2017 11th Asian Control Conference (ASCC)
Gold Coast Convention Centre, Australia
December 17-20, 2017

Nesterov Accelerated Gradient Descent-based Convolution Neural Network with Dropout for Facial Expression Recognition

Wanjuan Su, Luefeng Chen, Min Wu, Mengtian Zhou, Zhentao Liu and Weihua Cao

Abstract— **Nesterov accelerated gradient descent-based convolution neural network (NAGDCNN) with dropout is proposed for facial expression recognition, which fuses the convolution neural network (CNN) with Softmax regression to construct a deep convolution neural network (DCNN) that can excavate high-level expression features and classify them. The dropout layer is added after the sub-sampling layer which can effectively reduce overfitting and the network's training time, moreover, the Nesterov accelerated gradient descent (NAGD) is used to optimize the network weights that can predictably prevent the algorithm from going too fast or too slow and enhance the response capability of the network. To verify the effectiveness of the proposal, experiments on benchmark database are conducted, and the experimental results show that the proposal outperforms the state-of-the-art methods. Futhermore, the application experiment is also curried out and the results indicate the feasibility of the proposal in practical applications.**

Key Words—**Deep learning, facial expression recognition, Nesterov accelerated gradient descent, dropout, principal component analysis**

I. INTRODUCTION

With the development of various kinds of technologies, the level of social intelligence is also increasing, and people's requirements for human-robot interaction (HRI) experience are getting higher and higher. However, the existing machines are unable to interact with people emotionally [1]. Facial expression is one of the main channels for human to express emotion [2], so achieving facial expression recognition (FER) is conducive to the realization of the machine's recognition for human emotions or even understanding it. FER has a wide range of applications [3], such as fatigue driving test, remote nursing, HRI, etc. Therefore, the realization of more accurate FER can promote the development of social intelligence.

FER can be divided into expression feature extraction and expression feature recognition [4]. As for expression feature extraction, various methods are employed in the previous papers, e.g., active appearance models [5], scale-invariant feature transform [6], local binary pattern [7], Gabor wavelet transform [8] and so on. Especially, principal component analysis (PCA) [9] is also a commonly used feature extraction algorithm which can simplify data structure by reducing data dimensionality. In order to learn projection subspaces equipped with the ability of robustness and generalization, a new subspace learning algorithms based on the standard PCA, linear discriminant analysis, clustering based discriminant analysis (CDA) and their combinations is proposed [10], the combination of PCA and CDA has achieved better performance on facial expression database. Here, PCA is also chosen to extract expression feature to solve the problems of data redundancy and high dimension.

The facial expression feature recognition aims to design a suitable classification mechanism to recognize facial expression, of which common algorithms have hidden Markov model [11], support vector machines (SVM) [12], etc.

A framework for FER by using appearance features of salient facial patches (SFP) is proposed [13], which investigates the relevance of different facial patches, and experiments on benchmark databases show the effectiveness of SFP. Nevertheless, its process of obtaining the high-level expression feature is very complicated, deep learning (DL) has a strong ability of unsupervised feature learning which has brought about changes and leaps in various fields [14].

The DL aims at discovering the input data's high-levels of distributed representations which has been widely used in speech recognition, image recognition and other fields. Hinton [15] et al. used deep belief network (DBN) and deep automatic encoders to perform simple image recognition and dimensionality reduction task which proved the feasibility of the application of deep neural network (DNN) in image recognition. Based on this, many researchers begin to apply DL to FER. For instance, Liu et al. [16] proposed to adapt 3D convolutional neural networks (3DCNN) with deformable action parts (DAP) constraints, namely, a deformable parts learning component is incorporated into 3DCNN which can detect specific facial action parts under the structured spatial constraints, and obtain the discriminative part-based representation simultaneously. The deep convolution neural network (DCNN) is applied to perform feature learning and smile detection simultaneously [17], by using the learned features to train the SVM or AdaBoost classifier, which shows that the learned features have impressive discriminative ability. It can be seen that DL can effectively combine feature learning and classification into a single model. In CNN, convolution layers and sub-sampling layers are usually stacked iteratively to extract high-level semantic features. However, here we develop a DCNN for FER that fuses the CNN with Softmax regression (SR) to construct a DCNN. Besides, the dropout layer is employed in the DCNN procedure which can effectively alleviated overfitting problem [18] and reduce the network's training time to some extent.

This work was supported by the National Natural Science Foundation of China under Grants 61733016, 61603356 and 61210011, the Hubei Provincial Natural Science Foundation of China under Grant 2015CFA010, and the 111 project under Grant B17040.

W. J. Su, L. F. Chen, M. Wu, M. T. Zhou, Z. T. Liu, and W. H. Cao are with the School of Automation, China University of Geosciences, Wuhan 430074, China, and also with the Hubei Key Laboratory of Advanced Control and Intelligent Automation for Complex Systems, Wuhan 430074, China. (Corresponding author: chenluefeng@cug.edu.cn)

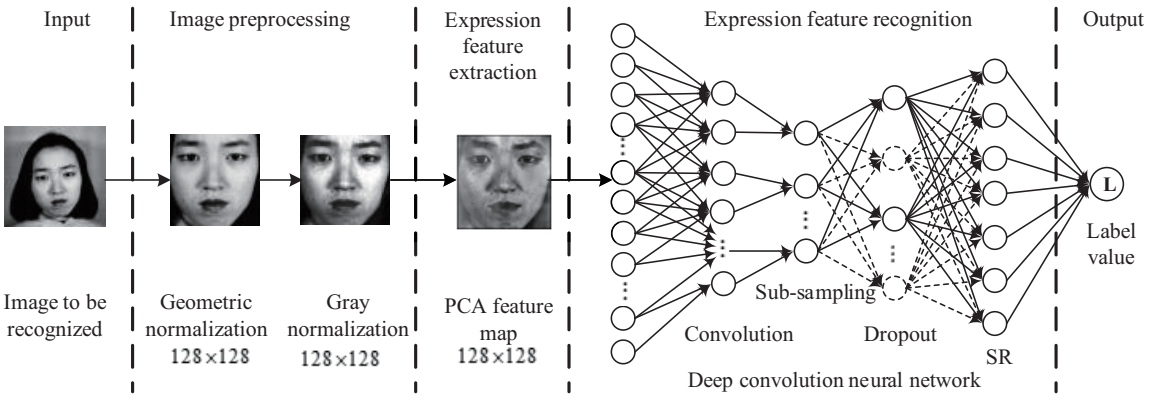

Fig. 1. The whole structure of the NAGDCNN for FER

Optimization algorithm plays a central role in DL [19], traditional optimization algorithms are easy to fall into poor local optimum, e.g., batch gradient descent, stochastic gradient descent, etc. Nesterov [20] introduced acceleration in the context of gradient descent, showing that it realizes an improved convergence rate with respect to gradient descent. Hence, the network weights are optimized by using Nesterov accelerated gradient descent (NAGD) where parameters are updated according to the previous momentum, and then the gradient is corrected to achieve the parameter updating. This pre-update method prevents large oscillations, misses the minimum, and is more sensitive to parameter updates. In sum, NAGD not only can predictably prevent the algorithm from going too fast or too slow, but also can enhance the response capability of the network.

In general, the Nesterov accelerated gradient descent-based convolution neural network (NAGDCNN) with dropout is proposed in this parper which first conducts expression image preprocessing, then extracts expression feature based on PCA, finally the DCNN is used to learn expression feature and classify it where the dropout can improve the efficiency of feature learning and alleviated overfitting. Moreover, the NAGD is used to optimize the network weights so that the proposed method can achieve good recognition results even if the number of available training samples is small.

The remainder of this article is structured as follows. In section II, the NAGDCNN for FER is introduced. Experimental results and analysis are presented in section III. In section IV, a summary of the proposal will be made.

II. Nesterov accelerated gradient descent-based convolution neural network with dropout

This section will describe the NAGDCNN for FER's specific implementation process, the structure of it is shown in Fig. 1.

A. Expression Image Preprocessing

Generally, the facial expression image contains a lot of noise which isn't conducive to the extraction of expression features. Thus, it's necessary to conduct the expression image preprocessing before the expression feature extraction, that is, geometric normalization and gray normalization.

Firstly, the eye and nose coordinates are acquired manually. Then the image is rotated according to the coordinate value so that the both eyes are on a horizontal line. After clipping the face, the rich distribution of expression feature is obtained. Finally, perform histogram equalization on facial expression sub-region to normalize the facial expression image's gray scale.

B. Expression Feature Extraction Based on Principal Component Analysis

After pretreatment, the facial expression image still has the problems of data redundancy and high dimension, while the PCA can extract the expression feature data which is easy to deal with, so the extraction of expression feature is based on PCA, whose process is shown as follows.

Step 1: Store the facial expression image data in matrix, that is, $\{x^{(1)}, x^{(2)}, ..., x^{(n)}\}$, $x^{(i)} \in R^{(n)}$, $n = 128$.

Step 2: Subtract the average brightness of the image data by (1).

$$x^{(i)} = x^{(i)} - \frac{1}{n}\sum_{i=1}^{n} x^{(i)} \tag{1}$$

Step 3: Compute the eigenvector U of the image data, so x can be expressed by the base vector $\{u_1, u_2, ..., u_n\}$ of the eigenvector U, then, the x_{rot} is obtained.

Step 4: Choose the k principal component to retain the variance of 99% by (2). And retain the k main components, set the rest to 0, namely, the $\tilde{x}$ is obtained which is an approximate representation of x_{rot},

$$\frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{n} \lambda_j} \geq 0.99 \tag{2}$$

Step 5: Rescale the data by PCA whitening which is shown in (3).

$$x_{PCAwhite,i} = \frac{\tilde{x}_i}{\sqrt{\lambda_i + \varepsilon}} \tag{3}$$

Step 6: Transform the covariance matrix into unit matrix by ZCA whitening which is shown in (4).

$$x_{ZCAwhite} = Ux_{PCAwhite} \tag{4}$$

After the steps above, the desired expression features are obtained. Compared with the original data, the extracted feature data's dimensionality and redundancy is greatly reduced.

C. Expression Feature Recognition Based on Deep Neural network

The concrete structure of DCNN is shown in Fig. 2, which is composed of a convolution layer and a sub-sampling layer, combined with dropout layer and SR to form a DCNN. The DCNN sets 100 convolution kernels with a dimension of 100×100 in the convolution layer, and the convolution kernel moves at a step size of 1. The sub-sampling layer uses the average pooling, the average pooling dimension is set to 4, and the moving step is 4.

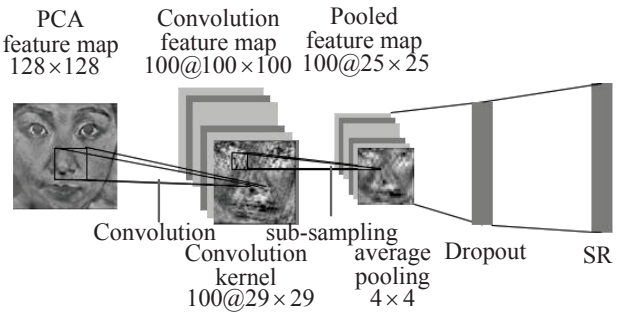

Fig. 2. Deep convolution neural network

1) Convolution Neural Network for Feature Learning: The CNN uses the convolution kernel to implement local connections between neurons in the neighboring layer to excavate local association information in the PCA feature maps, that is, excavating high-level expression features. The convolution process is implemented by

$$a_i^{(l)} = f(x_i^{(l)} * rot90(W_k^{(l)}, 2) + b_k) \tag{5}$$

where $x_i^{(l)}$ is the *ith* input expression feature map data, $x_i^{(l)} * W_k^{(l)}$ is the "valid" convolution operation of the *ith* input and the *kth* convolution kernel in the *lth* layer of the network, and b_k is the deviation corresponding to the *kth* convolution kernel. The $f(\cdot)$ is an activation function, which is given by

$$f(x) = \frac{1}{1+e^{-x}} \tag{6}$$

The convolution feature map generated by the convolution layer is then input to the sub-sampling layer for average pooling, thereby weakening the sensitivity of the network to the change of the expression feature position in the feature map. The average pooling implementation formula is

$$a_i^{(l)} = f(a_i^{(l-1)} * \frac{1}{p^2}) \tag{7}$$

where $a_i^{(l-1)}$ is the *ith* convolution feature map in the *l-1* layer, $p = 4$ is the average pooling dimension. And $a_i^{(l-1)} * (1/p^2)$ is also the "valid" convolution operation.

2) Dropout for Reducing Overfitting: The dropout layer can effectively solve the overfitting problem of the DNN and improve the efficiency of feature learning, so the dropout layer is added before the SR layer to alleviate the overfitting. In the training phase, it randomly lets all nodes of the input layer don't work at each iteration, but their weights will be preserved. This is a random process which means that each time the parameter is updated, the effective structure of the network will change, and the calculation process is shown in (8). In the test phase, all neurons of the network are activated to make the network into a complete structure, which is equivalent to the collection of multiple neural networks, of which calculation process is as shown in (9).

$$DropoutTrain(x) = RandomZero(p) \times x \tag{8}$$

$$DropoutTest(x) = (1-p) \times x \tag{9}$$

Dropout layer introduces the sparsity of randomization, each training optimization of the network structure are not the same, it weaks the joint adaptability between neurons, helps to improve the generalization ability of the network, this is similar to sexual reproduction in natural selection.

3) Softmax Regression for Feature Classification: SR is the output layer of the network to classify the learned expression feature. For training set $\{(x^{(1)}, y^{(1)}), ..., (x^{(m)}, y^{(m)})\}$, there is $y^{(i)} \in \{1, 2, ..., k\}$. In SR, it uses the hypothesis function $h_\theta(x)$ to calculate the probability $p(y = j|x)$ of the input *x* in each category *j*. The output of $h_\theta(x)$ is a *k*-dimensional vector, and the vector elements' values correspond to the probability values of the *k* classes respectively, the sum of vector elements is 1. The form of $h_\theta(x)$ is

$$h_\theta(x^{(i)}) = \begin{bmatrix} p(y^{(i)}=1|x^{(i)};\theta) \\ p(y^{(i)}=2|x^{(i)};\theta) \\ \vdots \\ p(y^{(i)}=k|x^{(i)};\theta) \end{bmatrix} = \frac{1}{\sum_{j=1}^{k} e^{\theta_j^T x(i)}} \begin{bmatrix} e^{\theta_1^T x(i)} \\ e^{\theta_2^T x(i)} \\ \vdots \\ e^{\theta_k^T x(i)} \end{bmatrix} \tag{10}$$

SR uses cost function to evaluate its classification effect in training phase, the definition of Softmax cost function is

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{k} 1\{y^{(i)}=j\} \log \frac{e^{\theta_j^T x^{(i)}}}{\sum_{l=1}^{k} e^{\theta_j^T x^{(i)}}}\right] + \frac{\lambda}{2}\sum_{i=1}^{k}\sum_{j=0}^{n}\theta_{ij}^2 \tag{11}$$

where $1\{y^{(i)} = j\}$ is the exponential function, the value of the rule is $1\{$ a true statement $\} = 1$ or $1\{$ a false statement $\} = 0$.

The derivation of (11) is

$$\nabla_{\theta_j} J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[x^{(i)}(1\{y^{(i)}=j\} - p(y^{(i)}=j|x^{(i)};\theta))] + \lambda\theta_j \tag{12}$$

When training the SR, the (12) is substituted into the NAGD to be minimized, and the concrete implementation steps will be described in detail in the next section.

4) Nesterov accelerated gradient descent for optimizing: The NAGD is the improvement of stochastic gradient descent based on momentum (SGDM), NAGD can see the gradient of the next update in advance, if the next step update gradient is larger than the current position gradient, the update will be set larger when updating; if the next step update gradient is smaller than the current position gradient, the current velocity vector v is set smaller when updating. That is to say, NAGD not olny retain the advantages of SGDM, but also has the ability to predict which predictably prevent the algorithm from going too fast or too slow, and can enhance the response capability of the algorithm.

NAGD uses the momentum term γv_{t-1} to update θ, and calculates $\theta - \gamma v_{t-1}$ to obtain the approximate values of the future positions of the θ. NAGD update is given by

$$v_t = \gamma v_{t-1} + \alpha \nabla_\theta J(\theta - \gamma v_{t-1}; x^{(i)}, y^{(i)}) \tag{13}$$

$$\theta = \theta - v_t \tag{14}$$

where $\nabla_\theta J(\theta; x^{(i)}, y^{(i)})$ is the gradient of the θ calculated from the training set $(x^{(i)}, y^{(i)})$, α is a learning rate that prevents a large offset in cost function. The v is the current velocity vector, which is the same as the dimension of θ. And $\gamma \in (0, 1]$ determines how many gradient iterations will be involved in the update.

III. Experiment on Facial Expression Recognition

A. Experimental Environment Setting

Topological structure of HRI system based on FER is shown in Fig. 3, which is mainly composed of a wheeled robot, an affective computing workstation, routers and data transmission equipment. The system first acquires facial expression image frame data through Kinect configured on the wheeled robot, then transmits the data to the affective computing workstation. Then, the workstation will input the data into the trained FER system to identify the expression, the result will be feedback to the wheeled robot, so that wheeled robots can achieve natural and harmonious interaction with people.

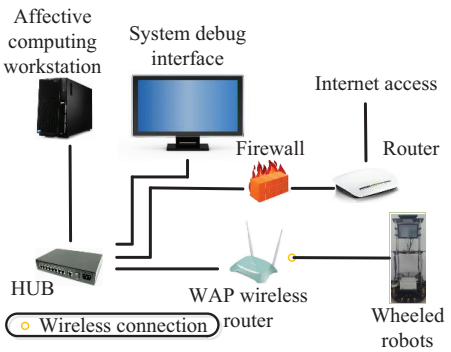

Fig. 3. Topological structure of HRI system based on FER

B. Experimen on JAFFE and CK+

1) Experimental Samples Setting: Ekman and Friesen [21] defined 6 basic expressions: happiness (HA.), angry (AN.), surprise (SU.), fear (FE.), disgust (DI.) and sadness (SA.), each of which can be used as the only expression to reflect a unique emotion, it has also become the standard of FER classification. In this paper, 7 facial expressions are recognized, namely, in addition to recognizing the 6 basic expressions, the neutral (NE.) is also recognized. Two benchmark databases, JAFFE [22] and the Extended Cohn-Kanade (CK+) [23], are chosen to evaluate the validity of the proposal.

The JAFFE contains 213-grayscale images depicting 10 Japanese female subjects. Each subject has 2 to 4 images of each facial expression. Select 80% of the database for the training set, another 20% of the database for the test set. The partial expression image samples of JAFFE are shown in the Fig. 4.

Fig. 4. The partial expression image samples of JAFFE

The CK+ consists of 18 to 50 years old adults in 210 different racial and gender, contains expression image sequences starting from the neutral state and finishing at the emotion apex. The images of emotion apex from each sequence are selected, and the neutral emotional state images are also selected, hence, the 399 images composed of seven basic expressions are selected. The CK+'s coverage of the age, gender and race is wider, in order to learn all kinds of people's various expression features better , the CK+ uses greater proportion of training set than JAFFE, that is, select 90% of the database for the training set, the remaining 10% for the test set. The partial expression image samples of CK+ are shown in the Fig. 5.

Fig. 5. The partial expression image samples of CK+

2) Experimental results and analysis: The experiment was carried out by an affective computing workstation that is shown in Fig. 3, which was configured to use the Intel Core i5-4590 CPU processor having a 3.3 GHz system clock and a RAM of 4.00 GB. The experimental software is MATLAB R2016a, and the experiment on benchmark database is designed to verify the effectiveness of the proposal.

One of the common problems in DL is that it requires a lot of data to learn at the training phase [24]. However, the amount of data available in the public databases is insufficient to satisfy the data required by DL. Therefore, in order to increase the number of training samples, the original image is symmetrically transformed, doubling the

number of database samples. We conducted two experiments on JAFFE, one by using the original data version and the other one by using the enriched versions. In the experiments, the p in dropout is 0. Table I gives the results of the experiments, showing that the enrich data can effectively improve accuracy.

TABLE I
RESULTS OF ORIGINAL DATA AND ENRICHED DATA

Index	Original data	Enriched data
Average accuracy (%)	80.85	87.23

Fig. 6 shows the FER accuracy and training time varying with the p in the dropout layer. It's obvious that with the increasing of p, training time is gradually shortened, while the accuracy has an increasing trend, which indicates that when training the DCNN, selecting an appropriate value in the dropout layer will help improve the generalization performance of the network and shorten the training time. Considering the influence of training time and accuracy, this paper chooses $p = 0.5$ as the optimal value.

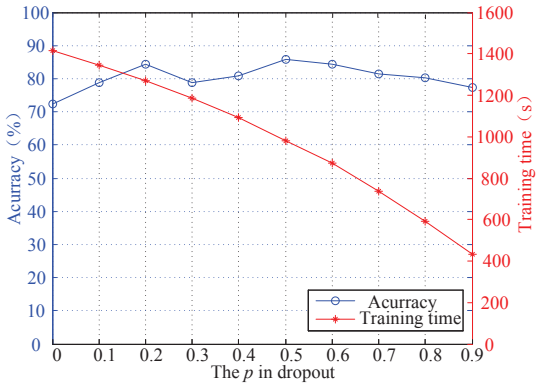

Fig. 6. The influence of the p on the accuracy and training time

In order to verify the effectiveness of the proposal, experiments are carried out on JAFFE and CK+ databases respectively. The experiments also use SGDM and NAGD to train the network respectively, and each training image increases one symmetrical transformation image. Table II lists the results of experiments, the proposal achieves better accuracy on both JAFFE and CK+ databases, with the accuracy of 97.62% on JAFFE and the accuracy of 95.12% on CK+.

It is worth mentioning that, Table III shows the utility time of our method and other method, in our method the average training time of each image is 0.6757 seconds, and the average test time of each image is 0.1258 seconds which is the half of SPF [13]. We can see from Fig.7 and Fig.8, the 6 kinds of expression in addition to SA. all achieved good accuracy in the both database. In the JAFFE database, the accuracy of SA. is 83%, 17% is wrongly classified as DI. In the CK+ database, the accuracy of SA. is 60% , 20% were misclassified as AN., and 20% were misclassified as NE. This may due to people are not good at expressing the emotion of SA., and the expression range is small which is difficult to be displayed in the image, so it is easy to be mistakenly classified into other categories.

TABLE II
COMPARISONS WITH STATE-OF-THE-ART APPROACHES

Method	JAFFE	CK+
NAGDCNN	97.62 %	95.12 %
SGDM+DCNN	92.86 %	90.24 %
PCA+CDA [10]	61.43 %	71.84 %
SFP [13]	91.8 %	94.09 %
3DCNN-DAP [16]	\	92.4 %

TABLE III
COMPARISONS OF THE UTILITY TIME

Method	Training time	Recognition time
The proposal	0.6757 s	0.1258 s
SFP [13]	\	0.2955 s

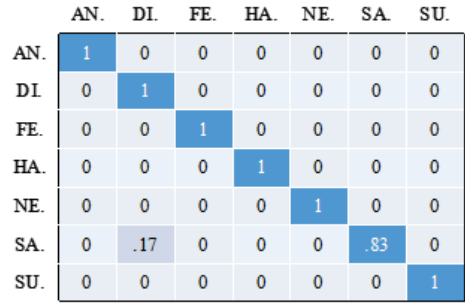

Fig. 7. The confusion matrix of JAFFE

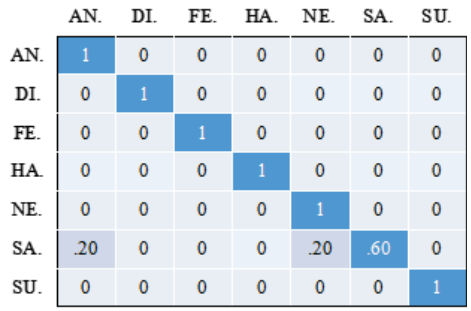

Fig. 8. The confusion matrix of CK+

C. Results and Analysis of the Application Experiment

The application experiment is carried out by the affective computing workstation, system debug interface and Kinect which are given by Fig. 3. In system debug interface, the GUI interface is built which is shown in Fig. 9. In the GUI interface, the system will call the Kinect's color camera to capture real-time image and the image will be displayed in the window of "image capture" when the image preview

button is clicked. When the expression recognition button is clicked, the current image frame will be captured and displayed in the window of "image frame to be recognized", after the manual for eyes and nose coordinates, the face is corrected and clipped, then the face image will be input to the trained FER system, the final recognition results will be displayed in the GUI interface.

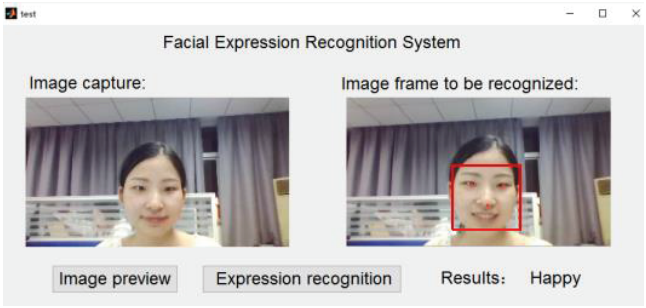

Fig. 9. GUI interface diagram

The 7 basic expression image frames of 2 groups of 3 individuals are collected as training sets and input to the proposal for training. After that, the Kinect captured real-time image frames are input to the trained network for recognition. Table IV gives the online recognition results, three groups of average recognition accuracy is 76.19%, showing the application prospect of the proposal. There are some ways to improve the experiment in the subsequent work, such as the training set can be lager to meet the shortcomings of the DL which requires a large amount of data for learning.

TABLE IV
THE RESULTS OF THE APPLICATION EXPERIMENT

	AN.	DI.	FE.	HA.	NE.	SA.	SU.
First group	AN.	SA.	FE.	DI.	NE.	SA.	SU.
Second group	AN.	DI.	FE.	DI.	DI.	SA.	SU.
Third group	AN.	NE.	FE.	HA.	NE.	SA.	SU.

IV. CONCLUSIONS

In this paper, the NAGDCNN is proposed for FER, which fuses the CNN with SR to construct a DCNN, and the dropout layer is added after the sub-sampling layer, where the NAGD is used to optimize the network weights. Compared to other method, the proposal has two advantages:

(1) The dropout layer can effectively alleviate overfitting and reduce the network's training time that improve the efficiency of the algorithm.

(2) The NAGD can predictably prevent the algorithm from going too fast or too slow and enhance the response capability of the network.

The results of the experiments on benchmark database and application experiment evaluate the validity of the proposal.

REFERENCES

[1] S. Poria, E. Cambria, R. Bajpai, *et al.*, "A review of affective computing: From unimodal analysis to multimodal fusion", *Information Fusion*, vol. 37, pp. 98-125, 2017.

[2] A. Mehrabian, "Communication without words", *Psychological today*, vol. 2, no. 4, pp. 53-56, 1968.

[3] Deshmukh, Shubhada, M. Patwardhan, *et al.*, "Survey on real-time facial expression recognition techniques", *IET Biometrics*, vol. 5, no. 3, pp. 155-163, 2016.

[4] X. M. Zhao, S. Q. Zhang, "A review on facial expression recognition: feature extraction and classification", *IETE Technical Review*, vol. 33, no. 5, pp. 505-517, 2016.

[5] M. N. Islam, M. Seera, C. K. Loo, "A robust incremental clustering-based facial feature tracking", *Applied Soft Computing*, vol. 53, pp. 34-44, 2017.

[6] T. Zhang, W. Zheng, Z. Cui, *et al.*, "A deep neural network-driven feature learning method for multi-view facial expression recognition", *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2528-2536, 2016.

[7] S. J. Wang, W. J. Yan, T. Sun, *et al.*, "Sparse tensor canonical correlation analysis for micro-expression recognition", *Neurocomputing*, vol. 214, pp. 218-232, 2016.

[8] W. Gu, C. Xiang, Y. Venkatesh, *et al.*, "Facial expression recognition using radial encoding of local Gabor features and classifier synthesis", *Pattern Recognition*, vol. 45, no. 1, pp. 80-91, 2012.

[9] X. S. Shi, Z. H. Guo, F. P. Nie, *et al.*, "Two-dimensional whitening reconstruction for enhancing robustness of principal component analysis", *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 38, no. 10, pp. 2130-2136, 2016.

[10] K. Papachristou, A. Tefas, I. Pitas, "Symmetric subspace learning for image analysis", *IEEE Transactions on Image Processing*, vol. 23, no. 12, pp. 5683-5397, 2014.

[11] Y. Sun, A. N. Akansu, "Facial expression recognition with regional hidden Markov models", *Electronics Letters*, vol. 50, no. 9, pp. 671-673, 2014.

[12] X. J. Fan, T. Tjahjadi, "A dynamic framework based on local Zernike moment and motion history image for facial expression recognition", *Pattern recognition*, vol. 64, pp. 399-406, 2017.

[13] S. L. Happy, A. Routray, "Automatic facial expression recognition using features of salient facial patches", *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 1-12, 2015.

[14] Y. Guo, Y. Liu, A. Oerlemans, *et al.*, "Deep learning for visual understanding: A review", *Neurocomputing*, vol. 187, pp. 27-48, 2016.

[15] G. E. Hinton, R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks", *Science*, vol. 313, no. 5786, pp. 504-507, 2006.

[16] M. G. Liu, S. X. Li, S. G. Shan, *et al.*, "Deeply learning deformable facial action parts model for dynamic expression analysis", *Proceeding of the 12th Asian Conference on Computer Vision (ACCV)*, pp. 143-157, 2014.

[17] J. K. Chen, Q. H. Ou, Z. R. Chi, *et al.*, "Smile detection in the wild with deep convolutional neural networks", *Machine Vision and Applications*, vol. 28, no. 1, pp. 173-183, 2017.

[18] N. Srivastava, G. Hinton, A. Krizhevsky, *et al.*, "Dropout: A simple way to prevent neural networks from overfitting", *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929-1958, 2014.

[19] A. Wibisono, A. C. Wilson, and M. I. Jordan, "A variational perspective on accelerated methods in optimization", *Proceedings of the National Academy of Sciences of the United States of America,* , vol. 113, no. 47, pp. 7351-7358, 2016.

[20] Y. Nesterov, "A method of solving a convex programming problem with convergence rate $O(1/k^2)$", *Soviet Mathematics Doklady*, vol. 27, no. 2, pp. 372-376, 1983.

[21] P. Ekman, W. V. Friesen, "Constants across cultures in the face and emotion", *Journal of personality and social psychology*, vol. 17, no. 2, pp. 124-129, 1971.

[22] The Japanese female facial expression (JAFFE) Database, 1998, http://www.kasrl.org/jaffe.html

[23] Cohn-Kanade (CK and CK+) database download site, 2000, http://www.consortium.ri.cmu.edu/data/ck/

[24] H. Chen, D. Ni, J. Qin, *et al.*, "Standard plane localization in fetal ultrasound via domain transferred deep neural networks", *IEEE Journal of Biomedical & Health Informatics*, vol. 19, no. 5, pp. 1627-1636, 2015.