Frontiers of Information Technology & Electronic Engineering
www.zju.edu.cn/jzus; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Using Kinect for real-time emotion recognition via facial expressions*

Qi-rong MAO†‡, Xin-yu PAN†, Yong-zhao ZHAN, Xiang-jun SHEN

(*Department of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang 212013, China*)

†E-mail: mao_qr@ujs.edu.cn; pxyz@vip.qq.com

Received June 12, 2014; Revision accepted Oct. 22, 2014; Crosschecked Mar. 9, 2015

Abstract: Emotion recognition via facial expressions (ERFE) has attracted a great deal of interest with recent advances in artificial intelligence and pattern recognition. Most studies are based on 2D images, and their performance is usually computationally expensive. In this paper, we propose a real-time emotion recognition approach based on both 2D and 3D facial expression features captured by Kinect sensors. To capture the deformation of the 3D mesh during facial expression, we combine the features of animation units (AUs) and feature point positions (FPPs) tracked by Kinect. A fusion algorithm based on improved emotional profiles (IEPs) and maximum confidence is proposed to recognize emotions with these real-time facial expression features. Experiments on both an emotion dataset and a real-time video show the superior performance of our method.

Key words: Kinect, Emotion recognition, Facial expression, Real-time classification, Fusion algorithm, Support vector machine (SVM)

doi:10.1631/FITEE.1400209 **Document code:** A **CLC number:** TP391.4

1 Introduction

Emotion recognition via facial expressions (ERFE) has attracted a great deal of interest with recent advances in artificial intelligence and pattern recognition, mainly because of its various applications such as video games, medicine, security, intelligent human-computer interaction, and affective computing. A human face contains most of the feeling information of a human, and facial expressions constitute the main channel which is used to communicate emotions. These facts highlight the importance of facial expressions in emotion recognition, and justify the interest that ERFE has attracted over the last two decades.

Recently, sensors used in most ERFE research are RGB cameras, which can capture only 2D images. Since human faces are 3D objects, the process representing 3D faces with 2D images is deficient in essential geometrical features.

Another major challenge in ERFE is the requirement of recognizing facial expressions in real time. Recognition approaches based on RGB cameras are in general computationally expensive (Zhu and Ramanan, 2012). Even though real-time identification can be realized by performing highly complex calculations in the pre-processing phase instead of the recognition phase, the fundamental problem still exists (Ma *et al.*, 2013). This problem must be solved before performing ERFE in video sequences.

In this paper, we propose a real-time ERFE approach based on animation units (AUs) and feature point positions (FPPs) tracked by a Kinect sensor to recognize six basic emotions (i.e., anger, disgust, fear, happiness, sadness, and surprise) and neutral. Specifically, FPPs refer to 3D feature point positions in this paper. In our system, automatic face

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China (No. 61272211) and the Six Talent Peaks Project in Jiangsu Province of China (No. DZXX-026)

ORCID: Qi-rong MAO, http://orcid.org/0000-0002-5021-9057

detection and expression feature extraction are conducted on a per-frame basis under slightly controlled conditions. Then emotions are recognized by support vector machine (SVM) classifiers using these two features respectively, and the recognition results of 30 consecutive frames are fused by the fusion algorithm based on improved emotional profiles (IEPs). The recognition results using different features are fused by the fusion algorithm based on maximum confidence. Finally, our approach provides an application for real-time ERFE. The contributions of this paper are:

1. Facial expression features (AUs and FPPs) extracted from both RGB and depth data are adopted in our approach, and they are significant for face detection and emotion recognition. Furthermore, to improve the speed of recognition, we reduce the number of FPPs and obtain better recognition accuracy.

2. We create a comprehensive application for real-time ERFE, including database creation and model training based on Kinect and Face Tracking SDK (Microsoft, USA). As far as we are aware, this is the first system that can recognize facial emotions in real-time video sequences from Kinect with multi-type features.

3. We first propose a fusion algorithm based on IEPs to fuse the recognition results of the latest 30 frames with each type of features (AUs or FPPs). Then the emotional estimation could be achieved by the fusion algorithm based on maximum confidence combining the results with both types of features.

2 Related work

ERFE is a growing active research field in computer vision because of several advantages it holds, especially in the areas of human-computer interaction. Compared to some other emotion channels, such as body actions and speech (Mao *et al.*, 2013), facial expressions have better expressive force and a larger application space.

As a pioneer of the earlier ERFE field, Ekman (1993) proposed six basic facial emotions which refer to anger, disgust, fear, happiness, sadness, and surprise. Most of the efforts in ERFE are with the purpose of recognizing these six basic emotions or a subset of them. In addition, a seventh class is often considered to model the neutral emotion.

Recently, efforts on ERFE turn to the recognition of complex and spontaneous emotions rather than those prototypical and deliberately displayed ones (Zeng *et al.*, 2009; Nicolaou *et al.*, 2011; Vinciarelli *et al.*, 2012). However, most of these works are highly sensitive to recording conditions such as poses and illumination. Fortunately, recent inventions in 3D technologies solve the problem perfectly by providing the means to measure and reconstruct 3D faces. Compared with 2D images, 3D faces hold more geometric shape data and are invariant to variables in recording conditions. For another, RGB cameras are normally superior to 3D equipment in many ways, such as scanning speed, resolution, and cost. Thus, it is necessary to use both 2D and 3D devices for ERFE.

As the software supporting 3D technologies, algorithms for 3D ERFE are mostly model-based and popular for capturing essential geometrical features. Ekman and Friesen (1978) proposed the facial action coding system (FACS), which was the first comprehensive technique for scoring all visually distinctive, observable facial movements called action units. Based on FACS and some other models, many 3D face databases containing expression data, such as Bosphorus (Savran *et al.*, 2008), ICT-3DRFE (Stratou *et al.*, 2011), and D3DFACS (Cosker *et al.*, 2011), have been created using high resolution but low scanning speed professional 3D equipment.

Common 3D capturing equipment has become even more popular in the field of ERFE, represented by Kinect sensor (Microsoft, USA) and Creative Interactive Gesture Camera (CIGC) (Creative, USA).

Kinect is a high speed sensor with the abilities of both RGB cameras and 3D scanning equipment. Typically, Kinect based face recognition systems use both color and depth data for localizing different feature points in 3D space. Compared with traditional 3D equipment, Kinect is cheap, fast in scanning speed, and compact in size. Even though the scanning accuracy is relatively low and serious noises exist, the affordable price means that there is potential for various applications. van den Hurk (2012) performed an experiment of gender determination using visual and depth imagery obtained with Kinect. Li BY *et al.* (2013) presented an algorithm using a low resolution Kinect sensor for robust face recognition under variations in poses, illumination, expression, and disguise. In Seddik *et al.* (2013),

facial expressions were recognized and mapped to a 3D face virtual model using depth and RGB data captured from Kinect. Li DX *et al.* (2013) realized a novel performance-driven real-time facial animation system with 3ds Max and Kinect. Breidt *et al.* (2011) presented a specialized 3D morphable model for facial expression analysis and synthesis with noisy depth data from Kinect. Their results demonstrate the feasibility of Kinect for robust facial analysis, and indicate its potential at a deeper level.

CIGC is another small, light-weight camera which is very similar to Kinect but specially tuned for near-range interactivity. Unlike Kinect, CIGC is made for only one user at a time, and it can get confused when multiple people or hands appear in front of it at the same time. Furthermore, the performance of facial recognition is not as accurate as it in recognizing gestures. Given the above disadvantages, together with the relatively short maximum range of just 1 m, Kinect is more suitable for real-time ERFE.

3 Recognition method for real-time emotion recognition via facial expressions (ERFE)

Based on 2D RGB and 3D depth data captured by Kinect, various features can be applied to facial expression analysis. Seddik *et al.* (2013) adopted the results of principal component analysis called 'EigenExpressions' for facial expression recognition. In Li DX *et al.* (2013), animation units and 2D key facial points were used. Also, a 3D morphable model is available (Breidt *et al.*, 2011). Facial expression analysis is normally per-frame based. However, the change of actual emotions is a dynamic procedure. To recognize the current emotion, image frames over a short period should be taken into consideration. In this paper, we adopt animation units and 3D facial points for timely estimation of the emotion of the latest 30 frames.

The architecture of our system is shown in Fig. 1. First, video sequences acquired from Kinect are input. Then face detection and feature extraction are performed on each frame. AU and FPP features are extracted by the Face Tracking SDK engine (Microsoft, USA) for model training and sent for real-time recognition. After that, a fusion algorithm based on IEPs is proposed to obtain the emotion of a video sequence by fusing the pre-recognition results of the latest 30 frames in AU and FPP channels, respectively. The final emotion is achieved by fusing the estimated emotions in both channels. Specifically, in the first stage of our process, output labels of each 7-way 1-vs-1 classification are produced and input into two memory buffers respectively, i.e., memory buffer of AUs (MB-AUs) and memory buffer of FPPs (MB-FPPs). Then emotional estimation in each channel is produced by the IEPs-based fusion algorithm for the recognition results of the latest 30 frames in the buffer. The fusion algorithm of both channels used in the final stage is based on maximum confidence. In the following, we will provide the

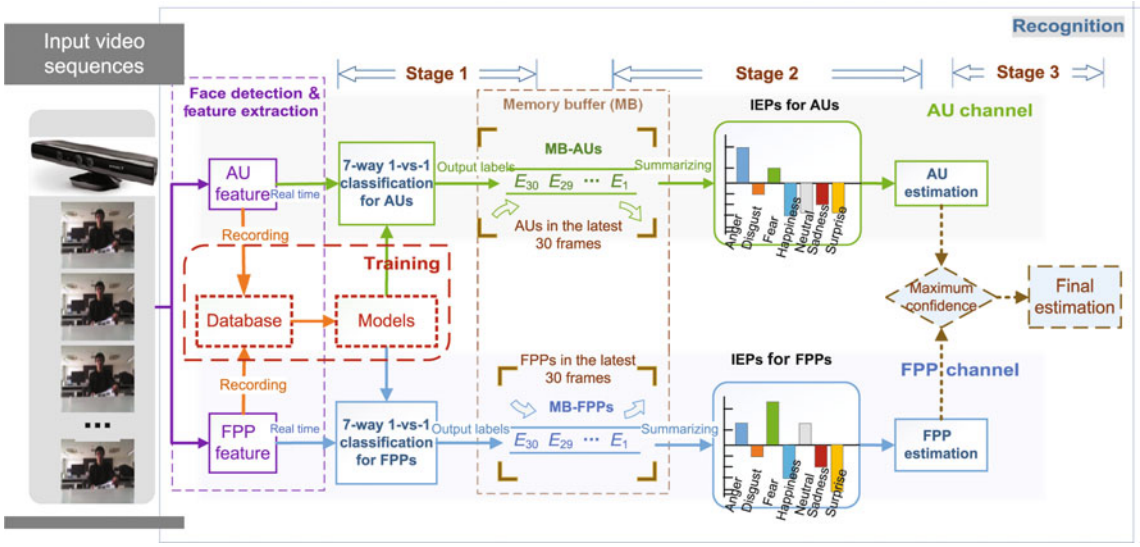

Fig. 1 Architecture for our system

pre-recognition and the fusion algorithm based on IEPs and maximum confidence. Feature extraction is discussed in Section 4.

3.1 Fusion algorithm based on improved emotional profiles in each channel

A fusion algorithm based on IEPs is proposed to fuse recognition results of the latest 30 frames in each channel, where IEPs express the confidence of each of the seven emotion-specific binary decisions (Fig. 1). Each of the sub-classifiers is trained using an emotion-specific feature set, and their outputs are stored in a memory buffer (type of queue) for the following summary.

In the pre-recognition stage (stage 1), the 7-way 1-vs-1 classification establishes an independent sub-classifier per emotion (six basic emotions and the neutral), and seven output labels representing membership in the class (denoted as '+1') or out of the class (denoted as '−1') are created to weight the confidence of each emotion. The output labels can be denoted as $L_{\mathrm{AUs}}(i,j)$ in AU channel and $L_{\mathrm{FPPs}}(i,j)$ in FPP channel respectively, where i $(i=1,2,\ldots,7)$ is the identifier corresponding to each emotion and j $(j=1,2,\ldots)$ is the frame index. For the jth frame, $C_{\mathrm{AUs}}(i,j)$ and $C_{\mathrm{FPPs}}(i,j)$ $(i=1,2,\ldots,7;\ j=1,2,\ldots)$ denote the cumulative confidence of each emotion in AU and FPP channels, respectively, and the initial value of them is set to 0. In the AUs channel, $C_{\mathrm{AUs}}(i,j)$ can be calculated as

$$C_{\mathrm{AUs}}(i,j)=\begin{cases}L_{\mathrm{AUs}}(i,j), & j=1,\\ C_{\mathrm{AUs}}(i,j-1)+L_{\mathrm{AUs}}(i,j), & 1<j\le 30,\\ C_{\mathrm{AUs}}(i,j-1)+L_{\mathrm{AUs}}(i,j) & \\ \quad -L_{\mathrm{AUs}}(i,j-30), & j>30.\end{cases} \tag{1}$$

Similarly, $C_{\mathrm{FPPs}}(i,j)$ in the FPP channel can be achieved by

$$C_{\mathrm{FPPs}}(i,j)=\begin{cases}L_{\mathrm{FPPs}}(i,j), & j=1,\\ C_{\mathrm{FPPs}}(i,j-1)+L_{\mathrm{FPPs}}(i,j), & 1<j\le 30,\\ C_{\mathrm{FPPs}}(i,j-1)+L_{\mathrm{FPPs}}(i,j) & \\ \quad -L_{\mathrm{FPPs}}(i,j-30), & j>30.\end{cases} \tag{2}$$

The memory buffer stores output labels from only the latest 30 frames, which equals a queue (whose length is 30) in principle. The set of labels comes into the queue while the queue length is not longer than 30, or else the element in front will be deleted first. After summarizing with IEPs, the confidence of each emotion will be achieved, and the emotion with the maximum confidence will be the estimated emotion in each channel.

$\mathrm{MC}_{\mathrm{AUs}}$ and $\mathrm{MC}_{\mathrm{FPPs}}$ refer to the maximum confidence in AUs and FPPs channels respectively, defined as

$$\mathrm{MC}_{\mathrm{AUs}}=\max\{C_{\mathrm{AUs}}(i,j):i=1,2,\ldots,7\}, \tag{3}$$

$$\mathrm{MC}_{\mathrm{FPPs}}=\max\{C_{\mathrm{FPPs}}(i,j):i=1,2,\ldots,7\}. \tag{4}$$

The emotion with the maximum confidence is regarded as the recognized emotion of this video sequence in one channel. Fig. 2 shows an example of pre-recognition and the IEPs-based fusion algorithm in the AU channel. We can see that the maximum confidence $C_{\mathrm{AUs}}(1)$ corresponds to the emotion of anger. Thus, anger is the current estimation in the AU channel.

3.2 Maximum confidence based fusion algorithm of both channels

The final estimation of current emotion can be obtained using the maximum confidence based fusion algorithm. The emotion with higher confidence in AUs and FPPs is the final estimation of this video sequence. The confidence corresponding to final emotional estimation can be denoted as

$$\mathrm{FE}=\max\{\mathrm{MC}_{\mathrm{AUs}},\mathrm{MC}_{\mathrm{FPPs}}\}. \tag{5}$$

Algorithm 1 lists the pseudo code of our real-time ERFE method.

4 Expression feature extraction

The Face Tracking SDK (Microsoft, USA) is part of Kinect for Windows Developer Toolkit. It can be used for tracking human faces with RGB and depth data captured from a Kinect sensor. The face tracking engine computes facial animation units and 3D positions of semantic facial feature points, which could be used to recognize emotions via facial expressions.

4.1 Animation units

The results of Face Tracking SDK can be expressed in terms of weights of six animation units,

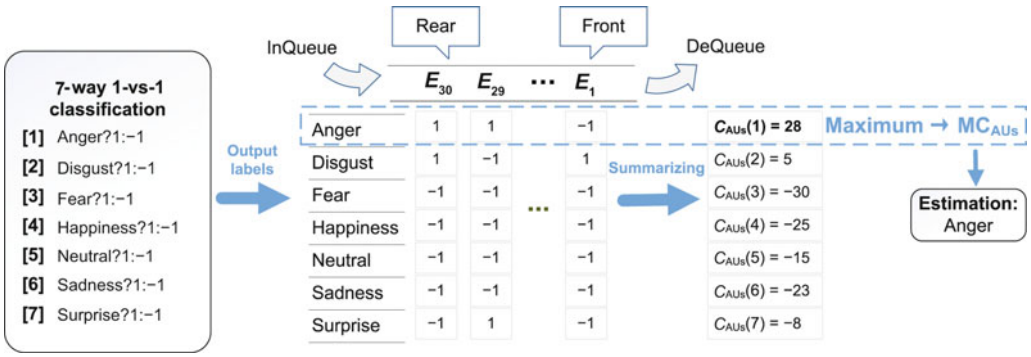

Fig. 2 Pre-recognition and fusion algorithm based on IEPs (AU channel for example). The 7-way 1-vs-1 classification in each channel contains seven sub-classifiers with the same name as corresponding emotions. Each sub-classifier has the ability of estimating whether or not input features belong to the corresponding emotion

Algorithm 1 Real-time ERFE method

$j \leftarrow 0$
while $j >= 0$ **do**
  $j \leftarrow j + 1$
  for $i \leftarrow 1$ to 7 **do**
    Get $L_{\mathrm{AUs}}(i, j)$ and $L_{\mathrm{FPPs}}(i, j)$
    Compute the emotional confidence $C_{\mathrm{AUs}}(i, j)$
    Compute the emotional confidence $C_{\mathrm{FPPs}}(i, j)$
  end for
  if $j >= 30$ **then**
    Compute confidence $\mathrm{MC}_{\mathrm{AUs}}$
    Compute confidence $\mathrm{MC}_{\mathrm{FPPs}}$
    Compute the confidence FE corresponding to final estimation
  end if
end while

which are a subset of what is defined in the Candide3 model (Ahlberg, 2001). 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. It tracks the AUs as visualized in Fig. 3.

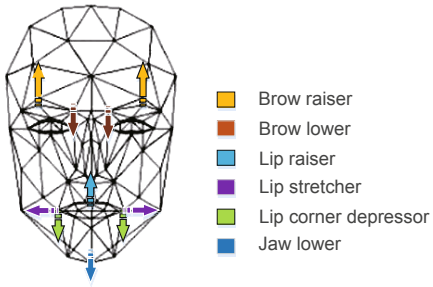

Fig. 3 Visualization of AUs. (references to color refer to the online version of this figure)

Each AU is expressed as a numeric weight varying between −1 and +1, and the neutral states of AUs are normally assigned to 0. The AUs' feature of each frame can be written in the form of a 6-element vector:

$$\bar{\boldsymbol{a}} = (A_1, A_2, A_3, A_4, A_5, A_6), \tag{6}$$

where A_1, A_2, A_3, A_4, A_5, and A_6 refer to the weights of 'lip raiser', 'jaw lower', 'lip stretcher', 'brow lower', 'lip corner depressor', and 'brow raiser', respectively. For example, $(0.3, 0.1, 0.5, 0, -0.8, 0)$ corresponds to a happy face, which means showing teeth slightly, lip corner raised and stretched partly, and the brows are in the neutral position.

4.2 Feature point positions

The Face Tracking SDK uses the Kinect coordinate system to output its 3D tracking results (Fig. 4). The measurement units are meter (m) for translation and degree (°) for rotation angles. In near mode, which is adopted in our system, the value of Z should range from 0.4 to 3.5 m.

The Face Tracking SDK tracks 121 3D feature points in the human face, and 71 typical ones are defined with descriptive names. However, not all

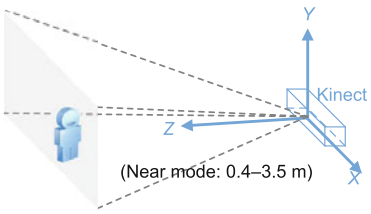

Fig. 4 Kinect coordinate system. The origin is located at the optical center of Kinect, Z-axis pointing towards a user, and Y-axis pointing up

of these points are strongly related to facial expressions. Fasel and Luettin (2003) proposed that the main areas involved in facial expression displays are the eyes, eyebrows, and the mouth. Xu *et al.* (2013) simplified the expression motion units as seven basic motion units based on eyes, eyebrows, and lips. Furthermore, given that it is not accurate to localize the details of eyes using Kinect, Xu *et al.* (2013) deleted all feature points localizing the positions of eyelids and pupils. As a result, 45 3D points on eyebrows, chin, mouth, eyes, and some other key positions were finally selected from the total 71 points to improve the recognition accuracy and speed. For each point, there is 3D information (X, Y, Z). Therefore, the FPP feature of each frame can be written in the form of a 135-element vector:

$$\bar{\boldsymbol{b}} = (X_1, Y_1, Z_1, X_2, Y_2, Z_2, \ldots, X_{45}, Y_{45}, Z_{45}), \quad (7)$$

where (X_i, Y_i, Z_i) $(i = 1, 2, \ldots, 45)$ are the 3D coordinates of each feature point.

5 Experiments and results

5.1 Datasets and experimental setup

Recently, many databases containing 3D facial expressions were captured by the Kinect sensor. Hg *et al.* (2012) compiled an RGB-D database containing 1581 RGB images and their depth data for 13 poses and 4 facial expressions by using Kinect. Cao *et al.* (2014) presented the FaceWarehouse, which is a Kinect based database of 20 3D facial expressions. As representatives of Kinect based facial expression databases, these two databases are associated only with a subset (happiness, anger, sadness) of the six basic emotions. Thus, we created the UJS Kinect emotion database (UJS-KED), which considers the six basic emotions.

Our approaches were evaluated on the UJS-KED database, the FaceWarehouse, and real-time video sequences, respectively. The UJS-KED consisted of two types of facial features (AUs and FPPs), and was recorded by 10 actors in our research group with variations in five poses ($-30°$, $-15°$, $0°$, $15°$, $30°$) and seven emotions (the six basic and neutral) using Kinect. It has over 6000 frames (640×480) of a Microsoft Kinect at 30 Hz. To prove the robustness of our recognition method, video streams from Kinect were captured in conditions of varying face appearance and illumination. The only requirement is that the subjects need to stay before the camera at appropriate distances and ensure their head yaw being less than 45°, in which condition their faces can be tracked and then facial features can be extracted successfully.

Experiments in this study were conducted on a computer with an Intel dual-core, 2.8 GHz CPU, and 4 GB RAM. For AU feature vectors (6227 in all, 5859 after deleting repeated ones), a subset which consists of 1500 instances served as the training set, yet the testing sets might be different according to specific conditions. Specifically, stratified sampling was provided to ensure the same class distribution in the subset. Furthermore, each experiment was performed 10 times, and an average of 10 experimental results was shown for analysis. Similarly, for FPP feature vectors (6227 in all), 3000 instances were selected as the training set, and testing sets differed according to specific conditions. Different numbers of instances were used (1500 vs. 3000). Note that, for complete training, 1500 instances are enough for AUs, but not for FPPs. In this study, we used 1500 instances for AUs and 3000 for FPPs.

The model of each sub-classifier can be trained using parameters (c, g, w_1, and w_{-1}) selected by the following methods and the corresponding training set.

5.2 Parameter selection

5.2.1 Parameter choice of support vector machine

Grid.py (Chang and Lin, 2011a) is a parameter selection tool for C-SVM classification using the radial basis function (RBF) kernel. It uses a cross validation technique to estimate the accuracy of each parameter combination in the specified range and helps decide the best parameters for our problem.

Parameters c and g are used in SVM. With 5-fold cross validation in selected training sets, the best parameters c and g used to train models for each sub-classifier are obtained. Table 1 shows the final values of parameters, which correspond to the best accuracy in 10 experiments.

5.2.2 Penalty factors for unbalanced data sets

Unbalanced data sets can be balanced using different misclassification penalties per class. Typically the smallest class gets the highest weight. A common

Table 1 Details of parameters c, g, and w_1

Sub-classifier	c		g		w_1	
	AUs	FPPs	AUs	FPPs	AUs	FPPs
[1] Anger	4	4	16	32	6.21	6.19
[2] Disgust	8	64	16	8	5.64	5.65
[3] Fear	16	2	16	64	6.89	6.87
[4] Happiness	32	64	32	8	5.30	5.38
[5] Neutral	4	128	32	8	5.55	5.52
[6] Sadness	4	512	32	4	6.81	6.73
[7] Surprise	128	8	4	4	5.79	5.86

approach is

$$N_1 w_1 = N_{-1} w_{-1}, \tag{8}$$

where N_1 and N_{-1} represent the numbers of positive and negative instances respectively, and w_1 and w_{-1} denote the weights of positive and negative instances respectively. LIBSVM (Chang and Lin, 2011b) allows us to do this using its -wX flags. In our experiment, the value of the penalty factor w_{-1} is fixed at 1, and thus w_1 is equal to N_{-1}/N_1. The details of w_1 are shown in Table 1.

5.2.3 FPP feature selection

To improve the time efficiency of recognition and the share of effective features, we reduced the number of feature points in FPPs from 71 to 45. Fig. 5 shows a comparison of recognition using 71 and simplified 45 feature points, and the recognition accuracy using simplified FPPs is generally higher with an accuracy over 99.8%, except for sub-classifier [5] which corresponds to the neutral state.

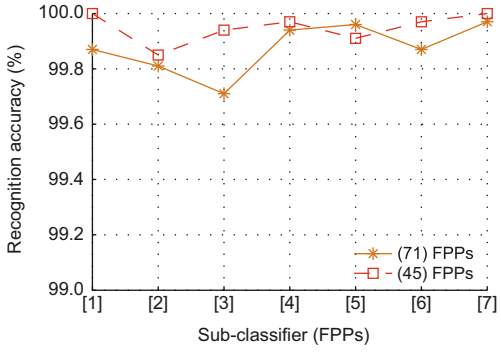

Fig. 5 Recognition accuracies using 71 feature points vs. simplified 45 feature points. [1]–[7] refer to the seven sub-classifiers in the FPP channel

5.2.4 Classifier selection

This experiment was conducted to choose the suitable classifier (1-vs-1 SVM or multi-class SVM). For 1-vs-1 classification, labels conforming to the subject of each sub-classifier were set to 1, and the others set to −1. For multi-class classification, emotions were labeled from 1 to 7. The applied classifier, LIBSVM (Chang and Lin, 2011b), is capable of both 1-vs-1 classification and multi-class classification.

AU feature data in UJS-KED was used in this comparative experiment. A subset which consists of 1500 instances served as the training set, and the remaining 4359 instances were stored as a testing set. The confusion table (Fig. 6) provides details of the performance of multi-class classification on each emotion. The values in the main diagonal give the recognition accuracy of each emotion, averaged over all poses. Fig. 7 is similar and shows the recognition accuracy of 1-vs-1 classification.

To analyze the data more explicitly, recognition accuracies of multi-class classification and 1-vs-1 classification are summarized in Fig. 8. The accuracies of 1-vs-1 classification are higher than those of multi-class classification for our dataset and tests. However, 1-vs-1 classification needs seven sub-classifiers and corresponding models, while multi-class classification needs only one classifier. In our work, the two methods can both meet the real-time requirement, and we selected 1-vs-1 classification for its higher recognition accuracy.

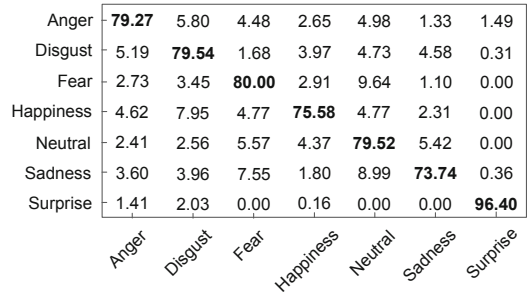

Fig. 6 Recognition accuracies (%) of multi-class classification on each emotion

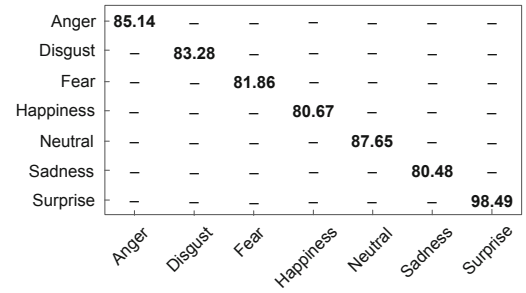

Fig. 7 Recognition accuracies (%) of 1-vs-1 classification

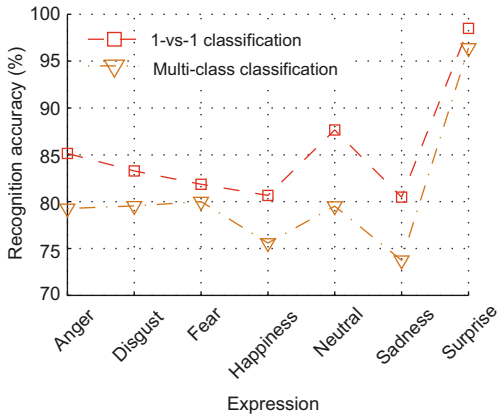

Fig. 8 Accuracy of recognition using 1-vs-1 classification and multi-class classification

5.3 Comparison of landmark localization using Kinect and RGB cameras

In this experiment, we compare the performance of recognition using Kinect and RGB cameras.

Face Tracking SDK captures FPP feature per frame. Since the highest refresh rate of Kinect is up to 30 Hz, the process of feature extraction can be completed within 0.033 s. Zhu and Ramanan (2012) achieved the state of the art in face detection and landmark localization using their approach and images captured by an RGB camera, we made several detections in three poses (0°, 15°, 30°), and these detections were repeated 50 times to obtain an average value. As a result, it took an average of 13.9 s per detection, which is over 420 times the time consumed using Face Tracking SDK (Table 2). On the other hand, Fig. 9 shows the mesh connections generated by Kinect and landmarks localized using the approach of Zhu and Ramanan (2012). By superimposing these results on RGB images, it is obvious that Kinect together with Face Tracking SDK had better accuracy on localizing facial landmarks, especially those localizing the mouth and some other details.

5.4 Performance evaluation on each frame using our database

5.4.1 Comparison of recognition accuracy using FPPs (2D) and FPPs

To examine the suitability of using depth data, we compared the accuracy of recognition using FPPs (2D) and FPPs. The FPP model of each sub-classifier was trained using 3000 135-element FPP instances, and the testing set consisted of 3227 instances. Specifically, FPPs (2D) means FPPs without the depth data. Fig. 10 shows the result of the comparison. Clearly, the dashed line is above the solid one with accuracy over 99.8%, which means that the accuracy of recognition using depth data is slightly higher on all sub-classifiers.

Table 2 Comparison of time consumed using Kinect and RGB cameras

Device and approach	Time (s)
Kinect & Face Tracking SDK	<0.033
RGB cameras & Zhu and Ramanan (2012)	≈13.9

5.4.2 Comparison of recognition accuracy using AUs and FPPs in different poses

AUs and FPPs are the two feature channels used in our recognition framework. Figs. 11a and 11b

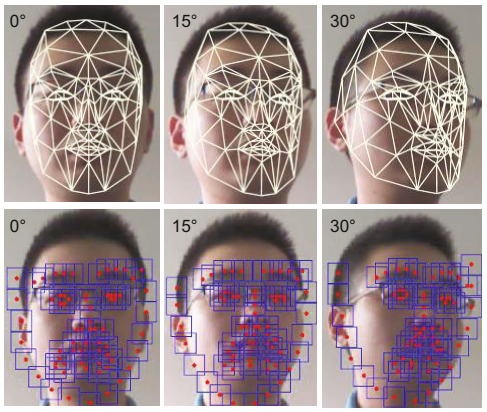

Fig. 9 Landmark localization using Kinect vs. RGB cameras. In the top row, mesh connections are generated by Kinect together with Face Tracking SDK. In the bottom row, landmarks are localized by the approach of Zhu and Ramanan (2012) using RGB images captured from RGB cameras. Detections in the same column use the same face

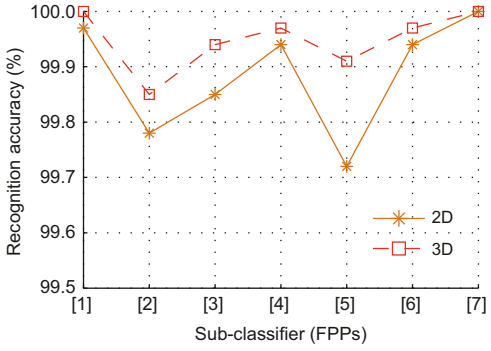

Fig. 10 Accuracy of recognition using FPPs (2D) and FPPs

present the recognition accuracies of sub-classifiers using AUs and FPPs in different poses (0°, ±15°, ±30°) over the whole UJS-KED. Fig. 11c shows the overall accuracies in different poses, which were almost equivalent. The average accuracies using AUs and FPPs were 95.22% and 99.96%, respectively. Fig. 12 shows the recognition accuracy of sub-classifiers averaged over all poses using AUs and FPPs over a testing set which consisted of 3227 instances. Clearly, the accuracy using AUs ranged from 92% to 94% except for the sub-classifier corresponding to surprise, while the accuracy using FPPs was always over 99%.

The features of AUs cannot be replaced by those of FPPs, even though the accuracy of using FPPs is much higher than that of using AUs. This is

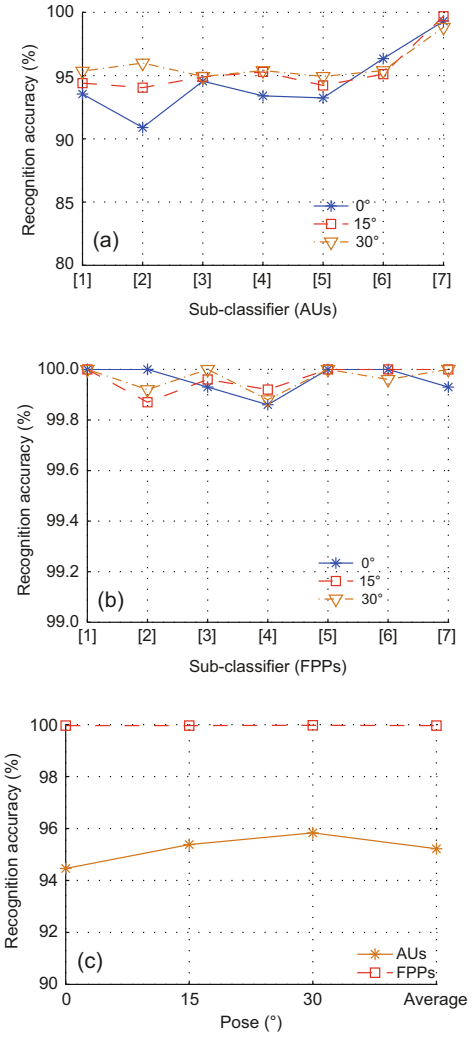

Fig. 11 Recognition accuracies using AUs (a) and FPPs (b) in different poses, and the overall accuracies in different poses (c)

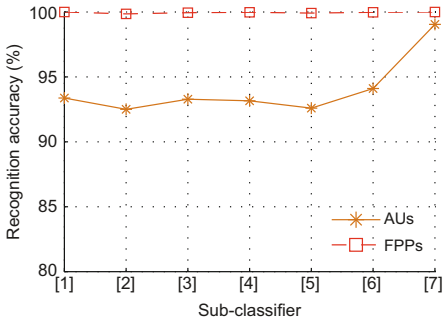

Fig. 12 Recognition accuracies of sub-classifiers using AUs and FPPs

because AUs can reach high accuracy only with 6-element features, while FPPs need 135-element features. Furthermore, as relative features are independent of poses and distances (PDs), AUs have an outstanding adaptability to different PDs. To reach similar adaptability, a training set with comprehensive PDs is necessary for FPPs, or the performance may be unsatisfactory in certain PDs. PDs in the training set and testing set of our dataset are almost the same, and this condition contributes to the high recognition accuracy using FPPs.

5.5 Performance evaluation using the FaceWarehouse database

FaceWarehouse (Cao *et al.*, 2014) is a database of 3D facial expression models. With Kinect, raw data sets from 150 individuals were captured. For each person, the neutral expression and 19 other expressions were captured, such as mouth-open, smile, and anger. For each expression data, 74 facial feature points (FPs) were localized and refined with user interaction. These 19 expressions corresponded only to three basic emotions (anger, sadness, and happiness). Thus, experiments on the FaceWarehouse database were aimed to recognize the neutral emotion and these three basic emotions.

To track AU and FPP features from RGB-D raw data in FaceWarehouse, some changes are necessary in Face Tracking SDK, to enable the Face Tracking engine to track features from specified buffers rather than the Kinect sensor. In addition, the skeleton data is not available from FaceWarehouse, so faces can be tracked using only color and depth data, which may affect the performance of feature extraction. As a result, only 347 of 600 instances were successfully tracked.

Fig. 13 presents the recognition accuracy of sub-

classifiers using the FaceWarehouse database. AU and FPP features were tracked using our approach, and FPs refer to feature points provided by FaceWarehouse. For the lack of a training set, models were not trained completely, which leads to some performance degradation in FaceWarehouse, but the information we needed was still presented. The accuracies using these three features were similar. However, compared with 45 automatically localized feature points used in FPPs, feature points in FPs were much more (74) and were refined with user interaction. Besides, the adopted instances of AUs and FPPs were much fewer than FPs (347 vs. 600), which shows the advantage of AUs and FPPs.

5.6 Real-time ERFE system

The real-time ERFE system accepts Kinect color and depth data as input. This system sets feature collection and real-time recognition all in one. Features can be easily captured and stored per frame, and emotions can be estimated by the fusion algorithm based on IEPs and maximum confidence.

For the defect (low resolution) of Kinect, the tracking quality may be affected by the image quality of these input frames (i.e., darker or fuzzier frames track worse than brighter or sharper frames). Also, larger or closer faces were tracked better than smaller faces.

Fig. 14 presents the main window design of our application. Three representative, simplified operating examples (0°, 15°, 30°), on continuous 30 frames are provided in Fig. 15.

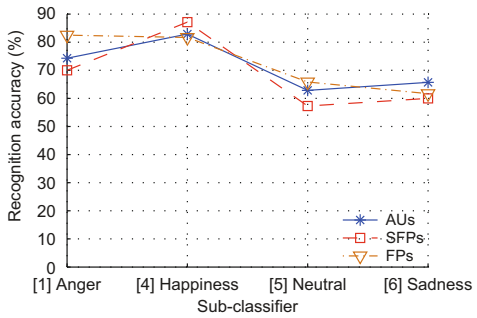

Fig. 13 Recognition accuracy of sub-classifiers using the FaceWarehouse database

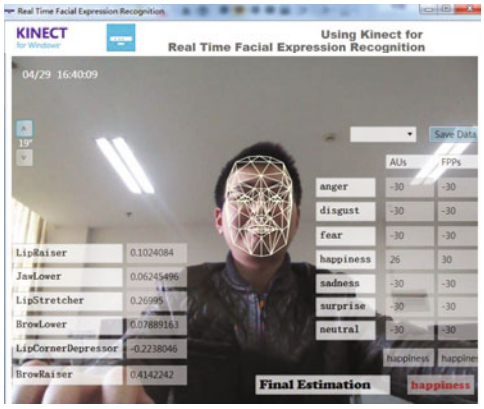

Fig. 14 Main application window. Real-time AU features are shown on the bottom left and a fusion algorithm based on IEPs and maximum confidence is displayed on the right

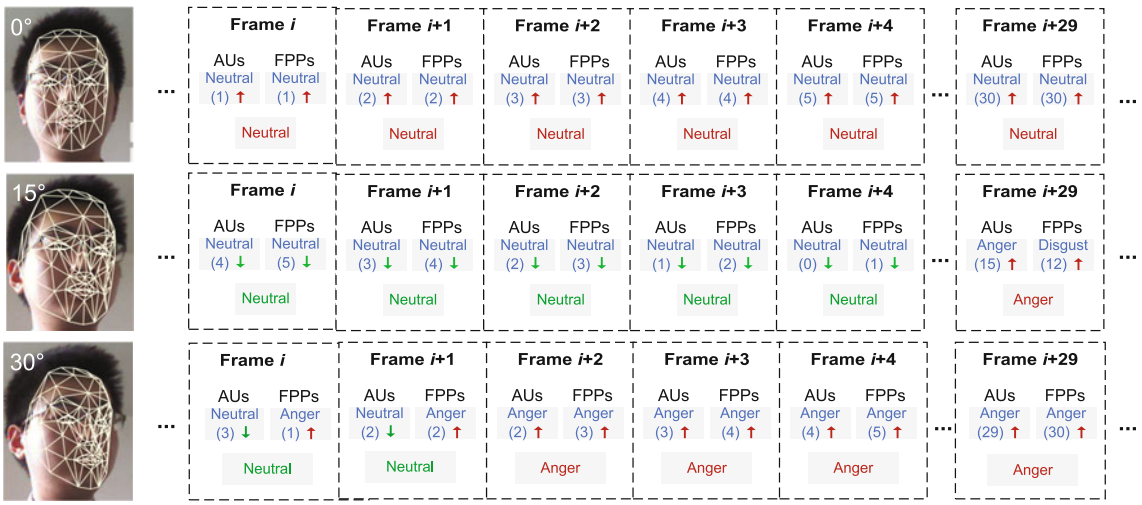

Fig. 15 Simplified operation examples on continuous 30 frames (frames i–$(i+29)$). The first row is a sample for neutral emotion in the pose of 0°. The cumulative confidences of neutral keep rising in the AU and FPP channels, respectively. The second and third rows are samples for emotion changing from neutral to anger, where the cumulative confidence of neutral keeps dropping and confidence of anger keeps rising

6 Summary and discussion

In this paper, we proposed a real-time ERFE approach based on both 2D and 3D features captured by Kinect. A fusion algorithm based on IEPs and maximum confidence functions is the kernel of our approach. Real-time AU and FPP features, together with this fusion algorithm, enable us to recognize real-time emotions via facial expressions. In the future, we plan to expand the UJS-KED, so that our model can be trained more comprehensively. Furthermore, a new method should be proposed to disengage from the limitations of Face Tracking SDK.

References

Ahlberg, J., 2001. Candide-3—an Updated Parameterised Face. Technical Report.

Breidt, M., Biilthoff, H.H., Curio, C., 2011. Robust semantic analysis by synthesis of 3D facial motion. Proc. IEEE Int. Conf. on Automatic Face & Gesture Recognition and Workshops, p.713-719. [doi:10.1109/FG.2011.5771336]

Cao, C., Weng, Y.L., Zhou, S., *et al.*, 2014. FaceWarehouse: a 3D facial expression database for visual computing. *IEEE Trans. Visual. Comput. Graph.*, **20**(3):413-425. [doi:10.1109/TVCG.2013.249]

Chang, C.C., Lin, C.J., 2011a. LIBSVM: a Library for Support Vector Machines. Available from http://www.csie.ntu.edu.tw/∼cjlin/libsvm/index.html.

Chang, C.C., Lin, C.J., 2011b. LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, **2**(3):27. [doi:10.1145/1961189.1961199]

Cosker, D., Krumhuber, E., Hilton, A., 2011. A FACS valid 3D dynamic action unit database with applications to 3D dynamic morphable facial modeling. Proc. IEEE Int. Conf. on Computer Vision, p.2296-2303. [doi:10.1109/ICCV.2011.6126510]

Ekman, P., 1993. Facial expression and emotion. *Am. Psychol.*, **48**(4):384-392. [doi:10.1037/0003-066X.48.4.384]

Ekman, P., Friesen, W.V., 1978. Facial action coding system: a technique for the measurement of facial movement. Palo Alto.

Fasel, B., Luettin, J., 2003. Automatic facial expression analysis: a survey. *Patt. Recog.*, **36**(1):259-275. [doi:10.1016/S0031-3203(02)00052-3]

Hg, R.I., Jasek, P., Rofidal, C., *et al.*, 2012. An RGB-D database using Microsoft's Kinect for windows for face detection. Proc. 8th Int. Conf. on Signal Image Technology and Internet Based Systems, p.42-46. [doi:10.1109/SITIS.2012.17]

Li, B.Y., Mian, A.S., Liu, W.Q., *et al.*, 2013. Using Kinect for face recognition under varying poses, expressions, illumination and disguise. Proc. IEEE Workshop on Applications of Computer Vision, p.186-192. [doi:10.1109/WACV.2013.6475017]

Li, D.X., Sun, C., Hu, F.Q., *et al.*, 2013. Real-time performance-driven facial animation with 3ds Max and Kinect. Proc. 3rd Int. Conf. on Consumer Electronics, Communications and Networks, p.473-476. [doi:10.1109/CECNet.2013.6703372]

Ma, X.H., Tan, Y.Q., Zheng, G.M., 2013. A fast classification scheme and its application to face recognition. *J. Zhejiang Univ.-Sci. C (Comput. & Electron.)*, **14**(7):561-572. [doi:10.1631/jzus.CIDE1309]

Mao, Q.R., Zhao, X.L., Huang, Z.W., *et al.*, 2013. Speaker-independent speech emotion recognition by fusion of functional and accompanying paralanguage features. *J. Zhejiang Univ.-Sci. C (Comput. & Electron.)*, **14**(7):573-582. [doi:10.1631/jzus.CIDE1310]

Nicolaou, M.A., Gunes, H., Pantic, M., 2011. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Trans. Affect. Comput.*, **2**(2):92-105. [doi:10.1109/T-AFFC.2011.9]

Savran, A., Alyüz, N., Dibeklioğlu, H., *et al.*, 2008. Bosphorus database for 3D face analysis. Proc. 1st European Workshop on Biometrics and Identity Management, p.47-56. [doi:10.1007/978-3-540-89991-4_6]

Seddik, B., Maâmatou, H., Gazzah, S., *et al.*, 2013. Unsupervised facial expressions recognition and avatar reconstruction from Kinect. Proc. 10th Int. Multi-conf. on Systems, Signals & Devices, p.1-6. [doi:10.1109/SSD.2013.6564032]

Stratou, G., Ghosh, A., Debevec, P., *et al.*, 2011. Effect of illumination on automatic expression recognition: a novel 3D relightable facial database. Proc. IEEE Int. Conf. on Automatic Face & Gesture Recognition and Workshops, p.611-618. [doi:10.1109/FG.2011.5771467]

van den Hurk, Y., 2012. Gender Classification with Visual and Depth Images. MS Thesis, Tilburg University, the Netherlands.

Vinciarelli, A., Pantic, M., Heylen, D., *et al.*, 2012. Bridging the gap between social animal and unsocial machine: a survey of social signal processing. *IEEE Trans. Affect. Comput.*, **3**(1):69-87. [doi:10.1109/T-AFFC.2011.27]

Xu, S.B., Ma, G.H., Meng, W.L., *et al.*, 2013. Statistical learning based facial animation. *J. Zhejiang Univ.-Sci. C (Comput. & Electron.)*, **14**(7):542-550. [doi:10.1631/jzus.CIDE1307]

Zeng, Z., Pantic, M., Roisman, G.I., *et al.*, 2009. A survey of affect recognition methods: audio, visual, and spontaneous expressions. *IEEE Trans. Patt. Anal. Mach. Intell.*, **31**(1):39-58. [doi:10.1109/TPAMI.2008.52]

Zhu, X.X., Ramanan, D., 2012. Face detection, pose estimation, and landmark localization in the wild. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, p.2879-2886. [doi:10.1109/CVPR.2012.6248014]