An Action Unit based Hierarchical Random Forest Model to Facial Expression Recognition

Jingying Chen[1,2], Mulan Zhang[1], Xianglong Xue[1], Ruyi Xu[1] and Kun Zhang[1,2]
[1]National Engineering Research Center for E-Learning, Central China Normal University, Wuhan, China
[2]Collaborative and Innovative Center for Educational Technology (CICET), Luoyu Street, Wuhan, China
{chenjy, zhk}@mail.ccnu.edu.cn

Keywords: Hierarchical Random Forests, Facial Expression Recognition, Facial Action Unit.

Abstract: Facial expression recognition is important in natural human-computer interaction, research in this direction has made great progress. However, recognition in noisy environments still remains challenging. To improve the efficiency and accuracy of the expression recognition in noisy environments, this paper presents a hierarchical random forest model based on facial action units (AUs). First, an AUs based feature extraction method is proposed to extract facial feature effectively; second, a hierarchical random forest model based on different AU regions is developed to recognize the expressions in a coarse-to-fine way. The experiment results show that the proposed approach has a good performance in different environments.

1 INTRODUCTION

Facial expressions convey a wealth of the emotion and behaviour information in interpersonal communication, hence, facial expression recognition (FER) is important in natural human-computer interaction (Jameel et al., 2015). Many works have been done in the field of FER and a lot of progress has been made (Kahraman, 2016; Li et al., 2013; Cao et al., 2014; Liu et al., 2006; Satiyan et al., 2010), good results have been reported, especially for high quality images. However, the performance could be decreased due to noisy and low resolution images. Hence, continuous efforts should be made to further improve the recognition accuracy for practical use.

FER usually includes two steps, i.e., feature extraction and expression classification. Based on different features, FER can be categorized into geometric-feature based approaches and appearance-feature based approaches. Geometric-feature based approaches use the location of facial feature points, e.g., eye corners, nostrils, mouth corners, or the shape of facial components, e.g., eyes, eyebrows and mouth. Approaches based on geometric features are available in (Kahraman, 2016; Li et al., 2013). These approaches can provide good results relying on accurate feature points and high quality images. Appearance-feature based approaches use facial texture, e.g., Local binary pattern (LBP) (Cao et al., 2014), Gabor features (Liu et al., 2006) and Haar features (Satiyan et al., 2010), which captures the intensity changes associated with different expressions, such as wrinkles, bulges and furrows, these approaches do not rely on accurate feature points and are robust to noisy and low resolution images.

Different classification approaches are proposed according to various applications, e.g. Neural Networks (NN) (Liu et al., 2006), (Satiyan et al., 2010), Convolution neural networks (CNN) (Levi et al., 2015), Support Vector Machine (SVM) (Valstar et al., 2012), (Cao et al., 2014), Adaboost (Bartlett et al., 2006) and Random Forest (RF) (El Meguid et al., 2014). Liu and Wang (Liu et al., 2006) used multiple Gabor features combined with Neural Network to recognize facial expressions, they combined different channels of Gabor filters and provided better performance than the original Gabor feature method. Levi and Hassner (Levi et al., 2015) applied an ensemble of multiple structure CNNs to mapped binary patterns in the wild challenges. Valstar and Pantic classified expressions based on Action Unit using SVM, they acquired good results (Valstar et al., 2012). Cao et.al. used SVM with the feature of LBP to classify six basic expressions from Cohn-Kanade (CK) database and provided good performance (Cao et al., 2014). Bartlett et al. (Bartlett et al., 2006) used a subset of Gabor filters selected by AdaBoost and trained SVM classifiers on the outputs of the selected filters. Their system can detect the facial action units

Chen, J., Zhang, M., Xue, X., Xu, R. and Zhang, K.
An Action Unit based Hierarchical Random Forest Model to Facial Expression Recognition.
DOI: 10.5220/0006274707530760
In *Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2017)*, pages 753-760
ISBN: 978-989-758-222-6

(AU) defined in Facial Action Coding System (FACS) (Ekman et al., 1978) automatically in spontaneous expressions during discourse. EI Meguid and Levine (El Meguid et al., 2014) combined a set of Random Forests paired with SVM labelers to classify multi-view facial expressions. Most of the approaches work well for clean database, however, practical use needs to recognize the facial expressions accurately and efficiently in various environment.

Random forest (RF) is an ensemble classifier that consists of many decision trees. It has been proven to be accurate and robust to solve computer vision problems (El Meguid et al., 2014; Dantone et al., 2012; Minka et al., 1999). Action Units (AUs) represent small visually discernible facial movements, they are independent of any interpretation and can be used as basis for any higher order decision making process including the recognition of basic emotions (Hager et al., 2002). AUs can describe the expression effectively. Hence, in this paper an AUs based hierarchical random forest model is presented to improve the efficiency and accuracy of the expression recognition in various (i.e., clean or noisy) environments. First, an AUs based feature extraction method is proposed to extract facial feature effectively; second, a cascaded hierarchical random forest model based on different AU regions is developed to recognize the expressions in a coarse-to-fine way. The experiment results show that the proposed approach has a good performance in different environments.

The outline of the paper is as follows. The proposed method is presented in Section 2. Section 3 gives the experimental results while Section 4 presents the conclusions.

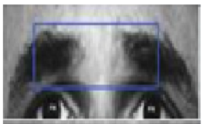
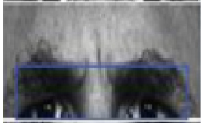
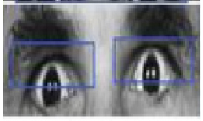
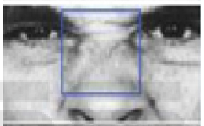
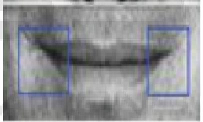
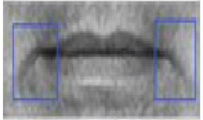
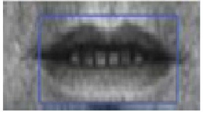
2 AN AUs BASED HIERARCHICAL RF MODEL

To obtain distinct facial features for expression recognition, firstly, appearance features are extracted within AU region, then, a hierarchical random forest model based on different AU regions is proposed to the recognize expressions in a coarse-to-fine way.

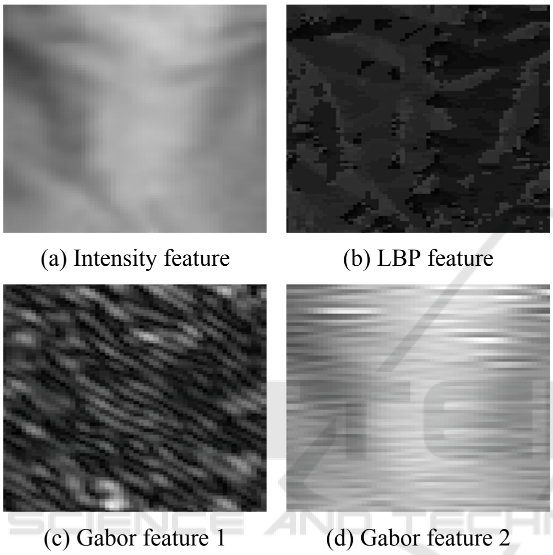
2.1 Feature Extraction based on AU Region

Facial actions code system (FACS) (Ekman et al., 1978) is first presented by Ekman and Friesen in 1978 to classify human facial movements. The FACS defines Action Units (AUs) to describe a contraction or relaxation of one or more muscles associated with facial expressions, which can be extracted from static images or dynamic sequences. AUs are extremely suitable to be used as midlevel parameters in an automatic facial behaviour analysis (Valstar et al., 2006), different facial expressions can be described with different combination of AUs, identifying the AUs is helpful to FER.

Table 1: Examples of AUs used in the proposed method.

AU Number	FACS Name	AU Region
1	Inner Brow Raiser	
4	Brow Lower	
5	Upper Lid Raiser	
9	Nose Wrinkler	
12	Lip Corner Puller	
20	Lip Stretcher	
25	Lips Part	

Based on the statistical analysis of a large number of facial expression datasets, certain AUs salient to describe the difference among facial expressions are chosen in this paper. However, the main challenge of AUs based expression recognition is to locate AUs accurately. To solve the problem, appearance features are extracted within AU regions (See Table 1) without precise AU location in this study. For example, among six basic facial expressions (i.e., anger, happiness, sadness, surprise, disgust and fear), first, AU9 (Nose Wrinkler) describes “disgust” effectively, appearance features (i.e., LBP, Gabor and intensity, see Figure 1) are extracted within AU9 region and used to distinguish the “disgust” from the other expressions; then, AU12+25 (Lip Corner Puller and Lips Part) are suitable to classify the other expressions into two groups, one contains “fear”, “happiness” and “surprise”, the other contains “anger” and “sadness”; next, AU12+20+25 (Lip Corner Puller, Lip Stretcher and Lips Part) are appropriate to separate “happiness” from “fear” and “surprise”; finally, AU1 (Inner Brow Raiser) is suitable to distinguish “sadness” from “anger”, AU4+5 (Brow Lower, Upper Lid Raiser) can tell the difference between “surprise” and “fear” well. According to the characteristic of AUs, each group of AUs focus on a binary classification and six expressions are classified in a coarse-to-fine way.


(a) Intensity feature (b) LBP feature

(c) Gabor feature 1 (d) Gabor feature 2

Figure 1: The different features from the AU9 region.

2.2 Random Forest

Random forests are an ensemble learning method which combines a family of weak classifiers to a strong classifier (Breiman, 2001). At training stage, each tree T in the forest $\mathcal{T} = \{T_t\}$ is built from a random subset of all the samples. The decision tree grows up by splitting the node, regarded as a weak classifier. The node stops splitting when a maximum depth is reached or the information gain is below a predefined threshold, named as leaf.

At the test stage, each sample $\mathcal{P}$ traverses all the trees and ends in a set of leafs l_t. The RF outputs the probability that the sample $\mathcal{P}$ belongs to the class C^n ($\mathrm{n} \in \{1, \dots, \mathrm{N}\}$) by averaging over all the decision trees, as described in (1).

$$\mathrm{p}(C^n|\mathcal{P}) = \frac{1}{T}\sum\nolimits_t p(C^n|l_t(\mathcal{P})) \tag{1}$$

The traditional RF method provides good performance for binary classification, but does not work well for multi-label classification of facial expression under noisy condition. Therefore, we propose a cascaded hierarchical RF to improve the efficiency and accuracy of the expression recognition in noisy environments.

2.3 Hierarchical RF Model based on AUs

A cascaded tree structure has been proposed and proven to be high accurate and efficient in (Minka et al., 1999). To improve the accuracy and efficiency of expression recognition, a cascaded tree structure is introduced to RF model based on different AU regions to classify the facial expression in a coarse-to-fine way.

As shown in Figure 2, the proposed model includes 4 layers:

- The first layer constructs a RF trained using the appearance features from the AU9 region and distinguishes the “disgust” expression from the other facial expressions;
- The second layer constructs a RF trained using the appearance features from the AU12+25 region and classifies the rest five facial expressions into two groups: {“happiness”, “fear”, “surprise”} and {“anger”, “sadness”};
- The third layer constructs two RFs corresponding to the two groups at the previous layer. One RF is trained using the appearance features from the AU12+20+25 region to distinguish “happiness” from “fear” and “surprise”, the other RF is trained using the appearance features from the AU1 region to distinguish “sadness” from “anger”;
- The last layer constructs a RF trained using the appearance features from the AU4+5 region and distinguishes “surprise” and “fear”.

Through all the layers, all the facial images are finally classified into the six basic categories. Each layer generates binary classification and AUs are selected with low correlation among the sub-classes.

According to the proposed cascaded hierarchical structure, the class distribution of the proposed model is determined by the product of the probability distributions leading to the class C^n, as described in (2).

$$\mathrm{p}(C^n|\mathcal{P}) = \prod_{i=1}^{K_n} \mathrm{p}(C_i^n|\mathcal{P}) \tag{2}$$

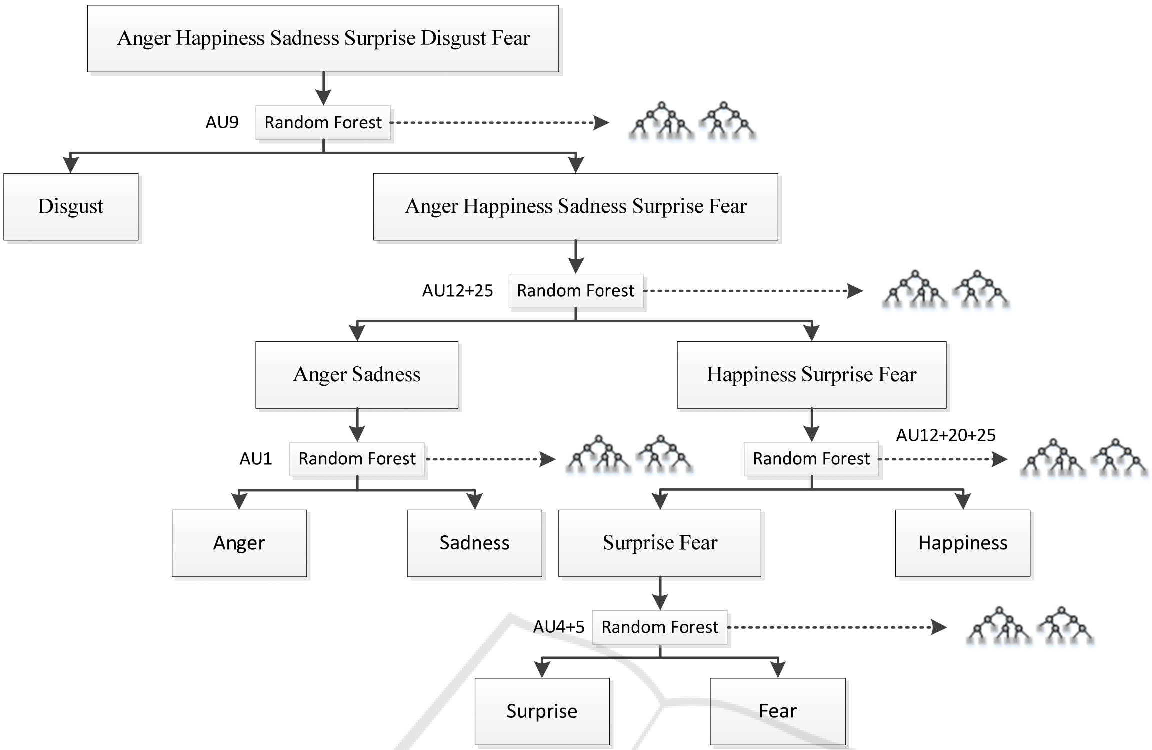


Figure 2: Hierarchical random forest model.

where $\mathrm{p}(C_i^n|\mathcal{P})$ is the probability that the patch belongs to the class C^n in the i-th layer of the proposed model, and the patch $\mathcal{P}$ needs to go through K_n layers of RFs. Substitute the Eq. (1) into Eq. (2):

$$\mathrm{p}(C^n|\mathcal{P}) = \prod_{i=1}^{K_n} \frac{1}{T_i} \sum\nolimits_t p(C_i^n|l_t(\mathcal{P})) \tag{3}$$

where T_i is the number of trees in the i-th layer RF. The training of a RF in the different layers of cascaded structure is given below:

- Label the patch with C_i^n that represents the patch of images from the n-th class used in the i-th layer of RFs. Labelling varies according to different layers as shown in Figure2. There are two kinds of labels for each layer because of the binary classification.
- A simple patch comparison feature is defined similar to (Dantone et al., 2012):

$$f_\theta(P) = |R_1|^{-1} \sum_{r \in R_1} F^q(r) - |R_2|^{-1} \sum_{r \in R_2} F^q(r) \tag{4}$$

where the parameter $\theta = \{R_{1,} R_{2,} q\}$ describes two rectangles R_1 and R_2 within the patch boundaries, and $q \in \{1,2,3\}$ denotes the different appearance feature channels.

- Generate a pool of splitting candidates $\phi = \{\theta, \tau\}$. Divide the extracted random patches $\mathcal{P}$ into two subsets $\mathcal{P}_L$ and $\mathcal{P}_R$ for each ϕ.

$$\mathcal{P}_L(\phi) = \{\mathcal{P}|f_\theta(P) < \tau\} \tag{5}$$

$$\mathcal{P}_R(\phi) = \{\mathcal{P}|f_\theta(P) \geq \tau\} \tag{6}$$

where τ is a predefined threshold.

- Select the splitting candidate ϕ which maximizes the evaluation function Information Gain (IG):

$$\phi^* = \underset{\phi}{\operatorname{argmax}}\, IG(\phi) \tag{7}$$

$$IG(\phi) = \mathcal{H}(\mathcal{P}) - \sum_{S \in \{L,R\}} \frac{|\mathcal{P}_S(\phi)|}{|\mathcal{P}|} \mathcal{H}(\mathcal{P}_S(\phi)) \tag{8}$$

where $\mathcal{H}(\mathcal{P})$ is the defined class uncertainty measure, which will be described for our case in (9). Selecting a certain split amounts to adding a binary decision node to the tree.

$$\mathcal{H}(\mathcal{P}) = -\sum_{i=1}^{N} \frac{\sum_i p(C^n|\mathcal{P}_i)}{|\mathcal{P}|} log\left(\frac{\sum_i p(C^n|\mathcal{P}_i)}{|\mathcal{P}|}\right) \quad (9)$$

- Create a leaf when a maximum depth is reached or the information gain is below a predefined threshold. Otherwise continue recursively for the two subsets $\mathcal{P}_L$ and $\mathcal{P}_R$ at the first step. A leaf of the RF stores the distribution that the patch belongs to the class C^n. Simply, the distribution is specified by a multivariate Gaussian:

$$\mathrm{p}(C^n|l) = \mathcal{N}(C^n; \overline{C_l^n}, \Sigma_l^n) \quad (10)$$

where $\overline{C_l^n}$ and Σ_l^n are the mean and covariance matrix of the estimations of the n-th class.

3 EXPERIMENTS

In this section, extensive experiments have been conducted on two widely used facial expression databases to evaluate the performance of the proposed AU based hierarchical RF model. The first database is the CK database (Kanade et al., 2000). It contains 486 sequences across 97 subjects. Each of the sequences contains images from neutral face to peak expression face. Six expressions (i.e. anger, disgust, fear, happy, sadness and surprise) have been labelled to peak expression by visual inspection from emotion researchers (see Figure 3). Because of its popularity, most researchers in the field have evaluated their improvements on the CK database (Mollahosseini et al., 2016). The second database is the JAFFE database (Lyons et al., 1998) which consists of 213 images from 10 Japanese female subjects with six expressions (see Figure 4).

Figure 3: Examples of images from the CK database.

Figure 4: Examples of images from the JAFFE database.

3.1 Training

As mentioned in the previous section, the proposed hierarchical random forest model consists of four layers. For training the RF, we fixed some parameters based on empirical observations, e.g., the forests at the first two layers have 15 trees, while the forests at the last two layers have 10 trees; the trees have a maximum depth of 20 and at each node we randomly generate 2000 splitting candidates.

Each tree grows based on a selected subset of 180 images with 30 images from each expression. The extracted patch size is 30×30 within approximate region of the specified AUs. There are three kinds of features to construct the candidate feature set, including the LBP feature with 58D, intensity feature with the size of 30×30 and the Gabor feature tuned with 7 orientations and 5 scales (with the size of 30×30×35).

3.2 Testing

To test the accuracy of the proposed model, a five-fold cross-validation is used, leaving 20% of the sample data as test data. Five sets are generated, each set contains 20% of sample data chosen randomly for each class as test data, the remaining data as training data. The training and test (i.e. classification) procedure is repeatedly five times, the classification accuracy is defined as the mean value of five procedures.

Experiments have been carried out on the clean databases and noisy databases with the salt-pepper noise and Gaussian noise respectively. The details of the testing are given as follow:

- Pre-processing: Detect the face and AU regions from the test images using a cascade of boosted classifiers working with haar-like features (Viola et al., 2004).
- Sampling: Randomly sample the patches with the same parameters as the training ones from the specified AU regions.
- Classification: All the patches go through the trees and end in different leafs. Each RF outputs the class probability by voting of leafs as described in Equation (10), we simplify the distribution by a multivariate Gaussian. The class of the test sample is finally determine using Equation (3).

3.3 Comparison with State of the Art

We compared our proposed algorithm with other state of the art algorithms, i.e., SVM (Chang et al., 2011) and RF (Dantone et al., 2012), using both CK and JAFFE databases.

Table 2: Comparison of our algorithm with SVM and RF.

	SVM		Our method		RF
	noise-free	noise	noise-free	noise	noise-free
Anger	0.400	0.240	0.935	0.872	0.567
Disgust	0.960	0.640	0.833	0.833	0.767
Fear	0.560	0.600	0.788	0.777	0.667
Happiness	0.960	0.800	0.914	0.883	0.433
Sadness	0.840	0.720	0.951	0.903	0.700
Surprise	0.840	0.760	0.851	0.822	0.767
AVG	0.760	0.627	0.889	0.848	0.650

To test the robustness and generalization performance of the proposed method, a mixed dataset with the salt-pepper noise or Gaussian noise is built from CK and JAFFE databases (see Figure 5). The comparison results of our method, SVM and RF are shown in Table 2. The average recognition rate of our method is 88.9% and 84.8% under clean and noise condition which are superior to the SVM and RF generally. The performance of RF under noise condition is much worse than the clean condition. For the expression of Disgust and Happiness, SVM provides a better performance than our method for the clean data.

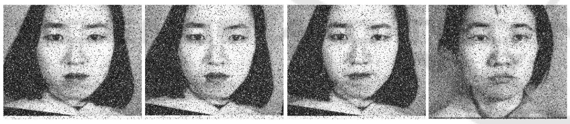


Figure 5: The samples with the salt-pepper and Gaussian noise.

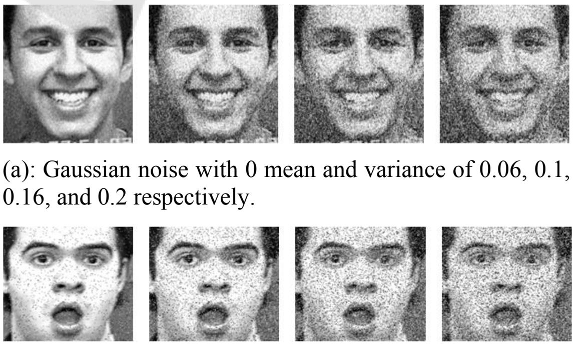
Table 3: The comparison results of the proposed method and RF (El Meguid et al., 2014) on CK database.

	Proposed Method	RF (El Meguid et al., 2014)
Anger	0.967	0.118
Disgust	0.967	0.704
Fear	0.983	0.536
Happiness	0.946	0.868
Sadness	0.880	0.667
Surprise	0.912	0.919
Average rate	0.943	——

Comparison has been made with the proposed method and the SVM labellers-paired RF (El Meguid et al., 2014) on CK database. From the comparison results in Table 3, one can see that the recognition rate of our method outperforms that of the method proposed in (El Meguid et al., 2014), which means that the proposed AU based hierarchical RF model helps to improve the accuracy of RF.

3.4 Degradation Analysis

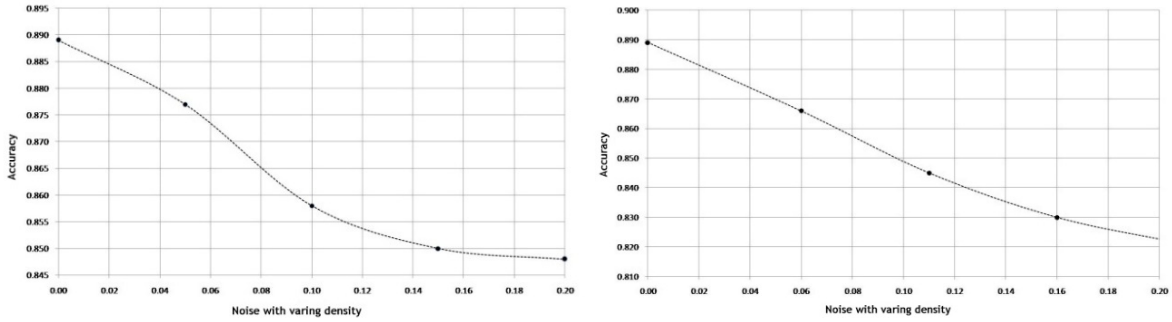
To assess the behaviour of the proposed method under varying degrees of noise. Images corrupted with different degrees of Gaussian or Salt-pepper noise are shown in Figure 6. The average recognition rates for different noise are given in Figure 7. The experiment results show that the proposed method degrades elegantly with good tolerance to varying degree of noise.



(a): Gaussian noise with 0 mean and variance of 0.06, 0.1, 0.16, and 0.2 respectively.

(b): Salt-pepper noise with density of 0.05, 0.1, 0.15 and 0.2 respectively.

Figure 6: Images corrupted with different degrees of noise.



(a): The recognition rate of Gaussian noise.

(b): The recognition rate of salt-pepper noise.

Figure 7: The recognition rate of different degrees of noise.

4 CONCLUSIONS

A hierarchical random forest model based on AUs has been proposed to improve the accuracy and efficiency for facial expression recognition. Firstly, appearance features (i.e., LBP, intensity and Gabor) are extracted within AU region, then, a cascaded tree structure is introduced to random forest model based on different AU regions to the recognize expressions in a coarse-to-fine way. The proposed approach has been evaluated with both CK and JAFFE databases and provides better performance than the SVM and RF method under both clean and noisy conditions. The experiment results show that the proposed method is robust to the noisy and degrades elegantly with good tolerance to varying degree of noise.

ACKNOWLEDGEMENTS

This work was supported by the National Social Science Foundation of China (Grant no. 16BSH107).

REFERENCES

Jameel, R., Singhal, A., & Bansal, A., 2015. A comparison of performance of crisp logic and probabilistic neural network for facial expression recognition. *In Next* Generation *Computing Technologies (NGCT), 2015 1st International Conference on* (pp. 841-846). IEEE.

Kahraman, Y., 2016. Facial expression recognition using geometric features. *In Systems, Signals and Image Processing (IWSSIP), 2016 International Conference on* (pp. 1-5). IEEE.

Li, Y., Wang, S., Zhao, Y., & Ji, Q., 2013. Simultaneous facial feature tracking and facial expression recognition. *IEEE Transactions on Image Processing*, 22(7), 2559-2573.

Cao, N. T., Ton-That, A. H., & Choi, H. I., 2014. Facial expression recognition based on local binary pattern features and support vector machine. *International Journal of Pattern Recognition and Artificial Intelligence*, 28(06), 1456012.

Liu, W., & Wang, Z., 2006. Facial expression recognition based on fusion of multiple Gabor features. *In 18th International Conference on Pattern Recognition (ICPR'06)* (Vol. 3, pp. 536-539). IEEE.

Satiyan, M., Nagarajan, R., & Hariharan, M., 2010. Recognition of facial expression using Haar wavelet transform. *Trans. Int. J. Electr. Electron. Syst. Res. JEESR Univ. Technol. Mara UiTM*, 3, 91-99.

Levi, G., & Hassner, T., 2015. Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. *In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* (pp. 503-510). ACM.

Valstar, M. F., & Pantic, M., 2012. Fully automatic recognition of the temporal phases of facial actions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(1), 28-43.

Bartlett, M. S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., & Movellan, J., 2006. Fully automatic facial action recognition in spontaneous behavior. *In 7th International Conference on Automatic Face and Gesture Recognition (FGR06)* (pp. 223-230). IEEE.

El Meguid, M. K. A., & Levine, M. D., 2014. Fully automated recognition of spontaneous facial expressions in videos using random forest classifiers. *IEEE Transactions on Affective Computing*, 5(2), 141-154.

Ekman, P., & Friesen, W. V., 1978. Manual for the facial action coding system. *Consulting Psychologists Press*.

Dantone, M., Gall, J., Fanelli, G., & Van Gool, L., 2012. Real-time facial feature detection using conditional regression forests. *In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on* (pp. 2578-2585). IEEE.

Minka, T., 1999. The dirichlet-tree distribution. *Paper available online at: http://www.stat.cmu.edu/minka/papers/dirichlet/minka-dirtree. pdf.*

Hager, J. C., Ekman, P., & Friesen, W. V., 2002. Facial action coding system. *Salt Lake City, UT: A Human Face.*

Valstar, M. F., & Pantic, M., 2006. Biologically vs. logic inspired encoding of facial actions and emotions in video. *In 2006 IEEE International Conference on Multimedia and Expo* (pp. 325-328).

Breiman, L., 2001. Random forests. *Machine learning*, 45(1), 5-32.

Kanade, T., Cohn, J. F., & Tian, Y., 2000. Comprehensive database for facial expression analysis. *In Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on* (pp. 46-53). IEEE.

Chang, C. C., & Lin, C. J., 2011. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3), 27.

Mollahosseini, A., Chan, D., & Mahoor, M. H., 2016. Going deeper in facial expression recognition using deep neural networks. *In 2016 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 1-10). IEEE.

Lyons, M., Akamatsu, S., Kamachi, M., & Gyoba, J., 1998. Coding Facial Expressions with Gabor Wavelets. *IEEE International Conference on Automatic Face and Gesture Recognition*, 1998. Proceedings (Vol.1998, pp.200--205).

Viola, P., & Jones, M. J. (2004). Robust real-time face detection. *International journal of computer vision*, 57(2), 137-154.