IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Digital Object Identifier 10.1109/ACCESS.2017.DOI

An Automatic Recognition of Tooth-Marked Tongue Based on Tongue Region Detection and Tongue Landmark Detection via Deep Learning

WENJUN TANG [1], YUAN GAO[2], LEI LIU[1], TINGWEI XIA[3], LI HE[2], SONG ZHANG [1], JINHONG GUO [4], (MEMBER, IEEE), WEIHONG LI[3], AND QIANG XU[2], (MEMBER, IEEE).

[1]Hospital of Chengdu University of Traditional Chinese Medicine, Chengdu610075, China
[2]College of Medical Information Engineering,Chengdu University of Traditional Chinese Medicine,Chengdu611130, China
[3]School of Basic Medical Science, Chengdu University of Traditional Chinese Medicine, Chengdu611130, China
[4]School of Information and Communication Engineering, Institute of Medical Equipment, University of Electronic Science and Technology of China, Chengdu611731, China

Corresponding authors: Qiang Xu (xuqiang@cdutcm.edu.cn), Weihong Li (lwh@cdutcm.edu.cn), and Jinhong Guo (guojinhong@uestc.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFC1703304, in part by Sichuan Science and Technology Department Project under Grant 2020YJ0496 and 2020YFS0386, and in part by Chengdu Science and Technology Department Project under Grant 2019-YF05-00289-SN and 2019-YF05-00496-SN. Wenjun Tang, Yuan Gao, and Lei liu contributed equally to this work.

ABSTRACT The tooth-marked tongue refers to the tongue with the edge featured in jagged teeth marks, which is a significant indicator for reflecting the conditions of patients' internal organs in Traditional Chinese Medicine (TCM). From the perspective of computer vision, due to the small variance in the global region (original image) but the large variance in the local region (tongue region), especially in the differential region (tongue edge region around landmarks), the recognition of the tooth-marked tongue is a naturally fine-grained classification task. To address this challenging task, a two-stage method based on tongue region detection and tongue landmark detection via deep learning is proposed in this paper. In the first stage, we introduce a cascaded convolutional neural network to detect the tongue region and tongue landmarks simultaneously for minimizing the redundancy information and maximizing discriminative information explicitly. In the second stage, we send not only the detected tongue region but also the detected tongue landmarks to a fine-grained classification network for the final recognition. Conclusively, our method is highly consistent with human perception. Moreover, to the best of our knowledge, we are the first attempt to manage the tooth-marked tongue recognition via deep learning. We conducted extensive experiments with the proposed method. The experimental results demonstrate the effectiveness of the proposed method.

INDEX TERMS Tooth-marked tongue, Tongue region detection, Tongue landmark detection, Deep learning, Traditional Chinese Medicine.

I. INTRODUCTION

AS a popular complementary and alternative medicine, traditional Chinese medicine (TCM) is effective in preventing and treating various diseases, which has been incorporated into the latest global medical outline (Ver.2019) by World Health Organization (WHO) [1, 2]. Tongue diagnosis is an effective, inexpensive, and non-invasive method to evaluate the conditions of patients's internal organs in TCM, which has been widely applied to clinical analyses and applications for thousands of years [3, 4, 5]. However, the popularization of traditional tongue diagnosis has been limited because the diagnosis process relies entirely on the experience and knowledge of the practitioners. In addition, environmental factors (i.e. illumination) and examinee factors (i.e. subjects' position) may also interfere the tongue diagnosis and lead to a unreliable and inconsistent tongue diagnosis result. Hence, it is imperative to utilize the image processing and pattern recognition technology in aid of the objective analysis of tongue image. Phenomenally, the tooth-marked tongue is the tongue with the edge featured in jagged

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3017725, IEEE Access

teeth marks that are attributable to a long-period compression of teeth due to the enlarged tongue body (see Fig.1), which is a sensitive indicator for the deficiency of YANG and QI [5, 6]. Meanwhile, in the TCM system, YANG and QI are the essential substances that constitute and maintain the life activities. Therefore, the recognition of the tooth-marked tongue plays a vital role in the TCM diagnosis process.

Since tongue diagnosis is not widely accepted by modern western medicine as well as the difficulties involved in collecting and labeling tongue image datasets, the computerized tongue image processing is not so popular as other mainstream visual tasks like face recognition, facial landmark detection, etc. Nevertheless, some researchers have been contributing to the computerized tongue image processing over the last few decades, including tongue segmentation [4, 7, 8, 9, 10], tongue image color analysis [11, 12], and tongue shape analysis [13], etc.

As to the aspect of automatic tooth-marked tongue recognition, Shao et al. [6] proposed an algorithm by taking advantage of the features of concave and change of brightness. Wang et al. [14]introduced a method to recognize the teeth-marked tongue by calculating the slope of the margin of the tongue and the length and degree of the concave regions. Normally, these methods above were primarily based on handcrafted features. It is well known that the handcrafted features must be very carefully defined prior to recognition and be extremely time-consuming and tedious. Recently, with the outstanding ability of feature learning and representation of deep learning, Wang et al. [5] fused a CNN varient into the recognition of tooth-marked tongue. The proposed method contained three stages which were very similar to R-CNN. First, with the method by which all convex hulls were generated. Then, a ConvNet was resorted to extract a fixed-length feature vector for each convex hull. Finally, feature vectors were grouped and an SVM was resoluted to classify the tongue. However, there were two shortcomings restricting the application: 1) As the digitally acquired tongue images not only encompass the tongue, but also involve the parts of the face region. However, the input in this paper was a manually segmented tongue image that reduced the degree of automation. 2) The stage of generation convex hulls involved a threshold based on prior knowledge to binarize the image that contributed to the algorithms more leading-prone to overfitting. Consequently, in this paper, we aim to establish a highly automatic method with more accuracy for the recognition of tooth-marked tongue.

For a better elaboration on the novelty of our work, we divide the digital images into three regions deliberately, i.e., global region (original image), local region (tongue region) and differential region (tongue edge region around landmarks). From the perspective of computer vision, due to the fact that tooth-marked tongue/nontooth-marked tongue is similar in the global region (see Fig.1-[a,b]) but differs from the local region (see Fig.1-[c,d,e,f]), especially in the differential region (see Fig.1-[g,h,i,j]), the recognition of the tooth-marked tongue is a naturally fine-grained classification

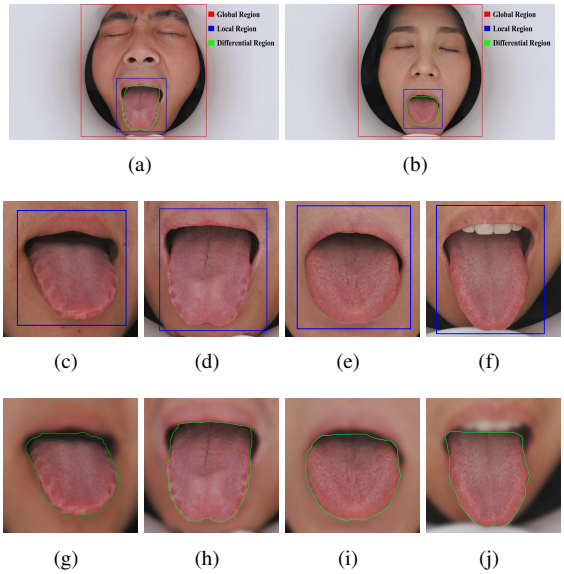


FIGURE 1: The illustrations of global region, local region and differential region, wherein Fig.(a)-(b) depict global region (original image), Fig.(c)-(f) depict local region (tongue region), and Fig.(g)-(j) depict differential region (tongue edge region around landmarks). It should be noted that Fig.(a), (c), (d), (g) and (h) correspond to the tooth-marked tongue and Fig.(b), (e), (f), (i) and (j) correspond to the nontooth-marked tongue, respectively.

task. Proverbially, fine-grained classification is a challenging task in computer vision fields. For all the fine-grained tasks, the key is to minimize redundancy features from the global region while maximizing discriminative features from the local region. As to the task of tooth-marked tongue recognition, the subtask, tongue region detection, is introduced to minimize the redundancy information from the global region. Meanwhile, another subtask, tongue landmark detection, is employed to maximize discriminative information from the local region. After the tongue region and tongue landmarks detected, both the two information are fed to a classifier for the final recognition. In our method, a cascaded convolutional neural networks (CNNs) is employed for the tongue region detection and tongue landmark detection. Meanwhile, a fine-grained classification network is introduced for the tooth-marked tongue recognition. The main contributions of this paper are as follows:

(1) As stated above, the recognition of the tooth-marked tongue is a naturally fine-grained classification task. To manage this task, we introduce two subtasks, i.e., tongue region detection and tongue landmark detection, to minimize the redundancy information and maximize discriminative information from original image. The experimental results show that our recognition method outperforms the methods without tongue region detection and/or tongue landmark detection.
(2) Given its outstanding ability of feature extraction, deep learning has shown remarkable performances in various

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3017725, IEEE Access

computer vision fields. In this paper, two deep learning-based networks are fused into our study, in which the cascaded CNNs is used for the tongue region detection as well as tongue landmark detection, while the fine-grained network is employed for the tooth-marked tongue recognition. To the best of our knowledge, we are the first to introduce deep learning for an end-to-end tooth-marked tongue recognition.

II. PROPOSED METHOD

In this section, we describe the proposed algorithm in detail. The whole framework of our approach (see Fig.2) contains two stages. First, a cascaded CNNs is used to detect the tongue region as well as 30 landmarks along the tongue edge. And then, a fine-grained classification network is employed to recognize the tooth-mark tongue by inputting the detected tongue region and landmarks simultaneously.

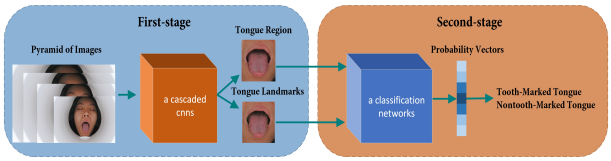


FIGURE 2: Overview of our framework

A. THE STAGE OF THE TONGUE REGION AND LANDMARKS DETECTION

As stated in the introduction section, both the tongue region detection and tongue landmark detection are essential for the recognition of the tooth-marked tongue. Inspired by the idea of Multi-task Cascaded Convolutional Networks (MTCNN) [15], we design a cascaded CNNs with multi-task learning to manage these two tasks simultaneously. By forming a cascaded architecture with three-stage deep convolutional networks and employing multi-task learning to predict tongue region and tongue landmarks in a coarse-to-fine way, the cascaded CNNs achieve excellent performance in both the tongue detection and tongue landmark detection.

Three interrelated subtasks are used to train the network in our method: bounding box regression (Subtask 1), tongue classification (Subtask 2), and tongue landmark localization (Subtask 3). Subtask 1 and Subtask 2 are employed for the tongue region detection, and Subtask 3 is employed for the tongue landmark detection, respectively. Specifically, the subtask of bounding box regression is used to locate the position of the tongue, the subtask of tongue classification is used to determine whether the target is a tongue, and the subtask of tongue landmark localization is used to locate the landmarks along tongue edge. The architecture of our cascaded CNNs (see Fig.3) is a combination of four modules, i.e., Image Pyramid, Coarse-Net, Fine-Net and Refine-Net. More details are as follows:

(1) Image Pyramid: In order to comprehensively extract feature information from pictures, we use the image pyramid for the image preprocessing in the input. Image pyramid is a common method for multi-scale representation in the field of computer vision. Combined with the subsequent modules, it can detect objects at different scales and ensure the original information wherever possible.

(2) Coarse-Net: Coarse-Net is designed to predict the candidate bounding boxes that contain the target tongue region as much as possible without high accuracies. by inputting the image at a small scale of 12×12, the network avoids paying attention to local details and extracts the bounding boxes from a global perspective. The Coarse-Net is a a fully convolutional network that can greatly reduce the amount of network parameters and increase the speed of the network. Meanwhile, we employ non-maximum suppression (NMS) to merge highly overlapped candidates.

(3) Fine-Net: After Coarse-Net, all the candidates are fed to Fine-Net, which further rejects a wealth of false candidates and performs calibration with bounding box regression. The image resolution of the input in this network is at a medium scale of 24×24. A higher resolution can provide more detailed feature information and enable the network to pick out more accurate candidate bounding boxes. Compared with Coarse-Net, Fine-Net increases the number of channels in every convolution layer. At the same time, a fully connected layer is added to Fine-Net, which can improve the understanding of the global information, thereby a more accurate prediction.

(4) Refine-Net: After Fine-Net, all the selected candidates are fed to Refine-Net. Compared with Fine-Net, we not only deepen the depth of the network, but also increase the width of the network. Moreover, the image resolution of the input in this network is at a large scale of 48×48. Through a larger scale input and a more complex network, the feature extraction capability of this network can be greatly improved. At the same time, the additional task, tongue landmark localization, is trained together with the bounding box regression task and tongue classification by the multi-task learning.

In the cascaded CNNs, the output of each network should include a $1 \times 1 \times 4$ vector to predict the bounding box coordinates of tongue region, and a $1 \times 1 \times 2$ vector to predict the confidence of tongue classification corresponding to the bounding box. In addition, compared with Coarse-Net and Fine-Net, Refine-Net further outputs a $1 \times 1 \times 60$ vector to predict the positions of 30 tongue landmarks, in which every tongue landmark is a 2D coordinate. The object functions are defined as follows:

- Bounding box regression: The bounding box are four coordinates, including left, top, height and width, thus $y_i^{box} \in R^4$. The objective function is defined as Eq.1, in which $\tilde{y}_i^{box}$ is predicted coordinate and y_i^{box} is the ground-truth coordinate for the i^{th} sample.

$$\mathcal{L}_i^{box} = \|\tilde{y}_i^{box} - y_i^{box}\|_2^2 \tag{1}$$

- Tongue classification: For every bounding box predicted by the network, the area containing the tongue is dif-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3017725, IEEE Access

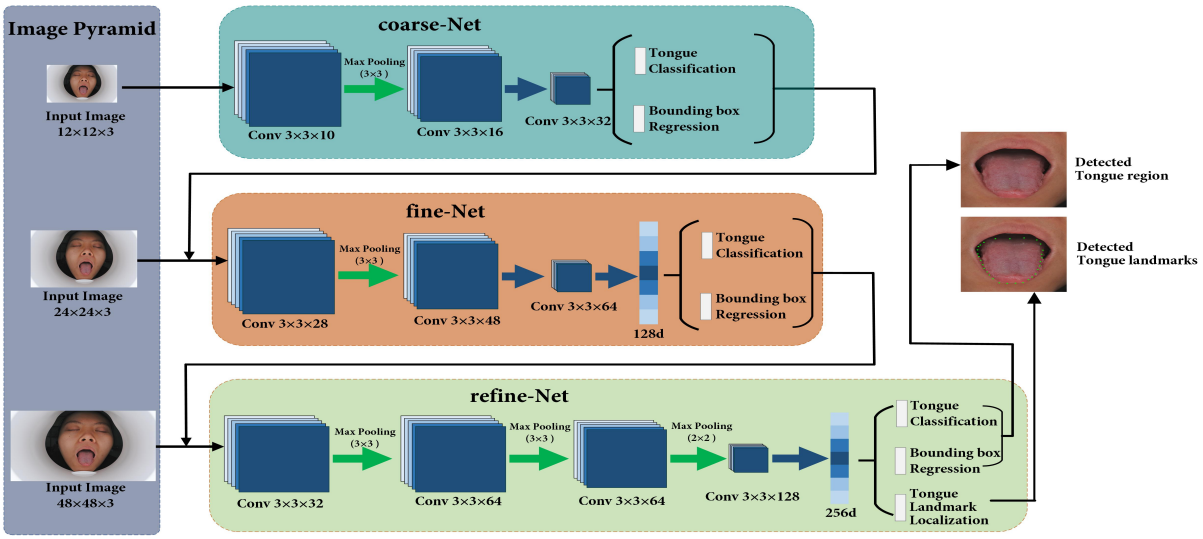


FIGURE 3: The architecture of the cascaded CNNs

ferent. Therefore, each of them has a confidence coefficient indicating the degree of whether the bounding box contains the tongue or not. Consequently, this task is formulated as a binary task. The objective function is defined as Eq.2, in which p_i is the confidence coefficient of the predicted bounding box, and $y_i \in \{0, 1\}$ is the ground-truth label, respectively.

$$\mathcal{L}_i^{det} = (y_i^{det} \log(p_i) + (1 - y_i^{det})(1 - \log(p_i))) \quad (2)$$

- Tongue landmark localization: Similar to bounding box regression task, tongue landmark localization is formulated as a regression task and the object function is defined as Eq.3, in which $\tilde{y}_i^{landmark}$ is the tongue landmarks' coordinates obtained from the network and $y_i^{landmark}$ is the ground-truth coordinate for the i^{th} sample. There are 30 landmarks, and thus $y_i^{landmark} \in R^{60}$.

$$\mathcal{L}_i^{landmark} = \|\tilde{y}_i^{landmark} - y_i^{landmark}\|_2^2 \quad (3)$$

- The overall objective function: Both the tongue images and none-tongue images are used in the training. For the none-tongue images, the $\mathcal{L}_i^{landmark}$ is set to 0. The overall objective is defined as Eq.4. in which N is the number of training samples and α_j denotes the weights of loss. $\beta_i^j \in \{0, 1\}$ is the image type indicator. Based on the empirical tests, we set the value of $\alpha_{det}/\alpha_{box}/\alpha_{landmark}$ to 1/1/1 in Coarse-Net and Fine-Net, while the values are set to 0.5/0.5/2 in Refine-Net.

$$J = min \sum_{i=1}^{N} \sum\nolimits_{j \in \{det, box, landmark\}} \alpha_j \beta_i^j \mathcal{L}_i^j \quad (4)$$

B. THE STAGE OF TOOTH-MARKED TONGUE RECOGNITION

As stated above, the recognition of tooth-marked tongue is a fine-grained classification task, of which the key is to maximize the useful information in the local region while minimizing the redundancy in the global region. Inspired by the idea of destruction and construction learning for image classification [16], we design a fine-grained classification network called Destruction and Construction Network (DCN) to extract the discriminative information from the detected tongue region.

DCN is composed of three units, i.e., Destruction Unit, Backbone, and Construction Unit. For the Destruction unit, we introduce the Region Shuffle Mechanism (RSM) to enhance local area differences and guide the network to learn more obvious differences in the image. While for the Construction unit, we introduce the Region Alignment Network (RAN) to reconstruct the destructed images and guide the network to focus on local differences and learn global information. The framework of DCN is shown in Fig.4.

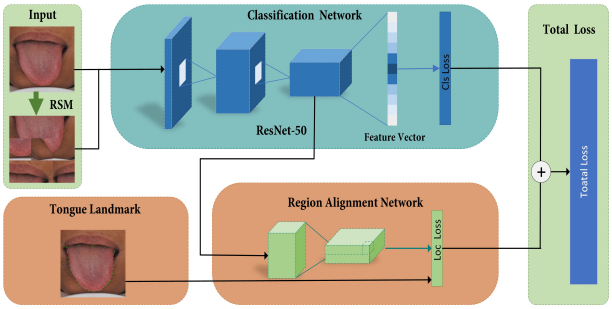


FIGURE 4: The architecture of the DCN

(1) Destruction Unit: In this unit, we introduce the RSM mechanism to disrupt the spatial layout of local image region, which can split the image into $N \times N$ non-repetitive patches. Notice that the difference between the destructed image and the original image should be neither too large nor too small. For the large difference, the network is likely to fail to learn the consistent features,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3017725, IEEE Access

while for the small difference, the network cannot learn enough distinctive information. Consequently, we have confined the split patches can only be exchanged with neighborhoods that prevent the destructed image from being vitally complicated.

(2) Backbone: The backbone of DCN is a classification network that maps the image of tongue region to a vector representing the probability values of different classes by a series of deep convolutions. In order to verify the generality of our method, we have selected Vgg-16 and ResNet-50 as the backbone of RAN. The objective function $\mathcal{L}_{cls}$ is formulated as Eq.5, in which $\mathcal{I}$ represents the image set, l represents the one-hot vector indicating the category, $C(I)$ and $C(\varnothing(I))$ represent the predicted probability vector of the original image and destructed image, respectively. Notice that the tongue region I, the destructed image $\varnothing(I)$, and the one-hot vector l are coupled as $\langle I, \varnothing(I), l\rangle$ for training.

$$\mathcal{L}_{cls} = -\sum_{I\in\mathcal{I}} l \cdot \log[C(I)C(\varnothing(I))] \tag{5}$$

(3) Construction Unit : In this unit, we introduce the RAN to measure the positional accuracy of different areas of the the tongue region to help locate the main object in image. The input of RAN is the feature map of the last convolutional layer of the backbone. A new feature map with the size of $2 \times N \times N$ is output by the convolution and average pooling operations. Two channels correspond to the location coordinates of rows and columns of image patches. It should be noted that in this unit, we send the tongue landmark information detected in the first stage, which enables the network to locate the target object more accurately. The objective function is formulated as Eq.6, where $P_{(i,j)}$ and $P_{\sigma(i,j)}$ represent the predicted coordinates of tongue landmarks located at(i, j) in original image and destructed image, $M_{(i,j)}(I))$ and $M_{\sigma(i,j)}(\varnothing(I))$ represent the predicted coordinates of tongue region located at(i, j) in original image and destructed image. The alignment loss is calculated by the L1 distance between the predicted locations and original locations.

$$\mathcal{L}_{loc} = \sum_{I\in\mathcal{I}}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(P_{\sigma(i,j)}\left|M_{\sigma(i,j)}(\varnothing(I)) - \begin{bmatrix} i \\ j \end{bmatrix}\right|_1 + P_{(i,j)}\left|M_{(i,j)}(I) - \begin{bmatrix} i \\ j \end{bmatrix}\right|_1\right) \tag{6}$$

(4) The overall objective function: The overall objective function of this stage is defined as Eq.7. in which α is the weight corresponding to the Backbone. Based on our empirical tests, we set the value of α to 0.7.

$$J = \alpha\mathcal{L}_{cls} + (1-\alpha)\mathcal{L}_{loc} \tag{7}$$

III. EXPERIMENTS

In this section, we use the proposed method described in Section II to perform the experiments. We conduct two experiments: one for tongue region detection and tongue landmark detection, and the other for tooth-marked tongue recognition.

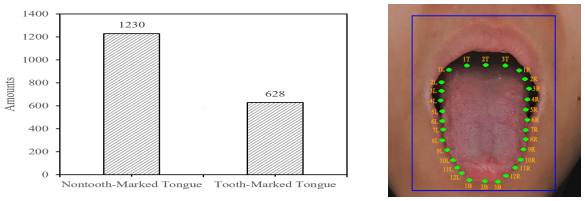
A. DATASETS AND ANNOTATIONS

Because an urgent need for tongue characterization exists, the National Key Research and Development Program of China supported a project titled "Research and development of TCM intelligent tongue diagnosis system". To date, 1,858 tongue images have been collected from the Affiliated Hospital of Chengdu University of Traditional Chinese Medicine since October 2018, of which 1,230 are nontooth-marked tongue images, and the rest are the tooth-marked tongue images. These images were captured by a specialized device with a high-end industrial CCD camera and the commonly used standard illumination (D50). All the tongue images are reliable, and their high quality is ensured. These images are randomly split into the training and testing sets with 1,538 and 300 images, respectively.

The data annotations consist of the following two aspects:

(1) Image annotation: Image annotation includes tongue landmark annotation and tongue region annotation. 30 tongue landmarks are annotated as follows (see Fig.5-b): 3 and 3 landmarks uniformly locate on the top and bottom of the tongue, while both sides of the tongue have 12 uniformly located landmarks. Tongue region is annotated by the bounding box, in which the upper boundary is horizontally passing through the top of the lip, the lower boundary is horizontally passing through the tongue tip, the left and right boundaries are vertically passing through the left and right corners of the mouth, respectively.

(2) Classification label: The binary classification, tooth-marked tongue/none tooth-marked tongue, is labeled by specialists. Two primary physicians conduct cross-validation, and a resident physician conducts final validation.



(a) (b)

FIGURE 5: The distribution of tongue images and the paradigm of image annotation

B. TONGUE REGION DETECTION AND TONGUE LANDMARK DETECTION EXPERIMENTS

In this section, we conduct the tongue region detection and tongue landmark detection experiments with the cascaded

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3017725, IEEE Access

CNNs described in Section II-A. A major breakthrough hitherto has been achieved for deep learning-based methods in object detection and landmark detection, especially in face detection and facial landmark detection. However, few algorithms have been applied to tongue region detection and tongue landmark detection so far. For comparison, we choose several typically non-deep learning-based methods widely used in face detection and facial landmark detection. The non-deep learning-based tongue detection methods are the Viola-Jones (VJ) [17], Histogram of Oriented Gradient (HOG) [18], and Deformable Part based Model (DPM) [19]. The non-deep learning-based tongue landmark detection methods are the Active Shape Model (ASM) [20], Active Appearance Models (AAM) [21], and Cascaded Pose Regression (CPR) [22]. In our experiments, our proposed model is trained on a Titan XP GPU and optimized by Adaptive Moment Estimation (Adam) [23] with a learning rate of 1e-4 and a termination at 4k iterations. We apply the early stopping strategy as our regularization method in the training phase. We use the model trained by the face datasets for transfer learning. At transfer learning, we set the output of prediction landmarks to be 30.

1) Evaluation metrics

For the tongue region detection, we use the Average Precision of 60% (AP_{60}) and the Intersection over Union (IoU) that are widely used in object detection as the evaluation indicators. For the tongue landmark detection, the performance is measured by Mean Error Rate (MER) and Failure Rate (FR). MER is calculated as Eq.8, in which S is the total number of test samples, (x, y) and $(x^{'}, y^{'})$ are the ground truth and the detected position corresponding to the s^{th} test sample, and l is the width of the bounding box returned by our model. When a sample error is larger than 5%, it is rated as a failed sample. Accordingly, FR is measured as Eq.9, in which M is the total number of the failed samples.

$$MER = \frac{1}{S}\sum_{s=1}^{S}\frac{\sqrt{(x-x')^2+(y-y')^2}}{l} \tag{8}$$

$$FR = \frac{M}{S} \tag{9}$$

2) Results and discussions

There have been no public tongue image datasets yet for evaluating the tongue region detection and tongue landmark detection performance. Our deep learning-based method and other non-deep learning-based methods are conducted in our private datasets for ease of comparison. The results of different methods applied in tongue region detection and tongue landmark detection are shown in Table 1, respectively. It could be seen that our deep learning-based method increases by 9.1%/6.3% in AP_{60}/IoU compared with the best non-deep learning-based method on the tongue region detection. Meanwhile, our deep learning-based method decreases by 1.2%/1.6% in MER/FR compared with the best non-deep learning-based method on the tongue landmark detection, which suggests that our method can better handle both the tongue region detection and tongue landmark detection. Fig.6 visualizes the result of our deep learning-based method, from which superb results are highlighted.

TABLE 1: The result of our deep learning-based method on tongue region detection and tongue landmark detection compared with other non-deep learning-based methods

Tongue region detection			Tongue landmark deteciton		
Model	AP_{60}(%)	IoU(%)	Model	MER(%)	FR(%)
VJ	42.6	80.4	ASM	4.6	6.2
HOG	43.0	82.5	AAM	4.2	5.6
DPM	50.4	86.9	CPR	3.7	4.5
Ours	59.5	93.2	Ours	2.5	2.9

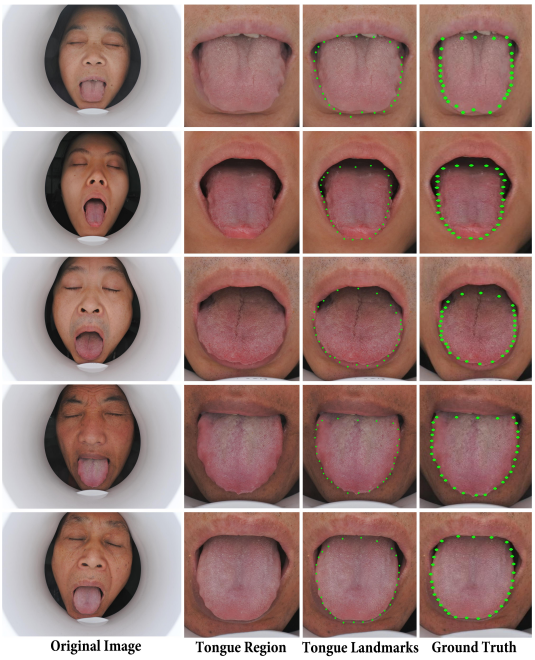


FIGURE 6: The visualization of our deep learning-based method on tongue region detection and tongue landmark detection

C. TOOTH-MARKED TONGUE RECOGNITION EXPERIMENTS

In this section, we conduct the tooth-marked tongue recognition experiments with the fine-grained classification network described in Section II-B. As introduced above, it is expected that an improvement in performance can be achieved by sending not only the detected tongue region but also the detected tongue landmarks. Comparatively, we conduct different experiments without sending the detected tongue region and/or the detected tongue landmarks. In these experiments, ResNet-50 and VGG-16 are employed as the backbone of the fine-grained classification networks, respectively. The

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3017725, IEEE Access

input images are resized to a fixed size of 224×224 and randomly cropped into 192×192. Random rotation and random horizontal flip are applied for data augmentation. Our proposed model is trained for 500 epochs, optimized by Adam with a learning rate of 1e-4.

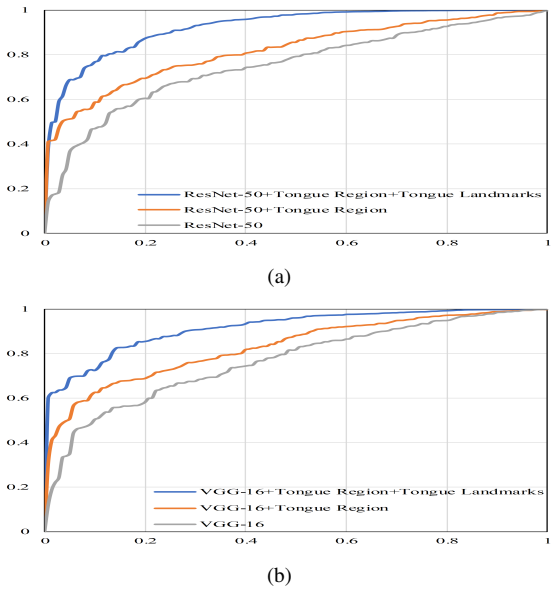
It should be noted that N in the RSM is a hyperparameter. The change of N will also affect the experimental results. However, the value of N pertains to the input image. More specifically, the width and height of the input image should be divisible by N. Through repeated experiments, we found that the best results are obtained when $N = 3$. Typically, in case of a minute N, the variance between the original image and the destructed image would not be significant, and the local discriminative information couldn't be learned. On the other hand, in case of a significant N, the destructed image would be excessively complicated. A loss of abundant global information results in a decrease in recognition performance.

1) Evaluation metrics

Precision, recall, and accuracy (acc) are taken as the indicators to evaluate the performance of tooth-marked tongue recognition, while F1-score is used as a comprehensive criterion for evaluation. The F1-score is calculated as Eq.10, wherein p_k and r_k indicate the precision and recall of the classes, respectively, and c indicates the number of class categories. In our study, the value of c is equal to 2.

$$F1 - score = \frac{1}{c}\sum_{k=1}^{c}\frac{2p_k \times r_k}{p_k + r_k} \tag{10}$$

2) Results and discussions



(a)

(b)

FIGURE 7: ROC curves of different recognition experiments based on the backbone of Resnet-50 and VGG-16

Table 2 provides the results of these experiments. It can be seen that without sending the detected tongue region and

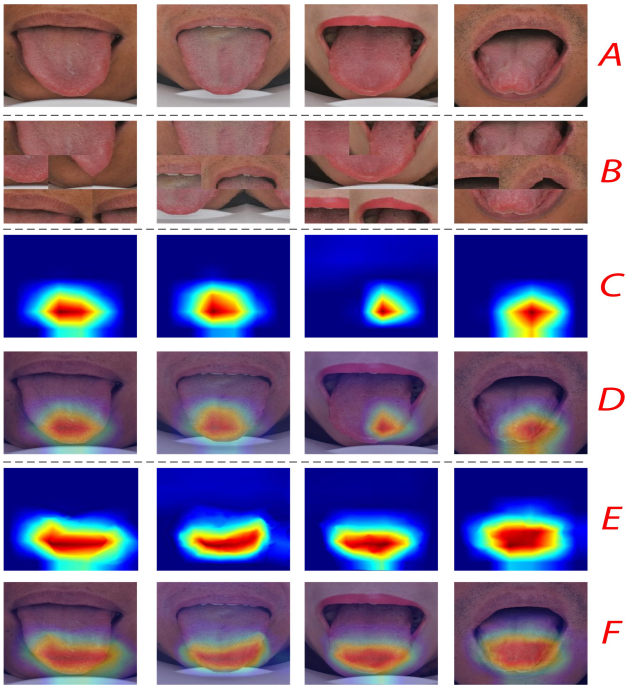


FIGURE 8: The heatmaps of different recognition experiments, in which (A) depicts the detected tongue region, (B) depicts the destructed images, (C, D) depict the heatmaps corresponding to the recognition experiment with the detected tongue region only, and (E, F) depict the heatmaps corresponding to the recognition experiment with the detected tongue region and the detected tongue landmarks together. (C, E) are the heatmaps directly output by Grad-CAM, and (D, F) are the heatmaps superimposed on the original image. The red parts of the heatmaps indicate a large impact on the recognition, while the blue ones indicate a little effect on the recognition.

detected tongue landmarks to the fine-grained classification networks, the value of F1-score based on the backbone of VGG-16/ResNet-50 is merely 0.709/0.727. In contrast, as we send the detected tongue region to the networks, the values rise to 0.874/0.862, which suggests that the detected tongue region could minimize the redundancy information from the original image. Importantly, when we send the detected tongue region and detected tongue landmarks together, the values increase up to 0.924/0.948, which suggests that the detected tongue landmarks facilitate a better recognition by maximizing discriminative information from the tongue region. Fig.7-a and Fig.7-b are the ROC curves of the above experiments based on ResNet-50 and VGG-16, respectively, from which conclusion could be reached that the networks scored the highest results by sending the detected tongue region and detected tongue landmarks together. Fig.8 visualizes the heatmaps of these experiments. It can be seen that when we send the detected tongue region only, the network is mainly concerned with the difference in the front of the tongue. However, when we send the detected tongue region

TABLE 2: The results of different recognition experiments based on the backbone of Resnet-50 and VGG-16

Backbone	Tongue region	Tongue landmark	Precision	Recall	Acc(%)	F1-score
VGG-16	×	×	0.705	0.715	67.8	0.709
Resnet-50	×	×	0.717	0.736	70	0.727
VGG-16	√	×	0.846	0.904	86.4	0.874
Resnet-50	√	×	0.846	0.88	85	0.862
VGG-16	√	√	0.897	0.952	91.7	**0.924**
Resnet-50	√	√	0.935	0.96	94.2	**0.948**

and the detected tongue landmarks together, the network's interest area is larger. Meanwhile, the tongue's edge has a greater impact on the network, which is in line with human perception and further proves the effectiveness of our proposed method.

IV. CONCLUSIONS

With the small variance in the global region and large variance in the local region, the recognition of the tooth-marked tongue is a naturally fine-grained classification task. To resolve this task, an automatic recognition method of tooth-marked tongue based on tongue region detection and tongue landmark detection via deep learning is proposed in this paper. We first introduce a cascaded CNNs to detect the tongue region and tongue landmarks simultaneously to minimize the redundancy information and maximize discriminative information. Then we send not only the detected tongue region but also the detected tongue landmarks to a fine-grained classification network for the final recognition. The experimental results demonstrate the effectiveness of our proposed method. Prospectively, work is in progress to collect more tongue image datasets and design outperformed network architectures to further improve the performance. Additionally, it is scheduled to investigate the relationship between tongue images and diseases, laboratory indicators, symptoms, and signs via computer vision to expand the practicability of tongue image analysis.

REFERENCES

[1] David Cyranoski. The big push for chinese medicine for the first time, the world health organization will recognize traditional medicine in its influential global medical compendium. Nature, 561(7724):448–450, 2018.

[2] Qiang Xu, Wenjun Tang, Fei Teng, Wei Peng, Yifan Zhang, Weihong Li, Chuanbiao Wen, and Jinhong Guo. Intelligent syndrome differentiation of traditional chinese medicine by ann: A case study of chronic obstructive pulmonary disease. IEEE Access, 7:76167–76175, 2019.

[3] Marzia Hoque Tania, Khin Lwin, and Mohammed Alamgir Hossain. Advances in automated tongue diagnosis techniques. Integrative Medicine Research, 2018.

[4] Zuoyong Li, Zhaochai Yu, Weixia Liu, and Zuchang Zhang. Tongue image segmentation via color decomposition and thresholding. In 2017 4th International Conference on Information Science and Control Engineering (ICISCE), pages 752–755. IEEE, 2017.

[5] Xiaoqiang Li, Yin Zhang, Qing Cui, Xiaoming Yi, and Yi Zhang. Tooth-marked tongue recognition using multiple instance learning and cnn features. IEEE transactions on cybernetics, 49(2):380–387, 2018.

[6] Qing Shao, Xiaoqiang Li, and Zhicheng Fu. Recognition of teeth-marked tongue based on gradient of concave region. In 2014 International Conference on Audio, Language and Image Processing, pages 968–972. IEEE, 2014.

[7] J. Guo, Q. Xu, Y. Zeng, W. Tang, W. Peng, T. Xia, Z. Li, F. Teng, and W. Li. Multi-task joint learning model for segmenting and classifying tongue images using a deep neural network. IEEE Journal of Biomedical and Health Informatics, 2020(in press),doi:10.1109/JBHI.2020.2986376.

[8] Li Chen, Dongyi Wang, Yiqin Liu, Xiaohang Gao, and Huiliang Shang. A novel automatic tongue image segmentation algorithm: color enhancement method based on l* a* b* color space. In 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pages 990–993. IEEE, 2015.

[9] Bingqian Lin, Junwei Xle, Cuihua Li, and Yanyun Qu. Deeptongue: Tongue segmentation via resnet. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1035–1039. IEEE, 2018.

[10] Yushan Xue, Xiaoqiang Li, Pin Wu, Jide Li, Lu Wang, and Weiqin Tong. Automated tongue segmentation in chinese medicine based on deep learning. In International Conference on Neural Information Processing, pages 542–553. Springer, 2018.

[11] Xingzheng Wang, Bob Zhang, Zhimin Yang, Haoqian Wang, and David Zhang. Statistical analysis of tongue images for feature extraction and diagnostics. IEEE Transactions on Image Processing, 22(12):5336–5347, 2013.

[12] Bo Pang, David Zhang, Naimin Li, and Kuanquan Wang. Computerized tongue diagnosis based on bayesian networks. IEEE Transactions on biomedical engineering, 51(10):1803–1810, 2004.

[13] Bo Huang. Tongue shape classification by geometric features. Information Sciences, 180(2):312–324, 2010.

[14] Hong Wang, Xinfeng Zhang, and Yiheng Cai. Research on teeth marks recognition in tongue image. In 2014

International Conference on Medical Biometrics, pages 80–84. IEEE, 2014.

[15] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 23(10):1499–1503, 2016.

[16] Yue Chen, Yalong Bai, Wei Zhang, and Tao Mei. Destruction and construction learning for fine-grained image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5157–5166, 2019.

[17] Paul Viola, Michael Jones, et al. Rapid object detection using a boosted cascade of simple features. CVPR (1), 1(511-518):3, 2001.

[18] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), volume 1, pages 886–893 vol. 1, June 2005.

[19] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(9):1627–1645, Sep. 2010.

[20] Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. Active shape models-their training and application. Computer vision and image understanding, 61(1):38–59, 1995.

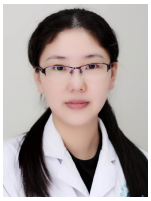
[21] T. F Cootes, G. J Edwards, and C. J Taylor. Active appearance models. Pattern Analysis Machine Intelligence IEEE Transactions on, 23(6):681–685, 2001.

[22] Piotr Dollar, Peter Welinder, and Pietro Perona. Cascaded pose regression. IEEE, 238(6):1078–1085, 2010.

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv: Learning, 2014.

WENJUN TANG received her B.S. degree and M.S. degree from Chengdu university of Traditional Chinese Medicine in 2006 and 2009, respectively. She currently works as a resident physician of TCM in Hospital of Chengdu University of Traditional Chinese Medicine. She conducted the data cleaning and data collection in this study.

YUAN GAO received his B.S. degree from Chengdu University of Information Engineering in 2006, and received his M.S. degree from University of Electronic Science and Technology of China in 2012. His research interests include medical image processing and biomechanics.



LEI LIU received her B.S. degree and M.S. degree from Chengdu university of Traditional Chinese Medicine in 2006 and 2009, respectively. She currently works as a resident physician of TCM in Hospital of Chengdu University of Traditional Chinese Medicine. She conducted image annotation in this study.

QIANG XU received his B.S. degree , M.S. degree and Ph.D. degre from Chengdu university of Traditional Chinese Medicine in 2012 , 2014 and 2017, respectively. He was a Post-Doctoral Fellow at College of Medical Information Engineering, Chengdu University of Traditional Chinese Medicine, from 2018 to 2020. His research interests include medical image processing and biomechanics.

WEIHONG LI has cultivated over 20 master's and doctoral students. she is currently a Professor and the Ph.D. Supervisor of the Chengdu University of Traditional Chinese Medicine,China. she conducted the final validation of the data in this study.

JINHONG GUO received the bachelor's degree in electronic engineering from the University of Electronic Science and Technology of China in 2010,and the Ph.D.degree in biomedical engineering from Nanyang Technological University in 2014. He was a Post-Doctoral Fellow of the pillar of engineering design with MIT, SUTD, Singapore, from 2014 to 2015. He is currently a Full Professor with the School of Information and Communication Engineering, University of Electronic Science and Technology of China. He has published over 100 publication in top journal, such as the IEEE TII, TBME, TBioCAS, Analytical Chemistry, and Biosensors and bioelectronics. His research interests include medical image processing and biomechanics. He served as a Lead Guest Editor and a Guest Editor for the IEEE TBioCAS, JBHI, TII, Micromachine, Electrophoresis, and Sensors.