A Multi-scale Attention-based Facial Emotion Recognition Method Based on Deep Learning

ZHANG Ning, ZHANG Xiufeng, FU Xingkui, QI Guobin

(*Dalian Minzu University*, *Dalian* 116600)

Abstract: Recently, people have been paying more and more attention to mental health, such as depression, autism, and other common mental diseases. In order to achieve a mental disease diagnosis, intelligent methods have been actively studied. However, the existing models suffer the accuracy degradation caused by the clarity and occlusion of human faces in practical applications. This paper, thus, proposes a multi-scale feature fusion network that obtains feature information at three scales by locating the sentiment region in the image, and integrates global feature information and local feature information. In addition, a focal cross-entropy loss function is designed to improve the network's focus on difficult samples during training, enhance the training effect, and increase the model recognition accuracy. Experimental results on the challenging RAF_DB dataset show that the proposed model exhibits better facial expression recognition accuracy than existing techniques.

Keywords: Mental Health, Facial Emotion Recognition, Deep Learning, Multiscale, Loss Function

1 Introduction

In today's increasingly competitive society, many people are under tremendous pressure, resulting in negative emotions prone to mental health problems without timely treatment. Especially in some rapidly developing cities, the incidence of mental illness is gradually increasing[1].

In the past decade, the rapid development of deep learning has led researchers to actively explore many models for facial expression recognition (FER)[2]. Although compared with traditional methods, deep learning methods have significantly improved recognition accuracy[3], there still exist many challenging problems, such as diverse light and angle deviations[4]. Accordingly, the existing FER models are not mature to be applied in practical applications, especially in terms of generality.

The lack of generality is mainly induced by a large amount of interference information in simple feature extraction from the images. Due to such undesirable effects of the interference information, the network does not focus on the key features. Psychological studies showed that facial expressions are mainly reflected in the regions of the eyes, nose, and mouth[5-6]. Irrelevant regions may provide not only useless feature information but also weaken the usefulness of effective features. Previous studies adopted local expression-based methods for FER systems[7], where different features were extracted from divided face regions and fused to obtain the final feature expression. However, those methods did not consider the association between each region[8]. In [9], an action unit detection-based method was proposed, which is then transformed into corresponding expressions according to the Facial Coding System (FACS)[10]. However, such methods used hand-crafted features for expression classification, which could not

bring deep semantic information to the recognition model.

On the other hand, previous studies used a small size of datasets (e.g., CK+ [11]), which were also collected in ideal laboratory environments. The lack of a large-size dataset with diverse images often leads to the low generality of the trained models, especially in complex practical applications. In order to alleviate this problem, data augmentation[12] was commonly used, increasing the amount of training data and adding various noises to enhance the robustness of the models. Recently, wild datasets, such as RAF_DB[13] and AffectNet[14], were constructed, where the images were collected from the web. The faces in these datasets have diverse poses, including occlusion, unprominent foreground, different angles, etc., which are more suitable for the actual complex environment.

This paper presents a deep multi-scale feature fusion network for emotion detection. The multi-scale feature fusion network locates the emotion region in the image through the emotion region detection module and obtains feature information at three scales, taking into account global feature information and local feature information. In addition, a focal cross-entropy loss function is designed to enhance the training effect and improve the recognition of the model, which is used to focus on difficult samples during training. The rest of the paper is organized as follows. Section 2 presents the related works. Section 3 describes the proposed pyramidal expansion convolutional network in detail. Section 4 presents the experimental results and analysis. Lastly, Section V concludes this paper with future works.

2 Related Works

Many studies were conducted on facial expression recognition based on facial features for mental state diagnosis [15]. Those methods can be categorized into three: face feature points-based models, facial geometric feature-based models, and facial action unit (AU)-based models.

Typical face feature points-based recognition models include the followings. Wang et al[16]. modeled the face by locating 58 feature points on the face and dividing the face into 28 regions, which were then fed as features into an SVM classifier for expression classification to observe the difference between the emotional states of patients with mental illness and normal people. Cohn et al[17]. modeled faces by the Active Appearance Model (AAM) and extracted coordinates of 68 feature points of faces as primary features. From the extracted features, distance, angle, and area features between feature points were obtained and fed into the classifier for expression recognition. These models can provide more interpretable features but require higher computational complexity. Also, they were mostly based on hand-crafted features and traditional machine learning methods. Only a few models were based on deep learning.

The Local Binary Pattern (LBP)-based models are the most widely used facial geometric features-based models. The advantages of LBP, including grayscale and rotation invariance, led to its great success in many applications. Saha et al[18]. combined digital wavelet changes and local binary patterns to recognize facial expressions. Jabid et al[19]. proposed Local Directional Pattern (LDP) features to represent facial geometry, where the edges in eight directions were encoded into an 8-bit binary pattern. The LDP features showed higher accuracy than the LBP on the CK+[11] dataset and Jaffe[20] dataset. Ahmed et al[21]. proposed the Compound Local Binary Pattern (CLBP), which considers the magnitude information of the difference in addition to the difference-pattern of LBP, effectively improving the robustness of expression recognition. However, such algorithms could not properly handle illumination variations.

The last category is based on facial AU for recognition. Ekman proposed to use facial muscle movement units to describe facial expressions under the fact that human expressions are innate[10] and basic human expressions have commonality across cultures. Based on the facial muscle movement units, the FACS was proposed to classify human faces into various AU, by which the types of expressions are discriminated. Giota Stratou et al[22]. conducted in-depth research on

some mental diseases (such as depression) through the characteristics of the AU combination, proving the significance and value of AU characteristics. The study in [23] also found that in patients with depression, some AUs with affinity (such as AU12 and AU15) appear less frequently, while some AUs with no affinity (such as AU14) appear relatively high. In addition, in many studies [24], the relevant characteristics of AU (such as the number of occurrences, duration, intensity, etc.) were used as the evaluation criteria for mental illness.

However, most mental disease recognition studies were conducted in a laboratory environment with a lack of consideration of the real environment, such as lighting, occlusion, head pose, and face ratio in the image. This resulted in a large gap between the trained model and the actual application. Thus, this paper designs a cascade pyramid network based on deep learning, which fully uses the extracted feature information to recognize the mental state by identifying various facial expressions.

3 Proposed Method

In this section, we first introduce the multi-scale feature learning network, and then describe the sentiment region detection module, which obtains the location of the face and the focal emotional organ from the original image to obtain the sentiment feature information. Finally, a focal cross-entropy loss function focusing on difficult samples is proposed, which focuses on difficult samples during training, allowing the model to learn more features of difficult samples and improve the recognition performance of the model.

3.1 Proposed Model

As the existing datasets are all labelled at the image level, there are no annotations of labels for sentiment regions. And to train the sentiment region detection module, region annotations are needed. To obtain pseudo-sentiment region labels, we used the Retina Face network to detect faces and focus on expressive organs, and used them as pseudo-sentiment regions to train our face and organ detection modules.

Our network model is a two-stage structure, where the first stage identifies local sentiment regions, and the second stage learns the same multiscale representation for sentiment classification for the identified sentiment regions. As shown in Fig.1, the backbone network is followed by two modules: the sentiment region detection module and the multi-scale feature learning module. The backbone network is a modified ResNet network, which is used to extract the feature information from the original image, the emotion region detection module is used to locate the emotion regions and key expression organs of the face that need attention from the whole image, and the multiscale feature learning module is used to fuse the feature maps at different scales and feed the fused feature information to the classifier for classification to obtain the final recognition results.

Since most of the existing publicly available datasets are only annotated with image-level sentiment, and do not have pixel-level face position labels. Therefore, in this paper, to train our sentiment region detection module, the RAF_DB dataset is run on the Retina Face network, and the obtained face positions and eye, nose and mouth corner positions are used as pseudo-targets for training our sentiment region detection module after manually removing the parts with large differences.

Considering the amount of computation and the recognition effect, we used ResNet as the base network to build the FPN network. Since P2 generated from C2 of ResNet takes more computational resources, we generate P3 directly from C3, and P6 is generated by 3×3 convolution with step size 2, and P7 is generated from P6 by 3×3 convolution with step size 2 through RELU activation. The specific network structure is shown in Fig.2. In each layer of the feature map, we generated a total of nine anchors at each pixel point with three different aspect ratios {1:1, 2:1, 1:2} and three different scales {20, 21/3, 22/3} and predicted these anchors.

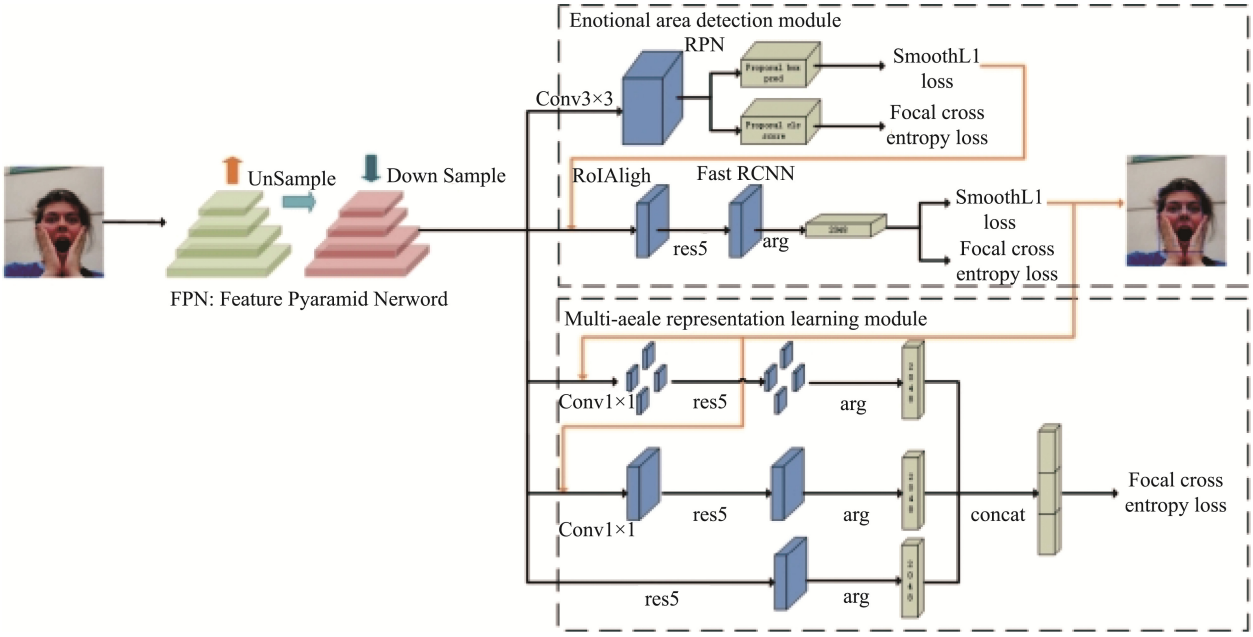

Fig.1 The Overall Structure of the Proposed Network

Fig.2 Samples of the RAF-DB Dataset

Our Emotional Region Detection module is based on Fast RCNN and is used to distinguish facial parts, facial focal organ regions and non-emotional regions from the original image. It first extracts features from candidate frame regions using the region of interest pair (ROIAlign) layer, and then converts the feature maps within the candidate frames into smaller feature maps of the same size. The reduced feature map is then sent to the residual block, followed by a global average pooling layer. Finally, the resulting feature maps are sent to two fully connected layers, classification and bounding box regression. In our network, the module is only used to detect emotional regions of faces and facial expression organs, and does not perform emotion classification.

3.2 Focal Cross-entropy Loss Function

In order to improve the efficiency of the network model in the training process and to focus the training on the samples that are not easy to distinguish, this paper designs a loss function to better train our network model. Since there is a target detection part and a classification part of our network, the loss function of the existing target detection and classification model is large as

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}}\sum_i L_{cls}(p_i, p_i^*) + \lambda\frac{1}{N_{reg}}\sum_i p_i^* L_{reg}(t_i, t_i^*)$$

Where the first half of the right hand side of the equation is the classification loss and the second half is the regression loss. Note that during the training process, it is the difficult samples in the training that determine the performance of the model, rather than the easy samples. Therefore, to make the value of the loss function more determined by the difficult samples, in this paper the classification loss part is changed by

$$L_{cls}(p_i, p_i^*) = -\log\left[p_i p_i^* + (1-p_i^*)(1-p_i)\right]$$

where p_i^* is the anchor label, positive samples are 1 and negative samples are 0, and p_i is the probability that the anchor is predicted to be the target. Replace with

$$L_{cls}(p_i, p_i^*) = -\left[(1-p_i p_i^* + (1-p_i^*)(1-p_i))\right]^{\gamma}$$
$$\log\left[p_i p_i^* + (1-p_i^*)(1-p_i)\right]$$

By adding parameters before the log function, where $p_i p_i^* + (1-p_i^*)(1-p_i)$ tends to 1, whether the sample is a positive sample or a negative sample, it is a simple sample, at which time the modulation parameter tends to 0, i.e. When the sample is not easy to distinguish, the probability of the positive sample being the target sample is relatively small, and the probability of the negative sample being the target sample is relatively large, the modulation parameter will be larger, at which time the contribution value of the loss will be larger; at the same time, the parameter γ can mediate the decay rate of the weights, thus achieving the purpose of focusing on difficult samples in the training process of the model.

3.3 Multi-scale Feature Learning Module

In this module, in order to focus on features in the emotional regions of the image, we extracted features at three scales for the final emotional classification. As shown in Fig.1, at the first scale, we extracted features using the facial organ region boxes obtained from the RoI Align layer classification, which were mapped onto the feature maps generated by the backbone network and based on

$$k = k_0 + \log(\sqrt{wh}/224)$$

Where k_0 is 4 and w and h are 1/224, the calculation gives k as 3. The feature map generated by the P3 layer of the mapping to the backbone network thus obtains feature information for each emotional organ. The resulting feature maps were transformed to a fixed size (14×14) by a 1×1 convolutional layer and fed to a residual block, followed by a global average pooling layer. At the second scale, the facial regions obtained from the RoI Align layer classification are mapped onto the P3 feature map to extract features and obtain feature information of the whole face of the person in the image. Similar to the feature map at the first scale, this part of the feature map is also converted to a fixed size (56×56) by a 1×1 convolution layer and sent to the residual block and the global average pooling layer. At the third scale, the feature maps generated by the P_3 layer of the backbone network are sent directly to the residual block and global average pooling layers. Finally, the feature information from these three scales is stitched together and fed into the classifier for sentiment classification.

4 Experimental Results

4.1 Dataset

In order to evaluate the proposed model, the RAF_DB [13] dataset is adopted. The RAF_DB dataset consists of 29,672 facial images annotated with basic or compound expressions by multiple professional annotators. The images have considerable diversity in subjects' age, gender, race, head pose, lighting conditions, masking (e.g., glasses, beard, or self-masking), and post-processing operations (e.g., various filters and special effects). The dataset includes two sub-sets: a single-labeled subset (basic emotions) and a two-labeled subset (compound emotions). The single-labeled subset was used in the experiments, which includes 14,186 images of six expressions as well as neutral emotions (Fig.2). The subset is divided into 11,822 training images and 2,364 testing images. Note that both the training and test sets are unbalanced.

4.2 Experiment Details

The image data was divided into a training set and a test set in a ratio of 5:1. We balanced the distribution of the expressions between the training and testing sets. The model was then trained with the training set by using the ADAM optimizer. The batch size was set to 32, the epoch was 50, and the drop value was 0.5. The performance of the proposed model was evaluated in comparison to the widely used classical networks. The models were implemented in Python, TensorFlow framework, and Keras library. The experiments were conducted on the operating system of Ubuntu with an NVIDIA GeForce RTX-3080Ti and 10GB RAM.

4.3 Emotional Region Detection Results

We used the Retina Face network for facial region recognition on the RAF_DB dataset and trained the Faster RCNN network as a pseudo-target to finally obtain the regions of the face and facial organs. As shown in Fig.3, after processing by the Retina Face network, the regions of the face, eyes, nose and mouth can be obtained from the original image, saving their respective categories and boundary coordinates, which are used to train our emotion region detection module.

Fig.3 Sample Diagram of Dataset Processing

4.4 Comparative Analysis with Other Models

Table 1 summarizes a comparison of the recognition accuracy of the proposed model with other models on the RAF_DB dataset. Since the distribution of the RAF_DB dataset is unbalanced, following the previous works, we use the mean accuracy to evaluate the performance, which is the mean on the diagonal of the confusion matrix. As shown in Table 1, the proposed model achieves an average accuracy of 87.63% and an overall recognition rate of 88.23% on the RAF_DB dataset, which outperforms the compared models. Fig.4 presents the confusion matrix obtained by our model on the RAF_DB dataset, showing the most considerable confusion between anger and surprise. The recognition error rate between fear and surprise is 8.16%, which is due to the more similar features between these two types of facial organs.

Table 1 The Overall and Average Accuracies of the Compared Models on the RAF-DB Test Set

Method	Overall Acc	Average Accuracy
DLP_CNN [25]	84.13%	74.20%
RAN [26]	86.90%	_
DACL [27]	83.11%	80.44%
ViT [28]	63.75%	_
CVT [29]	82.27%	_
Ours	88.23%	87.63%

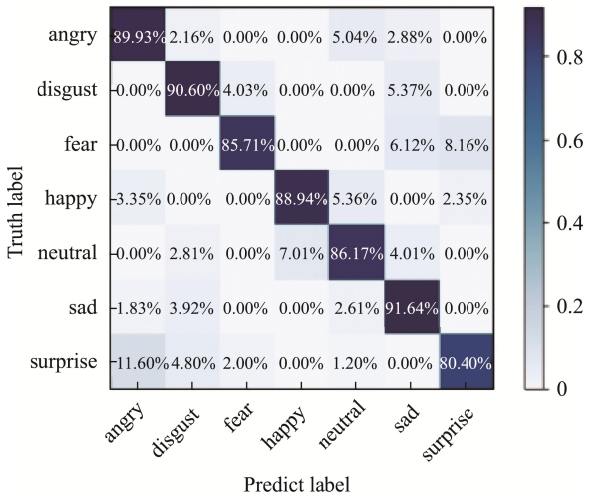

Fig.4 Confusion Matrix of this Model on RAF_DB

5 Conclusion

This paper presents a novel deep learning architecture for mental state diagnosis - a multi-scale focus detection network. The network incorporates features from different scales, fully considers the feature information in the original image, and focuses on the key emotional features of the face to solve the problem of insufficient focus on facial expressions. To increase the focus on difficult patterns and improve the training effect of the network, a focus loss function is designed to reduce the weight of easy patterns in the training process. The experimental results on the RAF_DB dataset demonstrate the effectiveness of the proposed model. However, the proposed model still cannot handle some extreme cases. For example, using only half or less of the faces can lead to a decrease in accuracy if the face deviation angle is too large. There may be situations where the same expression corresponds to different mental states in different poses or environments. Our future work will focus on addressing these issues.

References

[1] Szasz T S. The myth of mental illness[J]. American psychologist, 1960, 15(2): 113.

[2] Huang Y, Chen F, Lv S, et al. Facial expression recognition: A survey[J]. Symmetry, 2019, 11(10): 1189.

[3] Li S, Deng W. Deep facial expression recognition: A survey[J]. IEEE transactions on affective computing, 2020.

[4] Valstar M F, Mehu M, Jiang B, et al. Meta-analysis of the first facial expression recognition challenge[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2012, 42(4): 966-979.

[5] Tian Y I, Kanade T, Cohn J F. Recognizing action units for facial expression analysis[J]. IEEE Transactions on pattern analysis and machine intelligence, 2001, 23(2): 97-115.

[6] Tong Y, Liao W, Ji Q. Facial action unit recognition by exploiting their dynamic and semantic relationships[J]. IEEE transactions on pattern analysis and machine intelligence, 2007, 29(10): 1683-1699.

[7] Ghimire D, Jeong S, Lee J, et al. Facial expression recognition based on local region specific features and support vector machines[J]. Multimedia Tools and Applications, 2017, 76(6): 7803-7821.

[8] Yadav S P. Emotion recognition model based on facial expressions[J]. Multimedia Tools and Applications, 2021, 80(17): 26357-26379.

[9] Hassen O A, Abu N A, Abidin Z Z, et al. A new descriptor for smile classification based on cascade classifier in unconstrained scenarios[J]. Symmetry, 2021, 13(5): 805.

[10] Ekman P, Friesen W V. Facial action coding system[J]. Environmental Psychology & Nonverbal Behavior, 1978.

[11] Lucey P, Cohn J F, Kanade T, et al. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression [C]// 2010 IEEE computer society conference on computer vision and pattern recognition-workshops. IEEE, 2010: 94-101.

[12] Huang T R, Hsu S M, Fu L C. Data Augmentation via Face Morphing for Recognizing Intensities of Facial Emotions[J]. IEEE Transactions on Affective Computing, 2021.

[13] Li S, Deng W, Du J P. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2852-2861.

[14] Mollahosseini A, Hasani B, Mahoor M H. Affectnet: A database for facial expression, valence, and arousal computing in the wild[J]. IEEE Transactions on Affective Computing, 2017, 10(1): 18-31.

[15] Canedo D, Neves A J R. Facial expression recognition using computer vision: A systematic review[J]. Applied Sciences, 2019, 9(21): 4678.

[16] Wang P, Barrett F, Martin E, et al. Automated video-based facial expression analysis of neuropsychiatric disorders[J]. Journal of neuroscience methods, 2008, 168(1): 224-238.

[17] Cohn J F, Kruez T S, Matthews I, et al. Detecting depression from facial actions and vocal prosody[C] //2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops. IEEE, 2009: 1-7.

[18] Saha A, Wu Q M J. Facial expression recognition using curvelet based local binary patterns[C]//2010 IEEE International Conference on Acoustics, Speech and Signal

Processing. IEEE, 2010: 2470-2473.

[19] Jabid T, Kabir M H, Chae O. Robust facial expression recognition based on local directional pattern[J]. ETRI journal, 2010, 32(5): 784-794.

[20] Lyons M J, Akamatsu S, Kamachi M, et al. The Japanese female facial expression (JAFFE) database[C]// Proceedings of third international conference on automatic face and gesture recognition. 1998: 14-16.

[21] Ahmed F, Bari H, Hossain E. Person-independent facial expression recognition based on compound local binary pattern (CLBP)[J]. Int. Arab J. Inf. Technol., 2014, 11(2): 195-203.

[22] Stratou G, Scherer S, Gratch J, et al. Automatic nonverbal behavior indicators of depression and ptsd: the effect of gender[J]. Journal on Multimodal User Interfaces, 2015, 9(1): 17-29.

[23] Girard J M, Cohn J F, Mahoor M H, et al. Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses[J]. Image and vision computing, 2014, 32(10): 641-647.

[24] Lien J J, Kanade T, Cohn J F, et al. Automated facial expression recognition based on FACS action units[C]//Proceedings third IEEE international conference on automatic face and gesture recognition. IEEE, 1998: 390-395.

[25] Li S, Deng W, Du J P. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2852-2861.

[26] Wang K, Peng X, Yang J, et al. Region attention networks for pose and occlusion robust facial expression recognition[J]. IEEE Transactions on Image Processing, 2020, 29: 4057-4069.

[27] Farzaneh A H, Qi X. Facial expression recognition in the wild via deep attentive center loss[C]//Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2021: 2402-2411.

[28] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[29] Ma F, Sun B, Li S. Robust facial expression recognition with convolutional visual transformers[J]. arXiv preprint arXiv:2103.16854, 2021.Zheng, R. (2016). *Research on Solar Radiation Observation Instrument and the Calibration Technique*. D Sc Tech. Changchun University of Science and Technology.

Author Biographies

ZHANG Ning received B.Sc. degree from Dalian Minzu University in 2020. He is currently a M.Sc. candidate in Dalian Minzu University. His main research interest includes facial emotion recognition.

E-mail: zhangning946@163.com

ZHANG Xiufeng received his B.Sc. degree from Yanshan University in 2000, M.Sc. and Ph.D. degrees from Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences in 2005. He is currently an associate professor at the School of Mechanical and Electrical Engineering, Dalian University for Nationalities. His main research interests include artificial intelligence, intelligent medical image processing, intelligent detection technology and instruments.

E-mail: zhxf7710@dlnu.edu.cn