A Proposal of an Internal Organs Classification Model Based on Tongue Images and Tongue Descriptions

Chen Wang*, Lei Shi*, Qi-Wei Ge**, Quan Gan*,†

Tongue diagnosis is an effective non-invasive Chinese medicine diagnostic and treatment technique. In traditional Chinese medicine (TCM), the diagnostic process of evaluating a patient's comprehensive physical condition relies on the expert opinion of visual inspection, such as tongue color, shape, texture, and color of tongue coating. At the same time, TCM believes that there is a close connection between the tongue and the internal organs and that the distribution of pathological changes of the internal organs on the surface of the tongue summarizes a certain pattern. Therefore, the observation of the tongue can be used to determine the disease status of the patient's internal organs. In the actual diagnosis, doctors usually need to combine the image information of the tongue with the corresponding brief description to diagnose Based on this process we take both the tongue image and the tongue description into consideration and propose a multimodal fusion diagnosis and treatment method based on machine learning. In this method, the original image is first segmented to extract the tongue region. We use the tongue description generation model to generate the corresponding tongue description. Finally, the multimodal viscera diagnosis model is used to combine the tongue image with the tongue description to reach the diagnosis of the patient's five visceral lesions.

1. Introduction

Traditional Chinese Medicine (TCM) is a medical science that originated in ancient China. It uses the four diagnostic methods of “inspection, listening and smelling, inquiry, and pulse palpation” to diagnose illnesses. Inspection includes observing the skin, nails, hair, and tongue and has the advantages of easy accessibility and non-invasiveness [1]. Tongue diagnosis is an important part of inspection diagnosis. TCM believes the tongue is directly or indirectly connected to internal organs, and changes in the tongue reflect internal organ health [2], providing key insights for diagnosis and prognosis [3]. After a long period of development from ancient times to the present, tongue diagnosis has developed into a well-established theoretical system [4].

TCM diagnosis and treatment are highly dependent on the doctor's professional knowledge and clinical experience and are therefore highly subjective. At the same time, its diagnostic results often lack objective indicators and uniform norms [5]. These factors have caused a certain degree of hindrance to the modern development of tongue diagnosis and even TCM [6]. In recent years, with the development of artificial intelligence technology, many scholars have begun to use artificial intelligence methods to research the objectification of TCM diagnosis [7]. Current research on the objectification of tongue diagnosis mainly focuses on the classification of tongue image characteristics, and usually only uses image information or text information. TCM diagnosis not only requires the description of tongue image characteristics but also requires the judgment of the patient's physical condition based on a variety of information. Therefore, to solve the problem that previous methods only analyze tongue image characteristics and ignore the comprehensive analysis of a variety of information, this work aims to achieve a more realistic classification of internal organs based on tongue image and tongue image description by combining multimodal information of text and image.

2. Related work

TCM explores the relationship between the tongue, internal organs, meridians, qi, blood, and body fluids, believing in a close connection between the tongue and internal organs [8].

* School of Computer Engineer, Jiangsu Ocean University, Jiangsu Province, China;

** Faculty of Education, Yamaguchi University, Yamaguchi, Japan, Email: gqw@yamaguchi-u.ac.jp;

† Corresponding Author, Email: ganquan@jou.edu.cn

This link is supported by both ancient texts and clinical experience, as lesions of the five viscera and six bowels are often reflected on the tongue through the meridians [2], therefore doctors to diagnose internal organ issues via tongue observation.

The current research on tongue objectivization mainly includes tongue segment, tongue coating separation, tongue feature extraction, and tongue-based constitution recognition. And the methods used for various tasks are different. Clinically collected tongue pictures usually contain non-tongue region such as the face, lips, and teeth. Therefore, the first step in a tongue image study is usually to segment the tongue region. With advancements in machine learning, particularly deep learning, image segmentation has seen significant development across industries, especially in the medical field. Tongue image segmentation has similarly adopted these techniques, achieving notable success. P. Qu et al. [9] used SegNet with luminance statistics to determine the need for segmentation. J. Zhou et al. [10] applied U-Net with transfer learning for tongue segmentation, achieving excellent results.

In tongue feature recognition, H. Weng et al. [11] used an improved YOLO model for teeth-print and crack tongue detection, C. Song et al. [12] used Inception_V3 and ResNet network combined with transfer learning for the detection of multiple tongue features and obtained up to 94.88% accuracy on a self-built dataset. In addition to the identification of tongue features, automated tongue diagnosis has gradually become a focus of research. Research has focused on constitution recognition and recognition for specific diseases. G. H. Wen et al. [13] integrated reshaped tensor and wavelet attention mechanisms into a convolutional neural network-ResNet18 for constitution identification and tongue attribute prediction, enhancing the method's interpretability. Y. Yuan et al. [14] developed a constitution recognition method using histogram of gradients and support vector machines for tongue segmentation, followed by k-means segmentation in Lab color space to separate the tongue body and coating, and used their features for classification.

Previous studies usually focus on using a single feature of information for medical diagnosis, However, the four diagnostic methods used in TCM require multiple feature information for comprehensive judgment. Q. Ye et al. [15] fused knowledge graph techniques and used a natural language processing model BiLSTM trained on a large amount of medical record data to achieve multi-label classification. However, they only used text information and did not use image information.

Clinical diagnosis requires a combination of objective and subjective factors, mimicking the diagnostic process of a doctor. Through the observation of tongue images, the doctor will generate his subjective judgment and record the corresponding tongue description. The final diagnosis is made by combining the observation of the images and the summarized tongue descriptions. Therefore, unlike previous methods, we will use a deep learning method to simultaneously process the tongue image information and the corresponding tongue description information to achieve viscera classification based on tongue images and tongue descriptions that are more in line with the actual diagnostic process.

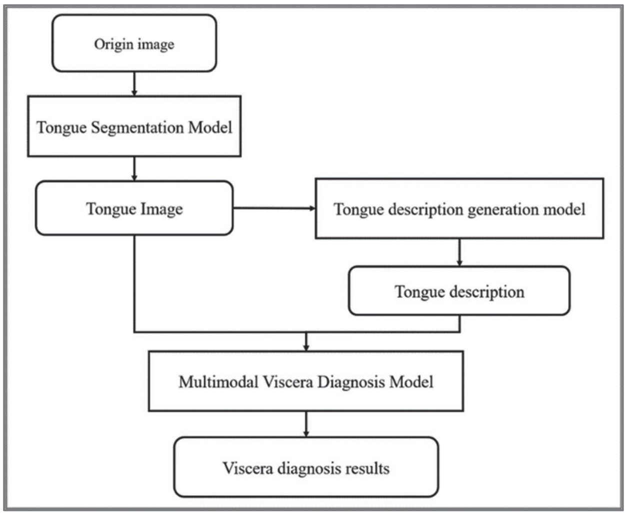

Origin image
Tongue Segmentation Model
Tongue Image
Tongue description generation model
Tongue description
Multimodal Viscera Diagnosis Model
Viscera diagnosis results

Figure 1 The overall research process of modal tongue diagnosis and viscera diagnosis

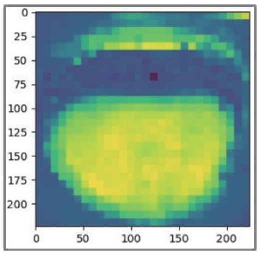


Figure 2 Heat map extracted from pre-experiment

3. Research Plan

To achieve multimodal tongue diagnosis of internal organs based on tongue images and tongue descriptions, the study will consist of the design of three models, namely, the design of the Tongue Segmentation Model, the Tongue Description Generation Model, and the design of Multimodal Viscera Diagnosis Model. The overall research flow of Multimodal Tongue Diagnosis and Viscera Diagnosis is shown in Figure 1.

3.1 Tongue Segmentation Model

As mentioned in Chapter 2, the original tongue image captured by any method often includes elements other than the tongue itself. These non-tongue elements significantly interfere with the neural network's ability to extract features. For instance, as shown in Figure 2, the attention feat ure map of the model from our preliminary experiments reveals that part of the model's attention is drawn to the teeth and lips. Therefore, segmenting the tongue image can effectively filter out irrelevant factors, enhancing the efficiency and accuracy of subsequent processes [16].

There are many basic models available for tongue image segmentation, and we choose the UNet model as the basic model. UNet is a deep convolutional neural network that is widely used in medical image segmentation tasks [12] and has excellent small-sample feature extraction capability. The model can extract global features during the downsampling process while implementing a residual- like approach to preserve local features of the image, and then use an upsampling process to obtain high-resolution image segmentation results. We will use several evaluation metrics to measure the segmentation performance of the model, including accuracy, Dice coefficient, and IoU (Intersection over Union).

We have collected more than one thousand tongue images from sources such as literature, books, partner hospitals, and publicly available datasets and manually segmented them using the annotation tool. The segmentation process is shown in Chapter 5 and thus has the database to conduct experiments.

3.2 Tongue description generation model

Although we ultimately aim to realize a multimodal tongue diagnosis of viscera with tongue images and tongue descriptions, we hope that the model can generate the required textual modalities solely based on image inputs. This is because in reality the tongue description is also given by the doctor's judgment and not information that is given in advance. Therefore, the second step of the research program is to design a Tongue description generation model for the automatic generation of tongue description text.

The field of machine learning already has the task of image caption generation, but it may not be suitable for our task. Tongue description involves detailed medical information, which requires an accurate description of the tongue's color, shape, texture, etc., and the common Image Caption task is usually used to generate common-sense, scenario-based descriptions, which is difficult to satisfy the needs of medical applications. At the same time, the medical field requires models to provide interpretable results as much as possible, but the Image Caption method may give unexpected descriptions. In terms of implementation difficulty, it is also difficult to generate descriptions for medical images. Existing Image Caption models need to be trained on large-scale datasets, and it may be difficult to achieve the expected results in the medical domain with less labeled data. Therefore, we simplify this task to a tongue feature classification task.

Based on the Chinese national standard [17] and our categorization of the tongue descriptions from the collected data, the tongue feature classification task was divided into three parallel parts: (1) Tongue color classification, this step will extract the tongue color features of the tongue, and the classification model will classify the tongue into three categories: pale red, reddish, purple black. (2) Tongue body classification, this step will extract the tongue body features and classify the tongue into four categories: big fat, thin, cracked, and tooth prints. (3) Tongue coating classification, this step will extract the tongue coating and tongue texture features and classify the tongue into six categories: thin white, thick white greasy, yellow greasy, few mosses, moist, and dry. The classification results are shown in Table 1. Finally, the classification results of the three models are combined to obtain the final tongue description. The processing flow is shown in the Figure 3.

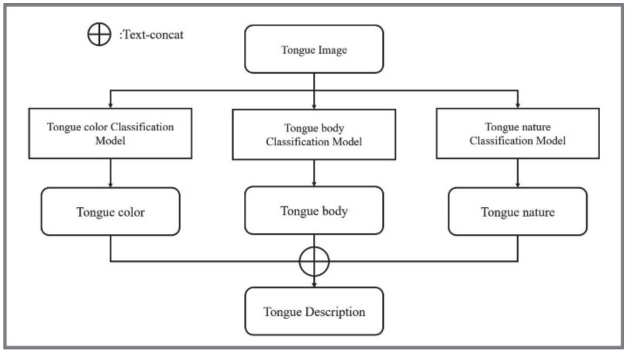

:Text-concat
Tongue Image
Tongue color Classification Model
Tongue body Classification Model
Tongue nature Classification Model
Tongue color
Tongue body
Tongue nature
Tongue Description

Figure 3 Tongue description generation model

Table 1 Tongue image attribute classification results

<table>
<tr><td colspan="12">Tongue color</td></tr>
<tr><td colspan="4">pale red</td><td colspan="4">reddish</td><td colspan="4">purple black</td></tr>
<tr><td colspan="12">Tongue body</td></tr>
<tr><td colspan="3">big fat</td><td colspan="3">thin</td><td colspan="3">cracked</td><td colspan="3">tooth prints</td></tr>
<tr><td colspan="12">Tongue nature</td></tr>
<tr><td colspan="2">thin white</td><td colspan="2">thick white greasy</td><td colspan="2">yellow greasy</td><td colspan="2">few coating</td><td colspan="2">moist</td><td colspan="2">dry</td></tr>
</table>

3.3 Multimodal Viscera Diagnosis Model

After obtaining the segmented tongue images and the generated tongue descriptions based on the tongue images, the third step of the research program is to construct a multimodal viscera diagnostic model that combines the tongue images with the generated descriptions to accurately identify the location of the patient's disease.

We have currently established a simple deep-learning model to implement the concept. We built our multimodal visceral diagnosis model framework using a series of Transformer-based variant models. The Transformer model is

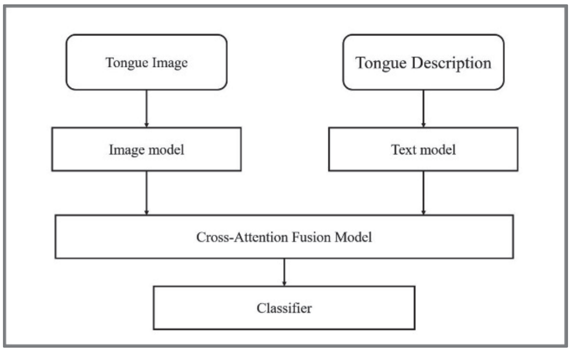


Figure 4 Multimodal Viscera Diagnosis Model

one of the core technologies in the field of deep learning in recent years, and they have demonstrated their power in several tasks such as natural language processing, computer vision, and multimodal fusion. The self-attention mechanism it uses can capture contextual dependencies in the input sequence, and its unified encoder-decoder architecture makes it capable of handling a wide range of tasks from text generation, and image classification, to visual question answering, etc. Meanwhile, a variant of the self-attention mechanism, cross-attention, can realize multimodal fusion by changing the weight matrix of Query and Key, Value. The overall framework of the model is shown in Figure 4, and the related experiments are described in Chapter 4.

First, we extracted features from the tongue image description text using a pre-trained text model, mapping them to a semantic representation suitable for multimodal fusion. A similar approach was applied to the tongue image, extracting its features using a pre-trained image processing model. In the fusion layer, we employed a cross-attention mechanism, allowing the image and text features to interact. The attention mechanism, which can be understood as 'selective attention,' enables the neural network to focus on the most relevant parts of the input data. This improves the model's performance. Cross-attention further enhances the model by enabling it to focus on content from multiple modalities simultaneously.

Through this multimodal fusion, our model can extract visual features from tongue images while combining the information about tongue features embedded in text descriptions to make more accurate diagnostic judgments. This process not only enhances the accuracy of diagnosis but also provides new insights for the objectification of tongue diagnosis.

4. Experiment

4.1 Dataset

We collected 300 samples from books, and collaborating hospitals containing pictures and descriptions of the tongue as well as classification labels for four categories, namely “Liver and Gallbladder”, “Spleen and Stomach”, “Heart and Lungs”, “Kidney”. The labels are provided by the original samples.

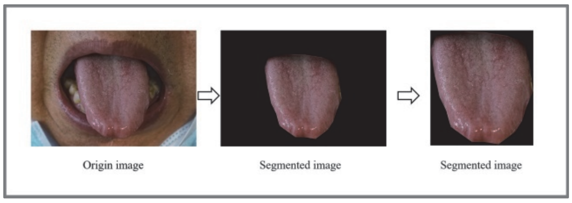


Figure 5 Tongue Image Processing Workflow

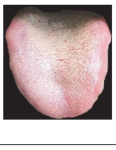
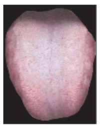
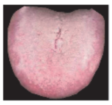
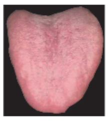
Organ	spleen and stomach	heart and lungs	kidney	liver and gallbladder
Tongue image				
Tongue description	Pale purple tongue, yellowish and greasy tongue coating	Dark purple tongue, ecchymosis and thin coating	Pale tongue, fat tongue body, white and greasy coating	Red tongue , little or no coating

Figure 6 Multimodal Data Example

For the tongue description information, due to the inconsistency of the data source, we standardized the format to better extract features as follows: “tongue texture, tongue coating, other features”, such as: “pale and swollen tongue, white and greasy coating, teeth print on the edges”.

For the tongue image, to ensure the model focuses on the effective region of the tongue, we use the LabelMe tool for tongue segmentation so that only the tongue region is retained and since the tongue region occupies a relatively small portion of some of the images, we crop the image size so that the tongue region fills the entire image. The process of tongue image processing is shown in Figure 5. In addition, Figure 6 shows the multimodal data containing tongue images and tongue descriptions used for the experiment. The number of individual labels in the dataset is shown in Table 2.

Table 2 Dataset Label Destribution

Organ	Number
Spleen and stomach	94
Heart and lungs	82
Kidney	64
Liver and gallbladder	60

4.2 Experiment settings

The text feature extraction and multimodal fusion modules are implemented using the BERT model, a bidirectional Transformer-based pre-trained language model that learns contextual representations of words from both the left and right directions, unlike the standard Transformer model. For image feature extraction, we use the ViT model, an image processing model also based on the Transformer architecture. ViT divides the image into multiple patches, transforming a long sequence of images into several sequential segments, which can then be processed by the Transformer model.

4.3 Experiment environment

We use the AdamW optimizer with weight decay set to 1e-2, batch size set to 64, we use bert-base and ViT pre-training parameters to initialize the text and image encoders in the model. Experiments were conducted on an NVIDIA GeForce RTX 3090 GPU for 200 epochs of training.

4.4 Result

We use accuracy as the criterion, which indicates the proportion of samples correctly predicted by the model to all samples. It is defined definition formula as Equation (1).

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \quad (1)$$

The results of the preliminary experiment are shown in Table 3, when the visual model was Vit-small, the accuracy was 57.87% on the self-built dataset. When we replaced the image model with ViT-base with more parameters, the accuracy dropped to 53.03%. This may be due to the overfitting caused by the strong feature extraction ability of the larger ViT model. When the image model was replaced with Resnet-152, the accuracy also dropped to 53.03%, which may be due to the fact that the extraction ability of the Resnet model is not as good as that of the transformer model.

Table 3 Comparison of various model experiments

Model	ViT_small+bert	ViT_base+bert	ResNet152+bert
Accuracy	57.87%	53.03%	53.03%

5. Conclusion

In this paper, we propose an automated tongue diagnosis method that integrates tongue segmentation, tongue feature extraction, and viscera classification. Unlike previous methods that rely solely on a single modality, our approach combines both image and text modalities for a more comprehensive and realistic diagnosis process. Preliminary experiments conducted on a self-built dataset have demonstrated promising results, highlighting the potential of our multimodal viscera classification method.

Acknowledgement

This work was supported by JSPS KAKENHI (20H04284), Project Haiyan-2023, the Research Start-up Fund of JOU under Grant KQ23032, Jiangsu Province College Students' Innovative Entrepreneurial Training Plan Program (SY202411641639007).

Reference

[1]. C. G. Li, D. Zhang, and S. X. Chen, "Research about tongue image of traditional Chinese medicine (TCM)based on artificial intelligence technology", 5th Information Technology and Mechatronics Engineering Conference, pp. 638-641. (2020)

[2]. A. C. Guo., Huangdi Neijing, China Traditional Chinese Medicine Publishing House. (2019) (in Chinese)

[3]. D. S. Liu, X. J. Han. "Theoretical Origin and Clinical Application of Traditional Chinese Medicine Inspection", Journal of Emergency in Traditional Chinese Medicine, vol.22, no.08, pp. 1345-1347(2013) (in Chinese)

[4]. D. Zhang, W. T. Pang, K. Y. Wang, et al. "The Present and Future of the Research on the Objectification of TCM Tongue Diagnosis Based on the Microscopic Angle", World Science and Technology-Modernization of Traditional Chinese Medicine, vol.24, no.11, pp. 4574-4579(2022) (in Chinese)

[5]. S. Q. Zhang, Y. H. Sun, N. X. Xian, et al. "Research progress on objectification and intelligence of the four diagnosis methods of traditional Chinese medicine", Guiding Journal of Traditional Chinese Medicine and Pharmacy, vol. 29, no.6, pp. 170-174(2023) (in Chinese)

[6]. L. Q. Zhang, M. H. Li, S. S. GAO, et al. "Summary of Computer-assisted Tongue Diagnosis Solutions for Key Problems", Computer Science, vol. 48, no.07, pp. 256-269 (2021) (in Chinese)

[7]. H. Y. Li, C. Li, X. F. Lang, et al. "Analysis of the Research Status and Hot Spot of Intelligent Four－Diagnosis in TCM", Nanjing Univ Tradit Chin Med Vol．38 No．2, pp.180-186, 2022 (in Chinese)

[8]. Wu Kunan (Qing Dynasty), annotated by Shao Xiangen (Qing Dynasty). Shanghan Zhizhang, Shanghai Science and Technology Press. (1959)

[9]. P. Qu, H. Zhang, L. Zhuo, et al. "Automatic tongue image segmentation for traditional chinese medicine using deep neural network", Proceeding of 9th International Conference on Intelligent Computing, pp. 247–259 (2017)

[10]. J. Zhou, Q. Zhang, B. Zhang, X. Chen. "TongueNet: A Precise and Fast Tongue Segmentation System Using U-Net with a Morphological Processing Layer", Applied Sciences, Vol. 9, no.15, pp. 3128 (2019)

[11]. H. Weng，L. LI，H. W. LEI, et al. "A weakly supervised tooth－mark and crack detection method in tongue image", Concurr Comput Pract Exp, vol. 33, no.16, pp. e6262 (2021)

[12]. C. Song, B. Wang, J.T. Xu. "A tongue image classification method based on deep transfer learning", Computer Engineering & Science, Vol. 43, no.08, pp. 1488-1496 (2021).

[13]. G. H. Wen, H. Z. Liang, H. H. Li, et al. "Interpretable Tongue Constitution Recognition via Reshaped Wavelet Attention". International Journal of Computational

Intelligence Systems, Vol.17, no. 31 (2024)

[14]. Y. Yuan and W. Liao, “Design and Implementation of the Traditional Chinese Medicine Constitution System Based on the Diagnosis of Tongue and Consultation”, IEEE Access, vol. 9, pp. 4266-4278 (2021)

[15]. Q. Ye, S. H. Zhang, C. L. Cheng, et al. “Multi-channel Chinese Medicine Syndrome Differentiation Model Integrating Knowledge Graph”, Science Technology and Engineering, Vol. 22, no. 21, pp. 9190- 9198 (in Chinese)

[16]. Q. Xu, Y. Zeng, W. Tang, et al. “Multi-task joint learning model for segmenting and classifying tongue images using a deep neural network”. IEEE J. Biomed. Health Inform. Vol. 24, no. 9, pp. 2481–2489 (2020)

[17]. China Association of Chinese Medicine. Classification and Judgment of TCM Constitution. China Association of Chinese Medicine, Beijing, China (2009)