Contents lists available at ScienceDirect

Digital Chinese Medicine

journal homepage: http://www.keaipublishing.com/dcmed

Establishing and validating a spotted tongue recognition and extraction model based on multiscale convolutional neural network

PENG Chengdong[a, b], WANG Li[c], JIANG Dongmei[d], YANG Nuo[b], CHEN Renming[b], DONG Changwu[c*]

a. School of Computer Science and Information Engneering, Hefei University of Technology, Hefei, Anhui 230009, China

b. Artificial Intelligence Laboratory, Hefei Yunzhen Information Technology Co., Ltd., Hefei, Anhui 230088, China

c. College of Chinese Medicine, Anhui University of Chinese Medicine, Hefei, Anhui 230012, China

d. Electronic Information Engineering College, Anhui Technical College of Water Resources and Hydroelectric Power, Hefei, Anhui 231603, China

ARTICLE INFO

Article history
Received 03 December 2021
Accepted 13 February 2022
Available online 25 March 2022

Keywords
Spotted tongue recognition and extraction
The feature of tongue
Instance segmentation
Multiscale convolutional neural network (CNN)
Tongue diagnosis system
Artificial intelligence (AI)

ABSTRACT

Objective In tongue diagnosis, the location, color, and distribution of spots can be used to speculate on the viscera and severity of the heat evil. This work focuses on the image analysis method of artificial intelligence (AI) to study the spotted tongue recognition of traditional Chinese medicine (TCM).

Methods A model of spotted tongue recognition and extraction is designed, which is based on the principle of image deep learning and instance segmentation. This model includes multiscale feature map generation, region proposal searching, and target region recognition. Firstly, deep convolution network is used to build multiscale low- and high-abstraction feature maps after which, target candidate box generation algorithm and selection strategy are used to select high-quality target candidate regions. Finally, classification network is used for classifying target regions and calculating target region pixels. As a result, the region segmentation of spotted tongue is obtained. Under non-standard illumination conditions, various tongue images were taken by mobile phones, and experiments were conducted.

Results The spotted tongue recognition achieved an area under curve (AUC) of 92.40%, an accuracy of 84.30% with a sensitivity of 88.20%, a specificity of 94.19%, a recall of 88.20%, a regional pixel accuracy index pixel accuracy (PA) of 73.00%, a mean pixel accuracy (mPA) of 73.00%, an intersection over union (IoU) of 60.00%, and a mean intersection over union (mIoU) of 56.00%.

Conclusion The results of the study verify that the model is suitable for the application of the TCM tongue diagnosis system. Spotted tongue recognition via multiscale convolutional neural network (CNN) would help to improve spot classification and the accurate extraction of pixels of spot area as well as provide a practical method for intelligent tongue diagnosis of TCM.

1 Introduction

In the objective research of tongue diagnosis, the study of the methods used to identify and judge the different features of tongue image is critical [1]. Spotted tongue refers to the pathological feature of swelling of fungiform papillae [2]. Tongue spots includes dots (red, white, and black), thorns (red spots like awn, which can be touched)

*Corresponding author: DONG Changwu, Professor, E-mail: dcw1018@aliyun.com.

Peer review under the responsibility of Hunan University of Chinese Medicine.

DOI: 10.1016/j.dcmed.2022.03.005

Citation: PENG CD, WANG L, JIANG DM, et al. Establishing and validating a spotted tongue recognition and extraction model based on multiscale convolutional neural network. Digital Chinese Medicine, 2022, 5(1): 49–58.

and petechia (bluish-purplish or bluish-darkish spots). According to the location of spots, the viscera of the evil heat can be inferred. The color and density of spots indicates the degree of heat evil, which has pathological significance for tradition Chinese medicine (TCM) tongue diagnosis.

The recognition and judgment of spotted tongue has always been valued by researchers. Some scholars, such as XU et al. [3], LI et al. [4], WANG et al. [5], have studied the computer recognition methods of spots on the tongue and achieved recognition results by capturing images with high-definition cameras under auxiliary light. This kind of method [3-5] is based on the principle of image spot detection, using color threshold, fuzzy clustering, and support vector machine (SVM) classification, however, due to the disadvantages of high sample dependency, challenges in establishing parameters, and poor algorithm robustness, so it is difficult to use in practice. With the development of artificial intelligence (AI), new technology has been provided for tongue image recognition and extraction [6-8]. In recent years, deep learning and convolutional neural network (CNN) have been applied to the TCM feature analysis method of tongue images [9-16] and achieved better recognition accuracy than image algorithms and machine learning.

However, the irregular distribution of spots on the tongue, the large difference between the size of spots and the tongue, and the small difference between the color of the spots and the tongue under natural light conditions, etc., are the difficulties and challenge of automatic recognition for spotted tongue [17]. In this study, a tongue image prickle recognition and extraction model based on image deep learning and case segmentation principle is established, which is suitable for the application of TCM tongue diagnosis system. This model can help improve spot classification and the accurate extraction of pixels of spot area as well as provide a practical method for the intelligent tongue diagnosis of TCM.

2 Designing the spotted tongue recognition model

The proposed model mainly includes multiscale feature map generation, candidate region search, and target region recognition. Firstly, deep convolution network is used to build multiscale low and high abstraction feature maps after which, target candidate box generation algorithm and selection strategy are used to select high-quality target candidate regions. Finally, classification network is used for classifying target regions and calculating target region pixels. As a result, the region segmentation of spotted tongue is obtained. The realization procedure is shown in Figure 1.

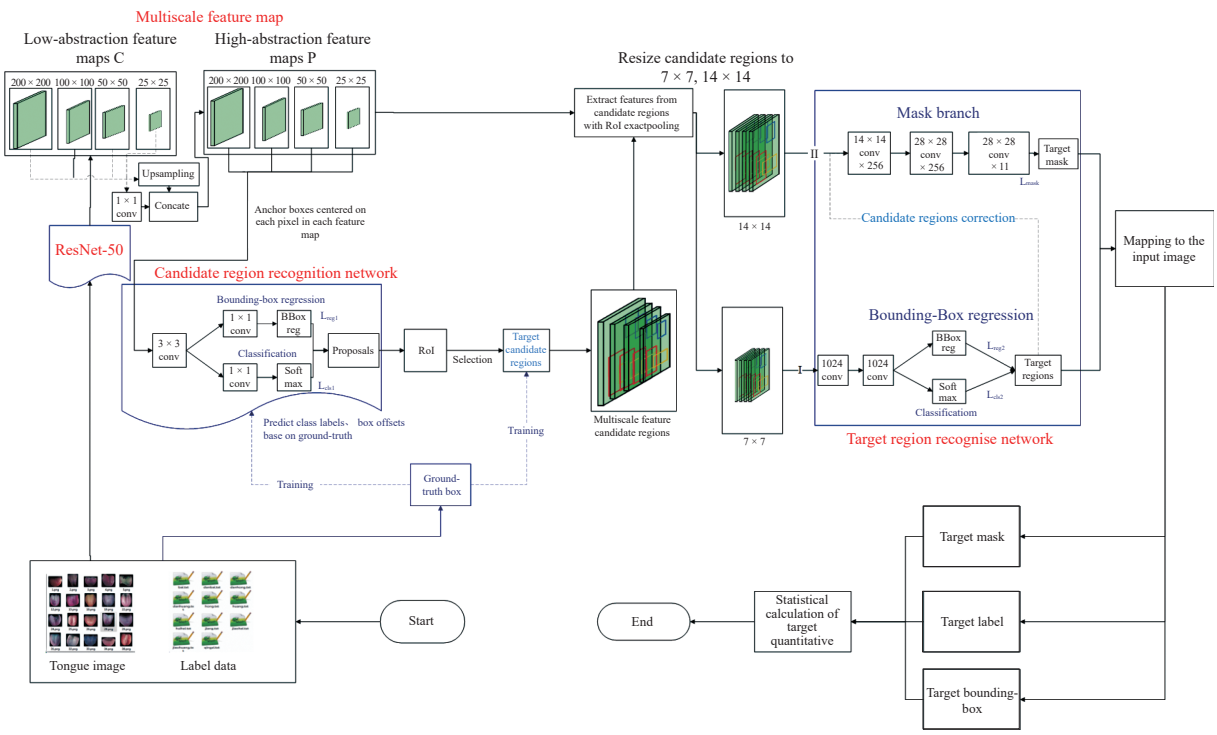

Figure 1 Network structure and computation of the spotted tongue recognition model

Conv represents the convolution; L_{cls1} represents loss of classify for the candidate region; L_{reg1} represents loss of bounding-box regression for the candidate region; L_{cls2} represents loss of classify for target candidate region; L_{reg2} represents loss of bounding-box regression for target candidate region; L_{mask} represents loss of pixel mask for target candidate region.

2.1 Multiscale feature map

2.1.1 Building low-abstraction feature map C This study uses the classic ResNet-50 [18] (PyTorch pre-trained ResNet-50 v1.5) as the backbone network. The output of residual module, including conv2, conv3, conv4, conv5, is regarded as low-abstraction feature map set and C {C2, C3, C4, C5} is used to indicate the map set. Their characteristic scales are 200 × 200, 100 × 100, 50 × 50, 25 × 25, which have a step size of {4, 8, 16, 32} pixels relative to the input image.

2.1.2 Building high-abstraction feature map P A feature map with higher resolution is obtained by twice upsampling on the high-level feature map of low-abstraction map set C, and then the feature map was laterally connected with the corresponding low-abstraction feature map by element-wise addition operations. This iteration continue until the shallowest convolution map is concatenated. Finally, the 3 × 3 convolution of all concatenated feature map eliminates the aliasing effect of upsampling. As a result, the high-abstraction feature map P {P2, P3, P4, P5} is obtained.

This multiscale high-abstraction feature map is more robust in the face of differently sized detection targets, especially in terms of not missing small targets.

2.2 Candidate region searching

2.2.1 Generating anchor boxes The window on the high-abstraction feature map P is first slid, after which the center point, which is mapped by each pixel base on feature map, of the receptive field on original image is taken as the reference point. Twelve anchor boxes of different aspect ratios and areas around the reference point are then selected. The initial anchor box contains four areas (4, 8, 16, 32) with each area containing three aspect ratios (1∶1, 1∶2, and 2∶1).

2.2.2 Anchor box classified prediction A classified training label for is generated each anchor box, with each label containing zero negative sample (target isn't covered) and one positive sample (target is covered). The strategy of label assignment is as follows: labels are labeled by an intersection over union (IoU) overlap between prediction box and ground-truth box, where the ground-truth box exists in the training set of manual calibration.

2.2.3 Selecting target proposal regions The score and bounding box regression of the candidate box with the prediction tag of one are converted to the proposal target, and the high-quality proposal regions are selected. To select proposal regions, the edges of proposal regions that exceed the image boundary are trimmed to ensure that the target proposal regions are within the range of the image. Next, all proposal regions are sorted from high to low by positive region scores to get the first 12 000 proposal regions. Finally, overlapping proposal regions are eliminated using non-maximum suppression to get the first 3 000 as target proposal regions.

2.2.4 Extracting features of target proposal regions According to its scale (width and height), the coordinate area is extracted from the feature map corresponding to the high-abstraction feature map P for each proposal region. The target proposal regions are pooled into proposal region feature maps of 7 × 7 and 14 × 14 using the local area average pooling algorithm region of interest (RoI) ExactPooling, which is the improved algorithm compared with RoI Pooling [19] and RoI Align [20]. Table 1 presents the RoI ExactPooling algorithm features.

2.3 Recognising the target region

2.3.1 Classifying the target region A classified training label is generated for each RoI with each label containing zero negative samples (target isn't covered) and k positive samples (target classes $k = 3$, indicating dot, thorn, and petechia). For tag assignment, there are ground-truth boxes of manually labeled spots on each image in tongue image training set. The IoU of the overlap ratio of prediction box and ground-truth boxes assists this process.

2.3.2 Correcting the border of target region The border position of the target region is regressed with high classified score after which, the horizontal translation offset and the height-width scaling offset of the corrected box relative to the prediction box is estimated. Predictive offset of candidate region $t_i = \{t_x, t_y, t_w, t_h\}$ (t_x, t_y represents the translation offset in x direction and y direction, and t_w, t_h represents the scaling offset of the width and height).

2.3.3 Calculating the target region mask A link is made to fully convolutional networks to generate a region binary mask for each target class to obtain pixel masks of each target, which are the $m \times m$ binary masks of k classes ($m \times m$ is the resolution of the target area, m = 14). Finally, the result is mapped to the original image to obtain the inspected target and region segmentation pixel.

2.4 Loss function for each module

2.4.1 Loss of proposal region generation network The loss of classify L_{cls1} and loss of bounding-box regression L_{reg1} for candidate region constitute loss of the proposed region generation network. The definition of this network loss function is shown in Table 2 .

Total loss L_{pro} of the proposed region generation network can be obtained as follows.

$$L_{pro}(p_i, t_i) = \frac{1}{256}\sum L_{cls1}(p_i, p_i^*) + \frac{1}{128}\sum p_i^* L_{reg1}(t_i, t_i^*) \quad (1)$$

Table 1 The improved local area mean-pooling algorithm

Algorithm	Feature of mapping in candidate region	Feature
RoI Pooling	Round the feature point to integer number before mean-pooling $\frac{\sum_{i=\lfloor x_1 \rfloor}^{\lceil x_2 \rceil}\sum_{j=\lfloor y_1 \rfloor}^{\lceil y_2 \rceil} w_{i,j}}{(\lceil x_2 \rceil - \lfloor x_1 \rfloor + 1)(\lceil y_2 \rceil - \lfloor y_1 \rfloor + 1)}$ $w_{i,j}$ is the pixel value of i, j position in the region from (x_1, y_1) to (x_2, y_2); $\lceil\ \rceil$ represents returning the ceiling of variable as an integral; $\lfloor\ \rfloor$ represents returning the floor of variable as an integral	After two rounds, the spatial deviation between the proposed region of the pooled feature map and of the original image is obvious; as a result, that we can't achieve accurate pixel prediction cannot be achieved
RoI Align	Sample four points for each mapped region using bilinear interpolation; conduct the mean pooling of the sampling points $\sum_{i=1}^{4}\frac{\text{bilinear}\ (a_i, b_i)}{4}$ where (a_i, b_i) is the coordinate of the sampling point	Retain the decimal; by sampling fitting, the feature map is aligned with the proposal region of the original image, enabling pixel-level prediction to be performed
RoI ExactPooling	Conduct the mean pooling of all feature points in the interval where each mapped region is $\frac{\iint_{\substack{x_1 \leqslant x \leqslant x_2 \\ y_1 \leqslant y \leqslant y_2}} \text{bilinear}(x, y)\mathrm{d}x\mathrm{d}y}{(x_2 - x_1)(y_2 - y_1)}$ $\iint$: double integral over rectangular region from the minimum point (x_1, y_1) to the maximum point (x_2, y_2); dx, dy represents the differentials of x and y	The pixel deviation between the feature map and the proposal region of the original image is reduced and the precision is high without increasing the calculation amount

Table 2 The definition of loss function for the proposed region generation network

Module	Training data	Loss function
Classifying the candidate region	Training set randomly takes 128 positive samples (IoU > 0.7) and 128 negative samples (IoU < 0.3); insufficient positive samples are supplemented with negative samples	$\mathrm{L}_{\mathrm{cls1}}(p_i, p_i^*) = -\log[p_i p_i^* + (1 - p_i^*)(1 - p_i)]$ $\mathrm{L}_{\mathrm{cls1}}$: loss of classify for the candidate region; p_i: probability of prediction as positive; p_i^*: 1 indicates the positive sample, and 0 indicates the negative sample
Bounding-box regression for the candidate region	Select 128 random positive samples	$\mathrm{L}_{\mathrm{reg1}}(t, t^*) = \sum_{i \in \{x,y,w,h\}} \mathrm{smooth}_{\mathrm{L1}}(t_i - t_i^*)$ $\mathrm{L}_{\mathrm{reg1}}$: loss of bounding-box regression for the candidate region; $t_i = \{t_x, t_y, t_w, t_h\}$: predictive offset of candidate region; t_i^*: offset between candidate box and ground-truth box; a smooth L1 function $\mathrm{smooth}_{\mathrm{L1}}(x)$ is defined as $\mathrm{smooth}_{\mathrm{L1}}(x) = \begin{cases} 0.5x^2, & \text{if } \lvert x \rvert < 1 \\ \lvert x \rvert - 0.5, & \text{if } \lvert x \rvert \geqslant 1 \end{cases}$

2.4.2 Loss of target candidate region recognition network The loss of classify $\mathrm{L}_{\mathrm{cls2}}$, loss of bounding-box regression $\mathrm{L}_{\mathrm{reg2}}$, and loss of pixel mask $\mathrm{L}_{\mathrm{mask}}$ for target candidate region constitutes a loss of the target candidate region recognition network. The definition of this network loss function is shown in Table 3.

Total loss $\mathrm{L}_{\mathrm{tar}}$ of target candidate region recognition network can be obtained as follows.

$$\mathrm{L}_{\mathrm{tar}}(p_i, u, t_i, t_i^*) = \sum \mathrm{L}_{\mathrm{cls2}}(p_i, u) + [u \geqslant 1] \sum \mathrm{L}_{\mathrm{reg2}}(t_i, t_i^*) + \sum \mathrm{L}_{\mathrm{mask}}(m, k) \quad (2)$$

2.5 Quantifying the spotted tongue indicator

2.5.1 Dividing the image of the tongue area The tongue is first divided according to the method of the tongue and viscera of LI et al. [21] which calculates the external rectangle of the tongue. The left one fifth area is the tongue left margin, the right one fifth area is the tongue right margin, the top one fifth area is the tongue root, the bottom one fifth area is the tongue tip, and finally, the middle area is the tongue center. See Figure 2 for a visual representation of how the tongue is divided.

2.5.2 Quantifying the degree of spots indicator For quantitative indicators of the whole tongue and the

Table 3 The definition of loss function for target candidate region generation network

Module	Training data	Loss function
Classifying the target candidate region	Label target class in the training set	$L_{cls2}(p_i,u)=-\log\left[\frac{e^{p_i}}{\sum_{i=1}^{k}e^{p_i}}\right]$ L_{cls2}: loss of classify for target candidate region; p_i : score of prediction as target category; u: tag encoding of the labeled target; e ≈ 2.718 28; $k=3$, indicating dot, thorn, and petechia
Bounding-box regression for the target candidate region	Label the target border in the training set	$L_{reg2}(t,t^*)=\sum_{i\in\{x,y,w,h\}}\text{smooth}_{L1}(t_i-t_i^*)$ L_{reg2}: loss of bounding-box regression for target candidate region; $t_i=\{t_x,t_y,t_w,t_h\}$: predictive offset of candidate region; t_i^*: offset between candidate box and ground-truth box; a smooth L1 function $\text{smooth}_{L1}(x)$ is defined as $\text{smooth}_{L1}(x)=\begin{cases}0.5x^2, & \text{if } \lvert x\rvert<1\\ \lvert x\rvert-0.5, & \text{if } \lvert x\rvert\geqslant 1\end{cases}$
Pixel mask of the candidate region	Label the pixel of the target region in the training set	$L_{mask}(m,k)=-\log\sum_{i=1}^{m}[p_ip_i^*+(1-p_i^*)(1-p_i)]$ L_{mask}: loss of pixel mask for target candidate region; m: pixel amount of region; p_i: probability of pixel prediction; p_i^*: pixel tag

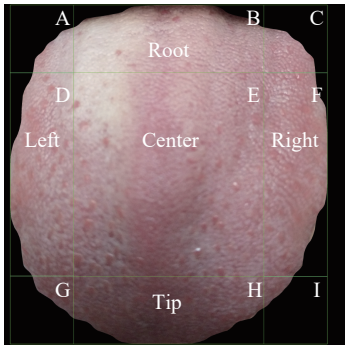

Figure 2 The method of dividing the tongue into five areas

A, D, and G regions are the tongue left margin; B region is the tongue root; C, F, and I regions are the tongue right margin; E region is the tongue center; and H region is the tongue tip.

various areas, (i) none: the number of spots is zero; (ii) a little: the number of spots is less than five; (iii) more: the number of spots is more than five and less than ten; and (iv) a lot: more than ten.

3 Training and testing the model

3.1 Spotted tongue image dataset

The spotted tongue image dataset contains 2 000 samples. These images are taken by nine kinds of phones and under different natural light or artificial light to record multiple tongue postures whilst also providing different backgrounds and resolutions. Data pre-processing is as follows.

3.1.1 Labeling spots in the sample Several of the TCM doctors, who were divided into in three batches, recognised spots on the tongue image. The recognition results were collected into the database through the same review method. A spotted tongue dataset of 2 000 samples was annotated and the statistical results of the labels were 33 378 dots, 11 649 thorns, and 1 219 petechiae.

3.1.2 Data augmentation in computer vision To make the training data set generic, data augmentation is implemented at random for samples with combinations of image rotation (random angle between − 90° and 90°), scaling (random multiple between 80% and 120%), flipping (up and down, horizontal mirror) and contrast (random multiple between 0.5 and 2). The dataset expands 50 times after data augmentation.

3.2 Training model

The ratio of training set∶validation set∶test set is 5∶3∶2. The training and validation sets are used to finish network model learning, optimize model parameters and identify the best network depth whereas the test set analyses the performance of model. Microsoft Common Objects in Context (MSCOCO) pre-training network parameters [22] are used for initialization and alternate training methods are used for the generation network of proposal regions and the calculation network of target regions. The training model is the first generation.

3.2.1 Initializing data According to the tongue dataset and the label of spots on tongue, the ground-truth box and object mask of target labeled on each image for calculating losses during the training phase is generated.

3.2.2 Training by fine tune method Fine tune training is used for the proposal region generation network and the target region recognition network to ensure that they share the convolutional layer and reduce the calculation

amount. The algorithm is trained and fine tune learning with the following four steps.

Step 1: the MSCOCO pre-training model is used for network initialization and training the proposal region generation network (for region recommendations).

Step 2: network parameters in step 1 are used for generating the proposal region and the MSCOCO pre-training model is used for network initialization and training the target region calculation network (for detection).

Step 3: the proposal region generation network is reinitialized using the target region calculation network after fine tuning in step 2. Only the layer unique to the proposal region generation network is fine tuned. This is done with the shared convolutional layer fixed after which the shared convolution layer will form.

Step 4: only the layer unique to the target region calculation network is fine tuned using subsequent region in step 3 with the shared convolutional layer fixed after which, the unified network will form.

3.2.3 Optimizing hyper-parameters of network First, the input image size is fixed and select different learning rates and RoI scaling parameters (Table 4) are selected. Then, experiments are conducted on the multiscale CNN and the loss changes of multiscale CNN trained by three groups of parameters are compared. The loss of No. 3 decreases faster than that of No. 1 and No. 2 and therefore, the parameters of No. 3 are selected, as shown in Figure 3. Choosing a set of the optimal hyper-parameters is as follows: (i) IoU is set to 0.5; (ii) the learning rate is 0.01 (learning rates decay by 10 times less every 60 epochs), the weight decay is 0.000 1, and the moment is 0.9; (iii) anchor ration is [0.5, 1.2] and anchor scales is [2, 4, 8, 16].

3.3 Testing model

The method of detecting spotted tongue target and parameter description in the inference stage of network model is described below.

3.3.1 Preprocessing the detected image The network model of tongue extraction is used to segment the tongue region on the image datasets. The obtained tongue image (black background) is then put into the spots recognition model for detection.

3.3.2 Inference process description (i) The detected tongue image is read first and then approximately 20 000 prediction boxes are generated through the proposal region generation network. (ii) The first border correction is performed on 20 000 prediction boxes to get a new frame after correction. (iii) The edges of box that exceed the image boundary are trimmed to make prediction box in the range of image. (iv) All prediction boxes are sorted from high to low by foreground score to get the first 1 000 boxes. (v) Overlapping prediction boxes are then eliminated using a network management system algorithm with a threshold of 0.7. (vi) For the remaining prediction box in the previous step, classification and second border correction is performed on the first 300 boxes with high scores. Pixel masks calculation is performed on the top

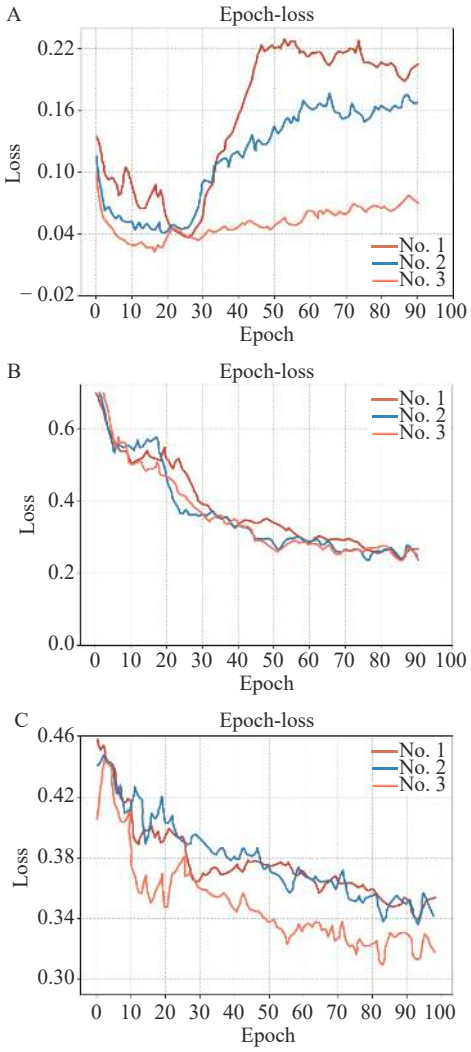

Figure 3 Analysis of loss change of multiscale CNN using different parameters

A, classifying target candidate region loss. B, bounding-box regression loss. C, pixel mask of candidate region loss.

Table 4 Experiment on multiscale CNN using different training parameters

No.	Image size (pixels)	Learning rate	Learning momentum	Weight decay	Anchor ration	Anchor scale
1	800 × 800	0.01	0.9	0.001	[0.5, 1.2]	[4, 8, 16, 32]
2	800 × 800	0.02	0.9	0.001	[0.5, 1.2]	[4, 8, 16, 32]
3	800 × 800	0.01	0.9	0.001	[0.5, 1.2]	[2, 4, 8, 16]

100 RoI. The inference efficiency is accelerated, and accuracy is improved.

3.3.3 Network parameter Take 1 000 candidate boxes of the candidate region generation network. Take 300 candidate boxes of the target region calculation network.

3.3.4 Time consuming analysis The experiments are conducted using PyTorch 1.7.0 and python 3.6.9. It spends 195 milliseconds per image using NVIDIA GeForce GTX 2080 Ti (11GB) GPU.

3.3.5 Model recognition results Spots recognition tests are performed on dots, thorns, and petechia. None of spots are tested using the first-generation training model. The color block is then used to indicate the spots' location and region pixel. Figure 4 shows the recognition results.

The test results show that the network model can correctly recognize dot, thorn, petechia and even dense irregularly shaped spots after learning. The model doesn't confuse the spots with the similar tongue papilla. Since too high a threshold will cause false detection and too low a threshold will cause missed detection, it is difficult for the traditional image spot detection algorithm to find balance between false detection rate and missed detection rate at the same time.

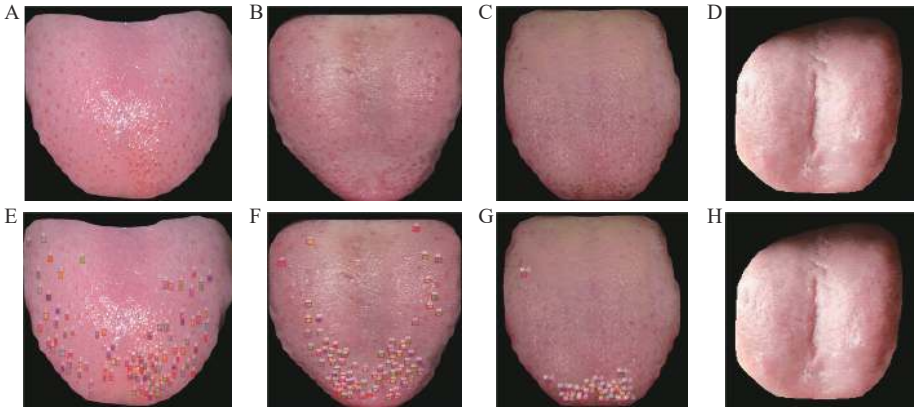

Figure 4 Recognition results of spotted tongue

A, original image with some dotted thorns. B, original image with dense dotted thorns. C, original image with some petechiae. D, original image with none of spots. E, recognition of some dotted thorns. F, recognition of dense dotted thorns. G, recognition of some petechiae. H, recognition of none of spots.

4 Results

The validation set was randomly selected from the tongue image datasets, including 500 spotted tongue images and 100 no-spotted tongue images. Additionally, four types of experimental datasets were constructed according to dot, thorn, petechiae, and no-spotted tongue. Three methods were used to detect the spots on the obtained tongue image (black background), namely image spot detection algorithm, Mask R-CNN ("R" is the abbreviation for "Region") [22] and multiscale CNN (first generation model).

The recognition performance of multiscale CNN model is better than image spot detection algorithm and Mask R-CNN. The spotted tongue recognition of multiscale CNN model achieved an area under curve (AUC) of 92.40%, an accuracy of 84.30% with a sensitivity of 88.20%, a specificity of 94.19%, and a recall of 88.20%. These results are presented in Figure 5. The statistical indicators of the spot area pixel accuracy (PA) of 73.00%, mean pixel accuracy (mPA) of 73.00%, IoU of 60.00% and mean intersection over union (mIoU) of 56.00% also present the highest performance of the three methods (Table 5).

The method in this study is sensitive to detecting three types of small-scale targets such as dots, thorns,

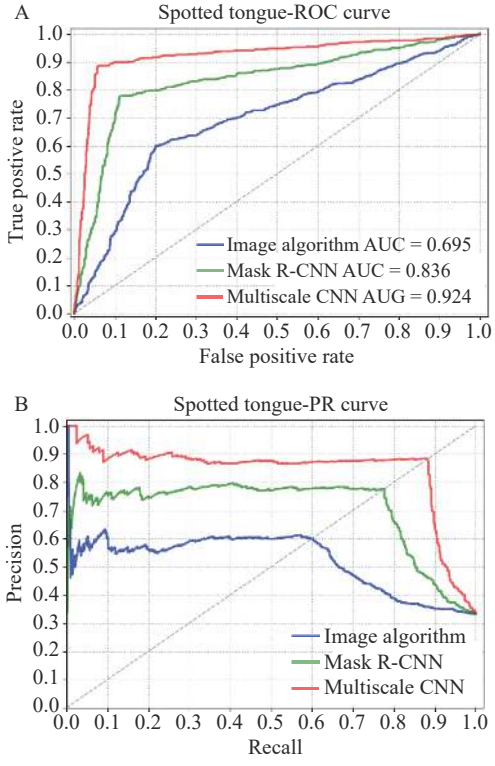

Figure 5 Quantitative analysis accuracy of the model

A, analysis of receiver operating characteristic (ROC) curve. B, analysis of precision recall (PR) curve.

Table 5 Evaluation indicators for pixel recognition of the model

Model	PA	mPA	IoU	mIoU
Spot detection algorithm	54.00%	49.00%	43.00%	39.00%
Mask R-CNN	64.00%	61.00%	51.00%	48.00%
Multiscale CNN	73.00%	72.00%	60.00%	59.00%

and petechiae. The comprehensive correction rate of spots recognition is 84.30%. The algorithm has strong anti-interference and a high degree of fit to the target boundary contour.

5 Discussion

In recent years, with the increasing demand for health, people pay more attention to TCM [23-28]. As a distinctive diagnosis and treatment method in TCM, tongue diagnosis is not only the main content of observation, but also one of the main tenets of clinical diagnosis of TCM. Traditional tongue diagnosis is affected by a variety of factors, such as the external environment, the doctor's knowledge, and the lack of quantitative indicators, and is subject to a certain degree of uncertainty and subjectivity [29, 30]. For a long time, one of the main goals in the objectification of tongue diagnosis is to solve the vagueness and uncertainty surrounding tongue diagnosis. The development of AI technology shows great advantages, improves the clinical application value of tongue diagnosis [31-35], and makes it tongue diagnosis scientific, specific, and objective [36]. Considering the shortcomings of most tongue diagnostic instruments, the research and development of the AI tongue diagnostic system as a new tongue diagnostic instrument can satisfy the needs of users to detect tongue images on mobile phones. The open application scenario is not easily disturbed by the external environment, and the collected information is richer. This is more in line with the collection characteristics of a tongue image, which is conducive to the improvement of the universality of tongue image feature recognition, so as to meet the objective research of tongue image under various conditions [37, 38].

This study established a model of spotted tongue recognition and extraction, multiscale CNN, which included multiscale feature map generation, region proposal searching, and target region recognition. Compared to the simple model, this model has several advantages such as multiscale RoI recognition, accurate boundary contour extraction, and a strong anti-interference ability. Firstly, this multiscale high-abstraction feature map is more robust in the face of different sizes of detection targets, especially in terms of not missing small targets. Secondly, the target candidate area is pooled into a smaller candidate area feature map through the local area average pooling algorithm RoI ExactPooling, which reduces the pixel deviation between the feature map and the original image candidate area without increasing the amount of calculation. Here, the pixel accuracy of the extracted area is also higher. Finally, the network is trained by means of the total loss of the proposal region generation network and the total loss of the target candidate region recognition network. Through the above several innovative means, the accuracy of the classification results of the tongue-piercing tongue and the extraction of edge pixels are improved.

6 Conclusion

In summary, the findings show that the multiscale CNN model has achieved ideal results in the classification of dots, thorns and petechia of different sizes and the pixel segmentation accuracy of the target regions. Soon, many spotted tongue datasets will be used iterative training. Multiscale CNN model will keep learning and spots recognition accuracy can be improved to adapt to the research of AI tongue diagnosis in TCM.

Fundings

Anhui Province College Natural Science Fund Key Project of China (KJ2020ZD77), and the Project of Education Department of Anhui Province (KJ2020A0379).

Competing interests

The authors declare no conflict of interest.

References

[1] ZHAO C, ZHANG XY, QIU RJ, et al. Application of artificial intelligence in tongue diagnosis of traditional Chinese medicine: a review. TMR Modern Herbal Medicine, 2021, 4(2): 24.

[2] LI F, DONG CW. Diagnostics of Traditional Chinese Medicine. 3rd ed. Beijing: Science Press, 2018.

[3] XU JT, ZHANG ZF, SUN Y, et al, Recognition of dotted-thorny and petechia features in tongue image analysis. Journal of Shanghai University of Traditional Chinese Medicine, 2004, 18(4): 38-40.

[4] LI NM, LI SW, LIU S, et al, Clinical study of brain fatigue of tongue picture. Lishizhen Medicine and Materia Medica Research, 2014, 25(10), 2424-2426.

[5] WANG S, LIU KH, WANG LT. Recognition and extraction of dotted-thorny and petechia in tongue diagnosis images. Computer Engineering and Science, 2017, 39(6): 1126–1130.

[6] KAN HX, ZHANG LY, DONG CW. A tongue image recognition method for TCM syndromes of type 2 diabetes mellitus. Chinese Journal of Biomedical Engineering, 2016, 35(6): 658–664.

[7] YANG SM. Study on automatic acquisition and feature recognition of TCM tongue image in chronic kidney disease. Chengdu: University of Electronic Science and Technology of China, 2018.
[8] LI R. Study on the diagnosis model of hyperactivity syndrome of liver fire in hypertension based on tongue diagnosis objectification. Beijing: China Academy of Chinese Medical Sciences, 2019.
[9] HUO CM, ZHENG H, SU HY, et al. Tongue shape classification integrating image preprocessing and Convolution Neural Network//Intelligent Robot Systems. IEEE, 2017. doi: 10.1109/ACIRS.2017.7986062.
[10] DONG JF, HUANG JC, WANG JY. Application of convolutional neural network in tongue image recognition of tumor patients. Beijing Journal of Traditional Chinese Medicine, 2020, 39(11): 1216–1219.
[11] ELHAM G, SEYED RKT, MARYAM K. Increasing the accuracy in the diagnosis of stomach cancer based on color and lint features of tongue. Biomedical Signal Processing and Control, 2021, 69: 102782.
[12] HAMED MOZAFFARI M, LEE W. Encoder-decoder CNN models for automatic tracking of tongue contours in real-time ultrasound data. Methods, 2020, 179: 26–36.
[13] ZHOU C, FAN H, LI Z. Tonguenet: accurate localization and segmentation for tongue images using deep neural networks. IEEE Access, 2019, 7: 148779–148789.
[14] LIU M, WANG XT, ZHOU L, et al. Study on TCM tongue image extraction and recognition based on deep learning and migration learning. Journal of Traditional Chinese Medicine, 2019, 60(10): 835–840.
[15] TRAJANOVSKI S, SHAN C, WEIJTMANS PJC, et al. Tongue tumor detection in hyperspectral images using deep learning semantic segmentation. IEEE Transactions on Biomedical Engineering, 2021, 68(4): 1330–1340.
[16] WANG X, LIU J, WU C, et al. Artificial intelligence in tongue diagnosis: using deep convolutional neural network for recognizing unhealthy tongue with tooth-mark. Computational and Structural Biotechnology Journal, 2020, 18: 973–980.
[17] WANG XM, WANG RY, GUO D, et al. Research on tongue image prick recognition method based on auxiliary light Source. Chinese Journal of Sensors and Actuators, 2016, 29(10): 1553–1559.
[18] REZENDE E, RUPPERT G, CARVALHO T, et al. Malicious software classification using transfer learning of ResNet-50 deep neural network. IEEE International Conference on Machine Learning & Applications, 2018, 738(4): 51–59.
[19] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6): 1137–1149.
[20] HE KM, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2020, 42(2): 386–397.
[21] LI NM, QU XF, LIU S, et al. Discussion on zonal method of tongue and viscera. Guangming Traditional Chinese Medicine, 2014, 29(5): 895–898.
[22] VINYALS O, TOSHEV A, BENGIO S, et al. Show and tell: lessons learned from the 2015 MSCOCO image captioning challenge. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2016, 39(4): 652–663.
[23] SONG AY, LOU YN, YANG QX, et al. Diagnosis of early esophageal cancer based on TCM tongue inspection. Biomedical and Environmental Sciences, 2020, 33(9): 718–722.
[24] FAN SY, CHEN B, ZHANG XR, et al. Machine learning algorithms in classifying TCM tongue features in diabetes mellitus and symptoms of gastric disease. European Journal of Integrative Medicine, 2021, 43: 101288.
[25] LI P, YI N, DING CS, et al. Research on classification diagnosis model of psoriasis based on deep residual. Digital Chinese Medicine, 2021, 4(2): 92–101.
[26] LEUNG YLA, GUAN BH, CHEN S, et al. Artificial intelligence meets traditional Chinese medicine: a bridge to opening the magic box of sphygmopalpation for pulse pattern recognition. Digital Chinese Medicine, 2021, 4(1): 1–8.
[27] LUO Y, LIU YN, LIN B, et al. Research on the correlation between physical examination indexes and TCM constitutions using the RBF neural network. Digital Chinese Medicine, 2020, 3(1): 11–19.
[28] HU Y, WEN GH, LUO MN, et al. Fully-channel regional attention network for disease-location recognition with tongue images. Artificial Intelligence in Medicine, 2021, 118: 102110.
[29] LI DX, GUAN J, LI F. Development of tongue diagnosis instrument and its current application in the objective study of tongue diagnosis. World Traditional Chinese Medicine, 2017, 12(2): 456–460.
[30] JIANG JP, YANG H, ZHANG HY. Identification of common tongue coating based on color features. Micromachines and Applications, 2017, 36(17): 102–105.
[31] XIANG LW. Optimization algorithm for image feature extraction of spleen deficiency and tongue image recognition. Nanchang: Jiangxi University of Science and Technology, 2017.
[32] XU L, CHEN W, ZHANG YJ. Common evidence type of chronic hepatitis B. Liaoning Journal of Traditional Chinese Medicine, 2014, 41(9): 1817–1819.
[33] HAO YM, ZHE RR, JIN MX, et al. Association analysis between objective parameters of tongue diagnosis and glycosylated hemoglobin in patients with type 2 diabetes. China Journal of Traditional Chinese Medicin, 2018, 33(4): 1520–1523.
[34] JIANG T, GUO XJ, TU LP, et al. Application of computer tongue image analysis technology in the diagnosis of NAFLD. Computers in Biology and Medicine, 2021, 135: 104622.
[35] WANG X, WANG XR, LOU YN, et al. Constructing tongue coating recognition model using deep transfer learning to assist syndromes diagnosis and its potential in noninvasive ethnopharmacological evaluation. Journal of Ethnopharmacology, 2021, 285: 114905.
[36] ZHANG K, JIN S, DU JQ, et al. Study on TCM tongue diagnosis and objectification based on image analysis. Science and Technology Square, 2013, 135(2): 9–11.
[37] ZHANG D, ZHANG JH, MENG SP, et al. Objective research prospect of TCM tongue diagnosis based on hyperspectral image technology. Chinese Journal of Chinese Basic Chinese Medicine, 2019, 25(9): 1324–1326.
[38] LV YT. Study on the objectification of the tongue diagnosis based on the auxiliary light source. Tianjin: Tianjin University, 2016: 60-61.

基于多尺度卷积神经网络的舌象点刺识别模型建立与验证

彭成东[a, b], 汪莉[c], 蒋冬梅[d], 杨诺[b], 陈仁明[b], 董昌武[c*]

a. 合肥工业大学计算机与信息学院, 安徽 合肥 230009, 中国
b. 合肥云诊信息科技有限公司人工智能实验室, 安徽 合肥 230088, 中国
c. 安徽中医药大学中医学院, 安徽 合肥 230012, 中国
d. 安徽水利水电职业技术学院电子信息工程学院, 安徽 合肥 231603, 中国

【摘要】目的 舌象中点刺所生的部位、颜色、分布的疏密可以推测邪热所在脏腑及其轻重。本研究聚焦于人工智能的图像分析方法研究中医点刺舌识别。**方法** 基于图像深度学习和实例分割原理，设计了舌象点刺识别与提取模型。该模型包括多尺度特征图生成模块、候选区域搜索模块和目标区域识别模块。首先使用深度卷积网络分别建立多尺度低、高抽象度的特征图谱，再在特征图上进行目标候选框生成算法和优选策略以精选出高质量目标候选区域，最后使用分类网络对目标区域分类、计算目标区域像素，最终得到舌象表面点刺的区域分割。在无辅助光源条件下手机拍摄的不同规格舌象，使用该方法进行实验。**结果** 实验结果表明，该点刺识别受试者工作特征曲线下的面积（AUC）值为 92.40%，精确度为 84.30%，灵敏度为 88.20%，特异度为 94.19%，召回率为 88.20%，区域像素准确率指标像素精度（PA）为 73.00%，均像素精度（mPA）为 73.00%，交并比（IoU）为 60.00%，均交并比（mIoU）为 56.00%。**结论** 本研究结果表明该模型适用于中医舌诊系统应用。基于多尺度卷积神经网络的点刺舌识别，有助于提高点刺分类和点刺区域像素的精准提取，为中医智能舌诊提供一种切实可行的方法。

【关键词】点刺识别提取；舌象特征；实例分割；卷积神经网络；中医舌诊系统；人工智能