International Journal of Innovative
Computing, Information and Control
Volume **13**, Number **6**, December **2017**

 ISSN 1349-4198
pp. 1843–1854

# 3D FACIAL EXPRESSION RECOGNITION BASED ON KINECT

Wei Wei and Qingxuan Jia

School of Automation
Beijing University of Posts and Telecommunications
No. 10, Xitucheng Road, Haidian District, Beijing 100876, P. R. China
{ wei_wei; qingxuan }@bupt.edu.cn

Received April 2017; revised July 2017

Abstract. *In order to endow robot with the ability of affective computing which is similar to person, we propose an approach that uses noisy depth data produced by the low resolution sensor for robust 3D facial expression recognition. In this paper, we introduce random forest (RF) to construct Kinect-based 3D facial expression recognition model for various facial expressions. We utilize RGB-D database and KinectFaceDB database which are captured by Kinect sensor in the experiment. RGB-D database acquired from 31 test participants in 17 different face poses and 5 different facial expressions three times. KinectFaceDB database acquired data onto 52 test participants in 6 different face poses and 3 different facial expressions twice. We extract facial expression feature vector combining facial feature points (FFPs) feature vector with action units (AUs) feature vector through Face Tracking SDK. Then RF model is trained with feature vectors extracted from two databases respectively. At last, the average recognition accuracies in two databases are 82.52% and 87.63% respectively.*

**Keywords:** Affective computing, Kinect, Facial expression recognition, Random forest classification

1. **Introduction.** Robot technology in the field of industry has been rapidly developed and applied to the walks of life since the 1960, which has led to a growth of the service robot. International Federation of Robotics (IFR) has given a preliminary service robot definition as follows: service robot is a kind of semi-autonomous or autonomous robots that perform useful tasks for humans or equipment excluding industrial automation application. People also expect robots have ability of nature harmonious interaction in the process of interaction. At present, the key technologies in human-computer interaction are mostly in the following aspects. Visual computing supports visual interaction, speech computing supports voice interaction, physiology computing supports physiological interaction and handwriting recognition supports handwriting interaction. Affective computing endows robot with the ability to observe, understand, and generate a variety of emotional characteristics which are similar to people. Thus, robot captures signals by a variety of sensors, recognizes signals by "emotional model", and then responds to people's behavior appropriately. M. Kudo combined WAN and LAN with a robot system for the realization of home care services [1]. Figure 1 depicts the system architecture of care-giving monitoring. The robot system uses image and sound sensors to capture various signals, and realizes function of face authentication.

Speech, physiological, and visual signals have been explored for affective computing applications recently. Many methods of affective computing based on various speech signal have been proposed to estimate complex emotions [2,3]. However, in real-world applications there are problems in the progress of acquirement of speech signals, since speech signals are discontinuous signals. Speech signals can be captured only when people are

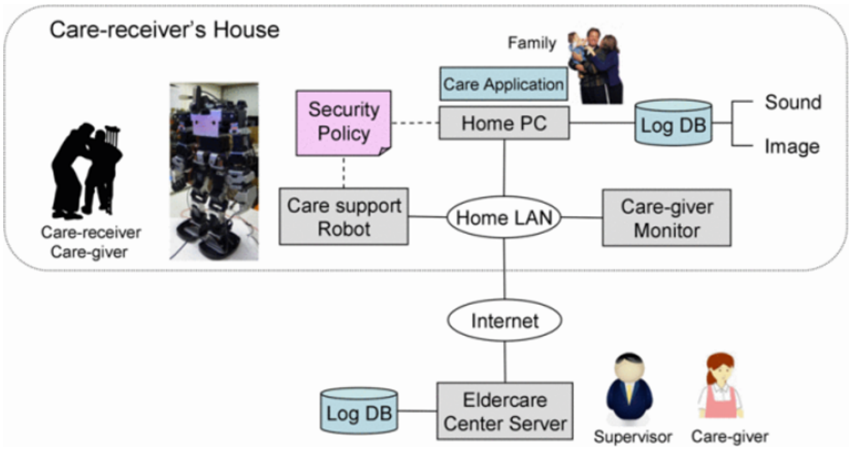


FIGURE 1. Component diagram for care-giving monitoring

talking. And some methods utilize various physiological signals for affective computing, such as brain activity, electrodermal activity, electromyogram activity, cardiovascular activity and some other signals [4,5]. However, changes in physiological signals cannot be observed directly. Unobtrusive, small, wearable, wireless physiological sensors [6,7] may be an ideal solution for physiology-based emotion recognition from laboratories to real-world applications.

A. Mehrabian's finding on emotion recognition has been known as the 7%-38%-55% Rule [8]. There are three elements accounting differently for the emotional meaning of a message: explicit verbal accounts for only 7%, tone of voice accounts for 38%, and facial expressions, postures, and gestures account for 55%. Besides, visual signals are continuous signals compared to speech signals. Visual signals can be observed directly compared with physiological signals.

Facial expressions arise owing to a person's internal emotional states which are important means of detecting emotions. As facial expression recognition being one of the applications of image classification, many effective algorithms have been proposed [9,10]. The function of facial expression recognition of smartphone in real time also has been developed [11]. All the above using RGB images are captured by 2D sensor, which is sensitive to head pose variations and surroundings, especially to illumination conditions. And 2D RGB images are not robust to human faces which are 3D objects.

While the main focus of face expression recognition algorithms is to 2D information, the work on facial expression recognition in 3D space has attracted numbers of attention and some techniques have been published in recent years due to the advantages of 3D data. 3D data, representing 3D physical coordinates, can capture essential geometrical features of the facial surface, and enables higher preservation of facial details insensitive to different conditions. Besides, 3D facial images are more robust to facial pose variations. The research of 3D human face recognition could further be used for facial expression recognition [12,13]. And varied strategies have been proposed for emotion recognition based on 3D facial expression [14,15].

Machine learning explores algorithms that can learn from and make predictions on data. Such algorithms are operated by building a model from example inputs in order to make data-driven predictions or decisions. Among them, random forest algorithm has emerged as machine community [16] and generated a great impact on signal and information processing since 1995. RF models achieve successful results in many challenging tasks, especially in speech and image domains [17,18]. It does not need prune to avoid over-fitting and can even handle small data set or asymmetric data set.

In this paper, we focus on investigating emotion recognition based on 3D facial expression in real time, which could be applied to application of various intelligent interactive platforms. For the proposed method, there are three main contributions and differences compared to the preliminary work. (1) A more advanced image processing strategy is used. In previous method, only texture and appearance information was exploited for feature extraction. Depth information of the object is introduced for improving efficiency and accuracy. (2) Both dynamic and static features are used for improving presentation of facial feature. (3) The proposed method has been evaluated in two databases which contain 5 and 3 kinds of emotions respectively. The rest of the paper is organized as follows. Section 2 gives an overview of related work on multi-modal feature extraction of facial expression and classification of emotion. Section 3 describes the selection of facial feature. Section 4 presents RF which is executed in classification algorithm. Section 5 verifies the proposed method by experiment and analyzes experimental results. Section 6 concludes with a brief discussion of the proposed approach.

2. **Related Work.** The fast development of 3D sensor techniques enables us to record and analyze the facial activities in a 3D environment. This development is leading to a new trend that integrates 3D facial expression with emotional factors. Figure 2 shows the flow of emotional recognition based on 3D facial expression, which consists of the following six main phases. First, users are exposed to designed or real-world stimuli according to the protocol. The emotional activities are recorded as 3D facial expression information. Then the raw data will be preprocessed to remove noise and artifacts. Some relevant features will be extracted and a classifier will be trained based on the extracted features. At last, user's current emotional states will be identified.

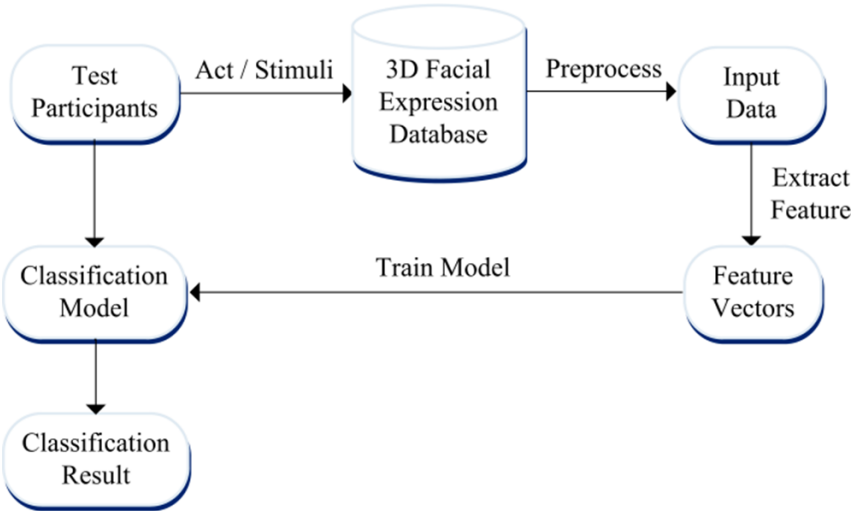


FIGURE 2. Flow of the emotional recognition based on 3D facial expression

The 3D facial expression images acquired from stereo imaging system require various preprocessing. And facial fiducial points of every face in the image should be manually annotated. The 3D facial expression data was captured with 3D face scanner system. However, the 3D capture system did not capture facial expressions dynamically, and they required the subjects to perform each expression for a short period of time. Recently, the emerging RGB-D cameras such as the Kinect sensor have been successfully applied to many applications based on 3D. With respect to the acquisition efficiency, the capturing of a high-resolution RGB image normally takes less than 0.05s, whereas the laser scanning of a face takes 9s on average [19]. And high-qualified 3D face scanning needs careful

cooperation of users. For biometrics, directly inheriting from its application in body parts segmentation and tracking, Kinect sensor face recognition systems can be implemented for real-time and online processing [20].

Facial expression analysis can be traced to the nineteenth century. Darwin had demonstrated the universality and continuity of facial expressions in man [21]. Facial expressions are specific inborn emotions, which are originated from serviceable associated habits. [22] postulated six primary emotions, each of which possesses distinctive content together with a unique facial expression. These prototypic emotional displays are referred to as so called basic emotions, and comprise happiness, sadness, fear, disgust, surprise and anger. [23] presented a preliminary investigation on automatic facial expression analysis from an image sequence. Afterwards, [24] proposed facial expressions are generated by contractions of facial muscles, such as eyelids, eyebrows, nose, lips and skin texture, often revealed by wrinkles and bulges. The changes of facial muscles are fleeting, lasting rarely more than 5s or less than 250ms.

[25] presented a facial point detection system based on improved multi-kernel learning to track 18 facial feature points automatically. The system has good performance at efficiency which is an important matter with real-time applications. [26] first divided face region into 20 regions of interest and applied Gabor feature based on boosted classifiers, to 20 facial feature points. However, their approach is not robust with respect to various head rotations and occlusions. [27] proposed a method based on a local and a global phase to automatic localization of 17 facial feature points that are salient facial points. And the system is robust with respect to self-occlusions and varying illuminations. [28] proposed unsupervised automatic facial feature point detector which is able to detect 54 facial points in images of faces with scaling differences and self-occlusions using Gabor filtering, binary robust invariant scalable keypoints (BRISK), an iterative closest point (ICP) algorithm and fuzzy c-means (FCM) clustering. They then conduct AU intensity estimation respectively using support vector regression and neural networks for 18 selected facial action units. [29] proposed a knowledge-driven prior model for AU recognition, which is totally learned from the generic domain knowledge without using any training data. And the model has no dependence on the data and can generalize well to different data sets.

There have been considerable efforts for 3D facial expression analysis so far [30]. Various studies in affective computing community try to build computational models to estimate emotional states using machine learning techniques. [31] employed neural network classifier to recognize six universal facial expressions. The approach relies on the feature vectors retrieved from the distribution of 11 facial feature points in a 3D facial model. [32] used a 3D facial surface descriptor and classified six facial expressions using hidden Markov models (HMMs). [33] utilized 96 distance and slope features extracted from a 3D face model with 87 points, and used multi-class support vector machines (SVMs) for the recognition of the six basic emotions. [34] proposed an approach relied on 19 AUs to recognize six basic emotions using multivariate logistic regression (MLR).

Although various approaches have been proposed for emotion recognition based on 3D facial expression, most of the experimental results cannot be compared directly to different setups of experiments. There is still a lack of publicly available 3D facial expression of emotional databases. To the best of our knowledge, the popular publicly available databases are RGB-D database [37] and KinectFaceDB database [38]. The databases used in this study are freely available to the academic community.

## 3. **Facial Expression Feature Extraction.**

With the fast development of 3D sensor techniques, Kinect sensor appears to person due to the high-performance and low-overhead

in terms of machine learning. Kinect sensor captures both 2D and 3D information with one RGB camera and two 3D depth sensors at the same time. The depth sensor includes an infrared laser projector and a monochrome CMOS sensor. Kinect sensor gets the texture and appearance information of the object through the RGB camera and the "Field" (Depth) information of the object through the depth sensor. "Field" indicates the distance from the object to the sensor. So Kinect sensor can obtain an RGB-D image that comprises a 2D color image (RGB) and a depth map (D) at the same time under any ambient light conditions. Though Kinect sensor is not the most accurate 3D sensor, it provides more information than 2D images alone. Besides, it provides good performance for real time, which meets our needs.

Emotion recognition software of 3D facial expression algorithm is mostly based on the model. P. Ekman et al. propose the first comprehensive visual numerical technology Facial Action Coding System (FACS). FACS has been established as a computed automated system that detects faces in videos, extracts the geometrical features of the faces, and then produces temporal profiles of each facial movement. It can encode movements of individual facial muscles from slight different instant changes in facial appearance. FAEC defined 44 facial movements called action units (AUs), which are a contraction or relaxation of one or more muscles and independent of any interpretation. So AUs can be used for any higher order decision-making process including the emotion recognition based on facial expression.

Mikael Rydfalk created the CANDIDE model at Linköping University in 1987. The motive of CANDIDE model is attempting to use animation for image compression. CANDIDE, as computer graphical face model, is a parameterized face mask, which is developed for model-based coding of human faces specifically. It is made up of a small number of polygons which includes a lot of vertices and triangles. And it is controlled by global and local action units. In view of the simplicity and public availability, CANDIDE has been widely used in many research fields around the world in the past decade. It is a popular tool in the field of image processing, especially. With the emergence of MPEG-4 standard, it is necessary to update the CANDIDE model to meet the demands which are set by MPEG-4. So far, CANDIDE-3 is the updated model which is compatible with the facial animation and definition parameters (FAPs and FDPs) defined in MPEG-4. Besides, it adds a few vertices to improve quality of crude mouth and eyes, which are closely related to facial expression.

The Microsoft Face Tracking Software Development Kit for Kinect for Windows (Face Tracking SDK) is part of Windows Software Development Kit (Kinect for Windows SDK). The Face Tracking SDK's face tracking engine aims at analyzing color and depth images from the Kinect sensor, and then it detects and selects the specific information of face depending on the requirement in real time. At last, it computes animation units (AUs) and 3D positions of facial feature points (FFPs), which could be used on emotion recognition based on facial expression.

3.1. **Animation units (AUs).** The Face Tracking SDK results are also expressed in terms of weights of six AUs, which are a subset of what is defined in the CANDIDE-3 model. The AUs are deltas from the neutral shape that you can use to morph targets on animated avatar models so that the avatar acts as the tracked user does. The Face Tracking SDK tracks 6 AUs. Each AU is expressed as a numeric weight varying between $-1$ and $+1$, and 0 represents the neutral states. Therefore, a 6-dimensional AUs feature vector can get from each input as follows.

$$\vec{A}_1 = (u_1, u_2, u_3, u_4, u_5, u_6),$$

where $u_1$, $u_2$, $u_3$, $u_4$, $u_5$, $u_6$ refer to the weights of 'upper lip raiser', 'jaw lower', 'lip stretcher', 'brow lower', 'lip corner depressor', and 'outer brow raiser', respectively.

3.2. **Facial feature points (FFPs).** The Face Tracking SDK uses the Kinect coordinate system to output its 3D tracking results. The tracking quality may be affected by the image quality of these input frames and some other factors. The Face Tracking SDK tracks 121 3D feature points of the face, which is a subset of what is defined in the CANDIDE-3 model. Each feature point is expressed as a 3-dimensional vector as follows: $(x, y, z)$. [35] proposed the features that appear temporarily in the face during any kind of facial expression are strongly linked to not the whole face but specific regions in the face, such as eyes, eyebrows, mouth, tissue textures and nose. Kinect sensor could not capture some detailed information of tiny region of face, such as eyelid and pupil. In order to improve speed of computation and accuracy of recognition, we get feature points down to a reasonable number if possible. Above all, we select 40 target feature points from the eyebrows, eyes, nose, mouth, jaw and some other key regions of face as facial expression points (FFPs). The $(x_i, y_i, z_i)$ $i = 1, 2, \ldots, 40$ are 3-dimensional coordinates of each FFP. Therefore, a 120-dimensional FFPs feature vector can be got from each frame of input as follows:

$$\vec{A}_2 = (x_1, y_1, z_1, x_2, y_2, z_2, \ldots, x_{40}, y_{40}, z_{40}).$$

Facial expression feature vector comes from the combination of AUs feature vector with FFPs feature vector. As a result, a 126-dimensional vector can get from each frame of input as follows:

$$\vec{A} = \left(\vec{A}_1, \vec{A}_2\right) = (u_1, u_2, u_3, u_4, u_5, u_6, x_1, y_1, z_1, x_2, y_2, z_2, \ldots, x_{40}, y_{40}, z_{40}).$$

4. **Classification Method Based on Random Forest.** Once a facial expression feature vector is extracted from the face, the next step is to use this vector in classification. Researchers have proposed several methods of this issue. For several years, ensemble learning algorithm has got more attention to the field of classification methods due to its good performance. Ensemble methods refer to the learning algorithms that produce collections of individual classifiers called weak learners which learn to classify by training learners individually and fusing their classified results. RF belongs to the category of ensemble learning algorithms, which correspond on combination of decision tree-type classifier.

During the step of training the model, the RF algorithm generates a set of decision trees. The sample subset $T_i$, got from bagging random, is different from each other. Each decision tree is trained on the sample subset $T_i$ independently. At each node, only a subset of variables will be selected randomly to determinate the best split. At last, all individual decision trees form RF model together. RF algorithm has only two parameters: (i) $n_{tree}$ the number of decision trees, (ii) $m_{code}$ the number of variables are selected at each node. According to L. Breiman [36], providing the total number of variable is $M$, then the ratio of $m_{code}$ can be sqrt($M$), 1/2sqrt($M$), 2sqrt($M$). The summary of the random forests algorithm for classification problem is presented in Algorithm 1. To avoid over-fitting problems in decision tree, pruning is the most common method. However, random sampling twice ensures randomness of data in RF algorithm. RF algorithm has good performances in avoiding over-fitting and anti-noise even without pruning.

During the step of classification, each decision tree in the RF gives an independent sub-classification result $O_i$ $(i = 1, 2, \ldots, n)$. The final classification result is determined by $C_j$ $(j = 1, 2, \ldots, k)$, which is the number of votes of all the decision trees of each facial expression. $MC$ refers to the most votes defined as $MC = \max\{C_j : j = 1, 2, \ldots, k\}$.

**Algorithm 1**. Random forests classifier

**For** $i = 1$ to $n_{tree}$ do

**Select** boostrap sample subset $T_i$ from the training data $T$ as training set for each decision tree

**Select** $m_{code}$ variables at each node, compute and choose the Highest GINI index on these $m_{code}$ variables

**end for**

**Aggregate** the predictions of the $n_{tree}$ trees

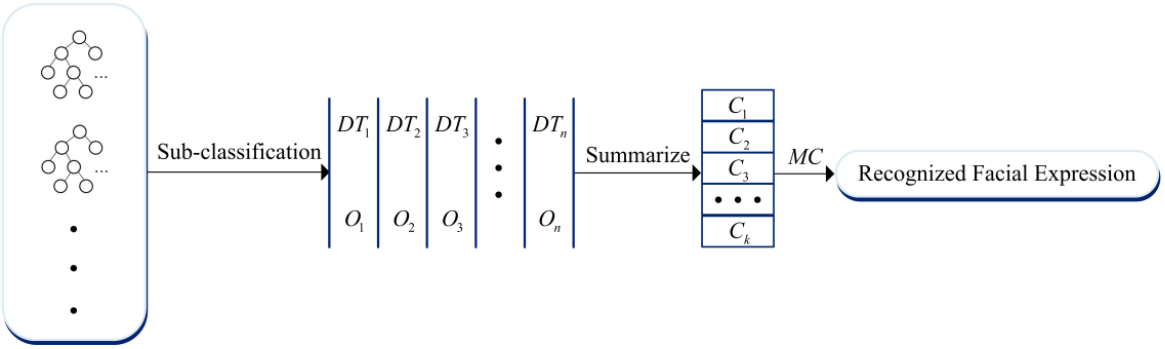


FIGURE 3. Flow of the RF classification algorithm

The facial expression with the most votes is regarded as the recognized facial expression. Figure 3 shows the flow chart of the classification of RF algorithm.

Recently, the desirable results of experiment have demonstrated RF algorithm works well in mining data. Firstly, RF algorithm is simple due to its non-parametric nature. Secondly, RF algorithm is fast in both steps of training and classification, especially when handling with high dimensional data as face recognition in large databases. It selects only a subset of the sample set and only a subset of variables each time which are not based on an exhaustive research through all the data. Thirdly, RF has a good adapting ability to discrete and continuous data even with greater noise. Therefore, it need not make data standardization in the stage of data processing firstly.

5. **Experiments.** In our experiment, collection tool of 3D facial expression information is Kinect sensor, and platform of data processing is a computer with Windows 7, Intel(R) Core(TM) i3-2120 CPU (3.30GHz), 4.00GB RAM. The specific step of the experiment is shown in Figure 4.

Face detection and facial expression recognition have been researched for many years. There are not many benchmark public databases which include both color and depth information of 7 facial expressions. There are some databases using Kinect sensor to capture 3D facial expressions of emotion to the public recently. Nevertheless, some available databases only have part of facial expressions of seven facial expressions. At last, we utilize two 3D facial expressions of emotion databases which are captured by Kinect sensor in this experiment.

R. I. Hg et al. [37] presented an RGB-D database containing 1581 RGB images and their depth counterparts. The process of data-gathering has been repeated three times from 31 test participants in 17 different face poses and 5 different facial expressions with a Kinect sensor. The 5 facial expressions are: neutral, smile, sad, yawn and angry. For each test participant in the database face the Kinect sensor when gathering facial expression

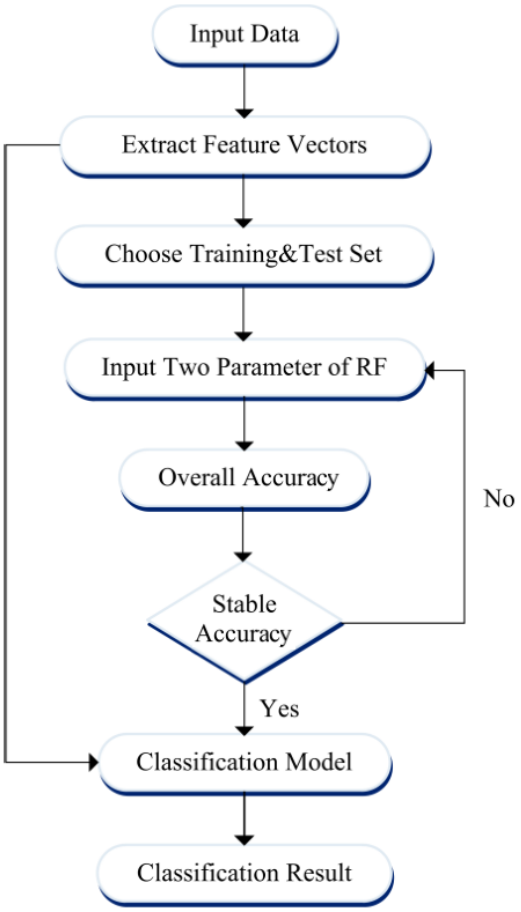


FIGURE 4. Flow of the experiment

TABLE 1. Detailed number of each set of RGB-D database

| | **Neutral** | **Smile** | **Yawn** | **Angry** | **Sad** | **Total** |
|---|---|---|---|---|---|---|
| **Primary Data** | 93 | 93 | 93 | 93 | 93 | 465 |
| **Sample Set** | 86 | 85 | 88 | 85 | 86 | 430 |
| **Training Set** | 57 | 57 | 59 | 57 | 57 | 287 |
| **Test Set** | 29 | 28 | 29 | 28 | 29 | 143 |

The first row of table is the number of original data, the second row is the number of feature vectors, and the last two rows are the size of training set and test set respectively.

data. The detailed number of images of each facial expression of each database is shown in Table 1.

R. Min et al. [38] compiled KinectFaceDB database containing 936 RGB images and their depth counterparts. The 52 test participants have attended two different sessions to gather data by Kinect sensor which include 6 different face poses and 3 different facial expressions. The 3 expressions are: neutral, smile and mouth open. For each test, participants in the database face the Kinect sensor when gathering facial expression data. The detailed number of images of each facial expression of each database is shown in Table 2.

With Face Tracking SDK, we can detect and track face, extract facial expression feature, and then get a 126-dimensional facial expression feature vector ultimately. However, the input data are gathered from Kinect sensor, not utilized directly. Only color and depth

TABLE 2. Detailed number of each set of KinectFaceDB database

| | **Neutral** | **Smile** | **Yawn** | **Total** |
|---|---|---|---|---|
| **Primary Data** | 104 | 104 | 104 | 312 |
| **Sample Set** | 97 | 96 | 99 | 292 |
| **Training Set** | 65 | 64 | 66 | 195 |
| **Test Set** | 32 | 32 | 33 | 97 |

The first row of table is the number of original data, the second row is the number of feature vectors, and the last two rows are the size of training set and test set respectively.

data of facial expression are available, and bone information is not stored. This will influence the result of feature extraction, and even lead to failure of feature extraction. Therefore, we need to remove the feature vectors which are extracted unsuccessfully or repeatedly. At last, we get two sample sets $S_1$ and $S_2$ from two databases independently. To be specific, sample set $S_1$ is composed of 430 different facial expression feature vectors which are extracted from the original 465 group data of RGB-D database. The detailed number of feature vectors of each facial expression is shown in Table 1. Similarly, sample set $S_2$ is composed of 292 different facial expression feature vectors which are extracted from the original 312 group data of KinectFaceDB database. The detailed number of feature vectors of each facial expression is shown in Table 2.

We use the method of stratification sampling to get training set $T$ and test set $V$ from sample set $S$. First, treat the sample set in disjoint layers on the basis of certain facial expressions. Then select feature vectors from each layer on a certain percentage independently and randomly. In the end, all these selected feature vectors come together to compose training set, while the rest feature vectors come together to compose test set. Stratification sampling ensures structural consistency of the overall sample set, which improves training set representation. The size of training set and test set in the sample set is on a ratio of 2:1 in this experiment. To be specific, we select 287 feature vectors from sample set $S_1$ with stratification sampling compose training set $T_1$, and the rest 143 feature vectors of sample set $S_1$ compose test set $V_1$. The detailed number of feature vectors of each facial expression is shown in Table 1. Similarly, we select 195 feature vectors from sample set $S_2$ with stratification sampling compose training set $T_2$, and the rest 97 feature vectors of sample set $S_2$ compose test set $V_2$. The detailed number of feature vectors of each facial expression is shown in Table 2.

There are different kinds of facial expressions of two databases. Thus, we experiment twice in RGB-D database and KinectFaceDB database respectively. We obtain training set $T_1$ and test set $V_1$ from RGB-D database. Then we use training set $T_1$ to train facial expression to recognize model $M_1$ which is based on RF algorithm. At last, we use test set $V_1$ to test model $M_1$ and get correct recognition rate of each facial expression. The detailed correct rate of each facial expression is shown in Figure 5. Similarly, we obtain training set $T_2$ and test set $V_2$ from RGB-D database. Then we use training set $T_2$ to train facial expression to recognize model $M_2$ which is based on RF algorithm. At last, we use test set $V_2$ to test model $M_2$ and get correct recognition rate of each facial expression. The detailed correct rate of each facial expression is shown in Figure 5.

According to the analysis based on the recognition results of the experiments in two databases, the facial expression of open mouth achieves the highest accuracy of all facial expressions as shown in Figure 5. In the RGB-D database, most neutral expressions are falsely recognized as smile expressions, while smile and angry expressions as neutral expressions, sad expressions as angry expressions as shown in Table 3. In the KinectFaceDB

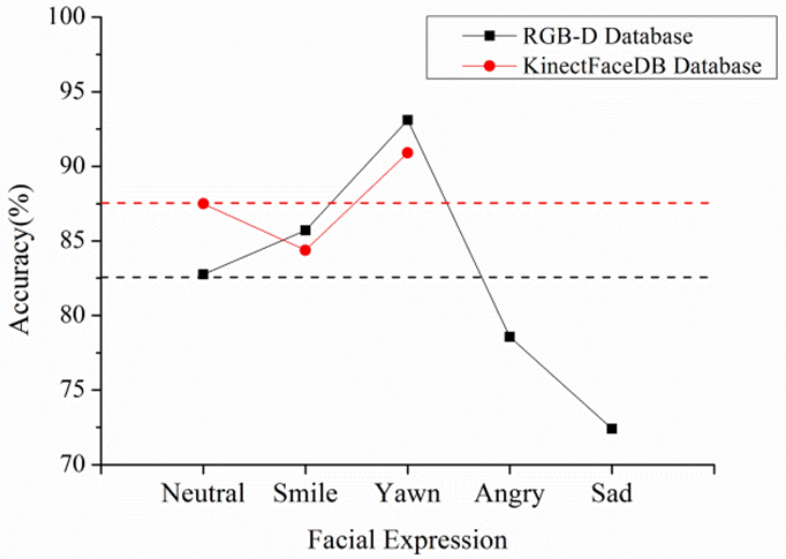


FIGURE 5. Recognition accuracy of each facial expression of each database

TABLE 3. Detailed recognition result of each facial expression of RGB-D database

| | **Neutral** | **Smile** | **Yawn** | **Angry** | **Sad** |
|---|---|---|---|---|---|
| **Neutral** | **24** | 4 | 0 | 4 | 2 |
| **Smile** | 4 | **24** | 0 | 0 | 1 |
| **Yawn** | 0 | 0 | **27** | 0 | 0 |
| **Angry** | 1 | 0 | 1 | **22** | 5 |
| **Sad** | 0 | 0 | 1 | 2 | **21** |

TABLE 4. Detailed recognition result of each facial expression of KinectFaceDB database

| | **Neutral** | **Smile** | **Yawn** |
|---|---|---|---|
| **Neutral** | **28** | 5 | 1 |
| **Smile** | 4 | **27** | 2 |
| **Yawn** | 0 | 0 | **30** |

database, most neutral expressions are falsely recognized as smile expressions, and smile expressions as neutral expression as shown in Table 4. Above all, there are few differences in eyes area of various facial expressions, and obvious mouth behaviors are essential for proper recognition of facial expression. According to the analysis based on falsely recognized samples in two databases, most test participants are wearing glasses. Refraction of spectacles lenses influences the feature extraction of eyes area, and influences the recognition result of facial expression. Besides, the lack of training samples leads to model training not fully enough. However, the last facial expression recognition model meets our demand and the recognition result is still good.

Solid line with square represents recognition accuracies of five facial expressions; lower dashed line represents average recognition accuracy of five facial expressions in RGB-D database. Solid line with circle represents recognition accuracies of three facial expressions; higher dashed line represents average recognition accuracy of three facial expressions in KinectFaceDB database.

6. **Conclusion.** In this paper, we propose an approach of 3D facial expression recognition based on features of FFPs and AUs captured by Kinect sensor. Kinect sensor is fast in data acquisition and high resolution of RGB images, which is not sensitive to head pose variations and surroundings, especially to illumination conditions. Stratification sampling ensures structural consistency of the overall sample set, and improves the representativeness of the training set. RF algorithm has good performances in avoiding over-fitting and anti-noise even when lacking of enough training samples, which can solve the imbalance and limitation of existing public 3D facial expression databases. The experimental results suggest that the approach based on random forests has good performance on the correct rate in facial expression recognition.

**Acknowledgment.** This work was supported by the National Natural Science Foundation of China (No. 61573066).

## REFERENCES

[1] M. Kudo, Robot-assisted healthcare support for an aging society, *Service Research and Innovation Institute Global Conference (SRII)*, San Jose, CA, USA, pp.258-266, 2012.

[2] J. B. Alonso, J. Cabrera, C. M. Travieso and K. López-de-Ipiña, New approach in quantification of emotional intensity from the speech signal: Emotional temperature, *Expert Systems with Applications*, vol.42, no.4, pp.9554-9564, 2015.

[3] W. H. Dai, D. M. Han, Y. H. Dai and D. R. Xu, Emotion recognition and affective computing on vocal social media, *Information and Management*, vol.52, no.7, pp.777-788, 2015.

[4] W. L. Zheng and B. L. Lu, Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks, *IEEE Trans. Autonomous Mental Development*, vol.7, no.3, pp.162-175, 2015.

[5] A. Greco, G. Valenza, L. Citi and E. P. Scilingo, Arousal and valence recognition of affective sounds based on electrodermal activity, *IEEE Sensors Journal*, vol.17, no.3, pp.716-725, 2017.

[6] T. Liang and Y. J. Yuan, Wearable medical monitoring systems based on wireless networks: A review, *IEEE Sensors Journal*, vol.16, no.23, pp.8186-8199, 2016.

[7] C. Habib, A. Makhoul, R. Darazi and C. Salim, Self-adaptive data collection and fusion for health monitoring based on body sensor networks, *IEEE Trans. Industrial Informatics*, vol.12, no.6, pp.2342-2352, 2016.

[8] A. Mehrabian, Communication without words, *Psychological Today*, vol.2, pp.53-55, 1968.

[9] M. Y. Liu, S. G. Shan, R. P. Wang and X. L. Chen, Learning expressionlets via universal manifold model for dynamic facial expression recognition, *IEEE Trans. Image Processing*, vol.25, no.12, pp.5920-5932, 2016.

[10] S. Sun, L. D. Li, G. Y. Zhou and J. He, Facial expression recognition in the wild based on multimodal texture features, *Journal of Electronic Imaging*, vol.25, no.061403, 2016.

[11] M. Suk and B. Prabhakaran, Real-time facial expression recognition on smartphones, *IEEE Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, pp.1054-1059, 2015.

[12] S. Elaiwat, M. Bennamoun, F. Boussaid and A. El-Sallam, 3-D face recognition using curvelet local features, *IEEE Signal Processing Letters*, vol.12, no.2, pp.172-175, 2014.

[13] Y. Lu and Q. J. Hui, Improved framework for 3D face feature points extraction method based on statistic deformable model, *Computer Engineering and Application*, vol.52, no.1660170, 2016.

[14] X. L. Li, Q. Q. Ruan, G. Y. An and Y. Jin, Analysis of range images used in 3D facial expression recognition systems, *Computing and Informatics*, vol.35, no.1, pp.203-221, 2016.

[15] X. L. Li, Q. Q. Ruan, G. Y. An, Y. Jin and R. Z. Zhao, Multiple strategies to enhance automatic 3D facial expression recognition, *Neurocomputing*, vol.161, no.5, pp.89-98, 2015.

[16] T. K. Ho, Random decision forests, *Proc. of the 3rd International Conference on Document Analysis and Recognition*, Montreal, Canada, pp.278-282, 1995.

[17] T. Zhao, Y. X. Zhao and X. Chen, Ensemble acoustic modeling for CD-DNN-HMM using random forests of phonetic decision trees, *Journal of Signal Processing Systems for Signal Image and Video Technology*, vol.82, no.2, pp.187-196, 2016.

[18] Y. Shi, L. M. Cui, Z. Q. Qi, F. Meng and Z. S. Chen, Automatic road crack detection using random structured forests, *IEEE Trans. Intelligent Transportation Systems*, vol.17, no.12, pp.3434-3445, 2016.

[19] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Min and W. Worek, Overview of the face recognition grand challenge, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, San Diego, CA, USA, pp.947-954, 2005.
[20] R. Min, J. Choi, G. Medioni and J. L. Dugelay, Real-time 3D face identification from a depth camera, *The 21st International Conference on Pattern Recognition (ICPR)*, Tsukuba, Japan, pp.1739-1742, 2012.
[21] C. Darwin, *The Expression of the Emotions in Man and Animals*, Oxford University Press, 1998.
[22] P. Ekman and W. V. Friesen, Constants across cultures in the face and emotion, *Journal of Personality and Social Psychology*, vol.17, no.2, pp.124-129, 1971.
[23] M. Suwa, N. Sugie and K. Fujimora, A preliminary note on pattern recognition of human emotional expression, *Proc. of the 4th International Joint Conference on Pattern Recognition*, Kyoto, Japan, pp.408-410, 1978.
[24] B. Fasel and J. Luettin, Automatic facial expression analysis: A survey, *Pattern Recognition*, vol.36, no.1, pp.259-275, 2003.
[25] T. Senechal, V. Rapp and L. Prevost, Facial feature tracking for emotional dynamic analysis, *The 13th International Conference on Advances Concepts for Intelligent Vision Systems (ACIVS)*, Belgium, pp.495-506, 2011.
[26] D. Vukadinovic and M. Pantic, Fully automatic facial feature point detection using Gabor feature based boosted features, *IEEE International Conference on Systems, Man and Cybernetics*, Waikoloa, pp.1692-1698, 2005.
[27] E. Sangineto, Pose and expression independent facial landmark localization using dense-SURF and the Hausdorff distance, *IEEE Trans. Pattern Analysis and Machine Intelllgence*, vol.35, no.3, pp.624-638, 2013.
[28] L. Zhang, K. Mistry, M. Jiang, S. C. Neoh and M. A. Hossain, Adaptive facial point detection and emotion recognition for a humanoid robot, *Computer Vision and Image Understanding*, vol.140, pp.93-114, 2015.
[29] Y. Li, J. Chen, Y. Zhao and Q. Ji, Data-free prior model for facial action unit recognition, *IEEE Trans. Affective Computing*, vol.4, no.2, pp.127-141, 2013.
[30] G. Sandbach, S. Zafeiriou, M. Pantic and L. Yin, Static and dynamic 3D facial expression recognition: A comprehensive survey, *Image and Vision Computing*, vol.30, no.10, pp.683-697, 2012.
[31] H. Soyel and H. Demirel, Facial expression recognition using 3D facial feature distances, *International Conference on Intelligent Automation and Robotics*, San Francisco, CA, USA, pp.831-838, 2007.
[32] Y. Sun and L. J. Yin, Facial expression recognition based on 3D dynamic range model sequences, *The 10th European Conference on Computer Vision*, Marseille, France, pp.58-71, 2008.
[33] H. Tang and T. S. Huang, 3D facial expression recognition based on properties of line segments connecting facial feature points, *The 8th IEEE International Conference on Automatic Face & Gesture Recognition*, Amsterdam, pp.1-6, 2008.
[34] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan and M. Bartlett, The computer expression recognition toolbox (CERT), *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops*, Santa Barbara, CA, USA, pp.298-305, 2011.
[35] C. P. Sumathi, T. Santhanam and M. Mahadevi, Automatic facial expression analysis: A survey, *International Journal of Computer Science & Engineering Survey*, vol.3, no.6, pp.47-59, 2012.
[36] L. Breiman, Random forests, *Machine Learning*, vol.45, no.1, pp.5-32, 2001.
[37] R. I. Hg, P. Jasek, C. Rofidal, K. Nasrollahi, T. B. Moeslund and G. Tranchet, An RGB-D database using Microsoft's Kinect for windows for face detection, *The 8th International Conference on Signal Image Technology and Internet Based Systems (SITIS)*, Naples, Italy, pp.42-46, 2012.
[38] R. Min, N. Kose and J. L. Dugelay, KinectFaceDB: A Kinect database for face recognition, *IEEE Trans. Systems, Man, and Cybernetics: Systems*, vol.44, no.11, pp.1534-1548, 2014.