Tooth-Marked Tongue Recognition Based on Mask Scoring R-CNN

Kong Xiangyong*, Rui Yingdi, Dong Xin, Cai Jian, Liu Yanan
University of Shanghai for Science and Technology, Shanghai, 200093, China
*kxy@usst.edu.cn

Abstract: Tooth-marked tongue is one of the important tongue features in Traditional Chinese Medicine tongue diagnosis. It reflects the dysfunction of liver, spleen and kidney, and contains rich pathological information. However, the recognition of tooth-marked tongue is challenging. Most existing methods only focus on the classification of tooth-marked tongue, and the exact location and number of teeth marks were not involved, which had no great practical significance for the subsequent treatment based on syndrome differentiation. In this paper, we try to solve these problems by proposing a method based on Mask Scoring R-CNN framework and transfer learning which can extract, visualize the teeth marks and identify the number of teeth marks. First, the tongue image is fed into residual network of depth 101 layers (ResNet-101) to generate regions of interest (RoIs) via region proposal network (RPN) and RoI feature via RoIAlign. Next, MaskIOU is then predicted using the predicted mask and RoI characteristics as inputs. The proposed approach achieved F1 score of 0.95. According to these experimental results, the approach can robustly detect and segment teeth marks, which can provide a basis for the severity analysis of tooth-marked tongue.

1. Introduction
The application of mobile health technology in Traditional Chinese Medicine (TCM) and Telemedicine is increasing, especially during the COVID-19. Tongue diagnosis is an important part of TCM diagnosis, because of its advantages of non-contact, effectiveness, non-invasive, real-time, convenience and low-cost, and other merits. For example, tongue diagnosis can be carried out at anytime and anywhere to meet the needs of the global primary health care system. The tongue can reflect the deficiency and excess of viscera, the rise and fall of Qi and blood, the depth of disease position, and the prognosis condition [1][2]. Objective studies on the tongue can observe the characteristics and changes of tongue image of COVID-19 patients and provide evidence for the treatment of patients [3]. Because the traditional tongue diagnosis is mainly based on the doctors’ personal experience, it is easily affected by the environment, and there are some problems such as subjectivity and inconsistent evaluation standards. Therefore, more and more computer researchers began to combine it with computer vision to establish an objective and quantitative assistant system for tongue diagnosis in TCM.

The objective study of tooth-marked tongue is important and challenging. Tooth-marked tongue is one of the important indexes in TCM diagnosis. The results of the clinical experiments showed that the tongue shape distribution of patients with different clinical types of COVID-19 was studied, and teeth marks were easy to be seen on the tongue of patients with common types. According to the research, 56% of the Chinese population has tooth-marked tongue, 11% of which are serious [4]. The recognition of tooth-marked tongue is more challenging. The main reasons are as follows: 1. There are various shapes of teeth marks, such as different colors, shapes and types of teeth marks; 2. Due to the limitation
of personal privacy and image acquisition, the number of tongue images is limited; 3. The number of images used in the gold standard is limited, and the tagging of tongue images requires the participation of many experts. However, experts can only classify the images within a limited time, and cannot mark the location of teeth marks on the images, they cannot provide information on the location, size and depth of teeth marks.

2. Related Work

Some progress has been made in tongue segmentation [5][6][7][8], tongue feature extraction and classification [9][10][11]. Li et al. [12] to solve the problem that the classification performance of a tooth-marked tongue based on the information of the indented area is not stable when it is not a concave area, a three-stage method using multiple instance learning and CNN is proposed. Sun et al. [13] proposed a visual interpretation method to solve the problem of feature visualization learned by neural network. This method takes the whole tongue as input, uses CNN to extract features, then classifies the tongue, and generates a rough localization map. Finally, the gradient weighted class is used to activate the mapped teeth mark area. Kanawong et al. [14] using Ada Boost and SVM of machine learning method and multiple-layer perceptual network used to establish the corresponding relationship between tongue image and symptoms, through the size, and distribution of ecchymosis spots on the tongue, to study its relationship with gastritis. Meng et al. [15] used CNN to construct the recognition of the cold disease, fever, and normal state in tongue diagnosis. The experimental results show that the recognition accuracy of the model on the problem of dichotomy is higher than that of the three classification problems. Hu et al. [16] constructed the neural network model corresponding to real-world tongue images and corresponding prescriptions, providing new ideas for medical services in the mobile medical system.

Deep networks are dramatically driving the development of computer vision, leading to a series of state-of-the-art in task including classification, object detection, semantic segmentation etc [17], which provides an opportunity to improve the accuracy of tongue feature recognition [18]. In this paper, a tongue characteristic recognition method based on Mask Scoring R-CNN [19] and transfer learning is proposed which can identify the teeth marks features, accurately locate the teeth mark position, calibrate the teeth mark size and extract the number of teeth marks. It can reduce the difficulty coefficient of applying deep learning technology on small sample data to a certain extent, and obtain higher classification accuracy.

3. Method

In this experiment, the flow chart of tongue image classification is shown in Figure 1. In the off-line training stage, first, the tongue image dataset is first constructed by selecting the tongue images that meet the requirements. Then, the picture is preprocessed and the tongue image is marked by many TCM experts. Finally, the model is trained with transfer learning. In the online testing stage, the model is tested through test pictures and its performance is evaluated.
3.1 Data Filtering and Preprocessing

In cooperation with Shu Guang Hospital affiliated to Shanghai University of Traditional Chinese Medicine, 1500 clinical tongue images were collected according to the collection standards by using professional collection equipment. In order to improve the accuracy of tongue image classification, several TCM professionals assisted in marking and completing the consistency evaluation.

In this experiment, 400 cases of the normal tongues and 756 tooth-marked tongues were selected. The resolution is 2816 * 2112. First, the selected images were uniformly named, and then the tongue image was clipped using the trained tongue segmentation network model, as shown in Figure 2. One is to reduce the impact of other facial regions in the image on the follow-up work; the other is that the resolution of the original image data is 2816 * 2112, which increases the difficulty and duration of network training. The final image resolution is 563 * 397, which can be directly inputted into the neural network.

![Original image and processed image](image)

3.2 Tongue Image Annotation and Dataset Partition

The training of Mask Scoring R-CNN (MS R-CNN) needs not only image but also corresponding mask information. The Labelme open-source image annotation software based on Python is used to label the tongue area and make a mask to obtain the segmentation information of the teeth marks in the tongue image, as shown in Figure 3. To ensure that there are sufficient data for model training, and reserve a certain amount of data for verification and testing, to avoid introducing other deviation due to data partition, the common method of reference dataset partition is Hold-out. We using the 693 images train split for training, 231 validation split for validation, 231 test-dev split for test. Finally, the annotation information of the training set and verification set is integrated into JSON format file according to COCO dataset format.

![Labelme was used to mark the tongue features](image)

3.3 Construction of Tongue Feature Extraction Model

3.3.1 Construction of Feature Extraction Network Model

MS R-CNN is the first framework to solve the problem of case segmentation hypothesis scoring, which is simple, effective, and universal. Basis on Mask R-CNN Huang et al. [19][20][21] added a network block to understand the quality of prediction mask, which can detect targets and evaluate the quality of prediction more accurately. The network combines Faster R-CNN for target detection and FCN for semantic segmentation. After Faster R-CNN detects the target, FCN is used for mask prediction, border
regression, classification, and mask scoring. MS R-CNN adds a network block based on Mask R-CNN to learn the quality of the instance mask to predict. The mask scoring strategy corrects the deviation between the mask quality and the mask score, improves the performance of case segmentation, and provides a new direction for the improvement of case segmentation.

The proposed method based on Mask scoring R-CNN consists of four stages, as shown in Figure 4. The first stage, called feature extraction, which uses backbone architecture with ResNet-101+FPN, extracts features from different levels of the feature pyramid according to their scales using RGB images. The second stage, called regions of interest (RoIs) generation, proposes candidate RoIs by RPN. regions of interest (RoI) are generated and candidate ROIs are extracted through RPN. The third stage extracts features using RoIAlign from each candidate RoI and performs bounding box regression, classification with softmax and a predicted mask for by FCN. The fourth stage called MaskIoU head, aims to regress operation between the predicted mask and the true ground mask [22].

![Figure 4. Network architecture of tongue feature extraction based on Mask Scoring R-CNN](image)

### 3.3.2. The Application of Transfer Learning

In this paper, the transfer learning method is used to reduce training time and prevent over fitting of neural networks. Although minimizing the parameters of the neural network model can avoid over fitting, the data of the medical image itself is difficult to collect, and the amount of data that can be used as a gold standard is small, which cannot be compared with hundreds of thousands of open-source datasets.

Transfer learning [23][24][25][26] is to apply the model trained in one task to another task through adjustment. In this experiment, the ResNet backbone network which is pre-trained on the coco dataset is used to delete the last layer of the pre-training model. Because the pre-training model is trained for different classification tasks, remove the full connection layer and delete the weight and bias related to classification score, bounding box prediction, and mask prediction layer. Replace the removed layers with untrained layers, adjust the categories into three categories, and finally input the dataset and related mask into the MS R-CNN model architecture, as shown in Figure 5 below:
3.4. Evaluation Standard of Model Effect

To verify the reliability and effectiveness of the proposed algorithm, we performed experiment using multiple images from the test dataset. In this study, the precision (P), recall (R) and F1 score as the evaluation indexes are computed. The combination of neural network prediction and real classification is divided into true positive (TN), false negative (FN), true negative (TN), and false-positive (FP). For example, if the real category is the tooth-marked tongue, the prediction category is truly positive, and the predicted category is the non-tooth-marked tongue, it is false negative; if the real category is non-tooth-marked tongue, the prediction category is true negative and the opposite is a false positive.

\[
\text{Precision} = \frac{TP}{TP + FP} \tag{1}
\]

\[
\text{recall} = \frac{TP}{TP + FN} \tag{2}
\]

\[
F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}
\]
4. Experiment and Discussion

4.1. Training Platform
The model training is completed offline in the Ubuntu 64bit operating system, the CPU was Intel Core i7-9700F, the memory was 32G, and the GPU was GeForce GTX 2080Ti. This paper selects PyTorch as the framework, compared with Tensorflow, one of the advantages of PyTorch is that PyTorch is a dynamic graph, which can implement tensors and dynamic neural networks basis on powerful GPU acceleration. Tensorflow and so on are static graphs, which are not conducive to expansion.

4.2. Training Details
The backbone network used in the experiment is a characteristic pyramid network derived from residual neural networks (ResNet-50, ResNet-101) with 50 or 101 layers. Using the computational complexity of the ResNet-50 backbone network model is lower than that of the ResNet-101. Using ResNet-101 can significantly improve the results without changing the model or training.

This experiment uses the ResNet-101 network which was originally trained on the COCO dataset and removes the full connection layer because the pre-trained model recognizes 88 samples. The full connection layer removes the weights related to Class Score, B-box prediction, and Mask prediction layers, and then replaces them with untrained layers. Then the dataset and the corresponding mask are input into the MS R-CNN model architecture.

In the transfer learning stage, the weights of the pre-trained ResNet-101 backbone network are frozen to enable the model to extract common low-level features from the tongue image. Then the newly added full connection layer is trained and the weight is adjusted according to the distribution of the new dataset. In model training, the network is updated and fine-tuned.

4.3. Test Results and Analysis
232 tongue test pictures (including normal tongue and tooth-marked tongue of varying severity) were used to evaluate the model. TP = 101, FN = 15, FP = 1, TN = 115, the F1 score of the model was 0.95, the precision was 0.99, and the recall was 0.914. According to the prediction results, the model has good generalization, and has good recognition effect for different severity (as shown in Figure 6) (mild, moderate, severe), tongue type (as shown in Figure 7) (circular, triangular, elliptical, square, etc.), and different tongue extension directions (as shown in figure 8). Among them, a total of 15 cases of the normal tongue were mistakenly identification error as a tooth-marked tongue. Through joint analysis with TCM experts, it was concluded that the normal tongue identified by the model contains congenital and non-dentate indentation caused by tooth deformity. Therefore, In the next step, we will add more information about the non-tooth-marked depression.

From the perspective of visualization, compared with only for tongue classification algorithm, the model provides the accurate position of high resolution visualization and tooth mark size, combined with the confidence and the number of teeth marks, the researchers can better understand the network and it can provide the basis for the severity analysis of tooth marks, which is more suitable for mobile medicine and telemedicine.

Figure 6. Recognition of tongue image with different degrees of tooth marks
5. Conclusions
In this experiment, the latest research results in the field of image processing are applied to the feature extraction and classification of tongue image of teeth marks in Traditional Chinese Medicine. It can solve the problem that the current tongue image recognition methods do not provide local high-resolution visualization of the location and size of teeth marks, and the deep learning technology is difficult to be applied to small sample medical data. This method, firstly, according to the characteristics of tongue feature recognition task and deep learning, an appropriate neural network is constructed. Secondly, in order to improve the classification accuracy of the algorithm and the robustness of tongue image classification under various scenes, the transfer learning method is adopted. Finally, the model is tested and analyzed, and the accuracy of the classification, the exact location and the number of tooth marks are obtained.

In the future, the research will continue to deepen in the following three directions. First of all, improve the accuracy of the model. Next, the model will be improved to make the model more lightweight, more specific for tongue diagnosis, and improve the F1. Secondly, through the analysis and experiment of MS R-CNN, it is known that MS R-CNN has good performance in the treatment of TCM tongue classification. Tooth-marked tongue is only one kind of abnormal tongue in Traditional Chinese Medicine tongue diagnosis. The tongue image can be divided into more detailed classifications in TCM. Studying the problem of multi-classification can better aid tongue diagnosis and treatment. Finally, the experiment focuses on the extraction and classification of tongue features. However, from the perspective of traditional Chinese medicine, to analyze a person's health, we should collect all the information of the four diagnostic methods. Therefore, in the next work, we will establish a more comprehensive depth neural network combined with the previous observation of the face and interrogation studies. It is hoped that this work can further promote the development of medical health.

References
[1] Pang B, Wang K, Zhang D, et al. On automated tongue image segmentation in Chinese medicine[C]/Object recognition supported by user interaction for service robots. IEEE, 2002, 1: 616-619.
[2] Zhou Y, Shen L, Yang J. Feature analysis method of tongue image by Chinese medical diagnosis
based on image processing[J]. Infrared and Laser Engineering, 2002, 31(6): 490-494.

[3] Liang K, Huang X, Chen H, et al. Tongue diagnosis and treatment in traditional Chinese medicine for severe COVID-19: a case report [published online ahead of print, 2020 Jul 19]. Ann Palliat Med. 2020; apm-20-1330. doi:10.21037/apm-20-1330

[4] Zhang J, Hu G, Zhang X. Extraction of tongue feature related to TCM physique based on image processing[C]//2015 12th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP). IEEE, 2015: 251-255.

[5] Zhai X, Lu H, Zhang L. Application of image segmentation technique in tongue diagnosis[C]//2009 International Forum on Information Technology and Applications. IEEE, 2009, 2: 768-771.

[6] Jian Z, Qi Z, Bob Z, et al. TongueNet: A Precise and Fast Tongue Segmentation System Using U-Net with a Morphological Processing Layer[J]. MDPI, 2019, 9(15).

[7] Li, X, et al. "Automatic Tongue Image Segmentation For Real-Time Remote Diagnosis," 2019 IEEE International Conference on Bio-informatics and Bio-medicine (BIBM), San Diego, CA, USA, 2019, pp. 409-414.

[8] Qu P, Zhang H, Zhuo L, et al. Automatic Tongue Image Segmentation for Traditional Chinese Medicine Using Deep Neural Network. International Conference on Intelligent Computing, Liverpool, UK, August 7-10, 2017.

[9] Hsu P C, Wu H K, Huang Y C, et al. The tongue features associated with type 2 diabetes mellitus[J]. Medicine, 2019, 98(19).

[10] Wu J, Zhang Y, Bai J. Tongue area extraction in tongue diagnosis of traditional Chinese medicine[C]//2005 IEEE Engineering in Medicine and Biology 27th Annual Conference. IEEE, 2006: 4955-4957.

[11] Dai, Y. and Wang G, "Analyzing Tongue Images Using a Conceptual Alignment Deep Autoencoder," in IEEE Access, vol. 6, pp. 5962-5972, 2018, doi: 10.1109/ACCESS.2017.2788849.

[12] Li XQ, Zhang Y, Cui Q, et al. Tooth-Marked Tongue Recognition Using Multiple Instance Learning and CNN Features[J]. Pubmed,2018.

[13] Sun Y, Dai, S, Li, Jide, et al. Tooth-Marked Tongue Recognition Using Gradient-Weighted Class Activation Maps[J]. Future Internet, 2019, 11(2): 45.

[14] Kanawong Ratchadaporn, Obafemi-Ajayi Tayo, Liu Dahai, et al. Tongue Image Analysis and Its Mobile App Development for Health Diagnosis. [J]. Pubmed, 2017, 1005.

[15] Xiao M, Liu G, Xia Y, et al. A Deep Learning Approach for Tongue Diagnosis[C]//International Conference on Advanced Machine Learning Technologies and Applications. Springer, Cham, 2019: 3-12.

[16] Hu Y, Wen G, Liao H, et al. Automatic construction of chinese herbal prescriptions from tongue images using CNNs and auxiliary latent therapy topics[J]. IEEE transactions on cybernetics, 2019.

[17] Wu X, Sahoo D, Hoi S C H. Recent advances in deep learning for object detection[J]. Neurocomputing, 2020.

[18] Tania MH, Lwin K, Hossain MA. Advances in automated tongue diagnosis techniques.[J]. Integrative Medicine Research, 2019, Vol.8(1): 42-56.

[19] Huang Z, Huang L, Gong Y, et al. Mask scoring r-cnn[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 6409-6418.

[20] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.

[21] Johnson J.W. (2020) Automatic Nucleus Segmentation with Mask-RCNN. In: Arai K., Kapoor S. (eds) Advances in Computer Vision. CVC 2019. Advances in Intelligent Systems and Computing, vol 944. Springer, Cham.

[22] Tu S, Liu H, Li J, et al. Instance Segmentation Based on Mask Scoring R-CNN for Group-housed Pigs[C]//2020 International Conference on Computer Engineering and Application (ICCEA).
IEEE, 2020: 458-462.

[23] Pan S J, Tsang I W, Kwok J T, et al. Domain adaptation via transfer component analysis[J]. IEEE Transactions on Neural Networks, 2010, 22(2): 199-210.

[24] Yosinski J, Clune J, Bengio Y, et al. How transferable are features in deep neural networks[C]//Advances in neural information processing systems. 2014: 3320-3328.

[25] Liu T, Yang Q, Tao D. Understanding How Feature Structure Transfers in Transfer Learning[C]//IJCAI. 2017: 2365-2371.

[26] Donahue J, Jia Y, Vinyals O, et al. Decaf: A deep convolutional activation feature for generic visual recognition[C]//International conference on machine learning. 2014: 647-655.