Benign and Malignant Solitary Pulmonary Nodules Classification Based on CNN and SVM

LIU Lu
Harbin University of Science and Technology
Harbin 150080,China
+8613796678312
liulu@hrbust.edu.cn

LIU Yapeng
Harbin University of Science and Technology
Harbin 150080,China
+8618845153910
1194184203@qq.com

ZHAO Hongyuan
Harbin University of Science and Technology
Harbin 150080,China
+8615303687500
616465660@qq.com

ABSTRACT
In order to assist the doctors to diagnose lung cancer and improve the classification accuracy of benign and malignant pulmonary nodules, this paper proposes a novel intelligent diagnosis model which is aiming at CT imaging features of pulmonary nodules . Specifically, this model uses the convolutional neural network to extract the features of the pulmonary nodules, then uses the principal component analysis to reduce the dimension of the extracted features, and finally classifies the final features with particle swarm optimization optimized SVM. With regard to the pulmonary nodules extracted from the LIDC-IDRI database, 400 pulmonary nodules are used for training and 310 pulmonary nodules are used for testing, the classification accuracy rate is 91.94%. This model can provide objective, convenient and efficient auxiliary method for solving the classification problem of benign and malignant pulmonary nodules in medical images.

CCS Concepts
• Applied computing → Life and medical sciences →Health care information system• Applied computing → Life and medical sciences → Health informatics•Computing methodologies → Machine learning → Machine learning approaches• Computing methodologies →Machine learning → Learning settings

Keywords
Benign and malignant nodules;convolutional neural network; lung cancer ;particle swarm optimization

1. INTRODUCTION
In the last few decades, lung cancer is a malignant tumor with high morbidity and mortality, which seriously endangers human life and health. In medical image processing, the early diagnosis of lung cancer is Solitary Pulmonary Nodule (SPN)[1], SPN is a single,circular, well-defined lung, with no pulmonary or mediastinal lymph nodes enlarged,no atelectasis,no pleural effusion,no mare than 30mm diameter nodules. As shown in Figure 1, four radiologists divide the SPN contour information delineated on the same computed tomography (CT) image. Clinically, if early can accurate judgement of SPN benign and malignant and benign SPNs avide surgery, malignant SPNs resection surgery, to a great extent, can give patients to reduce pain and improve the survival rate In this paper, the classification of benign and malignant solitary pulmonary nodules based on deep learning method and particle swarm optimization of the support vector machine method,used to assist doctors to diagnose and improve their work efficiency.

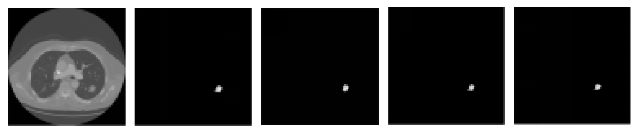

Figure.1 The lung CT image and four different radiologists outlined the solitary pulmonary nodul

2. RELATED RESEARCH
As computer-aided diagnostics are continually applied to the field of medical image research, domestic and foreign scholars have done a lot of valuable work on SPN benign and malignant classification research, the research focused on the feature extraction of pulmonary nodules and the design of the classifier.

2.1 Feature Extraction
In the traditional method, SPN classification is mainly based on morphological features classification diagnosis, growth rate classification diagnosis, texture feature classification diagnosis, and synthetic feature classification diagnosis. The literature [2] is based on the two-dimensional texture feature of gray symbiosis matrix to classify pulmonary nodules. Some researchers have studied the morphological features, texture features, and growth rates of solitary pulmonary nodules, the literature [3] based on the shape, size, brightness and texture of pulmonary nodules, four categories of 64 characteristics to determine its benign and malignant, the literature [4] based on the location , bias, size, growth rate, calcification and other characteristics of pulmonary nodules to determine the benign and malignant, and by comparison to get the size and calcification is its most effective features. However, there are some problems in the classification of pulmonary nodules based on artificial feature extraction: artificial extraction is influenced by human factors, and the process is complicated and difficult to generalize to other data.

In recent years, due to the deep learning methods has made a breakthrough in the image processing[5], medical image processing has begun the application of the deep learning methods [6-8]. In the literature [9], the deep belief network model and the deep convolution neural network model are respectively discussed to apply to the SPN benign and malignant classification, because the deep learning method can automatically train the image and can extract the multi-level features of the image, and does not need complex image processing steps, the process is simple, based

ICMVA 2018, April 23–25, 2018, Singapore, Singapore
© 2018 ACM ISBN 978-1-4503-6381-5/18/04…$15.00
DOI: http://dx.doi.org/10.1145/3220511.3220513

on the deep learning method of computer-aided diagnosis shows a good advantage.

2.2 Classifier

In the current classification of benign and malignant pulmonary nodules, commonly used classifiers are artificial neural network and support vector machine (SVM). For example[10] high-dimensional multi-resolution histogram extraction are used to get the 768 dimension spatial information characteristics of pulmonary nodules, combining with support vector machine classification algorithm, classification, isolated pulmonary nodules qualitative diagnostic accuracy was 69.58%. In the literature [11], 88 features were extracted based on the shape and texture of pulmonary nodules, then, the main features were analyzed by principal component analysis (PCA), and six main features were obtained, finally, the classification was conducted by artificial neural networks with accuracy of 90.63%.

In summary, computer-aided diagnosis based on the deep learning method is simpler and more effective than the traditional method, SVM in solving small sample,nonlinear and high dimensional pattern recognition shows the unique advantages and has wide application. In this paper, the pulmonary nodule data of lung CT images was used as the research object, and the convolution neural network was applied to the feature extraction of pulmonary nodule images, the extracted features were identified by particle swarm optimization support vector machines algorithm.

3. DATA AND METHODS

Firstly, obtain clinical CT images of lungs, division to get the benign and malignant SPN images from the lung CT image, set up training set and test set. Secondly, after normalizing and pre-processing these SPN images, SPN is extracted by convolution neural network. Then, in order to improve the efficiency of the operation, principal component analysis is used to reduce the dimension. Finally, the Particle Swarm Optimization (PSO) optimized SVM classifier is used to classify and identify the final features. The overall process shown in Figure 2.

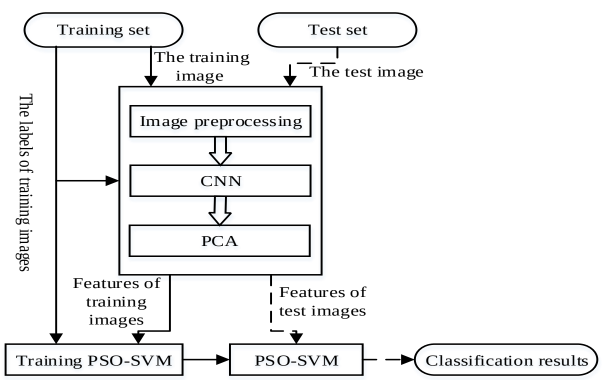

Figure.2 Process of pulmonary nodules classification with CNN-PSO-SVM

3.1 Data Set and Preprocessing

The Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) [12]is composed of chest medical image files and corresponding diagnostic results.

In order to improve the credibility of the study, the nodules selected in this paper are pulmonary nodules marked by three or four radiologists . Meanwhile, since the size of the lung nodes is less than 3mm,the features are not obvious,and they are dificult to identify,so the data of pulmonary nodules in the range of 3-30mm are selected for the study.

In the LIDC-IDRI database, a total of 620 pulmonary nodules were extracted: three or four radiologists agreed that there were 305 SPNs for malignancy<3, which we marked as benign; three or four radiologists agreed that malignancy>3 of the SPN has 315, we are marked as malignant. Part of the pulmonary nodule CT images shown in Fig.3.

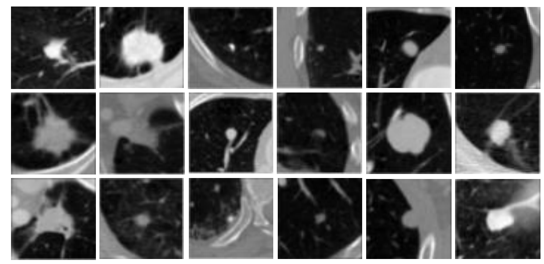

Fig.3 The CT image of SPNs.

The CT image data needs to be pretreated as a result of acquired ct images from different acquisition environments and different instrumentation. First, each original CT image was normalized. According to the SPN contour coordinate information drawn by the radiologist, we intersected the SPN area of interest. In order to reduce the interference of SPN tissues around the SPN, the grayscale value outside the SPN was set to 0. SPN as the center, intercept the size of 64 * 64 images for the next step. The LIDC-IDRI database has a total of four radiologists who interpret lung CT images, and each radiologist may have different contours for the same pulmonary nodule. Based on the study of the literature[13], this study group collected the different contour areas of the same pulmonary nodules delineated by four different radiologists, and obtained the integrated pulmonary nodule region as the follow-up study object. As shown in Figure 4.

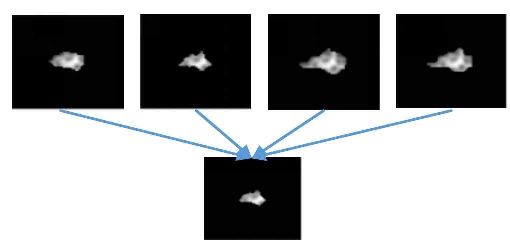

Figure.4 Extraction processing of area and boundary of the candidate nodules

3.2 Feature Extraction Based on Convolutional Neural Network

Convolutional Neural Network[14] is a learning algorithm for multi-layer neural networks. Illustration of convolutional neural networks as shown in Figure 5, CNN extracts the characteristics of the image through multiple serial convolution layers and pooled layers, and reduces the number of weights and makes it easier to train due to weight sharing.

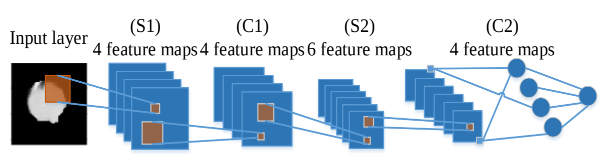

Figure.5 Illustration of convolutional neural networks

As shown in Figure 6, the convolution layer is convolved with the convolution of the target image, creating a feature graph,and realizing the partial feature perception and feature extraction[15].

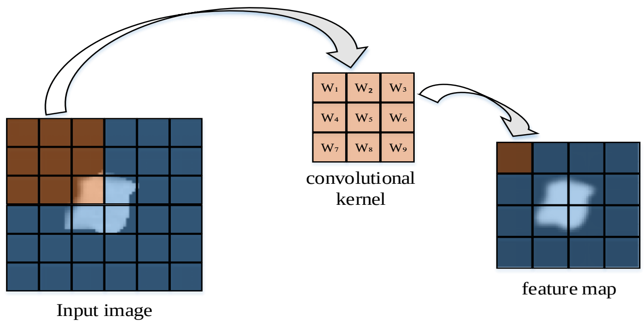

Figure.6 Convolutional layer

In the convolution neural network,the pooling sample is to use the reservoir to polymerize the characteristics of different locations,to reduce the dimensions of the features,and to achieve the translational invariance of the input image,and the improved robustness[16]. Figure 7 shows that the input image is pooled using the largest pooling method.

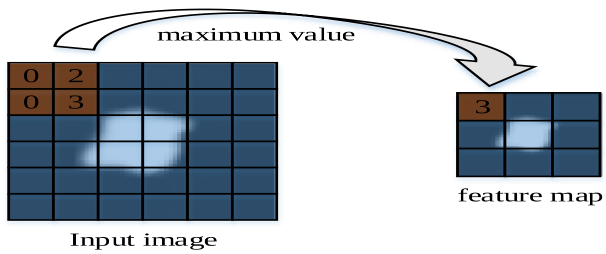

Fig.7 Pooling layer

A number of fully connected layers are typically preceded by a convolutional layer or a pooled sample layer of a depth convolution neural network. because the previous convolution layer and the pooled sampling layer have reduced the feature map to an acceptable size, the use of the fully connected layer does not result in a particularly large computational burden, which constitutes a shallow layer of the multi-layer perception machine, played a role as a classifier.

In this paper, the design of the convolution neural network structure is based on the network structure of literature[17]and after many experiments, the adaptive expansion is carried out: 3 convolution layers (C1, C3, C5), the convolution kernel sizes are 5*5, 5*5, 4*4 pixels, respectively, convolution step sizes are 1, 1 , 1 pixels, the number of features of the output were 4 , 8, 12; the three sub-sampling layers (S2, S4, S6) are in the form of maximum pooling, the pool size is 2*2 pixels, and the pooling step is 2 pixels; two full-connection layers (F7,F8), the F7 layer is a hidden layer containing 300 neurons, and the F8 layer is a hidden layer containing 50 neurons. Although the outputs of the two fully connected layers can be used as features of the image, The output data of the F7 layer contains more image information, so the output data of the F7 layer is used as the input data for the subsequent SVM classifier.

3.3 Particle Swarm Optimization Support Vector Machine Algorithm

SPN qualitative diagnosis based on CT image feature can be attributed to an image classification recognition problem. Support Vector Machine [18] is a kind of machine learning classification method based on statistical VC dimension theory and structural risk minimization principle. For the SPN classification problem, set the sample set $S=\{(x_i,y_i)\,|\,(x_i,y_i)\in R^m\times R, i=1,2,\cdots,n\}$, Where $x_i\in\chi=R^m$ is the sample vector, $y_i\in Y=\{-1,1\}$ is the category number, n is the number of samples. Assuming that there is a mapping $\varphi(x_i)$, the data x_i is mapped from the original feature space χ to the high-dimensional feature space F and the relaxation variable ξ_i is introduced, the original problem of SVM can be expressed as:

$$\min_{w,b,\xi}\frac{1}{2}\|\mathrm{w}\|^2+C\sum_{i=1}^{n}\xi_i$$

$$s.t.\; y_i(w\cdot\varphi(x_i)+b)\ge 1-\xi_i\,,\quad \xi_i\ge 0\,,\quad i=1,2,\cdots,n \quad (1)$$

The Lagrangian function can be used to derive the dual problem of the original problem:

$$\min_{\alpha}\frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{n}\alpha_i\alpha_j y_i y_j k(x_i,x_j)-\sum_{i=1}^{n}\alpha_i$$

$$s.t.\;\sum_{i=1}^{n}\alpha_i y_i=0\,,\quad 0\le\alpha_i\le C\,,\quad i=1,2,\cdots,n \quad (2)$$

Select the appropriate kernel function and parameter C, the classification decision function is:

$$f(x)=\mathrm{sgn}(\sum_{i,j=1}^{n}\alpha_i^* y_i k(x_i,x_j)+b^*) \quad (3)$$

Radial Basis Function (RBF) has better learning and classification performance for image classification. After selecting RBF as SVM kernel function, the parameters that need to be optimized in SVM are RBF parameter g and penalty factor C. If the parameters g and C are chosen empirically, the efficiency and outcome of the classification may not be satisfactory. Particle swarm optimization algorithm[19] is a heuristic optimization algorithm, it has a strong global search capabilities and optimization features. The particle swarm optimization algorithm can automatically optimize the SVM classification performance of the support vector machine.

3.4 Results Evaluation

The experimental results were measured by sensitivity (SEN), specificity (SPE) and accuracy (ACC). The expressions of SEN, SPE and ACC were:

$$SEN=\frac{TP}{TP+FN} \quad (4)$$

$$SPE=\frac{TN}{TN+FP} \quad (5)$$

$$ACC=\frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

Where: TP = true positive number, TN = true negative number, FP = false positive number, FN = false negative number.

4. EXPERIMENTS AND ANALYSIS

The experimental platform for the Win10 operating system, i7 quad-core 8-thread CPU, clocked at 4GHz, 32G memory, the software environment is MATLAB 2016b. In the following experiments: the training samples consisted of 400 64×64 pulmonary nodule CT images, including 200 benign pulmonary nodules and 200 malignant pulmonary nodules. The test samples consisted of 310 additional 64×64 pulmonary nodular CT images composition, of which 160 cases of benign pulmonary nodules, 150 cases of malignant pulmonary nodules.

To test the speed and stability of CNN network during training, the rl curve is defined and mapped:

$$rL(n+1) = 0.99 \times rL(n) + 0.01 \times e \tag{7}$$

Where: $n \in [0, \frac{num}{bnum} \times iterations + 1]$, num is the total number of training samples, $bnum$ is the number of samples per batch, $iterations$ is the total number of iterations for the sample, e is the mean square error of the current output and the actual result. The mean square error rL curve is a smoothing sequence with minimum mean square error, and its shape shows that with the increase of the number of iterations, the CNN model training process predicts the error. On the other hand also represents the speed and stability of the network convergence. As shown in Fig. 8, the rL curve gradually decreases with the increase of the number of iterations in the training process, show that the training network is stable and reliable,the CNN can be used as the subsequent recognition.of SPN.

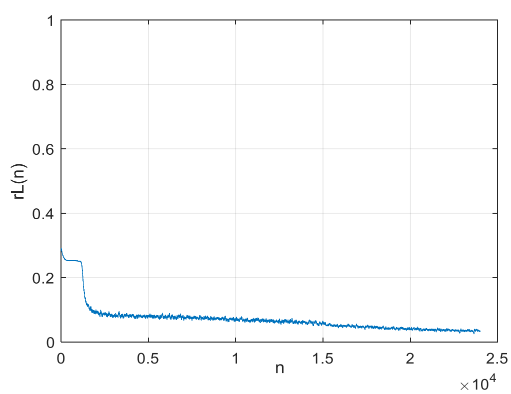

Figure.8 The convergence process of CNN structure in the process of training

After CNN completes the training, the output data of the F7 layer in CNN is taken as the image feature of SPN, which contains 300 neurons. In order to reduce the time complexity of SVM parameter optimization classification, the classical PCA algorithm is used to simplify the output data of the F7 layer, keep 98% of its energy, and obtain the input data of SVM classifier. The particle swarm optimization algorithm was used to optimize the SVM parameters. the initial population was 20, the evolutionary algebra was 120, the acceleration coefficient C1 = 1.4, C2 = 1.6. The fitness curve obtained from the training is shown in Fig. 9. As the evolutionary algebra of the particle population increases, the fitness changes, and when the evolutionary algebra reaches a high degree of fitness at about 39, the penalty factor c and the kernel function g were 34.1803 and 0.4036, respectively.

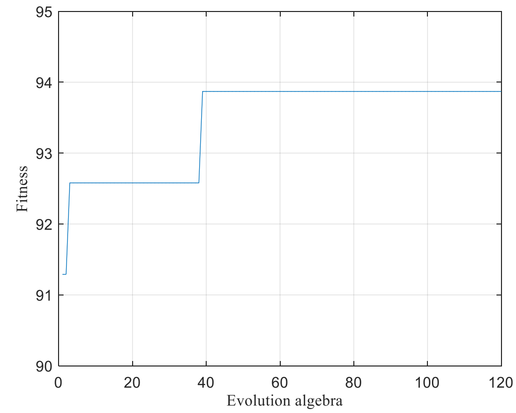

Figure.9 The fitness curve of particle swarm algorithm

By testing the test set, the convolution nerve network of this article on the benign and malignant classification of pulmonary nodules, as shown in Table 1.

Table 1 Classification results based on CNN method

	sample	correct	error	Accuracy (%)
malignant nodules	150	131	19	87.3
benign nodules	160	145	15	90.625
all nodules	310	276	34	89.03

Using the particle swarm optimization SVM as CNN classifier, CNN-PSO-SVM classification model for benign and malignant classification of pulmonary nodules, as shown in Table 2.

Table 2 Classification results based on CNN-PSO-SVM method

	sample	correct	error	Accuracy (%)
malignant nodules	150	137	13	91.3
benign nodules	160	148	12	92.5
all nodules	310	285	25	91.94

Since the optimization of the SVM parameters is continuous by the particle swarm optimization algorithm, the optimization parameters with better classification accuracy can be obtained. Compare the above data can be found that the accuracy of the whole classification was 91.94%, which was about 3%-4% higher than that of CNN alone, using the Particle Swarm Optimization Support Vector Machine algorithm as the classifier of CNN network to determine the benign and malignant of pulmonary nodules.

In order to further compare the superiority of CNN-SVM method in this paper, two other methods of this method are compared, the literature [2] based on the gray-level co-occurrence matrix and the

BP neural network classification algorithm, the literature [10]uses high-dimensional multi-resolution histogram and combined with SVM classification algorithm. Comparison of experimental results shown in Table 3, compared with the traditional method of this method, the classification accuracy is further improved. In this paper, we can automatically extract the feature of the extracted image, and the particle swarm optimization algorithm can optimize the support vector machine model parameters, which can avoid the randomness of artificial selection and have good adaptive learning and optimization ability.

Table 3 Performance comparison between proposed method and others

method	Accuracy(%)
MRH+SVM	72.58
GLCM+BP	88.71
CNN+PSO+SVM	91.94

5. CONCLUSION

Doctors in reading, by personal experience and knowledge, and sometimes can not effectively use the imaging features of pulmonary nodules, affecting the efficiency of diagnosis. In this paper, a novel CNN-PSO-SVM classification model for the classification of benign and malignant pulmonary nodules is proposed by a simple summary of the previous work. Firstly, the features of the pulmonary nodule are extracted by CNN network. Secondly, the extracted features are reduced by PCA. Finally, the parameters c and g in SVM are trained by PSO algorithm, and the performance of SVM classifier is improved. The overall classification effect is optimized. Convolution neural network can easily and effectively extract the characteristics of pulmonary nodules, particle swarm optimization algorithm to support vector machine model parameters, to avoid the random selection of man-made. The results show that the model has a certain improvement in the accuracy of pulmonary nodule benign and malignant classification, and the processing is simple, convenient, high efficiency, comprehensive identification ability is superior to the general classification model. This model can provide a theoretical reference for the young physician, especially the inexperience young physician, and have good application prospect.

6. ACKNOWLEGEMENTS

This work is supported by the Heilongjiang Province Natural Science Fund Project of China.(F201208)

7. REFERENCES

[1] Zhou Qinghua,Fan Yaguang,Wang Ying,et al.China national guideline of classification,diagnosis and treatment for lung nodules.Chinese Journal of Lung Cancer,2016, 19(12),793-798.

[2] Anand,S.K.V., Segmentation coupled textural feature classification for lung tumor prediction. 2010 IEEE International Conference on Communication Control and Computing Technologies, 2010,pp.518-524.

[3] Zinovev ,D., Raicu ,D., Furst, J., et al. Predicting Radiological Panel Opinions Using a Panel of Machine Learning Classifiers.Algorithms,2009, 2(4),1473-1502.

[4] Brader ,P., Abramson ,S. J., Price ,A. P., et al. Do characteristics of pulmonary nodules on computed tomography in children with known osteosarcoma help distinguish whether the nodules are malignant or benign.Journal of Pediatric Surger,2011, 46(4),729.

[5] Lecun ,Y., Bengio, Y., Hinton ,G.,et al. Deep learning.Nature,2015, 521(7553),436-444.

[6] Greenspan,H., Ginneken,B.V., Summers,R.M., et al.Guest Editorial Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique.IEEE Transactions on Medical Imaging,2016, 35(5),1153-1159.

[7] Khatami ,A., Khosravi, A., Nguyen, T., et al. Medical Image Analysis using Wavelet Transform and Deep Belief Networks.Expert Systems with Applications,2017, 86.

[8] Litjens ,G., Kooi ,T., Bejnordi, B. E., et al. A survey on deep learning in medical image analysis.Cell Reports,2017, 19(9),1953-1966.

[9] Hua, K. L., Hsu, C .H., Hidayati ,S. C., et al. Computer-aided classification of lung nodules on computed tomography images via deep learning technique.Oncotargets & Therapy,2015, 8,2015.

[10] Liu Lu,Liu Wanyu,Chu Chunyu,et al.Fast classification of benign and malignant solitary pulmonary nodules in CT image.Opticsand Precision Engineering,2009, 17(8),2060-2068

[11] Dand,E., et al. Artificial neural network-based classification system for lung nodules on computed tomography scans.2014 IEEE International Conference of Soft Computing and Pattern Recognition,2014,pp.382-386.

[12] Novo ,J., Gonçalves, L., Mendonça, A. M., et al. 3D lung nodule candidate detection in multiple scales.2015 IEEE Iapr International Conference on Machine Vision Applications,2015,pp.61-64.

[13] Han Fangfang,et al.Research on the detection and diagnosisi methods for pulmonary nodules based on multi-dimensional features from CT images.Northeastern University,2015

[14] Chang Liang,Deng, X.,Zhou, M.,et al.Convolutional neural network in image understanding.Acta Automatica Sinica,2016, 42(9),1300-1312.

[15] Lecun, Y., Kavukcuoglu, K ., Farabet, C., Convolutional networks and applications in vision. 2010 IEEE International Symposium on Circuits and Systems. 2010,pp.253-256.

[16] Kang, K., Wang, X.,et al. Fully Convolutional Neural Networks for Crowd Segmentation.Computer Science,2014, 49(1),25–30.

[17] Zhou, L., Li ,Q., Huo, G.,et al. Image Classification Using Biomimetic Pattern Recognition with Convolutional Neural Networks Features.Computational Intelligence and Neuroscience,2017.

[18] Chang ,C. C., Lin ,C. J., LIBSVM: A library for support vector machines. Acm Transactions on Intelligent Systems & Technology,2011, 2(3),1-27.

[19] Gao Ying,Xie Shengli,et al.Particle swarm optimization algorithms with immunity.Computer Engineering and Application,2004, 40(6):4-6