Cognitive Computation (2019) 11:587–597
https://doi.org/10.1007/s12559-019-09654-y

Facial Expression Recognition Based on a Hybrid Model Combining Deep and Shallow Features

Xiao Sun[1] · **Man Lv**[1]

Received: 10 July 2017 / Accepted: 17 May 2019 / Published online: 25 May 2019

Abstract

Facial expression recognition plays an important role in the field involving human-computer interactions. Given the wide use of convolutional neural networks or other neural network models in automatic image classification systems, high-level features can be automatically learned by hierarchical neural networks. However, the training of CNNs requires large amounts of training data to permit adequate generalization. The traditional scale-invariant feature transform (SIFT) does not need large learning samples to obtain features. In this paper, we proposed a feature extraction method for use in the facial expressions recognition from a single image frame. The hybrid features use a combination of SIFT and deep learning features of different levels extracted from a CNN model. The combined features are adopted to classify expressions using support vector machines. The performance of proposed method is tested using the publicly available extended Cohn-Kanade (CK+) database. To evaluate the generalization ability of our method, several experiments are designed and carried out in a cross-database environment. Compared with the 76.57% accuracy obtained using SIFT-bag of features (BoF) features and the 92.87% accuracy obtained using CNN features, we achieve a FER accuracy of 94.82% using the proposed hybrid SIFT-CNN features. The results of additional cross-database experiments also demonstrate the considerable potential of combining shallow features with deep learning features, and these results are more promising than state-of-the-art models. Combining shallow and deep learning features is effective when the training data are not sufficient to obtain a deep model with considerable generalization ability.

Keywords Facial expression recognition · Convolutional neural networks · Scale-invariant feature transform · Deep learning features · Support vector machines

Introduction

Recently, human facial expression recognition (FER) has attracted significant attention. Automated FER is presently used in many applications in the field of human-computer interaction. Researchers have proposed some methods to describe facial behaviors, in which the facial action coding system (FACS) [1] is the most popular one. FACS remains a standard method of systematically categorizing the physical expression of emotions. Some of the recent works have already gain good performances [2, 3]. However, facial expressions recognition in the wild is still a challenging problem [4–6].

Algorithms for FER usually involve feature extraction and classification. Feature extraction is a very important step in most FER systems, and this step aims to represent facial images as feature vectors [7]. The extraction of features from input data significantly influences the final classification accuracy. For face representation, most algorithms and methods adopt different hand-crafted features, such as Gabor wavelet coefficients [8], histograms of LBPs [9], HOGs [10], SIFT descriptors [11], or a combination of these features [12].

Despite the success of traditional shallow features and generic image descriptors, recent developments in convolutional neural networks (CNNs) has demonstrated significant success using automatically learned features. The CNN technique has recently yielded impressive

✉ Xiao Sun
sunx@hfut.edu.cn

Man Lv
lvxman@foxmail.com

[1] School of Computer and Information, Hefei University of Technology, No. 193, TunXi Road, BaoHe District, Hefei, China

performance across a wide variety of competitive tasks and challenges [13, 14]. Unlike traditional approaches, neural networks in deep learning methods are used to extract high-level features from training data [15]. CNNs can be used as a classifier and also represent an effective method of learning deep features automatically from image data.

CNNs have the ability to learn high-level features of images, other than traditional hand-designed features used in image classification. However, CNNs training requires many labeled training samples. This usually prevents the wild adoption of CNNs in FER because of the limitations of available databases [16, 17]. Therefore, overfitting becomes a serious problem because of the limited training size. While, SIFT and other hand-crafted methods do not require extensive databases for generalization. Meanwhile, such hand-designed methods are limited in that their modeling capacities.

In this paper, we proposed novel hybrid features extraction framework called CNN-SIFT that take advantage of the synergy between these two superior groups of features. First, deep learning features are extracted using a CNN and SIFT bag of features (BoF), respectively. Support vector machines are then trained and used to classify expressions using the proposed hybrid features. The method is evaluated using the well-known extended Cohn-Kanade facial expression database [18]. Moreover, we also perform experiments with cross-database configuration (i.e., training is performed with CK+, and testing is performed using JAFFE [19] and MMI [20]) to test the generalization of our proposed method; generalization is crucial in real-world applications. The experiment results show that the proposed method demonstrates considerable skill in capturing informative features from images showing different facial expressions.

Related Work

Many algorithm have been adopted to recognize facial expressions, and significant progress has recently been made. Before the appearance of deep learning method, most traditional feature extraction method extracted hand-crafted features locally from facial images. Shan [9] evaluated facial features based on local statistical features called LBPs. Experiments have shown that LBP features can perform better over a range of facial images and can yield good performance. Dahmane et al. [10] used dynamic grid-based HOGs for emotion recognition. The results show that these features perform better than the static uniform LBP implementation. Donato et al. [21] explored different methods of representing facial images for use in facial action recognition; these techniques included principal component analysis, LBPs and local schemes. Lowe [11] proposed SIFT, a local feature descriptor and is robust to substantial affine distortion. The SIFT features are wildly used in many image processing applications and research works. For image classification, researchers introduced the BOW (bag of words) method used in the field of character recognition into image representation, yielding the BoF approach [22]. Combining the BoF extracted from SIFT and SVM classifiers has been successfully implemented in various classification tasks, including classifying emotions using facial images [23].

Due to increases in available computational power and training data, neural network-based machine learning techniques have recently seen a resurgence in popularity. In particular, CNNs have recently become the most popular method for FER tasks [13, 15]. Lopes [13] proposed a solution for FER combining CNNs and image pre-processing method, and the proposed method achieved a high accuracy rate when applied to the CK+ database. Mollahosseini [15] proposed a deep learning model to address the problem of FER across seven public datasets. Their research work is the first to adopt the inception layer architecture in neural network for FER.

Convolutional network proposed by Krizhevsky et al. [14] achieved competition champion status when applied to the ImageNet dataset, which consists of more than one million labeled high-resolution images. Features extracted directly from AlexNet can be used as a powerful deep descriptor in a variety of recognition tasks [24, 25]. Chen et al. [24] used CNNs to learn image representations automatically through fully connected layers. Their experimental results show that deep learning-based methods outperform traditional methods based on hand-crafted features in classifying images according to emotions. Zhang [25] extracted a CK+ image feature vector from the pre-trained AlexNet and combined it with a simple SVM-based classifier. The results show that CNNs can extract the essential features of images that other methods cannot surpass. These examples show that CNNs can be used as a tool in extracting deep learning features.

In the classification step, many algorithms have been employed in static FER systems, such as SVM [26], KNN [27], and ANN-based approaches [28]. SVM is a discriminative algorithm proposed according to the principle of minimizing the Vapnik structural risk [29].

Proposed Method

Scale-Invariant Feature Transform

In this paper, we adopt SIFT descriptors as shallow feature representations of images. We chose SIFT features as the shallow features for combination instead of Hog

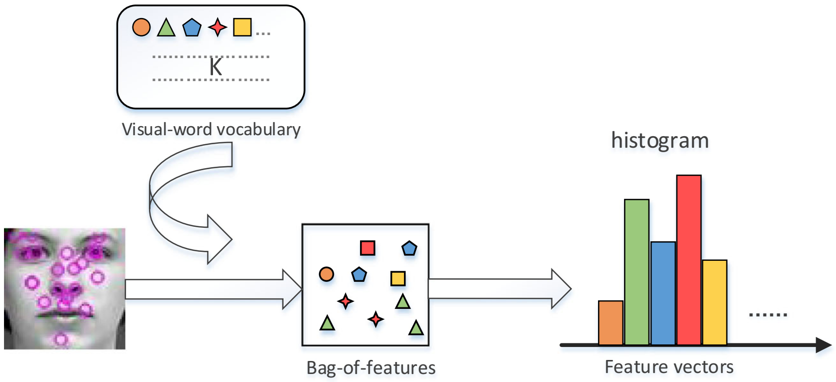

Fig. 1 Framework of an image built using BoF with SIFT

or LBP. When we use some feature extraction method to extract features from facial image, two points are usually considered: One point is that the algorithm can automatically identify potential points of interest; the other is that these points of interests must be unique, that is, the points of interest cannot change with the scale and rotation of the face or the illumination of the image. SIFT algorithm recognizes the points of interests of an image in the same-scale space through Gaussian differential function, and assigns descriptions of multiple directions to each points of interests based on the local gradient direction of the image. By considering the former characteristics, the above two points are well satisfied, so that the SIFT is chosen instead of HOG or LBP.

The area around the key points is divided into 4 × 4 subregions with 8 orientation bins each. Therefore, SIFT is defined as a $4 \times 4 \times 8 = 128$-dimension vector.

SIFT features have high dimensionality, and for image classification, the dimensionality of the feature space needs to be reduced. BoF is a visual descriptor that is frequently used in classifying visual data. There are four main steps in building a BoF with SIFT features in image classification [30].

- The image features are detected and described using SIFT.
- The descriptor is grouped to the set of clusters (vocabulary) using the K-means vector quantization algorithm. The K centers define the K ”visual words.”
- A BoF is constructed that represents the amount of features for each group or cluster.
- Classification is performed by training adopting the BoF as feature vectors, and each image is categorized.

The final representation of the BoF is a frequency histogram that represents the counts of the occurrences of each visual word in the image. Figure 1 is an example. The input image is represented as a vector of K. The size of the vocabulary K is specified as 300, 400, 500, and 600. The best value of K is selected by cross-validating the average recognition accuracy.

Deep Learning Features by CNN Model

CNNs were first introduced by Lecun et al. [31]. The main advantage of CNN-based models is that they employ raw images, rather than hand-crafted features, as inputs. CNNs could learn the deep features that best represents the desired classification [32].

CNN Architecture

The layer structure of the CNN adopted in this paper is shown in Fig. 2. The weights of the CNN are locally connected and shared spatially. Therefore, the CNN includes relatively few connections and parameters and is comparatively easy to train. The network receives a 64 × 64 grayscale image as input and outputs the confidence of each expression.

We use 5 × 5 filters with stride 1 in the first convolutional layer, and 3 × 3 filters are used in the second and third convolutional layers. Furthermore, 8, 16, and 32 filters are adopted for the first, second, and third convolutional layers, respectively. ReLu activation function is used for each convolutional layer to prevent the vanishing gradient problem. For pooling layers, we use max pooling and make the filter size and stride to 2 × 2 and 1, respectively.

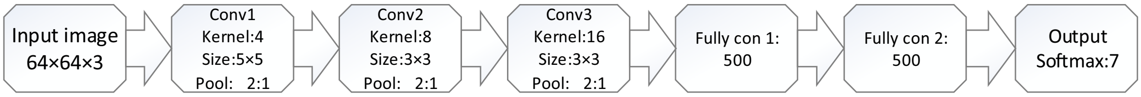

Fig. 2 CNN structure design

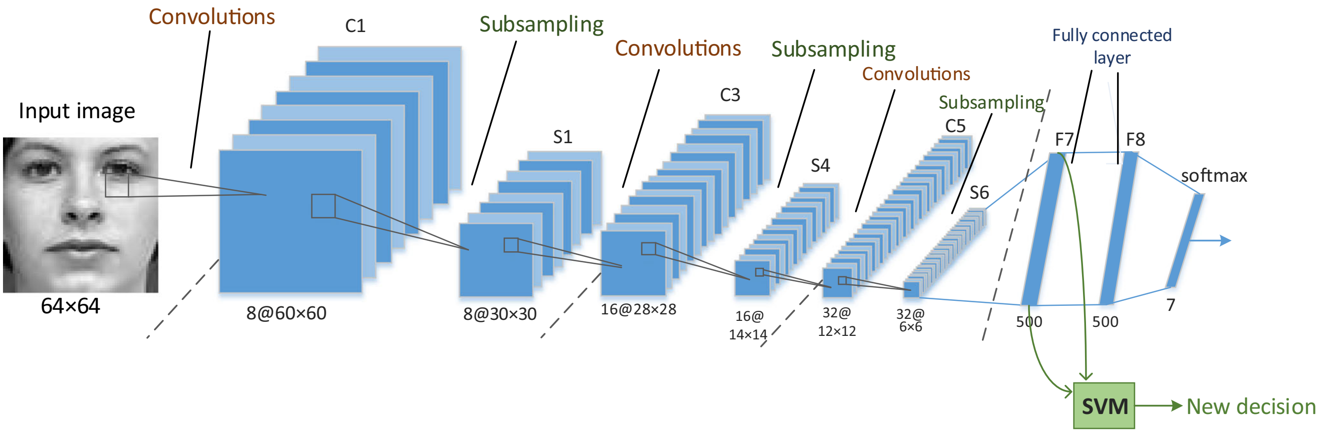

Fig. 3 Classification process based on deep learning feature extracted from the CNN model

The pooling layer adopts a down-sampling operation. This operation reduces the input image size by half, resulting in data with smaller dimensions. The CNN includes two fully connected layers, which have 500 nodes each. The number of nodes is determined empirically. Furthermore, these two layers are regularized using dropout to avoid overfitting.

Deep Learning Features

CNNs can be trained to learn higher-level features by using neural networks [15]. To achieve this goal, we extract the outputs from CNN's fully connected layer as a new feature vector. We then use them as training inputs for a simple classifier, such as an SVM-based classifier. Figure 3 shows the structure of our FER system based on the features extracted from the CNN. The output of every layer is accessible. We exam the effectiveness of using the first and the second fully connected layers respectively, which are referred to as layer 7 and layer 8, respectively. The dimensions of both layers are 500. We also assess how the performance compares to the classification accuracy of the original CNN. In Fig. 4, by visualizing the output of Conv1/Conv2/Conv3, it can be found that the output becomes more abstract as the number of convolution layers increases. However, it can also be found that the highlighted positions (the focus area) in the output of convolution were concentrated in the area of eyes and nose. This is consistent with the way humans judge expressions. Therefore, CNN has the ability to extract high-level semantics of images for facial expression recognition.

Hybrid CNN-SIFT Feature

Hybrid CNN-SIFT features are proposed to fully combine the advantages of SIFT and CNN, which can extract superficial invariant features of face images and high-level semantic features. In contrast to hand-crafted shallow features, CNN is fully data based end to end method. However, the CNN does not consider (at least as far as we have been able to determine) manually extracted features, such as the local feature descriptor SIFT, even though these points of interest may contribute substantially to the classification of emotions.

Although many experiments show that deep learning representations perform better in FER tasks, there is still considerable room for improvement. Much research has been carried out to find more powerful feature descriptions to improve the classification performance. Therefore, we speculate that the use of shallow features can augment the deep learning features in image representation in our FER task. For each image, we extract its BoF descriptor

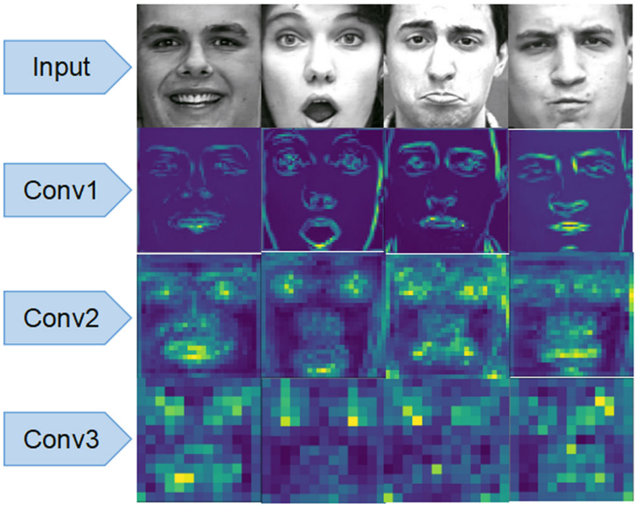

Fig. 4 Visualization of the different convolutional features extracted by our CNN model

based on SIFT and its deep learning features based on the CNN model, respectively. We then present a novel integrated feature called CNN-SIFT, which connects the SIFT-BoF features with deep learning features to produce new features. The size of the extracted CNN feature is 500. The dimension of the SIFT-BoF feature is the size of the bag vocabulary K. Therefore, the hybrid CNN-SIFT feature vector has the dimension of 500 plus K. Finally, we normalize the feature vector to (0,1) prior to the classification phase.

Classification

After obtaining the representation, the next step involves classifying the features extracted from the facial images. In our experiment, the SVM model is chosen because it can successfully address problems of spurious regression and pattern recognition [33]. SVMs [26] with different kernel functions can transfer a nonlinearly separable problem into a linearly separable problem. In our work, we use a RBF kernel in the following:

$$\mathrm{RBF}\left(x_i, x_j\right) = \exp\left(-\gamma \left\|x_i - x_j\right\|^2\right), \gamma > 0 \qquad (1)$$

SVMs were firstly proposed to solve two-class problems. Later, two extensions to this method, the one-against-all and one-against-one approaches, were sugggested to solve multiclass problems. We use LIBSVM [34] package in our experiments.

Experiment and Results

To test the effectiveness and performance of the proposed hybrid CNN-SIFT features extraction method, we conduct experiments by adopting public facial expression datasets, such as the CK+ database. Furthermore, to test the generalization performance of our method, we also perform a cross-database experiment on the Japanese Female Facial Expressions (JAFFE) and MMI databases. We use the features discussed previously to classify images from the databases according to the seven basic emotions.

Databases

The CK+ database includes 593 sequences acted out by 123 participants [18]. The subjects include 15% African Americans and 3% Asians and Latinos. Of the 593 sequences, only 327 have corresponding emotion labels, whereas the remainder are unlabeled. The labels correspond to seven expressions: happiness, anger, sadness, contempt, disgust, fear, and surprise. We select final frame of each sequence with peak expression, and this procedure results in 503 images. All of the images are set to 640 × 480 pixel arrays. We also convert all of the images to grayscale. Some sample images from CK+ database are shown in Fig. 5.

In cross-database experiment, the proposed model is trained and validated on CK+, and the accuracy of the evaluation is calculated using the JAFFE and MMI. The JAFFE dataset [19] consists of 213 images of different facial expressions. For every subject, we select one example of each of the seven expressions in our test set. All the images are set to 256 × 256 pixels, and the grayscale values have 8-bit precision. Figure 5 shows some sample images from JAFFE.

MMI database [20] contains more than 20 subjects of both genders range in age from 19 to 62. Some examples from MMI are shown in Fig. 5. The original images in MMI are color images, which were converted into 8-bit grayscale images in our experiments. We select 50 image sequences from total dataset from MMI. The neutral face and one peak

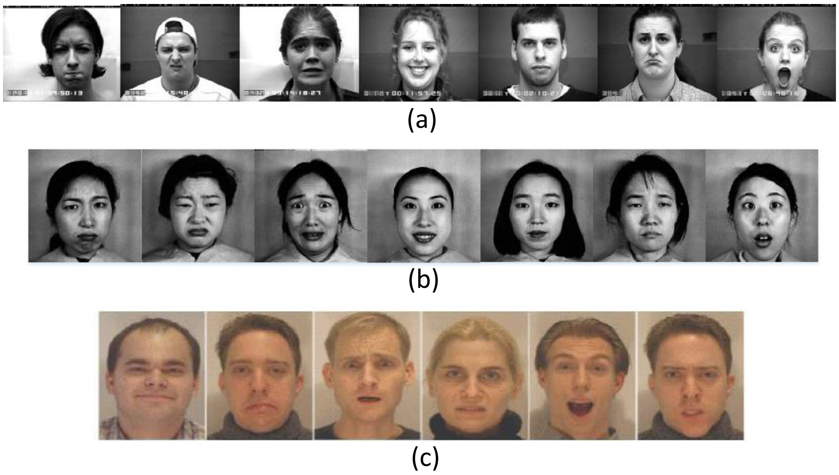

Fig. 5 **a** Examples of seven basic emotions in CK+. **b** Examples of seven basic emotions in JAFFE. **c** Examples of six prototypic emotions in MMI

frame in each sequence (100 images in total) are used in the test set.

Image Preprocessing

The pre-processing of facial images has been shown to be a key step in obtaining high recognition performance. The images in the databases vary in rotation and size. These variations are not corresponding to the facial expressions. Lopes [35] reported that combining multiple pre-processing step could increase the performance.

To address this problem, pre-processing methods that include image cropping, facial spatial normalization, and facial intensity normalization are adopted in our experiment. As shown in Fig. 6, the original images contain substantial background information that is irrelevant to expression recognition. We detect and crop the faces from the original images based on Haar-like features and the AdaBoost classifier and resize the faces image to 70 × 70. We then use the OpenCV library to carry out face alignment. All the face images are adjusted to a horizontal direction according to the eye location. To remove the effects of variations in brightness and contrast, intensity normalization is applied [36].

Data Augmentation

Deep learning with CNNs always needs a large number of training set to obtain good classification results. Given that our database is relatively small, overfitting becomes a serious problem. The method that is most commonly used to reduce overfitting is enlarging the database using artificial label-preserving transformations. Therefore, before training the CNN model, we augment the database with various transformations to generate various small changes in appearance and orientation.

Simple data augmentation methods are used to make the network more robust to the various situations. For each 70 × 70 training image, we extract five 64 × 64 patches from the four corners and one from resizing an original 70 × 70 image. We also generate their horizontal reflections, which increases the size of the training examples by a factor of 10. In every experiment, the validation and test sets are not augmented. During the test phase, we identify the test images directly.

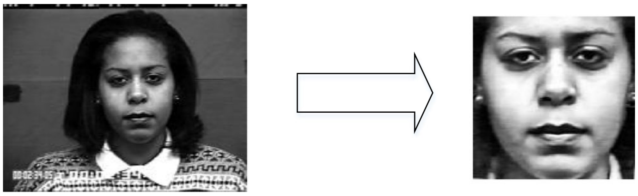

Fig. 6 Original (left) and pre-processing image (right)

Table 1 Results obtained using different values of the vocabulary size *K* in SIFT

Bag of words	Accuracy (%)
$K = 300$	72.28
K = 400	70.63
$K = 500$	76.57
$K = 600$	74.96

Results

Experiments Using the CK+ Database

To perform evaluations in CK+ dataset, CK+ database is split into eight groups that do not include the same subjects. Each group includes 12 subjects. Six groups are used for the training set, one group is used for cross-validation, and one group is used for testing. One group is selected for use as the test set each time, and our experiment is repeated eight times. Finally, the average recognition rate for the eight repetitions is taken as the final recognition performance.

We use Keras library to implement CNN model, and the SVM package LIBSVM is used in our experiments. SIFT-BoF feature extraction is carried out using OpenCV library [37].

The number of vocabularies used to build the clustering model has a substantial impact on the classification accuracy. We want to investigate how many clusters will fit on the grayscale images using SIFT. The number of clusters is set to 300, 400, 500, and 600. We use training and cross-validation data to obtain the set of BoFs, which is then used to create the BoF descriptor for all images. We train SVM on training data and validation data to identify the good matches (c, g). The results are summarized in Table 1.

As shown in Table 1, the best accuracy is obtained when the vocabulary size *K* is set to 500. The experiments show that average accuracy is 76.57%. Therefore, we select the SIFT-BoF features with $K = 500$ as our final shallow features. The confusion matrices of recognition performed using the RBF kernel and $K = 500$ are shown in Table 2, where the correct labels are arranged along the vertical axis and the predicted labels are arranged along the horizontal axis. The emotions with labels of happy, neutral, and surprise can be recognized with high accuracy, while the recognition performance for sadness and fear is lower.

We randomly choose eight groups to train our neural networks and one group for validation each time. We adopts the learning rate of $c = 0.01$ and decrease it by ten.

Table 2 Facial expression recognition by SIFT-BoF ($K = 500$) features

	Anger	Disgust	Fear	Happy	Neutral	Sadness	Surprise
Anger	51.7%	23.3%	0	0	4.6%	20.4%	0
Disgust	9.7%	82.6%	0	0	7.7%	0	0
Fear	8.3%	0	80.6%	6.9%	0	0	4.2%
Happy	0	0	12.5%	87.5%	0	0	0
Neutral	16.0%	0	0	0	78.9%	5.1%	0
Sadness	16.7%	6.3%	5.7%	0	10.8%	60.5%	0
Sadness	0	0	4.4%	0	0	9.1%	86.5%

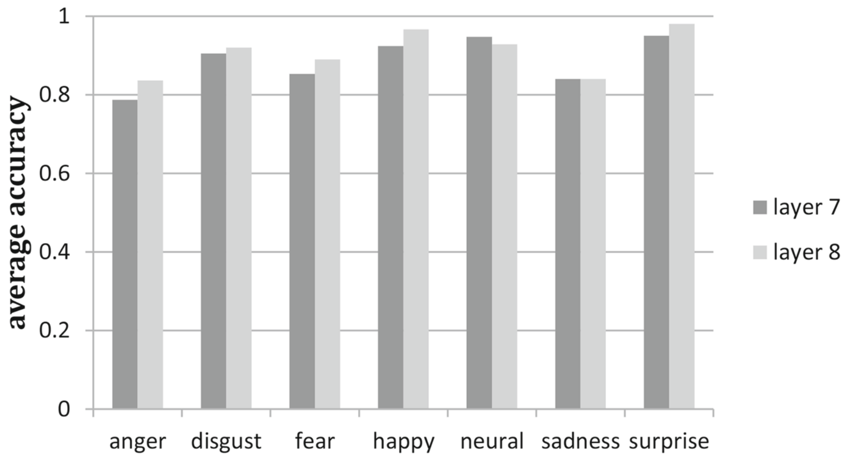

Fig. 7 Average recognition accuracy for the different categories present in the dataset. The plot compares the performance of the features extracted from the first and second fully connected layers

Table 3 Confusion matrix for the CNN using the features from layer 8

	Anger	Disgust	Fear	Happy	Neutral	Sadness	Surprise
Anger	83.7%	4%	0	0	0	10.7%	1.6%
Disgust	4.0%	87.0%	0	0	6.9%	2.1%	0
Fear	2.8%	0	89.3%	2.1%	0	5.8%	0
Happy	0	0	3%	95.8%	1.2%	0	0
Neutral	5.3%	0	0	0	92.2%	2.5%	0
Sadness	7.9%	0	0	0	5.4%	84.0%	2.7%
Surprise	0	1.3%	0	0	1.2%	0	97.5%

Table 4 Comparison of several facial expression recognition models

Method	Recognition accuracy (%)
SIFT-BoF + SVM	76.57
CNN	89.32
Deep learning feature + SVM	91.58
CNN-SIFT + SVM	94.13

In each epoch of training, the layer parameters of the networks are updated based on the loss. The CNN method obtains an average classification rate of 89.32% when applied to the test set. We also compare the results of an experiment involving the training set without augmentation. The average accuracy for all expression labels is 86.8%. An increase is noted in the average accuracy of about 3% when the proposed data augmentation method is used.

As mentioned previously, CNN is adopted as a feature extractor. We extract deep learning features from whole images using the trained CNN model. We then train a SVM using the training data and the cross-validation data and predict the classification of the testing data. The results show that combining the CNN features with the SVM is highly competitive with or even superior to the original CNN model for our FER tasks. However, performance is strongly affected by the layer from which the features are output, as shown in Fig. 7. We compare the performance of the two different extracted features on each category. The features extracted from layer 7 achieve an accuracy of about 90.02%, while the layer 8 features achieve an accuracy of 91.58%. Using the features from the eighth layer outperforms those from the seventh layer at about 1.5% overall. Table 3 shows the confusion matrix of our final deep learning feature representation method. The performance is 91.58% for an SVM trained with an RBF kernel and features taken from layer 8.

Table 6 Average facial expression accuracy in cross-database experiments

Method	Test: JAFFE (%)	Test: MMI (%)
SIFT-BoF + SVM	38.27	39.13
CNN	44.86	47.50
Deep learning features + SVM	46.73	52.22
CNN-SIFT + SVM	48.90	53.81

After these basic steps, we select the best shallow and deep learning features described above. The dimension of the SIFT-BoF feature is the size of the optimal bag vocabulary ($K = 500$). The dimension of the CNN feature is the number of the last connected layer node 500. We then merge these features to make a representation for each image. The dimension of the hybrid feature is 1000. For solving the final classification problem, we also apply the SVM for the training set and validation set with a RBF kernel.

The average recognition results obtained by three different facial expression features and a 7-class SVM classifier are shown in Table 4. Compared with other methods, the average accuracy of these new hybrid representation methods outperforms the deep learning features based on a CNN by at least 4.8%. Our experimental data show that the power of these feature descriptors, which represent our SIFT-BoF features, can help the CNN features to achieve improved accuracy in the FER task.

Using the novel hybrid features we propose above, the performances are shown in Table 5. The recognition performance of emotions with label disgust, happy, and surprise, which human beings can easily recognize, achieves an accuracy rate that exceeds 95%. On the other hand, the anger and sadness expressions are associated with lower recognition rates. The anger expression is confused with sadness a majority of the time. This result shows that the extracted features of two expressions extracted by the model are not well separated.

Cross-Database Experiments

Despite the high FER accuracy obtained for the individual databases, cross-database FER remains challenging. The cross-database generalization ability is usually poor. Shan et al. [10] trained selected LBP features using SVMs on CK+ dataset and tested it on the JAFFE database or

Table 5 Facial expression recognition using our proposed CNN-SIFT feature

	Anger	Disgust	Fear	Happy	Neutral	Sadness	Surprise
Anger	85.1%	2.2%	0	0	4.1%	8.6%	0
Disgust	1.1%	95.3%	0	0	0.9%	2.7%	0
Fear	0	0	90.3%	3.5%	0	6.2%	0
Happy	0	0	2.1%	96.4%	0	0	1.5%
Neutral	1.2%	0	0	0	95.5%	3.3%	0
Sadness	4.1%	0.8%	0	0	6.1%	89.0%	0
Sadness	0	0	0.5%	0	0.7%	0	98.8%

Table 7 Facial expression recognition obtained by using CNN-SIFT features on JAFFE in cross-dataset experiment

	Anger	Disgust	Fear	Happy	Neutral	Sadness	Surprise
Anger	31.9%	6.3%	0	8.1%	17.2%	25.7%	10.8%
Disgust	13.5%	45.1%	6.7%	0	15.2%	19.5%	0
Fear	10.9%	2.5%	27.4%	7.8%	0	18.3%	33.1%
Happy	6.2%	0	14.2%	61.3%	5.3%	13.0%	0
Neutral	11.9%	0	0	12.7%	47.3%	19.3%	8.8%
Sadness	17.1%	9.3%	3.5%	0	15.3%	54.8%	0
Sadness	2.7%	0	6.3%	0	4.2%	9.6%	77.2%

the MMI database. The experiments show that the across datasets performance is much poor, around 40% on JAFFE database and 50% on MMI database. These results reflect the difficulty of generalizing a facial expression recognition system trained with limited training set [38].

To construct a cross-dataset experiment, seven groups from CK+ database are adopted to train the proposed method, and one group is adopted as a cross-validation set. JAFFE and MMI databases are used to test the cross-dataset performance. In the cross-dataset experiments, no images from the JAFFE and MMI databases are used during the training of the network.

First, the number of vocabularies we choose to implement is 500. We use all of the CK+ data to obtain the set of BoFs, and we then use these BoFs to create the BoF descriptor for all of the images in the training and testing sets. We then train an SVM on CK+ database and test the classifier on the JAFFE database. Second, we choose seven groups of images to train the CNN and use the one remaining group for validation each time. The CNN architecture and training procedure are similar to those used in the experiment involving CK+ database. We save the model that yields the best performance on the validation set from the cross-validation. We implement the model to predict an expression in the JAFFE and MMI databases. We then adopt the trained CNN model to extract features from the second fully connected layers of the training and testing sets. Similarly, we train the SVM on CK+ database and make predictions using the JAFFE and MMI databases. Finally, we connect the SIFT-BoF features and deep learning features of the training and testing images. Using the novel hybrid feature, we also train the SVM on CK+ database and test it using the JAFFE and MMI databases. Table 6 presents the average cross-database accuracy when classifying the seven expressions.

Table 6 shows that the average recognition accuracy in the cross-database experiment using the CNN-SIFT features is much lower compared with the results obtained within the CK+ database experiment. Specifically, accuracies of approximately 48% and 53% are obtained for the JAFFE and MMI databases, respectively. The low accuracies reported in Table 6 can be explained in terms of the cultural differences. While the MMI database includes some subjects with the same ethnicity as those in the CK+ database, the same statement does not apply to the JAFFE database. Compared to the SIFT-BoF features and the deep learning features, our method yields a slight improvement in the cross-database experiment. The confusion matrix obtained through the use of our proposed novel hybrid features in the above experiments involving the JAFFE and MMI databases can be seen in Tables 7 and 8, respectively. The cross-database experiments convincingly prove the utility of the hybrid features, which would be effective when the training data are not sufficient to obtain a deep model with considerable generalization ability.

Table 8 Facial expression recognition obtained by using CNN-SIFT features on MMI database in cross-database experiment

	Anger	Disgust	Fear	Happy	Neutral	Sadness	Surprise
Anger	39.8%	16.3%	0	3.8	20.7%	8.2%	11.2%
Disgust	9.5%	46.5%	6.7%	0	18.5%	11.0%	7.8%
Fear	0	16.7%	37.1%	3.3%	10.3%	9.0%	23.6%
Happy	0	4.7	1.2%	69.0%	10.2%	0	14.9%
Neutral	14.0%	0	5.1%	7.9%	62.8%	10.2%	0
Sadness	12.5%	6.0%	8.7%	0	8.6%	44.6%	19.6%
Sadness	1.9%	0	0	5.8%	11.8%	0	80.5%

Table 9 Comparisons of expression recognition performance

Method	Lope et al. [35]	Mollahosseini et al. [15]	Hasani et al. [39]	BDBN [2]	AUDN [40]	LBP + SVM [9]	Proposed
CK+	95.79	93.2	93.04	96.7	93.7	91.4	94.13
JAFFE	37.36	—	—	68.0	—	41.3	48.90
MMI	—	55.6	—	—	72.2	51.1	53.81

Experimental Analysis

Furthermore, the performance of proposed method are compared with some state-of-the-art methods. Table 9 shows the results of applying other approaches on CK+ dataset and cross-database experiments in the JAFFE and MMI databases. Note that the methods shown in Table 9 do not use the same experimental method as we proposed (i.e., they use different *k*-fold cross-validation techniques and different criteria for selecting the datasets). Table 9 shows that the proposed method obtains better performance and show high results when applied to CK+ dataset. The bilinear deep belief network (BDBN) method, which features the highest accuracy of the methods shown in this table, is not comparable to our method because it identifies six expressions. Other experiments are performed using seven expressions, which is more challenging than the problem with six categories.

Some studies in the literature also include cross-database experiments using the JAFFE and MMI database. Note that these two databases do not represent realistic scenarios regarding the acquisition of samples. However, such databases are publicly available and have been used by many researchers for evaluation and benchmarking.The accuracy is computed using a binary classifier approach in [2], whereas the other methods employ the *n*-class classifier approach. In [15], the networks for the cross-database cases are tested on one of the CK+; Multiple Pose, Illumination and Expression (Multi-PIE); MMI; Facial Expression Recognition and Analysis; Denver Intensity of Spontaneous Facial Action; Static Facial Expressions In The Wild; and FER2013 databases after training on the rest. Compared with the method based on shallow features [9] and other methods based on deep learning, our experimental results show a convincing demonstration of the generalization ability of the proposed method.

Conclusion

The advantage of CNNs is that they can learn high-level representations of images. Thus, CNNs perform well in flexible learing tasks, than hand-designed features. However, learning CNNs requires abundant training samples. Although we apply some transformations to images to augment the databases, the augmented databases still cannot satisfy the data requirements of CNNs. On the other hand, hand-crafted shallow features do not require extensive databases to generalize. Thus, we use SIFT features to increase the whole performance on small datasets. Combining shallow and deep learning features would be effective when the training data are not sufficient to obtain a deep model with considerable generalization ability.

In this paper, a hybrid feature representation is proposed for the recognition of facial expressions that adopts a combination of SIFT-BoF features and deep learning features extracted from a CNN model, and we then train features and classify expressions using SVMs. The hybrid CNN-SIFT approach has the strengths of both methods. Several experiments have been performed using CK+ database to validate the usefulness of the proposed method. Compared with the accuracy of 76.57% obtained using the SIFT-BoF features and the accuracy of 92.87% obtained based on the CNN features, we achieve a FER accuracy of 94.82% using the hybrid SIFT-CNN features. The results indicate the considerable potential of combining shallow features with the deep learning features, and this technique displays greater promise than the state-of-the-art CNN method.

Funding The authors would like to thank Zhou for the use of the NLPCC 2017 Shared Task Sample Data: Emotional Conversation Generation as training dataset. The work was supported by the State Key Program of the National Natural Science of China (61432004, 71571058, 61461045). This work was partially supported by the China Postdoctoral Science Foundation funded project (2017T100447). This research has been partially supported by the National Natural Science Foundation of China under Grant No. 61472117. This work was also supported by the foundational application research of Qinghai Province Science and Technology Fund (No. 2016-ZJ-743). This work was also supported by the Open Project Program of the National Laboratory of Pattern Recognition (NLPR).

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflicts of interest.

Ethical Approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Ekman P, Friesen WV. Facial action coding system (FACS): a technique for the measurement of facial actions[J]. Rivista Di Psichiatria. 1978;47(2):126–38.
2. Liu P, Han S, Meng Z, et al. Facial expression recognition via a boosted deep belief network[C]. In: IEEE Conference on computer vision and pattern recognition. IEEE Comput Soc; 2014. p. 1805–12.
3. Liu Z, Wang H, Yan Y, et al. Effective facial expression recognition via the boosted convolutional neural network[C]. In: CCF Chinese conference on computer vision. Berlin: Springer; 2015. p. 179–88.
4. Bosse T, Duell R, Memon ZA, et al. Agent-based modeling of emotion contagion in groups[J]. Cogn Comput Springer. 2015;7:111.
5. Chen Y-w, Zhou Q, Luo W, et al. Classification of Chinese texts based on recognition of semantic topics[J]. Cogn Comput Springer. 2016;8:114.
6. Xu R, Chen T, Xia Y, Lu Q, et al. Word embedding composition for data imbalances in sentiment and emotion classification[J]. Cogni Comput Springer. 2015;7:226.
7. Fan H, Cao Z, Jiang Y, et al. Learning deep face representation[J]. Eprint Arxiv. 2014.
8. Zhang Z, Lyons M, Schuster M, et al. Comparison between geometry-based and Gabor-wavelets-based facial expression recognition using multi-layer perceptron[C]. In International conference on face & gesture recognition. IEEE Computer Society; 1998. p. 454.
9. Shan C, Gong S, Mcowan PW. Facial expression recognition based on local binary patterns: a comprehensive study[J]. Image Vis Comput. 2009;27(6):803–16.
10. Dahmane M, Meunier J. Emotion recognition using dynamic grid-based HoG features[C]. In IEEE International conference on automatic face & gesture recognition and workshops. IEEE; 2011. p. 884–88.
11. Lowe DG. Distinctive image features from scale-invariant keypoints. Int J Comput Vis. 2004;60(2):91–110.
12. Luo Y, Wu CM, Zhang Y. Facial expression recognition based on fusion feature of PCA and LBP with SVM[J]. Optik - Int J Light Electron Opt. 2013;124(17):2767–70.
13. Lopes AT, Aguiar ED, Oliveira-Santos T. Facial expression recognition system using convolutional networks[C]. Graphics, patterns and images. IEEE; 2015. p. 273–80.
14. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks[C]. In International conference on neural information processing systems. Curran Associates Inc. 2012; p. 1097–5.
15. Mollahosseini A, Chan D, Mahoor MH. Going deeper in facial expression recognition using deep neural networks[J]. Comput Sci. 2015; 1–0.
16. Lv L, Zhao D, Deng Q. A semi-supervised predictive sparse decomposition based on task-driven dictionary learning[J]. Cogn Comput. 2017;9(1):1–0.
17. Liu P, Li H. Interval-valued intuitionistic fuzzy power Bonferroni aggregation operators and their application to group decision making[J]. Cogn Comput. 2017;9(1):1–9.
18. Lucey P, Cohn JF, Kanade T, et al. The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression[C]. In Computer vision and pattern recognition workshops. IEEE; 2010. p. 94–101.
19. Kamachi M, Lyons M, Gyoba J. The Japanese female facial expression (JAFFE) database[J].
20. Pantic M, Valstar M, Rademaker R, et al. Web-based database for facial expression analysis[C]. In IEEE international conference on multimedia and expo. IEEE; 2005. p. 5.
21. Donato G, Bartlett MS, Hager JC, et al. Classifying facial actions[j]. IEEE Trans Pattern Anal Mach Intell. 1999;21(10):974.
22. Filliat D. A visual bag of words method for interactive qualitative localization and mapping[C]. In IEEE International conference on robotics and automation. IEEE; 2007. p. 3921–26.
23. Jorda M, Miolane N. Emotion classification on face images, Stanford University, CS229: Machine Learning Techniques project report.
24. Chen M, Zhang L, Allebach JP. Learning deep features for image emotion classification[C]. In IEEE International conference on image processing. IEEE; 2015. p. 4491–95.
25. Zhang SX. CNN deep learning model for facial expression feature extraction. Modern Comput: Professional Edition. 2016;2:41–4.
26. Burges CJC. A tutorial on support vector machine for pattern recognition. JData Mining Knowl Discov. 1998;2(2):121–67.
27. Yousefi S, Kehtarnavaz N, Cao YCY. Facial expression recognition based on diffeomorphic matching[J]. 2010;119(5):4549–52.
28. Tian Y, Kanade T, Cohn JF. Recognizing action units for facial expression analysis[J]. IEEE Trans Pattern Anal Mach Intell. 2001;23(2):97.
29. Asian O, Yildiz OT, Alpaydin E. Calculating the VC-dimension of decision trees[C]. In International symposium on computer and information sciences. IEEE. 2009; p. 193–8.
30. Azhar R, Tuwohingide D, Kamudi D, et al. Batik image classification using SIFT feature extraction, bag of features and support vector machine[C]. In: Information systems international conference; 2015. p. 24–30.
31. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE. 1998;86:2278–2324.
32. An DC, Meier U, Masci J, et al. Flexible, high performance convolutional neural networks for image classification[C]. In: IJCAI 2011, proceedings of the, international joint conference on artificial intelligence, Barcelona, Catalonia, Spain, July. DBLP; 2011. p. 1237–42.
33. Ouellet S. Real-time emotion recognition for gaming using deep convolutional network features[J]. Eprint Arxiv. 2014.
34. Chang CC, Lin CJ. LIBSVM: a library for support vector machines software available at http://www.csie.ntu.edu.tw/cjlin/libsvmS. 2001.
35. Lopes AT, Aguiar ED, Souza AFD, et al. Facial expression recognition with convolutional neural networks: coping with few data and the training sample order[J]. Pattern Recogn. 2016;61:610–28.
36. Wandell BA. Foundations of vision, 1st ed. Sunderland: Sinauer Associates Inc; 1995.
37. Bradski G, Kaehler A. Learning OpenCV: computer vision with the OpenCV library. Cambridge: O'Reilly; 2008.
38. Zhu R, Zhang T, Zhao Q, et al. A transfer learning approach to cross-database facial expression recognition[C]. In: International conference on biometrics. IEEE; 2015. p. 293–8.
39. Hasani B, Mahoor MH. Spatio-temporal facial expression recognition using convolutional neural networks and conditional random fields[J]. 2017.
40. Liu M, Li S, Shan S, et al. AU-inspired deep networks for facial expression feature learning[J]. Neurocomputing. 2015;159(C):126–6.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.