastrophic fusion, so we remove the corresponding classifiers from ensemble directly. Then we analyze the pairwise diversity, especially taking the Kuncheva's bound as reference to decide the performance of classifier according to if classifier pairs are closer to the bound curve k in kappa-error diagram. The classifier pairwise will be fare better for ensemble than that far away.

Base on the experiment results in **Table 2**, we make analysis of the error first. We could find that the pairwise error of SMO and MLP is smallest in the four different cases with different feature or in different databases. This indicates SMO and MLP plays comparatively well in facial expression recognition and can be considered to be adopted. The similar results are displayed visually in **Fig. 2**, where the points of MLP-SMO pairwise classifiers marked with the rectangle symbol '□'.

From the perspective of diversity criteria, the pairwise classifiers of MLP and 1-NN in three out of four cases shown in **Table 2** have lowest kappa. They are also the points marking with the small star symbol '*' in **Fig. 2**, which are closer to the bound having better ensemble performance with tradeoff of lower pairwise error and diversity.

On the other hand, three out of four of pairwise: 1-NN and NB has highest error shown corresponding column in **Table 2**. Actually, analyzing the NB in facial expression recognition, its performance drops because "High dimensionality and small size samples" is widely encountered in facial expression recognition, whilst the Bayes works well depending on the training with big samples. So avoiding the catastrophic fusion, we don't count the NB as base classifier.

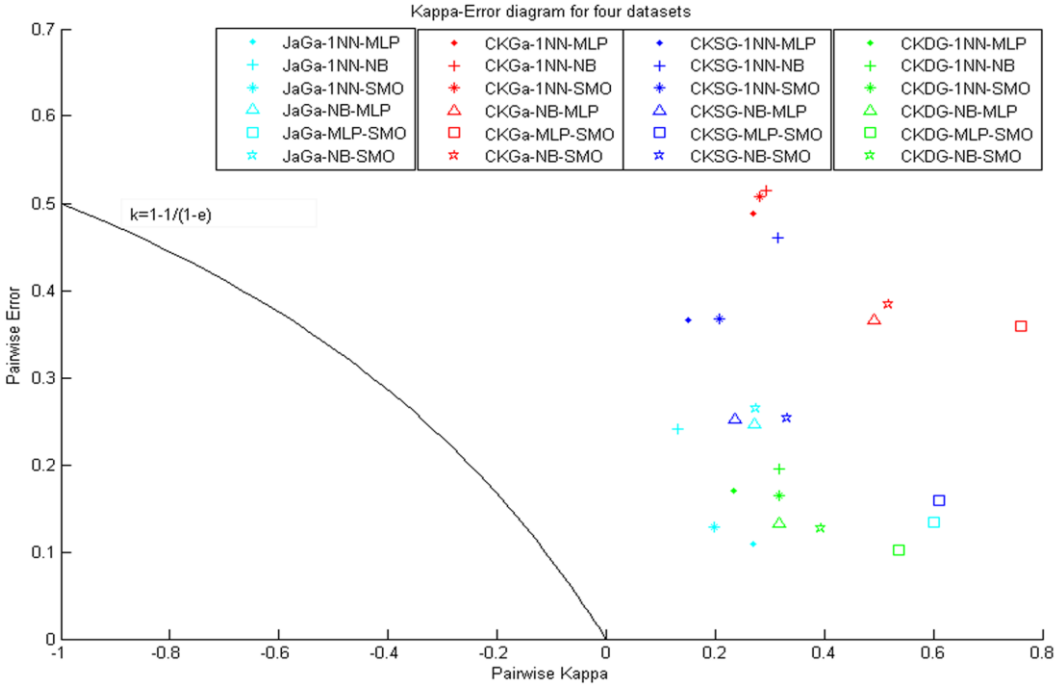

Fig. 2. Kappa-error diagram about six pairs of classifiers. The points obtained by experimenting on Gabor feature of JAFFE are shown cyan, the points obtained by experimenting on Gabor feature of Cohn-Kanade are shown red, the points obtained by experimenting on static geometry feature of Cohn-Kanade are shown blue, and the points obtained by experimenting on dynamic geometry feature of Cohn-Kanade are shown green.

Table 2. Information of Six Pairwise Classifiers

		1-NN-MLP	1-NN-NB	1-NN-SMO	NB-MLP	MLP-SMO	NB-SMO
JAFFE-Gabor Feature	*error*	0.1093	0.2404	0.1284	0.2459	0.1339	0.2650
	kappa	0.2705	0.1313	0.1988	0.2719	0.6007	0.2752
CK-Gabor Feature	*error*	0.4887	0.5141	0.5070	0.3662	0.3592	0.3845
	kappa	0.2692	0.2926	0.2819	0.4917	0.7617	0.5180
CK-Static Geometry Feature	*error*	0.3662	0.4606	0.3676	0.2521	0.1592	0.2535
	kappa	0.1515	0.3159	0.2071	0.2368	0.6106	0.3310
CK-Dynamic Geometry Feature	*error*	0.1704	0.1958	0.1648	0.1324	0.1014	0.1268
	kappa	0.2347	0.3175	0.3175	0.3172	0.5365	0.3945

According to the above analysis, 1-NN, MLP and SMO are selected as candidates of base classifiers for stacking ensemble. In making the final determination, the performance of the pairs of classifiers SMO and the related MLP versus the unrelated 1-NN are considered. We can find from the **Table 2**, the error and kappa of those pairs are located at the middle with moderate average error and lowish kappa. From this perspective, fusion of SMO and 1-NN wouldn't cause unacceptable error and their diversity is also big enough for the ensemble to enhance expression recognition. Therefore MLP, 1-NN and the SMO implementation of SVM are selected as base classifiers in the paper.

We also make some further explanation in relation to some patricular points in **Fig. 2**. Because we don't use elaborate preprocessing filters, such as illumination normalization, all points in red in **Fig. 2** are far from the bound curve. Here, all these points labeling in red

Copyright ⓒ 2014 KSII

symbols represent the performance of all pairwise classifiers based on the Gabor feature in Cohn-Kanade database. It is usually caused by other factors such as unstable of illumination. We prefer to make comparative comparison across all filters, features and databases.

3. 3 Determining of Meta-classifier

The role of meta-classifier is to fuse each individual classifier's recognition results to obtain more robust decision by retraining approach. The input of the meta-classifier is each individual classifier's results and output is fusion recognition results of six basic expressions. We can't have clear idea about their distribution to determine discriminate form. As the sample size in our system is not large, we consider SMO as our meta-classifier. SMO is able to provides a solution for small sample problem and training dataset. It has the good ability to solve the non-linear classification problem by transform into the high dimension feature space to construct the hyper-plane to do the linear classification in hyper-space.

4. Experiments and Analysis

To verify the robustness and effectiveness of stacking fusion strategy, we did experiments on two public databases and with three different expression features. Full comparison was done among our proposed stacked ensemble system with KNN, MLP and SMO. To get the further understanding of stacking-based multi-classifier ensemble in facial expression recognition system, further comparison were made among the results between stacking based approach and vote-based [20] and bagging-based [21] ensemble methods respectively.

4.1 Preprocess of Facial Expression Images in Databases

The evaluation of our proposed multi-classifier fusion approach is done on the public databases JAFFE with 183 samples and Cohn-Kanade with 355 samples. Samples in JAFFE are independent static expression images, while samples in Cohn-Kanade database are a series of frames of six basic expressions. In every expression sequence, the first frame is normally neutral state and the last frame is the extreme emotional state when making expression by face, which we called peak frame. So we use this database to extract the dynamic feature to verify our stacking fusion strategy. The images in two databases are preprocessed to extract the corresponding features respectively as follows.

For images in JAFFE database are only static facial expression images, we only extract the holistic feature Gabor as the representation. First, we use the Viola-Jones face detection algorithm to locate and partition the rectangular area of face part, as shown **Fig. 3(a)**. Then resize the face images to 80×100 pixels. Second, process the resized image with the Gabor wavelet filter. To get the low dimension feature, PCA is utilized to reduce dimension. Here we select the first 300 dimensions as features. In below experiments, we marked the preprocessed result as "JAFFE-Gabor Feature".

On Cohn-Kanade database, we used AAM to do the point location, where we select 70 points referring to the FDP in MPEG-4 on the peak frame, as shown in **Fig. 3(b)**. Based on the point location, the face part is segmented as shown in **Fig. 3(c)** and resize into 90×96. Similar as the "JAFFE-Gabor Feature", we used Gabor and PCA to extract features for a total of 300 dimensions. The result was named "CK-Gabor Feature".

(a) (b) (c) (d)

Fig. 3. (a) Face sub-image gotten by using Haar-wavelet. (b) 70 key points on the peak frame. (c) Normalization face. (d) 47 key points used as geometry feature.

Based on the extracted 70 key points on the Cohn-Kanade database, we generate a kind of geometric feature. Considering the areas of mouth, nose and eyes contribute most information to facial expression [22], we select 46 key points from above 70 points around these areas as feature points. In addition, we select the point of the lower jaw to keep global information as well. So, there are 47 feature points, as shown **Fig. 3(d)**, which are selected as key points for geometric features. To alleviate the influence of head movement, we adjust the coordinate of the above key points by taking the point at the tip of the nose as reference to do the normalization. Then the location of these 47 points are concatenated as the static geometry features, which were called "CK-Static Geometry Feature" in following experiments, with a total of 94 dimensions.

Considering the Cohn-Kanade database contains multi-frame sequences and dynamic information, we sought to take the dynamic features into account. Dynamic feature is represented by computing the difference between locations of above 47 key points of the first frame (the neutral expression) and that of the last frame (the peak expression). The derived dynamic geometry features were called "CK-Dynamic Geometry Feature" in following experiments, with a total of 94 dimensions.

4.2 Evaluation Metric

Accuracy is a usual performance measure criterion for evaluating the classifier's effectiveness. However, it's highly likely that classifier will get a significant proportion of its classification correct by chance. Therefore, to get more objective and comprehensive knowledge of classifier's performance, we adopt three common evaluation criteria, which are Cohen's kappa and informedness besides accuracy.

Table 3. The Statistic Result for Recognition

predict / original	C_1	C_2	C_3	C_4	C_5	C_6	*Sum*
C_1	a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a_{16}	X_1
C_2	a_{21}	a_{22}	a_{23}	a_{24}	a_{25}	a_{26}	X_2
C_3	a_{31}	a_{32}	a_{33}	a_{34}	a_{35}	a_{36}	X_3
C_4	a_{41}	a_{42}	a_{43}	a_{44}	a_{45}	a_{46}	X_4
C_5	a_{51}	a_{52}	a_{53}	a_{54}	a_{55}	a_{56}	X_5
C_6	a_{61}	a_{62}	a_{63}	a_{64}	a_{65}	a_{66}	X_6
Sum	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	N

Copyright ⓒ 2014 KSII

For convenience to introduce accuracy, kappa and informedness, the paper gives **Table 3** which shows the statistic result for recognition. C_i represents one expression set(6 basic expressions in all used in the paper), N is the amount of total samples in a database and a_{ij} represents the amount of samples, which belong to C_i expression set and was classified to C_j expression set. X_i, Y_i ($X_i = \sum_{i=1}^{6} a_{ij}$, $Y_i = \sum_{j=1}^{6} a_{ij}$) are derived values indicating the total amounts of sample in original and predicting set respectively of each expression.

Accuracy is obtained as Eq. (6), indicating the proportion of right prediction amount from total samples.

$$accuracy = \sum_{i=1}^{6} a_{ii} / N \tag{6}$$

Cohen's kappa [23] is a more conservative metric since it cancels off the chance component (and renormalizes to the form of a probability). Eq. (7) gives the method to calculate kappa.

$$kappa = \frac{P_o - P_c}{1 - P_c} \tag{7}$$

Where P_o is the observed probability, and P_c is the hypothetical probability of chance. Further, $p_o = \sum_{i=1}^{6} a_{ii} / N$ and $p_c = \sum_{i=1}^{6} X_i Y_i / N^2$.

The last metric called Informedness [24] which corresponds to the probability that you are making an informed decision versus guessing. Informedness is calculated by Eq. (8).

$$Informedness = \frac{winloss}{N} \tag{8}$$

Where $winloss = \sum_{i \neq j} \left(a_{ij} \bullet bias[j] / (prev[j] - 1) \right) + \sum_{i=j} \left(a_{ij} \bullet bias[j] / prev[j] \right)$, $prev[i] = X_i / N$, $bias[i] = Y_i / N$.

4.3 Experiment Results and Analysis

After preprocessing of the facial expression images in JAFFE and Cohn-Kanade databases respectively, we train and test our stacking fusion expression recognition approach by using 10-fold cross-validation (CV). This effectively makes up the shortage of insufficient samples with running classifiers 10 times in round by partitioning the sample into 10 with one "fold" containing 10% of the data reserved for testing and the remainder of the data (90% data) used for training the classifiers in each cycle.

In order to have further understanding of our approach's performance, we make comprehensive comparison with the several individual classifiers and other common multi-classifier fusion methods: vote and bagging which are used into the expression recognition on the same samples. The comparisons with the existing facial expression recognition are also given in the paper.

4.3.1 Comparison with Each Individual Base Classifier

To verify if the fusion approach improves the recognition result comparing with the individual classifiers, we use the three base classifiers and our proposed stacking fusion approach to do the facial expression recognition respectively. The results are counted up by computing the corresponding performance evaluation metric: accuracy, kappa and informedness in two databases and with different features. The results of the experiments are shown in **Table 4**. The best recognition result is emphasized with writing values in boldface. The numerals in brackets represent the standard error (SE) of the mean.

Table 4. Evaluation of Stacking Ensemble Approach Comparing with a Single Base Classifier

	Classifier	Accuracy	Kappa	Informedness
JAFFE-Gabor Feature	MLP	88.45(±0.36)	86.12(±0.43)	89.11(±0.36)
	SMO	84.65(±0.69)	81.56(±0.82)	84.90(±0.69)
	1-NN	89.56(±0.45)	87.47(±0.54)	89.26(±0.46)
	Stacking	**92.31**(±0.44)	**90.76**(±0.53)	**91.95**(±0.50)
CK-Gabor Feature	MLP	65.91(±0.42)	58.36(±0.51)	61.89(±0.52)
	SMO	62.27(±0.51)	54.11(±0.62)	56.71(±0.59)
	1-NN	36.35(±0.41)	21.83(±0.49)	25.84(±0.67)
	Stacking	**67.04**(±0.36)	**59.60**(±0.44)	**64.34**(±0.43)
CK-Static Geometry Feature	MLP	84.19(±0.29)	80.70(±0.36)	83.99(±0.30)
	SMO	83.92(±0.36)	80.40(±0.44)	83.33(±0.42)
	1-NN	42.53(±0.33)	30.36(±0.39)	30.36(±0.55)
	Stacking	**85.32**(±0.34)	**82.05**(±0.42)	**84.99**(±0.36)
CK-Dynamic Geometry Feature	MLP	89.31(±0.18)	86.98(±0.22)	89.52(±0.17)
	SMO	90.40(±0.24)	88.26(±0.30)	90.29(±0.24)
	1-NN	76.62(±0.32)	71.47(±0.39)	75.63(±0.43)
	Stacking	**90.68**(±0.19)	**88.64**(±0.23)	**90.78**(±0.17)

From the experiment results, we can find our proposed stacking fusion expression recognition approach has the best performance in all different cases and with different evaluation metric. Here the three evaluation parameters show the similar rules of each classification approach and provide the further proof of reliability of our experiment results. Now we neglect the fusion results and only analyze the individual performance of each classifier first, we could find their performance varies in different databases and features. As **Table 4** shows, 1-NN displays the best performance in the first case, MLP is the best one in the second and third cases, and SMO outperforms the other two in the last cases. By fusion their complementary contribution in different cases, the stacked ensemble system outperforms the individual classifiers. Analyzing the SE of the mean shown in brackets in **Table 4**, we could find that SE of stacking approach displays good performance being or approximate to lowest SE. It verifies the effectiveness of our proposed stacking fusion facial expression recognition approach again from another perspective.

Nevertheless, we still find that stacking ensemble system is higher but not surpasses the best single classifier very much according to the each evaluation value under different cases as in **Table 4**. For example, all the classifiers perform well in the JAFFE database. Using the dynamic feature, the recognition results are better than the corresponding ones in the static feature. So, performance of multi-classifier ensemble system relies on individual classifiers. It is important to select individual classifiers which have good individual performance and diverse contribution as base classifiers to improve the fusion result.

4.3.2 Comparison with Bagging and Vote Fusion Approaches

Bagging, that is bootstrap aggregating, involves having each model in the ensemble vote with equal weight. In order to promote model variance, bagging trains each model in the ensemble using a randomly drawn subset of the training set. According to the existing study, bagging with a decision tree is able to achieve good classification accuracy [25]. So, our experiments adopt REPTree as one of base classifiers for bagging to do the comparison besides base

Copyright ⓒ 2014 KSII

classifiers: 1-NN, SMO and MLP, which are the base classifiers for our stacking fusion approach. **Table 5** shows the experiment results about bagging with different classifiers under corresponding cases. We could find in the Cohn-Kanade database, the bagging with SMO as the base classifier has best recognition results and the bagging with 1-NN displays best performance in JAFFE database.

Vote is another ensemble method which fuses several classifiers by voting algorithms. In our experiments we use four common voting combining rules respectively, viz: "average probabilities", "product probabilities", "maximum probability" and "majority voting". The results of vote which fuses 1-NN, SMO and MLP are shown in **Table 6**. It shows that majority based voting rule outperforms the others.

Based on the above results from **Table 5** and **Table 6**, we compare our stacking approach with the best ones under each case. As shown in **Table 7**, our stacking fusion approach outperforms the others under the first three cases. Although the value is slightly lower than that of vote (majority voting), the gap is less than 0.35% under the last case. It demonstrates that dynamic information plays important roles in representing the emotion expressing. Most classifiers can play good recognition results with the dynamic feature. So the majority displays good performance. However, the performance of our approach still approximates the best one under this case. In general, our proposed approach shows robustness performance in facial expression recognition with several different features and in two different common databases.

Table 5. Recognition of Bagging

		Bagging (1-NN)	Bagging (SMO)	Bagging (MLP)	Bagging (REPTree)
JAFFE-Gabor Feature	Accuracy	**89.56**	84.61	39.57	64.36
	Kappa	**87.47**	81.54	26.90	57.22
	Informedness	**89.26**	84.47	36.50	65.09
CK-Gabor Feature	Accuracy	38.03	**64.49**	44.25	54.06
	Kappa	23.99	**56.63**	31.97	43.35
	Informedness	28.33	**60.51**	36.81	52.25
CK-Static Geometry Feature	Accuracy	44.19	**83.33**	18.11	72.93
	Kappa	32.37	**79.66**	0.00	67.06
	Informedness	33.10	**82.94**	0.00	70.28
CK-Dynamic Geometry Feature	Accuracy	77.18	**90.12**	17.46	85.62
	Kappa	72.13	**87.93**	5.77	82.47
	Informedness	77.05	**90.26**	7.51	85.57

Table 6. Recognition of Vote

		Vote (average probabilities)	Vote (product probabilities)	Vote (maximum probability)	Vote (majority voting)
JAFFE-Gabor Feature	Accuracy	91.11	90.09	90.58	**91.70**
	Kappa	89.33	88.09	88.70	**90.03**
	Informedness	90.89	90.06	90.33	**91.73**
CK-Gabor Feature	Accuracy	56.30	53.48	45.88	**63.64**
	Kappa	46.38	42.88	33.36	**55.55**
	Informedness	50.77	47.78	38.58	**59.53**
CK-Static Geometry Feature	Accuracy	69.27	63.91	56.60	**82.79**
	Kappa	62.54	56.13	47.22	**78.98**
	Informedness	66.59	58.21	49.32	**82.73**
CK-Dynamic Geometry Feature	Accuracy	86.17	83.92	82.52	**90.97**
	Kappa	83.15	80.41	78.70	**88.98**
	Informedness	85.87	85.52	81.55	**90.86**

Table 7. Recognition of Ensemble Method

		Accuracy	Kappa	Informedness
JAFFE-Gabor Feature	Vote (majority voting)	91.70	90.03	91.73
	Bagging(1-NN)	89.56	87.47	89.26
	Stacking	**92.31**	**90.76**	**91.95**
CK-Gabor Feature	Vote (majority voting)	63.64	55.55	59.35
	Bagging(SMO)	64.49	56.63	60.51
	Stacking	**67.04**	**59.60**	**64.34**
CK-Static Geometry Feature	Vote (majority voting)	82.79	78.98	82.73
	Bagging(SMO)	83.33	79.66	82.94
	Stacking	**85.32**	**82.05**	**84.99**
CK-Dynamic Geometry Feature	Vote (majority voting)	**90.97**	**88.98**	**90.86**
	Bagging(SMO)	90.12	87.93	90.26
	Stacking	90.68	88.64	90.78

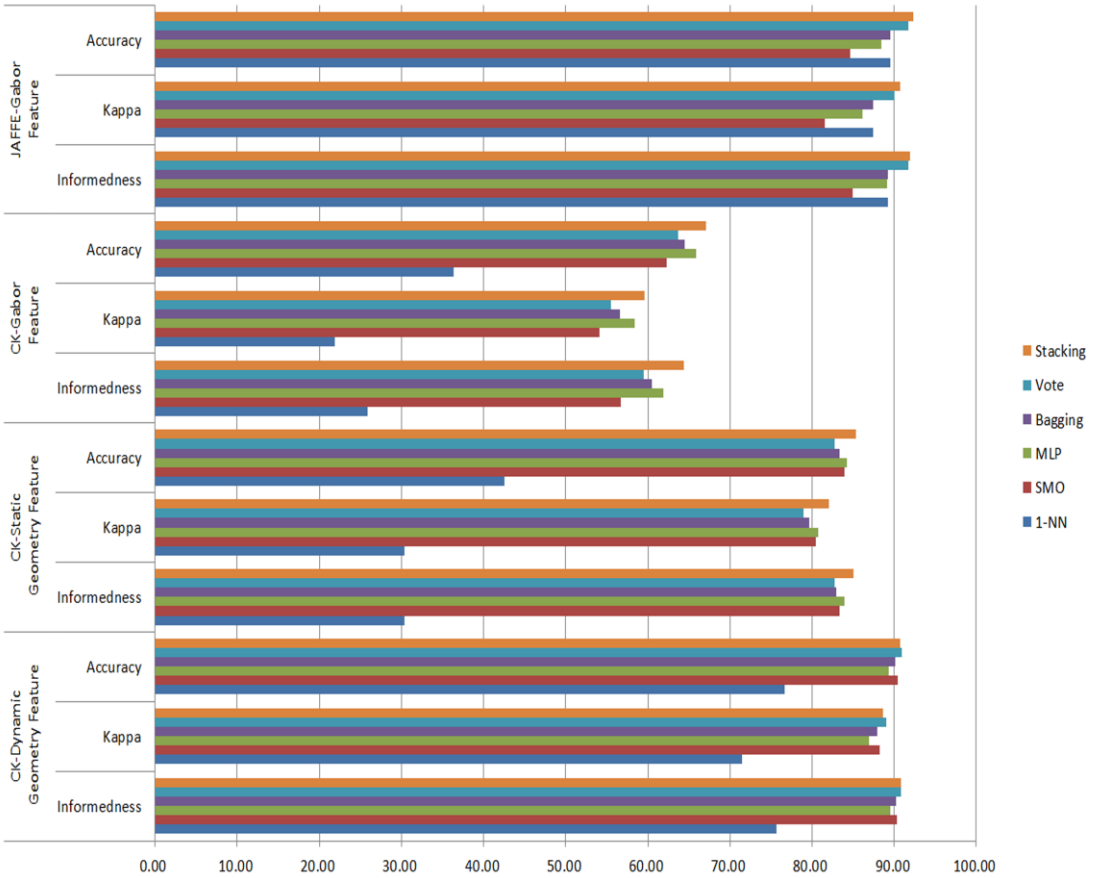

Fig. 4. Comparison among all the methods

To give a clearer overall idea of the above comparison, we draw the bar graph in **Fig. 4**. Obviously, stacking outperforms the others including both the individual and the ensemble classification approaches for both the standard databases using either Gabor or simple geometric features. Although the evaluation values appear slightly lower for the CK database

Copyright ⓒ 2014 KSII

with the dynamic geometric features, stacking is not significantly worse than the best one. Overall, our proposed stacking approach provides stable facial expression recognition.

4.4 Comparison with Existing Expression Recognition Methods

We compared our method with existing works that using either the Cohn-Kanade or the JAFFE database with recognition accuracy shown in **Table 8** and **Table 9**. Note, all the results come from the original published papers. Compared with the results available, our proposed stacking ensemble fusion expression recognition method achieve relatively good results achieving the highest recognition accuracy with the dynamic geometry feature in CK database in **Table 8** and Gabor feature in JAFFE database in **Table 9**. It is noted that we do not lay emphasis on the feature extraction, but focus on the multi-classifier integration by using the complementary contribution to improve the overall effectiveness. So, the performance could potentially be improved with more discriminative features and normalizing for factors such as unstable illumination.

5. Conclusion and Future Work

Classifier plays important roles in facial expression recognition, but individual classifier shows some extent of bias in different databases or with different representation as features. To overcome this shortage, stacking ensemble system is employed in our paper to integrate the performance of multi-classifiers. We especially propose to use the kappa-error diagram in selection of base classifiers from frequently-used classifiers with different mechanics. The experiment results show that the stacking always outperforms the others either in different databases or with different features. Our proposed ensemble stacking overcomes the bias of individual classifiers. Using the learning way to estimate and weight the contribution of base classifiers outperforms the voting-based multi-classifier fusion algorithm and bagging ensemble method in most cases.

In this paper, we used only one kind of feature respectively on each case. However, several kinds of feature may perform better to describe expression. So, to fuse different sets of features – in particular the static and dynamic features of Cohn-Kanade, and potentially classifiers on the cross product of features space and classifier choice is the future work.

Table 8. Comparison with several existing Methods on Cohn-Kanade Database

Reference	Accuracy (%)	Method
Cohen et al. (2003)[26]	73.2	Geometric feature + Tree-augmented-NB
Shan et al. (2005)[27]	88.4	LBP+SVM
Bartlett et al. (Exp. II) (2005)[8]	89.1	Gabor filter +Adaboost + SVM
Shan et al. (2009)[6]	88.9	Boosted-LBP + SVM
Thiago et al. (Exp. II) (2013)[10]	88.9	Ensemble based on Gabor and LBP
Our proposed approach	90.68	Dynamic geometry feature + Stacking

Table 9. Comparison with several existing Methods on JAFFE Database

Reference	Accuracy (%)	Method
Bashyal et al.(2008)[28]	90.2	Gabor filters + LVQ
Koutlas et al. (2008)[16]	92.3	Gabor filters + Artificial Neural Networks
Yu et al. (2013)[29]	85.7	WLD + Pool-based active learning with SVM
Our proposed approach	92.31	Gabor filters + Stacking

References

[1] Tariq, Usman, and Thomas S. Huang, "Features and fusion for expression recognition—A comparative analysis," in *Proc. of IEEE Computer Society Conf. on Computer Vision and Pattern Recognition Workshops*, pp.146-152, June 16-21, 2012. Article (CrossRef Link).

[2] Ekman, Paul, and Wallace V. Friesen, "Constants across cultures in the face and emotion," *Journal of personality and social psychology*, vol. 17, no.2, pp.124-129, February, 1971. Article (CrossRef Link).

[3] P. Ekman, W. Friesen, and J. Hager, "The Facial Action Coding System: The Manual on CD ROM," *A Human Face*, Salt Lake City, 2002. Article (CrossRef Link).

[4] Lanitis, Andreas, Christopher J, Taylor, and Timothy F. Cootes. "Automatic interpretation and coding of face images using flexible models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, no.7, pp. 743-756, July, 1997. Article (CrossRef Link).

[5] Zhang, Yongmian, and Qiang Ji, "Active and dynamic information fusion for facial expression understanding from image sequences," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol.27, no.5, pp.699-714, May, 2005. Article (CrossRef Link).

[6] Shan, Caifeng, Shaogang Gong, and Peter W. McOwan. "Facial expression recognition based on Local Binary Patterns: A comprehensive study," *Image and Vision Computing*, vol.27, no.6, pp. 803-816, May, 2009. Article (CrossRef Link).

[7] Yang P, Liu Q, Metaxas D N, "Exploring facial expressions with compositional features," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp.2638-2644, June 13-18, 2010. Article (CrossRef Link).

[8] Bartlett, M., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., & Movellan, J. "Recognizing facial expression: Machine learning and application to spontaneous behavior," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp.568–573, June 20-25, 2005. Article (CrossRef Link).

[9] Koelstra Sander, Maja Pantic, and Ioannis Patras, "A dynamic texture-based approach to recognition of facial actions and their temporal models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.32, no.11, pp.1940-1954, November, 2010. Article (CrossRef Link).

[10] Thiago H.H. Zavaschi, Alceu S. Britto Jr., Luiz E.S. Oliveira, Alessandro L. Koerich, "Fusion of feature sets and classifiers for facial expression recognition," *Expert Systems with Applications*, vol.40, no.2, pp.646-655, February, 2013. Article (CrossRef Link).

[11] David H. Wolpert, "Stacked generalization," *Neural Networks*, vol.5, no.2, pp.241-259, March, 1992. Article (CrossRef Link).

[12] Sulzmann J N, Fürnkranz J, "Rule Stacking: An approach for compressing an ensemble of rule sets into a single classifier," *Discovery Science*, Springer Berlin Heidelberg, pp.323-334, 2011. Article (CrossRef Link).

[13] D.D. Margineantu and T.G. Dietterich, "Pruning Adaptive Boosting," in *Proc. of 14th Int'l Conf. Machine Learning*, vol.97, pp.211-218, July, 1997. Article (CrossRef Link).

[14] J. Kittler, M. Hatef, R.P.W. Duin, J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.20, no.3, pp.226-239, March, 1998. Article (CrossRef Link).

[15] Li Shoushan, Huang ChuRan, "Chinese Sentiment Classifi-cation Based on Stacking Combination Method," *Journal of Chinese Information Processing*, vol.24, no.5, pp.56-61, November, 2010. Article (CrossRef Link).

[16] Koutlas, A., & Fotiadis, D. I., "An automatic region based methodology for facial expression recognition," in *Proc. of IEEE Conf. on Systems, Man and Cybernetics*, pp. 662–666, October 12-15, 2008. Article (CrossRef Link).

[17] Yong Xu, Qi Zhu, Zizhu Fan, Minna Qiu, Yan Chen, Hong Liu, "Coarse to fine K nearest neighbor classifier, " *Pattern Recognition Letters*, vol.34, no.9, pp.980–986, July, 2013. Article (CrossRef Link).

[18] Ludmila I. Kuncheva, "A Bound on Kappa-Error Diagrams for Analysis of Classifier Ensembles," *IEEE Transaction on Knowledge and Data Engineering*, vol.25, no.3, pp.494-501, March, 2013. Article (CrossRef Link).

Copyright ⓒ 2014 KSII

[19] Platt, John, "*Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines*", April, 1998. Article (CrossRef Link).
[20] Kuncheva L I, "Combining Pattern Classifiers: Methods and Algorithms (Kuncheva, LI; 2004)[book review]," *Neural Networks, IEEE Transactions on*, vol.18, no.3, pp.964-964, May, 2007. Article (CrossRef Link).
[21] Ting, K. M., Witten, I. H., "Stacking Bagged and Dagged Models," in *Proc. of 14th Conf. on Machine Learning*, pp.367-375, March, 1997. Article (CrossRef Link).
[22] Zhong L, Liu Q, Yang P, et al, "Learning active facial patches for expression analysis", in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp.2562-2569, June 16-21, 2012. Article (CrossRef Link).
[23] Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol.20, no.1, pp.37- 46, April, 1960. Article (CrossRef Link).
[24] Powers, D. M. W. "Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation," *Journal of Machine Learning Technologies*, vol.2, no.1, pp.37-63, February, 2011. Article (CrossRef Link).
[25] Breiman, L., "Bagging Predictors," *Machine Learning*, vol.24, no.2, pp.123-140, August, 1996. Article (CrossRef Link).
[26] Cohen, I., Sebe, N., Garg, A., Chen, L., & Huang, T. S. "Facial expression recognition from video sequences: Temporal and static modeling," *Computer Vision and Image Understanding*, vol.91, no.1, pp.160–187, July, 2003. Article (CrossRef Link).
[27] Shan, C., Gong, S., McOwan, P.W., "Robust facial expression recognition using local binary patterns," in *Proc. of IEEE Conf. on Image Processing*, pp.370–373, September 11-14, 2005. Article (CrossRef Link).
[28] Bashyal, S., & Venayagamoorthy, G. K., "Recognition of facial expressions using gabor wavelets and learning vector quantization," *Engineering Applications of Artificial Intelligence*, vol.21, no.7, pp.1056–1064, October, 2008. Article (CrossRef Link).
[29] Kaimin Yu, Zhiyong Wang, Li Zhuo, Jiajun Wang, Zheru Chi, Dagan Feng, "Learning realistic facial expressions from web images," *Pattern Recognition*, vol.46, no.8, pp.2144–2155, August, 2013. Article (CrossRef Link).

Xibin Jia is currently Associate Professor of computer college, Beijing University of Technology. Member of IEEE and CCF. She received the Ph.D. in Computer science and technology from Beijing University of Technology in 2007. Her main research interest is visual information perception, multi-source fusion. She now especially engages in expression recognition and behavior cognition based on multi-information.

Yanhua Zhang is a Master candidate in Beijing University of Technology. Her main research direction is visual facial expression recognition.

David M W Powers is currently Professor of Cognitive and Computer Science, Associate Dean (International) and Director of the Centre of Knowledge and Interaction Technologies, in the School of Computer Science, Engineering and Mathematics, Flinders University, Adelaide, South Australia, as well as Visiting Professor at the Bejing University of Technology, with support from the Importation and Development of High-Caliber Talents Project of Beijing Municipal Institutions.

Humayra Binte Ali is a PhD candidate in Flinders University, South Australia. Her current research broad area is machine learning and pattern recognition. And her research interests are in machine learning, pattern recognition, computer vision, image analysis and 3D image analysis. She had research experience in mobile autonomous ground robotic vehicle (DSTO project) and early heart rate detection using machine learning in different university projects.