Subject Independent Facial Expression Recognition: Cross-Connection and Spatial Pyramid Pooling Convolutional Neural Network

Leilei Li[#,1,2], Yue Yuan[#,1,2], Mi Li[*,1,2], Hongpei Xu[1,2], Richeng Li[1,2] , Shengfu Lu[1,2]
[1]Department of Automation, Faculty of Information Technology, Beijing University of Technology
[2]The Beijing International Collaboration Base on Brain Informatics and Wisdom Services,
100 Ping Le Yuan, Chaoyang District, Beijing 100024, China
+86-10-67396464
Email: limi@bjut.edu.cn

ABSTRACT
Facial expression recognition is still a problem at present, especially in the case of individual independence. On the one hand, due to the influence of morphological changes, ethnic differences and other factors, the expression of individual expressions varies greatly. On the other hand, there is currently no publicly available large-scale dataset that can support deep neural networks. To this end, this paper proposes cross-connection and spatial pyramid pooling convolutional neural network. The model not only uses spatial pyramid pooling for high-level feature enhancement, but also combines cross-connection and spatial pyramid pooling to extract important low-level features. Finally the different levels of features are connected to improve the generalization performance of the model. We validate our approach in four widely used public expression datasets (CK+, JAFFE, MMI, NimStim). Compared to other facial expression recognition methods, our proposed method achieves comparable or superior results. In the case of subject independence, the model achieved a good result with 97.41% accuracy on the CK+ dataset.

CCS Concepts
• **Computing methodologies → Artificial intelligence → Computer vision → Computer vision problems → Object recognition**

Keywords
Convolutional Neural Networks; Facial Expression Recognition; Label Smoothing; Spatial Pyramid Pooling; Cross-Connection

1. INTRODUCTION
Facial expression is one of the most important features of human emotional recognition [1]. Facial expression recognition is a task that can be easily handled in human daily life. Li and Jain defined facial expressions as facial changes corresponding to people's internal emotional state, intention or social interaction [2]. American psychologists Ekman and Friesen define happy, sad, anger, fear, surprise, disgust and neutrality as seven basic facial expressions [3]. Darwin first introduced it into the research field in his book "Emotional Expressions of Man and Animals"[4]. Facial expression recognition involves many disciplines such as computer science, psychology, behavioural science, emotional computing, artificial intelligence theory, etc. It has important significance and practical value in the fields of character animation, human-computer interaction system, safe driving and distance education.

Facial expression recognition is a process in which a computer performs pre-processing and feature extraction on obtained facial expression data, and performs facial expression classification. Facial expression recognition systems can be divided into two categories: static image recognition and dynamic image sequence recognition. The static image based method does not use time information, that is, the feature vector only includes information about the current input image; the sequence-based method uses the captured time information of one or more frames of the expressed image to identify the expression. An automated facial expression recognition system receives desired input (static image or dynamic image sequence) to give an expression output, the output usually gives one of six basic expressions (such as anger, sad, surprise, happy, disgust, and fear), and some systems also recognize neutral expressions. Although recent expression recognition methods have an accuracy of more than 95% in the case of a frontal, controlled environment, and high-resolution image, it is still difficult for a computer to realize expression recognition in a natural scene. On the one hand, many of the work in the literature does not implement consistent evaluation methods (for example, there is no subject overlap in the training set and test set), so it presents misleading high precision, but cannot solve most of the expression recognition problems in real scenes. On the other hand, recognition accuracy is low in an uncontrolled environment and cross database evaluations.

Recently, with the development of deep learning techniques, especially convolutional neural networks, facial expression recognition technology has further developed, such as deep convolutional neural networks. Neural network technology combines image acquisition, feature extraction and selection, and

IVSP 2019, February 25–28, 2019, Shanghai, China

ACM ISBN 978-1-4503-6175-0/19/02…$15.00
DOI: https://doi.org/10.1145/3317640.3317662

[#]These authors contributed equally to this work.

[*]Corresponding author.

image classification of facial expression recognition system to form an end-to-end algorithm model. The convolutional neural network takes the original image or image sequence as input, and abstracts the original data layer into the required feature representation through the algorithm, and finally ends with the feature-to-category mapping. From the original data to the final expression classification is not mixed with any human operation. The convolutional neural network was proposed in 1998 by Yann LeCun [5] et al., which generally consists of several convolutional layers, pooled layers, and fully connected layers. The convolutional layer is the core layer of the convolutional neural network, and the convolutional layer is determined by the size of the kernel and the number of feature maps. The convolutional layer improves the machine learning system by sparse connections, parameter sharing, and isomorphic representations, and the convolution kernel moves over the active area of the input image to generate a feature map. The pooling layer reduces the feature map by the pooling function to provide spatial variance so that the input representation is approximately constant when a small amount of displacement is input. The pooling function replaces the model's output at that location by using the overall statistical characteristics of the adjacent outputs at a location. The common pooling types are maximum pooling and average pooling. The neurons of the fully connected layer are completely connected to the previous layer (generally a convoluted layer, a pooled layer or a fully connected layer).

Aiming at the problem of facial expression recognition on small-scale datasets, this paper proposes a classification model combining image data expansion and convolutional neural network. Compared with other work, this paper constructs a compact network structure by using spatial pyramid pooling algorithm [6] and cross-connection. In this paper, the expression recognition method of the proposed deep convolution model is tested on four common small datasets (CK+ [7], Jaffe [8], MMI [9] and NimStim [10]), and compared with other expression recognition models. The experimental results show that the model obtains satisfactory classification accuracy in the recognition of multiple small facial expression datasets.

2. RELATED WORK

In the past decades, a variety of facial expression recognition methods have been proposed by domestic and international researchers, and the performance of these methods is getting better and better. Liu et al. [11] proposed a boosted deep belief network (BDBN) composed of multiple weak classifiers. Each weak classifier is responsible for the classification of one facial expression. Based on the given eye coordinates, this algorithm aligns and cuts all facial expression images, and iterates continuously in the three learning stages of feature learning, feature selection and classifier fusion. Their algorithm achieved 96.7% accuracy in cohn-kanade Plus (CK+) datasets and 91.8% accuracy in JAFFE datasets, respectively, and achieved an average recognition accuracy of 68% in cross-database evaluation (CK+ dataset training, JAFFE testing). On the 6-core 2.4Ghz PC used by the author, the network needs about eight days of training time, and the recognition time of a single facial expression image is about 0.21 seconds.

Song et al. [12] developed a facial expression recognition system based on deep convolutional neural network that can be run on smart phones. The author applies data augmentation technology to increase the scale of training data and use dropout technology to prevent overfitting [13]. The authors evaluated the proposed network model on the CK+ dataset and three other datasets created by themselves, and achieved an average classification accuracy of 99.2% on the five basic expressions (anger, joy, sadness, surprise and neutral) of the CK+ dataset.

Burkert et al. [14] proposed a facial expression recognition method based on convolutional neural network for automatic data preprocessing and feature extraction. The convolutional neural network structure proposed by the author has a total of 15 layers, including 7 convolutional layers, 5 pooling layers, 2 full-connection layers and 1 regularization layer. The extracted features are classified through the final fully connection layer. The model achieves 99.6% and 98.63% recognition accuracy on CK+ dataset and MMI dataset respectively.

Liu [15] proposed a deep network model (AUDN) inspired by facial action unit (AU). Their model can learn three aspects of information: local appearance change, optimization method combining local change and expression of high-level features of facial expression. The model uses the method of cross validation to classify six basic expressions for model verification, and the recognition accuracy of CK+, MMI and SFEW has achieved an average of 93.70%, 75.85% and 30.14%, respectively.

Ali et al. [16] proposed an improved neural network ensemble algorithm for the recognition of multiple facial expressions. The authors combined JAFFE, TFEID, and RadBoud faces database (RaFD) to obtain a multiracial facial expression dataset containing Japanese, Taiwanese, Caucasian, and some Moroccan. Using five basic expressions (anger, happiness, sadness, surprise and fear) to evaluate the model, the authors achieved an average recognition accuracy of 93.75% on the dataset of multiple facial expressions, and an average recognition accuracy of 48.67% across the dataset (using Jaffe test and other two datasets for training).

Lopes et al. [17] proposed a solution to the problem of few data based on convolutional neural network. The author studied the effect of different data preprocessing on classification effect and proposed a data augmentation method based on eye position information. The convolutional neural network established by the author has 6 layers, including 2 convolutional layers, 2 pooling layers and 2 fully connected layers. Through data preprocessing and 70-fold data amplification, 96.7% recognition accuracy was achieved on CK+ dataset without subject overlap between training set and test set.

3. DATASETS AND EVALUATION SCHEME

This section will provide a detailed introduction to the datasets used in the paper and the evaluation scheme of the model.

3.1 Datasets

We validated our approach on four public datasets: CK+, Jaffe, NimStim, and MMI, and classified six basic expressions (happiness, anger, disgust, fear, sadness, and surprise) across all datasets.

CK + dataset is an extended dataset of Cohn-Kanade (CK) dataset, which contains 593 video sequences of 123 subjects, 327 of which have explicit tags. There are seven types of tags: contempt, happy, anger, disgust, fear, sad and surprise. Each expression sequence represents the expression of neutral to tagged emotion. In order to compare with other methods, we removed the image sequence of contempt expressions, and obtained a total of 309 expression sequences of 106 different subjects. Using the 3 most expressive frames of each sequence, 927 images were used.

The Jaffe dataset consists of 213 images of 10 Japanese female participants, containing data on six basic expressions (happy, anger, disgust, fear, sad, and surprise) and neutral expressions. Among them, each expression of each subject has data of not less than 3 images. All image sizes in the dataset are 256*256 pixels, and the grayscale value is 8-bit precision. We used six expressions in addition to the neutral expression, a total of 183 images for model validation.

The MMI dataset contains more than 20 participants, including Europeans, Asians, and South Americans. The subjects were both female and male. Its 237 picture sequences all transition from neutral to peak expressions and back to neutral expressions, each sequence corresponding to an expression. In order to compare with other methods, we select the data series of six basic expressions as input to our model. For each sequence, we selected 2 images, a total of 213 expression sequences, and removed the image sequence with glasses. A total of 125 image sequences from 21 people were left, and 250 images were used.

The NimStim dataset is a public dataset for expression recognition that is opened by the US Research Network on Early Experience and Brain Development. It is also a typical facial psychological expression dataset. Researchers often use happy, neutral and sad emotional face images as stimuli to examine the emotional attention bias of depressed patients in different emotional environments. It has a total of 646 emoticons, each corresponding to an expression. After retaining 6 basic expressions, 491 pictures of the remaining 42 participants were used for training and testing.

3.2 Evaluation Program

In the field of facial expression recognition, subject independent grouping and non-subject independent grouping are two common dataset grouping strategies for model evaluation. The grouping strategy description is shown in figure 1. The same subject's data will only appear in the training set (or test set), while in the non-subject independent group, different photos of the same person exist in the training set and Test set. Studies by Girard et al. [53] and Lopes et al. show that subject independent grouping strategies are more able to ensure the ability of the model to be compared than non-subject independent grouping. The subject independent grouping strategy is used as the evaluation scheme of the model.

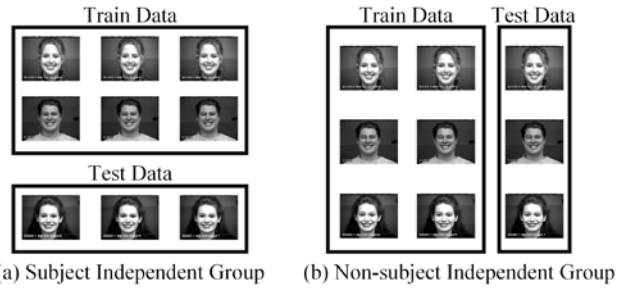

Figure 1. Subject independent and non-subject independent grouping strategy.

In order to validate the proposed model, the experiment uses K-fold cross-validation method, and divides the whole data set into K groups according to subjects. K-1 group is used as training set and the remaining one group is used as test set. Each time, different training sets and test sets are selected, K times are executed repeatedly, and the average of K times accuracy is taken as the accuracy of the dataset. Because the number of participants in different datasets is different, we choose different folds for different data sets. The sample size used for each data set and the cross-validation fold used are shown in Table 1.

The y tag of the facial expression dataset has certain uncertainty, and the wrong y is not conducive to maximizing log p(y|x). To prevent over-fitting of the model, we add noise to the output target of the model. Label smoothing [20] is a typical label noise model. It assumes that the correct probability of label y in training set is ϵ and that other labels are also $(1-\epsilon)/(k-1)$. And k is the number of categories. Label smoothing prevents the model from pursuing the exact probability without affecting the correct classification of the model learning. We set ε to 0.9 and other tags to 0.02.

Table 1. Dataset

Dataset	Subject number	Sample size	Fold number(K)
CK+	106	927	8
JAFFE	10	213	10
MMI	21	250	7
NimStim	42	491	8

4. PROPOSED METHOD

In this section, we describe the data preprocessing methods used in this paper and proposed several models. We study the performance of different convolutional neural network models on CK+ dataset in depth and width, and propose a cross-connection and spatial pyramid pooling convolutional neural network.

4.1 Data Preprocessing

First, we use the Haar feature-based cascade classifier and SDM algorithm provided by Intraface [18] toolbox to locate the face and the center of the human eye. Secondly, based on the obtained eye center coordinates, we use the data augmentation method proposed by Lopes et al. [17] to expand the training set data by an additional 70 times to increase the training samples of the model and to perform face image clipping and intensity normalization processing.

4.2 Convolutional Neural Network

We use the model proposed by Lopes et al. as our benchmark model. The specific information of the model is shown in Figure 2. This model contains 2 convolutional layers, 2 maximum pooling layers, and a hidden layer of 256 neurons. Unlike Lopes et al., our input image has a resolution of 48x48 instead of 32x32. Compared to 32x32 emoticons, 48x48 emoticons have more detail information. The model evaluation scheme is also different from the research by Lopes et al., which divides the data set into training set, test set and validation set, and runs the test set multiple times to select the optimal model to apply to the validation set. While we use K-fold cross-validation. Based on this model, we performed an 8-fold cross-validation on the CK+ dataset using the method described in Section 3.2, yielding an average classification accuracy of 94.50%.

Theoretical and experimental results show that depth and width are the core factors affecting the expression ability of neural networks. VGG [23] model shows that depth is wider than width in improving model complexity. Zeiler and Fergus [24] used deconvolution techniques to visualize the features of

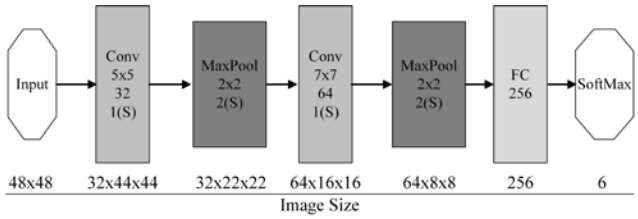

Figure 2. The structure of the benchmark convolution neural network, in which Image size represents the output size of each layer's feature graph.

convolutional neural networks and found that features in deep neural networks are hierarchical. Convolutional networks learn basic patterns such as edges in the shallow level, and learn high-level semantics such as faces at the top level. To further improve the model's ability to express on CK+, we first replace the 5x5 convolution kernel in the reference model with two 3x3convolution kernels, and replace the 7x7 convolution kernel with one 3x3 and one 5x5 convolution kernel. , keep the number of feature maps unchanged. Using the same evaluation program as the benchmark model, we obtained a classification accuracy of 94.92% on this model. We continue to add two 3x3 convolution layers and one pooling layer between the second pooling layer and the fully connected layer, and use 128 feature maps to retain more feature information. We named this model Plain7CNN. This model achieves 95.47% accuracy, and reduces 16% classification error rate compared with the benchmark model. We use three 3x3 convolution cores instead of the 7x7 convolution cores in the benchmark network to continue to increase the depth of the network and name this model Plain8CNN. The structure of Plain7CNN and Plain8CNN models is shown in Figure 3. Our test results show that the results of the model are no longer improved, and even show a downward trend. As shown in Figure 4, this model achieves an average classification accuracy of 94.71% on CK +. In machine learning, over-fitting exists universally. When the number of network layers is increased, the number of model parameters and the complexity of the model are increased. Because the amount of data used for facial expression recognition is small, the model learns the noise data in the image while learning the effective features, which leads to the decline of generalization ability, and thus obtains a lower accuracy in the test set.

Model fusion is an effective strategy to improve network width and help deep network training. Wang et al. [25] argues that the integration of different networks in the middle layer (add or concat) can generate many potential shared networks and optimize the flow of information. The ResNet and Inception structures are the most commonly used structures. The Inception structure splicing multiple different channels for different levels of information fusion, but because of more training parameters in the Inception model, it is not applicable in the case of small datasets. The spatial pyramid pooling model has a multi-branch structure similar to the Inception structure. The difference from the Inception structure is that it is all performed through the pooling operation. The structure is shown in Figure 5. The spatial pyramid pooling model is a combination of multiple pooling operations with difference step sizes, which can perform multiple levels of calculation on the prominent data while losing a small amount of output data, thereby playing the role of attention mechanism. The large output is adjusted by multiple neuron nodes while the smaller points are adjusted using a single node. In order to make the model learn better feature representation and reduce the interference of noise information, we use the spatial pyramid pooling model to pool the output of the last convolution layer and the second convolution layer at different scales, and connect the two outputs. The structure is shown in Figure 6. We named this model Cross-Connection and Spatial Pyramid Pooling Convolutional Neural Network (C-SPP). CK + dataset was used to test the model, and 97.41% classification accuracy was obtained.

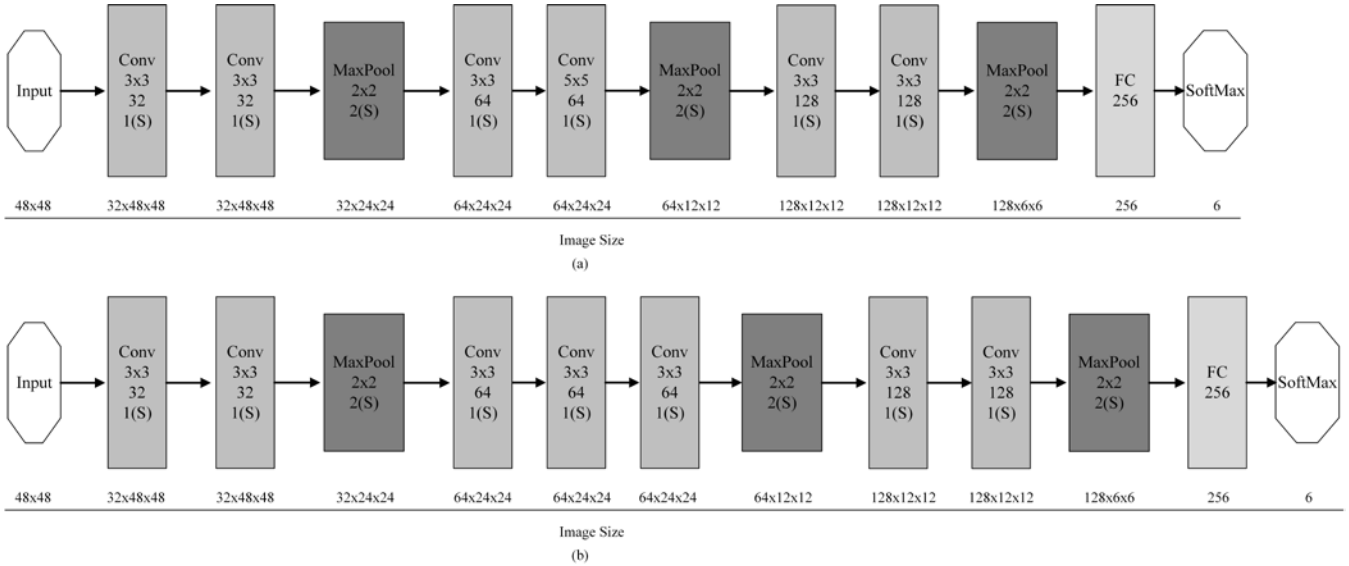

Figure 3. Plain7CNN and Plain8CNN structure, (a) is Plain7CNN, and (b) is Plain8CNN. Compared to the benchmark model set Plain7CNN, Plain8CNN has a deeper structure.

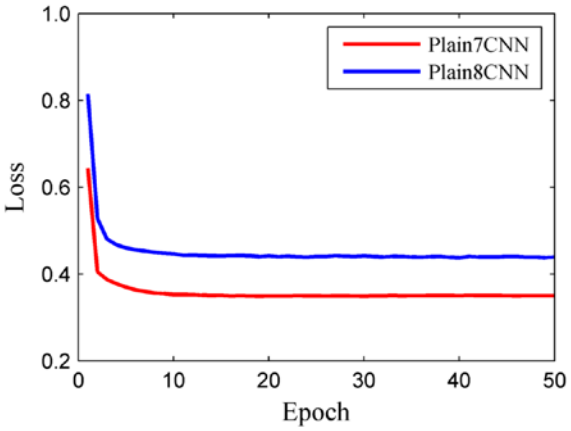

Figure 4. Plain7CNN and Plain8CNN error comparison, Plain8CNN has a deeper structure but shows higher error.

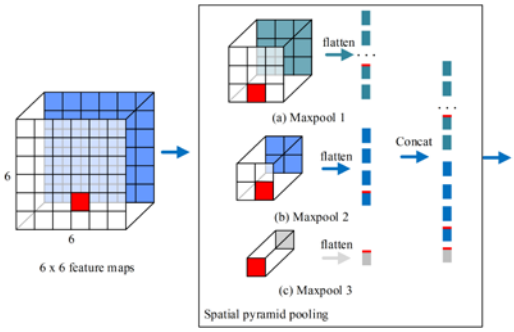

Figure 5. Spatial pyramid pooling (SPP). The structure in the graph with SPP (3, 2, 1). The 6x6 feature graph is given the specified size through three pooling layers with different steps. A feature map in which (a), (b), and (c) represent characteristic maps of 3x3, 2x2, and 1x1 obtained by non-overlapping pooling operations of steps 2, 3, and 6, respectively. The red block in the 6x6 feature map represents the maximum value of the first feature map, and the subsequent red block indicates the flow of information in the SPP, which performs forward prediction and reverse adjustment through multiple neural nodes.

5. Experiments and Discussion

In order to verify the model's ability to express on other datasets, we used Jaffe, MMI, and NimStim datasets to validate the model outside of the CK+ dataset, and all experiments selected the same network parameters. To further validate the generalization capabilities of the model, we used the CK+ dataset as the source dataset for training and cross-dataset testing on Jaffe, MMI, and NimStim. We used python, tensorflow [21] (GPU version) for data preprocessing and model building. All experiments were run on Intel Core i7, Nvidia GeForce GTX960 and window7 system platforms. The model is adjusted using the momentum gradient descent algorithm during training. The training parameters used in this experiment are shown in Table 2.

Table 2. Training parameters

Parameters	Values
Learning Rate	0.0009
Batch size	128
Number of Iterations	50 epoch
Momentum	0.9
Label Smoothing Factor(ϵ)	0.9
Dropout	0.5

5.1 Same Dataset Experiment

The same dataset experiment is an experiment in which a data set is divided into a training set and a test set. We conducted experiments on CK+, Jaffe, MMI, and Nimstim, and tested six expressions of Anger, Disgust, Fear, Happy, Sad, and Surprise. The correct rate and error curve of the training set and test set are shown in Figure 7, and the test results are shown in Table 3. As can be seen from Table, the happy expression has higher classification accuracy than other expressions in each data set, which may be due to the fact that the happy expression has a relatively obvious morphological representation. In addition, the classification accuracy of Jaffe and MMI is significantly lower than that of CK + datasets, which is mainly affected by the size of

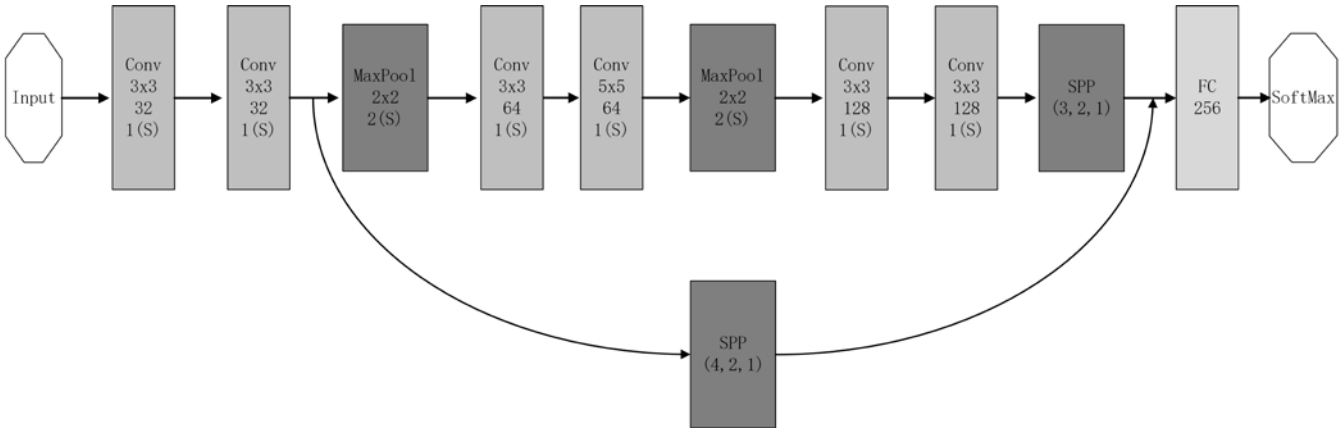

Figure 6. Cross-Connection and Spatial Pyramid Pooling Convolutional Neural Network, which contains two SPP modules (see Figure. 5). The numbers in parentheses in the SPP module indicate the specified feature map output size.

Table 3. The accuracy of classification of each data set and each expression

Dataset	Anger	Disgust	Fear	Happy	Sad	Surprise	Average
CK+	96.30%	97.18%	96.00%	100%	91.67%	98.39%	97.41%
Jaffe	76.67%	51.72%	46.88%	90.32%	64.52%	83.33%	68.85%
MMI	56.82%	33.33%	6.67%	87.5%	69.44%	75.00%	59.20%
NimStim	77.65%	76.25%	82.05%	96.75%	79.01%	79.55%	83.30%

Table 4. Confusion matrix of six expression classifications in CK+ dataset.

	Anger	Disgust	Fear	Happy	Sad	Surprise
Anger	130	3	0	0	1	1
Disgust	2	172	0	2	0	1
Fear	0	0	72	2	0	1
Happy	0	0	0	207	0	0
Sad	3	0	4	0	77	2
Surprise	0	0	1	2	1	245

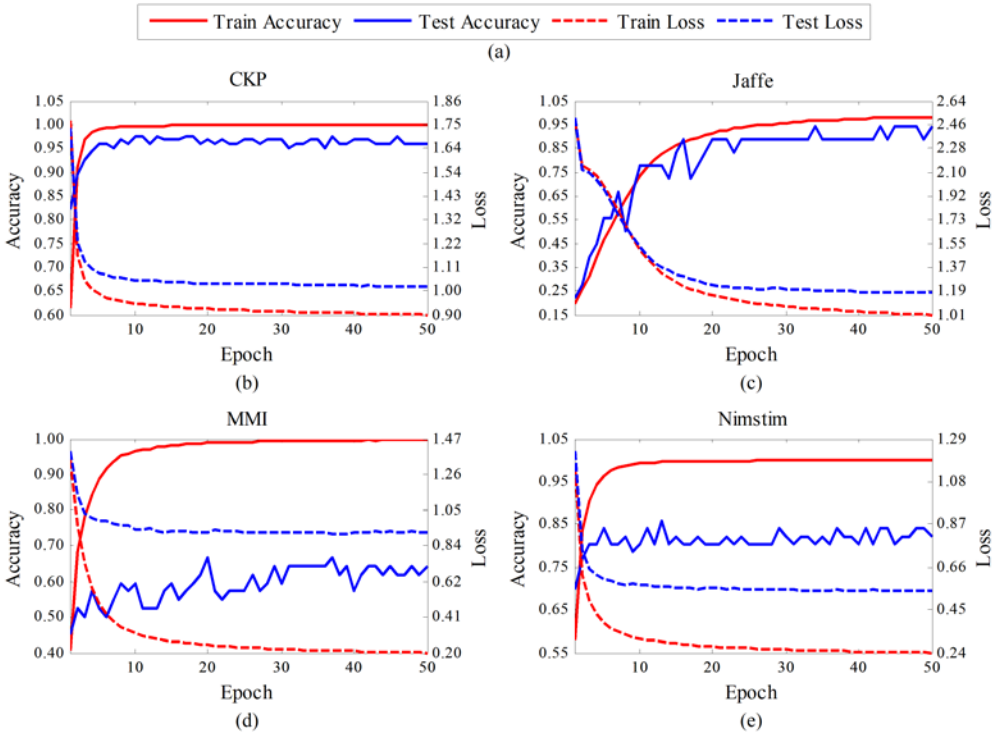

Figure 7. Training and test curves on CK+ (b), Jaffe (c), MMI (d), and NimStim (e). (a) is the legend.

data, and the reduction of accuracy also occurs in the results of other papers. [4, 5]. Compared to the Jaffe dataset, the MMI dataset has more subject numbers, but at the same time the MMI dataset has a more complex source of participants, and the differences in cultural background and ethnicity make it more complicated than Jaffe.

In order to observe the difference between different expressions, we made the confusion matrix of CK+, as shown in Table 4. The columns of the confusion matrix represent the prediction categories, the rows represent the actual categories of data, and the diagonals are the correctly classified data. Observing the confusion matrix, the angry expression is confused with the disgusting expression, and the sadness and fear and the angry part are easily misclassified. This shows that there is a close relationship between anger and fear and disgust in the feature space, and there is no good segmentation.

5.2 Cross-dataset Experiment

To test the generalization capabilities of the model, we performers cross-datasets validation. We trained the model on the CK+ dataset and tested it in the Jaffe, MMI and NimStim datasets. The

model was trained on the CK+ dataset using a 8-fold cross-validation to obtain 8 trained models. Running test data sets on these eight models yields correct rates across data sets. Taking into account the impact of the combination of training and test sets, we use the mean of 8 correct rates as the model result of cross-dataset validation. The results of Cross-data set are shown in Table 5. Compared with Jaffe's cross-dataset results, NimStim and MMI have higher results because NimStim, MMI and CK+ participants have higher similarities in cultural background and ethnicity.

Table 5. Cross-dataset training results.

Train Dataset	Test Dataset	Accuracy (%)
CK+	Jaffe	39.41
CK+	MMI	51.60
CK+	Nimstim	53.23

5.3 Comparison of Results

Table 6 shows the results of the Cross-Connection and Spatial Pyramid Pooling Convolutional Neural Network on the CK+ dataset, in which we compare the results of other work that also uses cross-validation and subject independent. As can be seen from the table, we achieved 97.41% of the results in CK+, which is significantly higher than other results, indicating the superiority of our model in small data sets.

Table 6. Comparison of the CK+ data set results.

Reference	Data Type	Classifier	Accuracy (%)
Gu et al.[21]	Image	Gabor+SVM	91.51
Liu et al.[26]	Video	UMM+STM	94.19
Fan et al.[27]	Video	PHOG+SVM	83.70
Liu et al.[11]	Image	BDBN	96.70
Mollahosseini [28]	Image	GoogLeNet	93.20
Lopes et al.[17]	Image	CNN	96.76
Proposed	Image	S-SPP	97.41

Table 7 shows the results of the cross-dataset validation. We compared the models that were trained using CK+ and tested on other datasets. Compared with the results of Gu et al., our model has poor results. Because the convolutional neural network performs feature learning based on CK+ data, many features are applicable to the expression pattern of CK+ and poor performance on other datasets. Gu et al. use local Gabor features to extract facial regions. Compared with convolutional neural networks, their features are more versatile, but with the end-to-end structure of convolutional neural networks, feature engineering requires the participation of prior knowledge, and performs poorly in data classification. Compared with the work of Lopes et al., our model has a higher accuracy rate on Jaffe, indicating that the proposed C-SPP model is more capable of extracting common features.

Table 7. Comparison of test results cross data sets (Use CK+ training).

Method	Test	Classifier	Accuracy
Gu et al.[22]	Jaffe	Gabor+SVM	55.87%
Lopes et al.[17]	Jaffe	CNN	38.80%
Proposed	Jaffe	S-SPP	39.41%
	MMI	S-SPP	51.60%
	NimStim	S-SPP	53.23%

6. CONCLUSION

In this study, the convolutional network model based on spatial pyramid pooling proposed by individual differences has better representation ability, and achieves 97.41% recognition accuracy on CK+. The comparison results show that the performance of the convolutional neural network model needs to be improved in ultra-small data sets such as Jaffe and MMI. In order to obtain data with clear subject markers, we use an expression dataset in a controlled environment. Next, we will try different network structures, especially the role of the attention mechanism in expression recognition and the expression recognition of expression datasets in different datasets, especially in the natural environment. On the other hand, cross-dataset accuracy is lower than the same dataset experiment, so we will further improve the generalization performance of the expression recognition model through data mix and network structure.

7. ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China (61602017), the National Basic Research Programme of China (2014CB744600), 'Rixin Scientist' Foundation of Beijing University of Technology (2017-RX(1)-03), the Beijing Natural Science Foundation (4164080), the Beijing Outstanding Talent Training Foundation (2014000020124G039), the National Natural Science Foundation of China (61420106005), the International Science & Technology Cooperation Program of China (2013DFA32180), the Special fund of Beijing Municipal Science and Technology Commission (Z171100000117004 and Z151100003915117), Beijing Municipal Administration of Hospitals Clinical Medicine Development of Special Funding Support (ZYLX201607), and Beijing Municipal Administration of Hospitals Ascent Plan (DFL20151801).

8. REFERENCES

[1] Wu, Y., Hong, L. and Zha, H. 2005. Modeling facial expression space for recognition. *IEEE/RSJ International Conference on Intelligent Robots & Systems*.

[2] Li, S. Z. and Jain, A. K. 2005. Handbook of face recognition (2005).

[3] Ekman, P. and Friesen, W. 1978. Investigator's guide to the facial action coding system. Palo Alto. CA: Consulting Psychologists Press.

[4] Darwin, C. 2012. The expression of the emotions in man and animals. *Portable Darwin*, 123, 1 (2012), 146.

[5] Lécun, Y., Bottou, L., Bengio, Y. and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86, 11 (1998), 2278-2324.

[6] Kaiming, H., Xiangyu, Z., Shaoqing, R. and Jian, S. 2014. Spatial Pyramid Pooling in Deep Convolutional Networks

for Visual Recognition. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 37, 9 (2014), 1904-1916.

[7] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z. and Matthews, I. 2010. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. (2010), 94-101.

[8] Lyons, M. J., Akamatsu, S., Kamachi, M., Gyoba, J. and Budynek, J. 1998. The Japanese female facial expression (JAFFE) database. *Proceedings of third international conference on automatic face and gesture recognition*(1998). 14-16

[9] Pantic, M., Valstar, M., Rademaker, R. and Maat, L. 2005. Web-based database for facial expression analysis. *IEEE International Conference on Multimedia & Expo*, 2005.

[10] Nim, T., Tanaka, J. W., Leon, A. C., Thomas, M. C., Marcella, N., Hare, T. A., Marcus, D. J., Alissa, W., Casey, B. J. and Charles, N. 2009. The NimStim set of facial expressions: judgments from untrained research participants. *Psychiatry Res*, 168, 3 (2009), 242-249.

[11] Ping, L., Han, S., Meng, Z. and Yan, T. Facial Expression Recognition via a Boosted Deep Belief Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1805-1812.

[12] Song, I., Kim, H. J. and Jeon, P. B. 2014. Deep learning for real-time robust facial expression recognition on a smartphone. *IEEE Conference on Computer Vision & Pattern Recognition*, 2014.

[13] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15, 1 (2014), 1929-1958.

[14] Burkert, P., Trier, F., Afzal, M. Z., Dengel, A. and Liwicki, M. 2015. DeXpression: Deep Convolutional Neural Network for Expression Recognition. *arXiv preprint arXiv*:1509.05371.

[15] Liu, M., Li, S., Shan, S. and Chen, X. 2015. AU-inspired Deep Networks for Facial Expression Feature Learning. *Neurocomputing*, 159, C (2015), 126-136.

[16] Ali, G., Iqbal, M. A. and Choi, T. S. 2016 Boosted NNE collections for multicultural facial expression recognition. *Pattern Recognition*, 55, 2016 (2016), 14-27.

[17] Xiong, X. and Torre, F. D. L. 2013. Supervised Descent Method and Its Applications to Face Alignment. Proceedings of the *IEEE conference on computer vision and pattern recognition*, 532-539.

[18] Denton, E., Zaremba, W., Bruna, J., Lecun, Y. and Fergus, R. 2014. Exploiting Linear Structure Within Convolutional Networks for Efficient Evaluation. In Advances in neural information processing systems, 1269-1277.

[19] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z. 2016. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818-2826.

[20] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G. and Isard, M. 2016. Tensorflow: a system for large-scale machine learning. *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI}* 16), 265-283.

[21] Gu, W., Xiang, C., Venkatesh, Y. V., Huang, D. and Lin, H. 2012. Facial expression recognition using radial encoding of local Gabor features and classifier synthesis. *Pattern Recognition*, 45, 1 (2012), 80-91.

[22] Zhan, W., Ruan, Q. and An, G. 2016. Facial expression recognition using sparse local Fisher discriminant analysis. *Neurocomputing*, 2016, 174 (2016), 756-766.

[23] Simonyan, K. and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv*:1409.1556 (2014).

[24] Zeiler, M. D. and Fergus, R. 2013. Visualizing and Understanding Convolutional Networks. *European conference on computer vision* (pp. 818-833).

[25] Wang, J., Wei, Z., Zhang, T. and Zeng, W. 2016. Deeply-Fused Nets. *arXiv preprint arXiv*:1605.07716.

[26] Liu, M., Shan, S., Wang, R. and Chen, X. 2014. Learning Expressionlets on Spatio-temporal Manifold for Dynamic Facial Expression Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1749-1756.

[27] Fan, X. and Tjahjadi, T. 2015. A spatial-temporal framework based on histogram of gradients and optical flow for facial expression recognition in video sequences. P*attern Recognition*, 48, 11 (2015), 3407-3416.

[28] Mollahosseini, A., Chan, D. and Mahoor, M. H. 2016. Going Deeper in Facial Expression Recognition using Deep Neural Networks. *2016 IEEE Winter conference on applications of computer vision (WACV).* 1-10.