Tooth-Marked Tongue Recognition Using Multiple Instance Learning and CNN Features

Xiaoqiang Li, *Member, IEEE*, Yin Zhang, Qing Cui, Xiaoming Yi, and Yi Zhang

***Abstract*—Tooth-marked tongue or crenated tongue can provide valuable diagnostic information for traditional Chinese Medicine doctors. However, tooth-marked tongue recognition is challenging. The characteristics of different tongues are multiform and have a great amount of variations, such as different colors, different shapes, and different types of teeth marks. The regions of teeth mark only appear along the lateral borders. Most existing methods make use of concave regions information to classify the tooth-marked tongue which leads to inconstant performance when the region of teeth mark is not concave. In this paper, we try to solve these problems by proposing a three-stage approach which first makes use of concavity information to propose the suspected regions, then use a convolutional neural network to extract deep features and at last use a multiple-instance classifier to make the final decision. Experimental results demonstrate the effectiveness of the proposed method.**

***Index Terms*—Convolutional neural network (CNN), multiple-instance learning, tooth-marked tongue.**

I. Introduction

TONGUE diagnosis is one of the most important diagnostic methods of traditional Chinese Medicine (TCM). According to TCM, the tongue is closely related to people's health. Different features (such as colors, shapes, etc.) of the tongue can reflect the internal state of the body and the health of the organs. Thus, tongue diagnosis is widely applied to clinical analyses and applications by TCM for thousands of years [1]. However, clinical effectiveness of the diagnosis heavily depends on the TCM practitioner's experience. So, it is important to build an objective and quantitative diagnostic standard for tongue diagnosis. Tooth-marked tongue, as one important symptom, can provide variety of valuable information for the TCM doctors [1]. Tooth-marked tongue is one appearance of the tongue when there are teeth marks along the lateral borders [2], [3]. Based on our observation and Shao's [1], tooth-marked regions are usually some indentations of which the color tends to be darker surrounded by some white area. The detailed characteristics will be described in Section II-A. The recognition of tooth-marked tongues can be viewed as a fine-grained classification problem since tooth-marked tongue is a symptom of the tongue. But it is more challenging than distinguishing between subcategories, such as different species of dogs or vehicles due to the large number of variances of tongue characters and scare amount of available data. Moreover, the tooth-marked tongue data is coarsely labeled. A tongue image is labeled as a tooth-marked tongue or a nontooth-marked tongue, no further information (such as the location of the teeth marks region) is available. In this paper, we try to solve these problems by multiple-instance learning and deep learning.

Multiple-instance learning is first introduced in the middle of the 1990s by Dietterich *et al.* [4]. It was successfully applied in classification of molecules in drug design. In multiple-instance learning, a classifier classifies bags instead of feature vectors, where each bag contains multiple feature vectors (also called instance). The instances are implicitly labeled since class labels are associated with bags instead of instances. A bag is considered positive if at least one instance of this bag is positive; a bag is considered negative if all instances of this bag are negative. The recognition of tooth-marked tongues is naturally a multiple-instance problem since it shares very similar assumption with the multiple-instance binary classification that a tongue is considered as a tooth-marked tongue if there is at least one tooth-marked region on the tongue. In this paper, we present a multiple-instance representation of the tongue in which a tooth-marked tongue is treated as a positive bag containing both tooth-marked regions (positive instances) and healthy regions (negative instances) and a healthy tongue is treated as a negative bag containing only healthy regions. Unlabeled regions are called as suspected tooth-marked regions. Then we successfully applied a multiple-instance support vector machines (SVM) to classify the tooth-marked tongues.

Since the successful usage in 2012 imagenet competition, the convolutional neural network (CNN or ConvNet) has significantly improved the performance of many computer vision tasks, such as image classification [5], object detection [6], semantic segmentation [7] and so on. In object detection, many works like OverFeat [8] and R-CNN [6] show that CNNs can be powerful feature extractors. [9] shows that even without fine-tuning, features extracted from ConvNet can perform well in many visual tasks. R-CNN also proposed a fabulous

Manuscript received May 5, 2017; revised September 14, 2017 and October 23, 2017; accepted November 2, 2017. This work was supported by the Shanghai Innovation Action Plan Project under Grant 16511101200. This paper was recommended by Associate Editor S. Ventura. *(Corresponding author: Xiaoqiang Li.)*

X. Li is with the School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China, and also with the Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China (e-mail: xqli@shu.edu.cn).

Y. Zhang, Q. Cui, X. Yi, and Y. Zhang are with the School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2017.2772289

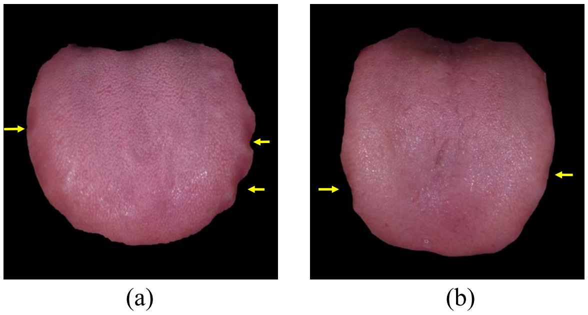


Fig. 1. Two examples of tooth-marked tongue. The tooth-marked regions on (a) are very obvious while the tooth-marked regions on (b) are difficult to identify.

workflow for object detection. R-CNN first uses selective search to generate region proposals. Then it uses ConvNet to extract a feature vector for each region. At last, it uses an SVM to classify regions. The ConvNet used by R-CNN is pretrained on a large auxiliary dataset (ILSVRC), and then fine-tuned on a small dataset (PASCAL). In this paper, we learn from experience of R-CNN and use a fine-tuned ConvNet to extract feature vectors of the tooth-marked regions (instances of the bags).

The computerized tongue diagnosis is not so popular as the other mainstream visual tasks like object detection due to the fact that tongue diagnosis is not widely accepted by the modern western medicine and there is no clear connection between tongue appearances and health status. However, some researchers have been contributing to computerized tongue diagnosis over the last few decades. In 2000, Chiu [10] proposed a computerized tongue examination system to quantize the tongue properties. In 2004, Pang *et al.* [11] utilized chromatic and textural features and Bayesian networks to explore the connection between tongue abnormal appearances and diseases. Many works also propose some novel techniques for tongue segmentation [12]–[14], tongue image color analysis [15], and tongue shape analysis [16], etc.

Many researchers use handcrafted features based on colors and concavity of the tooth-marked region to classify the tooth-marked tongue. Wang *et al.* [17] proposed a method which uses the information of the convex defects of the tongue. First, they do some image processing to separate the tongue body from the background. Then they calculate information like slope and length of the convex defects of the tongue body and use threshold value to identity the teeth marks on the tongue. The threshold values are statistically selected from their experiment with 300 tongue images. Shao *et al.* [1] defined some features of tooth-marked tongues which concentrated on features of convex, the change of brightness. They also classify the tooth-marked tongues by thresholding feature values. As described above and shown in Fig. 1, there will be indentations in the teeth marked regions. Intuitively, we want to make use of the information which provided by these indentations as what most methods do. However, the indentations is not very clear when teeth marks are not very serious, as shown in Fig. 1. And threshold values obtained from experiments make the algorithms easily suffer from overfitting.

Our method makes use of indentation information to select suspected tooth-marked region (like region proposals in R-CNN) instead of direct identification and use ConvNet to extract a feature vector for every tooth-marked region. Then we apply the idea of multiple-instance learning to our task that we group feature vectors from each tongue into bag and train a multiple-instance SVM to do the identification.

This paper is organized as follows. In Section II, we will describe our dataset and summarize our observation of the dataset in detail. Based on our observation of the tooth-marked tongues, Section III will present the specific steps of the proposed method. And the experiment results of our methods will be presented in Section IV. Finally, we will make a conclusion and discuss the future work in Section V.

II. Dataset and the Observation

In this section, we will describe the dataset used in this paper in detail and summarize the main characteristics of the tooth-marked tongue. We will also state the motivation of the proposed method based on our observation.

A. Dataset

As stated in [15], the dataset should include as many tongue images as possible and the acquired images should be of high and consistent quality. The tongue image dataset used in this paper is provided and acquired by Shanghai Daosh Medical Technology Company, Ltd. It is a small dataset which contains 641 tongue images (both healthy tongues and tooth-marked tongues) in total acquired in three different time. 297 tongue images are tooth-marked tongue images and the other 344 images are nontooth-marked tongue images. Though image acquisition device is consistent and very similar to [15], illumination conditions may vary in three different time. The image size is 4896 × 3672 pixels. The tongue images are labeled by TCM practitioners. However, whether a tongue is a tooth-marked tongue is not absolute, the boundary can be very blur when the symptoms is not serious which is the reason why this is a challenging task. So the label of a tongue image is voted by multiple TCM practitioners. The bounding boxes of tooth-marked regions are also provided represented by 4-element integer vectors (the top-left coordination, the width and the height). The tooth-marked region is about 180–300 pixels wide and 240–400 pixels high. We are also provided with the segmentation version of tongue images which manually separate the tongues from the background for better analysis.

B. Observation

As described above, a tooth-marked tongue is a tongue with indentations on its edge due to the compression of the teeth. By observation of our dataset, we summed up the following characteristics of the tooth-marked tongue: the tooth-marked region normally appears along the lateral borders of the tongue, as mentioned in other paper [1], [17], but we also find that, though very rare, it can appear in the apex (top) of the tongue. Due to the compression of the teeth, most tooth-marked regions are concave. However, an image is usually 2-D, it only records the flatten version of the tongue which loses some

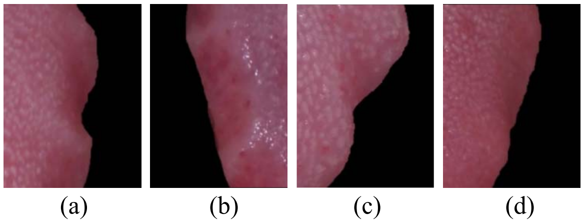


Fig. 2. Four types of tooth-marked regions. (a) Region which is dark red and concave. (b) Region which is dark red but not concave. (c) Region which is concave but not dark red. (d) Region is neither concave nor dark red.

spatial information. Thus, some tooth-marked regions are not concave. Instead, each of them appears to be a dark red scalloped region surrounded by a white ring, as shown in Fig. 1. At the same time, long-period compression of teeth makes the texture of the tooth-marked region smooth.

We organize tooth-marked regions into four categories, as is shown in Fig. 2(a), the region which is dark red and concave. Fig. 2(b) the region which is dark red but not concave. Fig. 2(c) the region which is concave but not dark red. Fig. 2(d) the region is nether concave nor dark red.

Some key insights we grained from the above analysis are as follows.

1) Most teeth marks appear along the lateral borders, and each tooth-marked region can be separated from the tongue, the rest of the tongue can be ignored. Thus, a whole tongue can be represented by multiple tooth-marked regions, which inspires us with the use of multiple-instance learning. A feature vector extracted from the tooth-marked region is the instance and a tongue is a bag consists of multiple feature vectors (instances).
2) Though some tooth-marked regions may not be concave, the color and brightness of the tooth-marked regions differ from the normal region. These differences can be utilized to make every tooth-marked region concave, which will be described in Section III-A in detail. By making every tooth-marked region concave, we are able to locate every suspected tooth-marked region based on the concavity.
3) From the observation, we also find that it is hard to describe the tooth-marked region. Many algorithms [1], [17] utilize concavity information. Nevertheless, as we stated above, not all tooth-marked regions are concave. More robust features are needed to describe the tooth-marked region which leads us to seek help from the CNN. A CNN can learn directly from the images and thus provides robust features [6], [8], [9].

III. Proposed Method

In this section, we will describe proposed algorithm in detail. The proposed method regards a tongue as a bag in multiple-instance learning and tooth-marked regions as instances.

The proposed method contains three stages which is very similar to R-CNN. First, the method generates all suspected tooth-marked regions. Then, we use a ConvNet to extract a fixed-length feature vector for each region. At last, we group feature vectors to bags and use a multiple-instance SVM to classify the tongue.

A. Generating Suspected Tooth-Marked Regions

The suspected tooth-marked regions generated from a tooth-marked tongue can include both real tooth-marked regions and healthy nontooth-marked regions, but at least one tooth-marked region should be included, just like the assumption of multiple-instance binary classification that a positive bag should contains at least one positive instance. The suspected tooth-marked regions generated from a healthy nontooth-marked tongue should include only healthy regions. We generate a bounding box for every suspected tooth-marked region.

To generate the suspected tooth-marked region, the steps are as follows.

Step 1: Apply a simple color reduction method ($I_{\text{new}} = \lfloor I_{\text{old}}/\text{div} \rfloor \times \text{div}$, where I is the RGB value, div is a hyperparameter. we use div $= 64$ in this paper) to a tongue image so that the dark red can be separated from the other colors. Then we convert the image from RGB color space to gray to further reduce the number of colors. The pixel value of the original dark red regions will be very close to 0 after this step.

Step 2: Choose the second smallest value of the gray image as the threshold value to binarize the image so that original dark red becomes black and there will be indentations along the edge of the tongue.

Step 3: Calculate the convex hull of the tongue using the binary image generated in step 2 and find convexity defects of it.

Step 4: Calculate bounding boxes for every convexity defects. The center point of the bounding box is the center point of the start point, the end point and the farthest point of the convexity defect

$$\begin{bmatrix} x_{\text{center}} \\ y_{\text{center}} \end{bmatrix} = \frac{1}{3}\left(\begin{bmatrix} x_s \\ y_s \end{bmatrix} + \begin{bmatrix} x_e \\ y_e \end{bmatrix} + \begin{bmatrix} x_f \\ y_f \end{bmatrix}\right). \quad (1)$$

The height and width of the bounding box are calculated as follows:

$$\text{height} = 1.5 \times |y_s - y_e| \quad (2)$$

$$\text{width} = 0.8 \times \text{height}. \quad (3)$$

Step 5: Get regions of interest (ROIs) of the original tongue image using the bounding box calculated in step 4 which then serves as suspected tooth-marked regions.

The first four steps are illustrated in Fig. 4 and some examples are showed in Fig. 5 The goal of this stage is to find as many suspected tooth-marked regions as possible so that at least one definite tooth-marked region is included. We also try to simply generate bounding boxes along the edge of the tongue with a fixed interval, but the computation will be much larger when we extract feature vectors of these regions. Most function implementations in this stage can be found in the OpenCV toolbox.

B. Feature Extraction

In this stage, we use a CNN to extract feature vectors of the tooth-marked regions. As stated in Section II, robust features are needed to describe the tooth-marked region which can combine color, shape and texture information instead of just concavity information. In this paper, we use a high-capacity CNN to extract a fixed-length feature vector of the tooth-marked region. The CNN aims at extracting feature vectors of the tooth-marked regions (the ROIs) instead of the whole tongue images.

1) Architecture: We use the pretrained VGG-16 in this paper. The detailed architecture is well described in [18]. It has 16 weight layers 13 of which are convolutional layers and the rest 3 of which are fully connected layers. There are 4096 units in the second fully connected layer and the outputs of this layer is used as features. Thus, we can extract a 4096-D feature vector for each region. We drop the last 1000-way fully connected layer and replace it with a 2-way fully connected layer (whether it is a tooth-marked region or not) during the network training. We also try to replace the softmax function with the logistic function, however, the difference is relatively small.

2) Network Training: The network is first pretrained on ILSVRC2012 dataset and then followed by fine-tuning on tooth-marked region images. All region images are generated from the tongue images using the bounding boxes either provided by TCM practitioners or generated using the algorithm described in Section III-A. The regions generated using the bounding boxes provided by TCM practitioners are considered tooth-marked regions (the positive instances) and regions generated from the healthy tongue using the algorithm described in Section III-A are considered nontooth-marked regions (the negative instances). These region images are only used for fine-tuning the network. There are around 4000 tooth-marked region images in total which are not enough to train such a high-capacity network. The network fails to converge if it is not pretrained. The pretrained weights are obtained from the Keras deep learning models [19]. The tooth-marked region image is resized to 256 × 256, and a fixed-size 227 × 227 subimage is randomly cropped and horizontally flipped from it using the same strategy described in [5]. We use stochastic gradient descent to fine-tune our network with a batch size of 128, learning rate of 0.0001, momentum of 0.9, and weight decay of 0.0005. We stop our training after 20 epochs since the accuracy stops increasing.

3) Testing: At test time, the network serves as a fixed feature extractor. We drop the last classification layer and only use the output of the 15th layer as the features. The center crop of the suspected tooth-marked region generated using proposed algorithm described in Section III-A is used as a input. Thus, we can extract a 4096-D feature vector for every suspected tooth-marked region.

It should be noted that prior knowledge like the bounding boxes provided by the TCM practitioners and whether a region is a tooth-marked region or a healthy region is only needed the first time we fine-tune the network. After the network is fine-tuned, it works as a fixed feature extractor like we described in Section III-B *Testing*, no prior knowledge is needed. What is more, the consistency of tongue image acquisition devices and the fine generality of the deep ConvNet features according to [6], [8], and [9] mean that there is no need to fine-tune the ConvNet every time new tongue images are acquired.

The input of the network is a region image instead of a tongue image. Thus, the network can produce a prediction for each suspected region image (whether it is a tooth-marked region or not). Though the classification layer of the network is dropped and only the outputs of the intermediate layer is used in our proposed method, the prediction of each tooth-marked region by the network can be utilized. We can actually classify the tongue directly by aggregating the predictions of the suspected regions. The comparisons and the analysis are shown in Section IV.

C. Classification

In this stage, we train a multiple-instance SVM to classify the tongue images.

The multiple-instance SVM used in this paper was proposed by Andrews *et al.* [20]. They introduce two forms of multiple-instance SVM in their work, mi-SVM and MI-SVM. In this paper, we describe the basic idea of these two multiple-instance SVMs, the detail is well introduced in [20]. The input of a multiple-instance SVM is a bag B_I which in our case represents the Ith tongue. And the instances in the bag are the feature vectors $\{x_i : i \in I\}$ we extracted from the tooth-marked regions of the tongue. Instead of explicitly associating a label y_i with each instance, a label Y_I is associated with a bag B_I. If $Y_I = -1$, then $y_i = -1$ for all $i \in I$; if $Y_I = +1$, then at least one instance $x_i \in B_I$ is a positive instance.

MI-SVM: In MI-SVM, the function margin of a bag is defined as follows:

$$\gamma_I = Y_I \max_{i\in I} (< w, x_i > +b). \tag{4}$$

The MI-SVM aims at maximizing the bag margin. The MI-SVM is defined as follows:

$$\min_{(w,x,\xi)} \frac{1}{2}\|w\|^2 + C\sum_I \xi_I$$
$$\text{s.t.} \quad \forall I : Y_I \max(<w, x> + b> \geq 1 - \xi_I, \xi_I \geq 0. \tag{5}$$

As can be see from the definition, instead of taking every instance into account, MI-SVM only looks at parts of the instances. For a positive bag, the margin is defined by the most positive instance, while the margin of a negative bag is defined by the least negative instance of the bag.

MI-SVM is considered as an instance-based multiple-instance method [21], since the classifier is trained at instance level. Andrews *et al.* [20] also proposed a heuristic approach to optimize the formulation (5) of which the essential is that it iteratively trains a standard SVM and in each iteration, SVM trained in last iteration is used to choose the positive instance in the positive bag in the current iteration. We use the implementation of MI-SVM by Doran and Ray [22] in this paper.

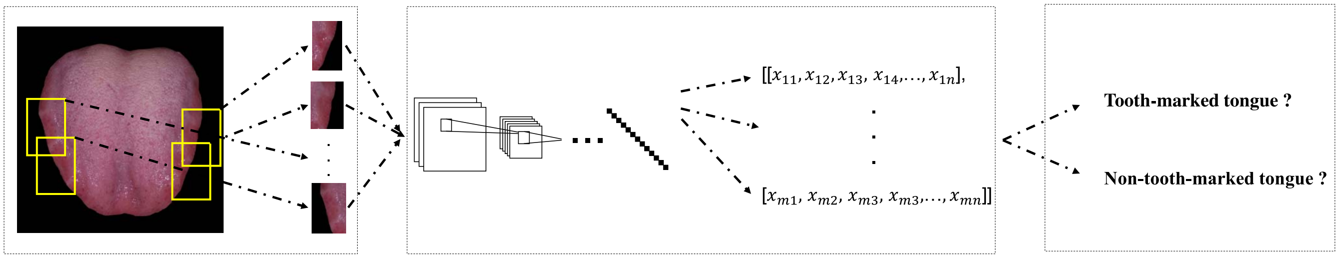

$[[x_{11}, x_{12}, x_{13}, x_{14},\ldots, x_{1n}],$
$[x_{m1}, x_{m2}, x_{m3}, x_{m3},\ldots, x_{mn}]]$
Tooth-marked tongue ?
Non-tooth-marked tongue ?

Fig. 3. Framework of the proposed method. The proposed method consists of three stages. The first stage generates the suspected regions of a tongue. These suspected regions are fed to a ConvNet in the second stage. In the last stage, the outputs of the ConvNet are treated as a bag (a 2-D matrix) and a multiple-instance classifier are used to make the final decision.

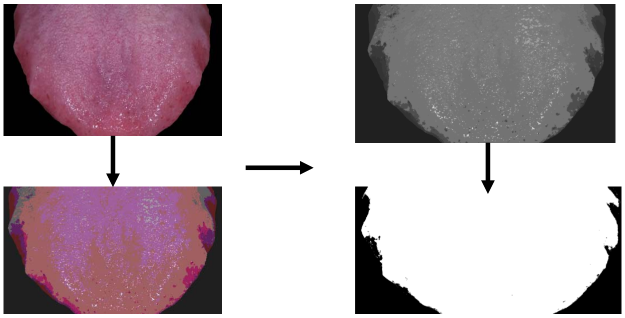


Fig. 4. First four steps of generating the suspected regions. As can be seen from the picture, the tooth-marked regions of the tongue become concave after the first four steps. As is described in Section II, not all tooth-marked regions are concave, we hope all tooth-marked regions or at least most of the tooth-marked regions are concave after these four steps.

We do not use mi-SVM since it takes every instance into account and is computationally intensive. However, mi-SVM and any other multiple-instance methods like diverse density [23] is effective to make the final classification. In Section IV, we also present the experiments results using an SVM with multi-instance kernels proposed by Gärtner [24].

At test time, as shown in Fig. 3, suspected regions are generated using the algorithm described in Section III-A first. Suspected regions from one tongue image are grouped into one bag and a feature vector is extracted from each suspected region using the fixed fine-tuned ConvNet. Then, a multiple-instance SVM is used to make the final decision.

The training samples of the multiple-instance SVM are bags (consist of feature vectors) and labels associated with bags. In this paper, bags are the same as we described in test time and the label for each bag is whether the tongue is tooth-marked tongue or not. Concisely speaking, only coarsely labeled images (whether it is a tooth-marked tongue or not) are needed to train a multiple-instance SVM. Finely labeling tongue images (such as locating the tooth-marked regions) takes a lot efforts and only professional TCM practitioners can participate in labeling. Since we can fix the ConvNet as we described in Section III-B, it means if new tongue images are acquired and they are temporarily coarsely labeled or some tongue images are coarsely labeled due to the historical reasons, they can still be utilized in training. This is also a practical reason why we choose the multiple-instance representation of the tongue image.

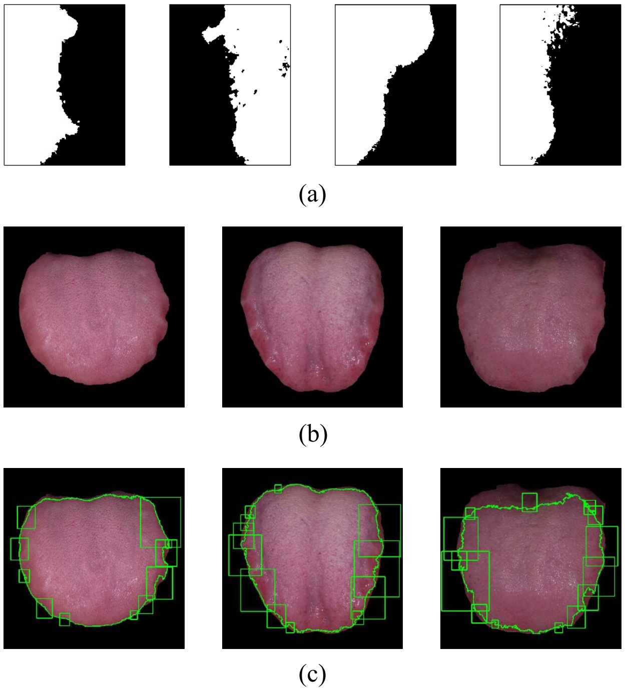


Fig. 5. Example of the first stage of the proposed method As can be seen from the picture. (a) Four types of tooth-marked regions become concave after the steps 1 and 2. (b) and (c) Most of the tooth-marked regions can be founded by the first stage. The goal of this stage is to find as many suspected tooth-marked regions as possible.

IV. Experiments

In this section, we present three different experiments results of the proposed method. The first is the result on fivefold cross-validation which is used to evaluate the performance of the proposed method. The second is the comparison with other works, such as Shao *et al.* [1]. And the last is the comparison between with and without multiple-instance SVM. The experiments results are evaluated by the following three metrics: 1) accuracy; 2) true positive rates (TPR); and 3) true negative rate (TNR)

$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}+\mathrm{TN}} \tag{6}$$

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{7}$$

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}. \tag{8}$$

TABLE I
FIVEFOLD CROSS-VALIDATION RESULTS

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Average
Accuracy	75.0%	68.8%	72.7%	71.9%	75.0%	72.7%
TPR	72.1%	62.1%	73.1%	70.4%	68.1%	69.1%
TNR	77.6%	74.3%	72.4%	73.0%	83.9%	76.2%

A. Fivefold Cross-Validation

To prove the effectiveness of our proposed method, we perform fivefold cross-validation on our tooth-marked tongue dataset including 641 tongue images. The detail of the dataset is described in Section II.

The fivefold cross-validation setting: since the dataset includes tongue images acquired in three different periods, the tongue images are randomly shuffled and then divided into five partitions. Each partition contains 128 tongue images except that the last partition contains 129 tongue images. Each time, we use four partitions for training and the other one for testing. It should be noticed that the MI-SVM is trained on the original tongue images while the ConvNet is trained on the region images. The region images are generated by both manual labeling and the first stage of the proposed method.

From Table I we can conclude that: 1) the performance of the proposed method is relatively stable which proves the effectiveness of the proposed method; 2) though the proposed method is effective, the overall accuracy is still struggling around 70%; and 3) the TNR exceed TPR by 10%, we investigate the reason in Section IV-C.

B. Comparison With Other Works

We conduct the experiments on the dataset which is described in Section II using another two methods, the one proposed by Shao *et al.* [1] and another one proposed by Wang [17]. Since [17] does not release their code to the public, we carefully reimplement their method according to [17]. Wang's method recognizes teeth marks on the tongue images, so if one teeth mark is recognized in the tongue image, the tongue is identified as a tooth-marked tongue. Wang's method manually selects parameters d, e, and f. We use the best setting stated in [17].

We also record the experiments results which obtained by replacing the MI-SVM and VGG16 of the proposed method. We choose another multiple-instance SVM which is a standard SVM using multiple-instance kernels proposed by [24]. We use linear minimax kernel (LMMK) and linear unnormalized set kernel (LSK) in the experiments. The implementation of LMMK and LSK is also by Doran and Ray [22] in this paper. And we also experiment on AlexNet [5] features to test if VGG16 can be replaced by a shallower ConvNet in the proposed method.

All results are obtained using fivefold cross-validation and the setting is the same as Section IV-A. The average accuracy, TPR and TNR are recorded in Table II.

The results are shown in Table II. As we can see from the results, though Shao's method is effective when teeth marks are very obvious, it loses its effectiveness when facing our dataset. Due to the difference between the datasets and the segmentation technologies and the fact that only the concavity information is utilized, Wang's method can easily misjudge the concave regions on the healthy tongue. Thus, the overall accuracy is low. Conversely, the multiple-instance learning with ConvNet features have the most stable performance. As can be seen from the results that VGG16 features perform better than AlexNet features since the better capability of VGG16. And the impacts of different multiple-instance SVM seem relatively small.

TABLE II
COMPARISON WITH OTHER WORKS

Method	Accuracy	TPR	TNR
Wang's[17]	49.8%	97.4%	8.68%
Shao's[1]	60.4%	42.5%	76.6%
MI-SVM with AlexNet features	66.6%	62.6%	73.9%
MI-SVM with VGG16 features	72.7%	69.2%	76.2%
LMMK with VGG16 features	71.4%	73.3%	75.4%
LSK with VGG16 features	73.1%	72.7%	73.9%

TABLE III
TEST ACCURACY ON REGION IMAGES OF THE CONVNET

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5
Accuracy	89.0%	85.3%	85.9%	87.6%	87.7%

C. Comparison Between MI-SVM With ConvNet Features and Using ConvNet Directly

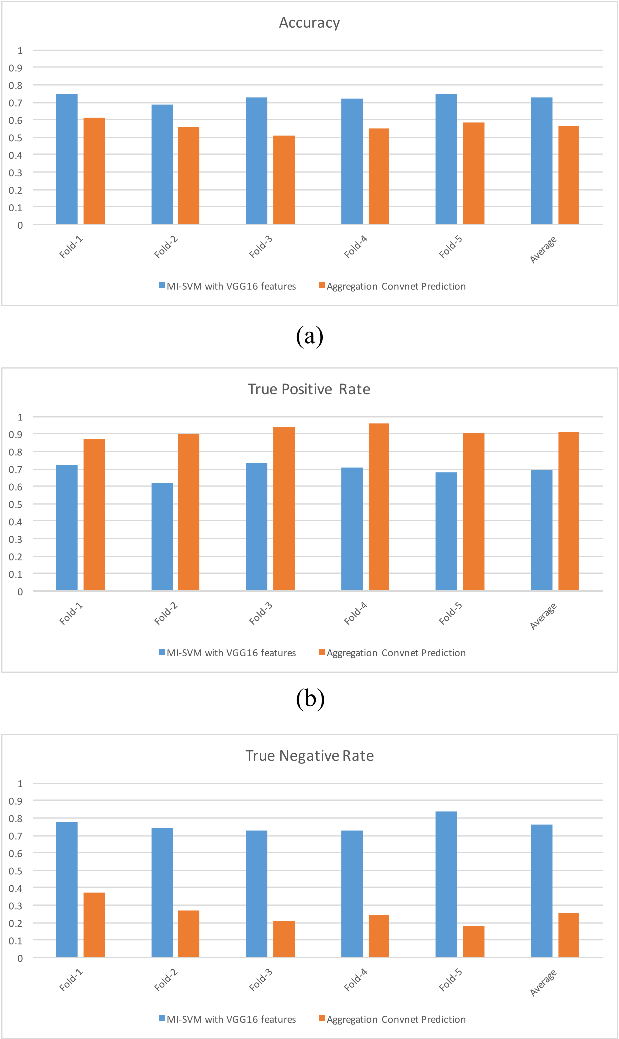
After fine-tuning the CNNs for tooth-marked regions, we can actually classify the tooth-marked tongue directly by aggregating the prediction of the ConvNet. Suppose p_i is the prediction of the ith suspected tooth-marked region (instance) by the ConvNet and there are m suspected tooth-marked regions(instances) in a tongue (bag), then the prediction of a tongue can be write as follows:

$$P_B = \max(p_0, \ldots, p_m). \tag{9}$$

The formula also fits the multiple-instance learning assumption that if a bag is a positive bag then at least one instance of this bag is positive. We call this method aggregation ConvNet prediction in this paper. We also evaluate the performance of this method by fivefold cross-validation using the same setting describe in Section IV-B. And the results are recorded in Fig. 6. We also record the test accuracy on region images of ConvNet after we fine-tune it in Table III.

We can see from Fig. 6 that the accuracy and TNR drop significantly without using multiple-instance SVM. The multiple-instance SVM works as a balancer which significantly increases the TNR while not overly reduces the TPR.

However, the Table III shows that the network can successfully recognize tooth-marked regions (the test accuracies of our ConvNet are relatively high), but Fig. 6 shows that aggregation ConvNet prediction performs poorly on tongue images (especially on nontooth-marked tongue images). Actually, the relatively high accuracy on region images of ConvNet showed on Table III means that the ConvNet does not suffer from overfitting. The reason aggregation ConvNet prediction performs



(a)

(b)

(c)

Fig. 6. Comparison between MI-SVM with ConvNet features and using CNN directly. (a) Comparison of accuracy. (b) Comparison of TPR. (c) Comparison of TNR.

poorly is because the variance of tooth-marked regions is relatively small while the types of nontooth-marked regions is almost infinite (all other regions are considered as nontooth-marked regions). Since it is important to train a machine learning model with a balanced training set, we cannot cover every type of nontooth-marked regions. We intendedly increase the number of nontooth-marked regions when we generate the suspected region samples using the algorithm described in Section III-A to train the ConvNet so that the number of nontooth-marked region is about 1.5 times more than tooth-marked regions. But it helps very little when we use aggregation ConvNet prediction instead of using multiple-instance SVM with ConvNet features. What is more, the ConvNet is trained solely on instance level, it may lack the information from a complete bag.

On the other hand, when we freeze the ConvNet as a feature extractor and train the multiple-instance SVM, (thought the underlying SVM is still trained on instance level) each training data is a bag (a tongue image). It means: 1) since the MI-SVM is decided only by the least negative instance in negative bag and the most positive instance in positive bag, we can take more negative instances into account and the algorithm will automatically select the most positive instance and the least negative instance and 2) MI-SVM will obtain some bag information since it will force that at least one instance (the most positive one) in a positive bag is positive and every instance in negative bag is negative during training.

V. Conclusion

In this paper, we have presented a multiple-instance method for the recognition of tooth-marked tongues. There are three stages of the proposed method. First, suspected regions are generated. Second, a deep ConvNet is used to extract the feature of every region. At last, a tongue is represented by a bag of feature vectors and a multiple-instance SVM is used to make the final classification. The experiments show that the proposed method greatly improve the accuracy than Shao's and show its effectiveness even when the dataset is noisy and the teeth marks are not obvious.

However, the proposed method still does not achieve a very high accuracy. There are still a lot works to improve the proposed method.

1) More new tongue samples will be acquired. Since we use a deep ConvNet as a feature extractor, the proposed model can always benefit a lot from a larger dataset.
2) Since illumination conditions may vary when acquiring new images, a more aggressive data augmentation will be adopted when we train our ConvNet so it can learn a more robust feature.
3) The region generating algorithm will be improved so that it can include more true tooth-marked regions and less nontooth-marked regions.
4) VGG-16 is chosen in this paper. However, VGG-16 is computationally intensive than some stat-of-art architectures. We will improve the architecture of the ConvNet to further improve the accuracy and reduce the computation cost.

References

[1] Q. Shao, X. Li, and Z. Fu, "Recognition of teeth-marked tongue based on gradient of concave region," in *Proc. Int. Conf. Audio Lang. Image Process. (ICALIP)*, Shanghai, China, 2014, pp. 968–972.

[2] G. Laskaris, "Color atlas of oral diseases," in *Perative Denti Try*. New York, NY, USA: Thieme, 2003, p. 213.

[3] A. G. Ghom and S. A. L. Ghom, *Textbook of Oral Medicine*. New Delhi, India: JP Med., 2014.

[4] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artif. Intell.*, vol. 89, nos. 1–2, pp. 31–71, 1997.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, 2014, pp. 580–587.

[7] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 3431–3440.

[8] P. Sermanet *et al.*, "OverFeat: Integrated recognition, localization and detection using convolutional networks," *CoRR*, vol. abs/1312.6229, 2013.

[9] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2014, pp. 806–813.

[10] C.-C. Chiu, "A novel approach based on computerized image analysis for traditional Chinese medical diagnosis of the tongue," *Comput. Methods Programs Biomed.*, vol. 61, no. 2, pp. 77–89, 2000.

[11] B. Pang, D. Zhang, N. Li, and K. Wang, "Computerized tongue diagnosis based on Bayesian networks," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 10, pp. 1803–1810, Oct. 2004.

[12] W. Zuo, K. Wang, D. Zhang, and H. Zhang, "Combination of polar edge detection and active contour model for automated tongue segmentation," in *Proc. 3rd Int. Conf. Image Graph. (ICIG)*, Hong Kong, 2004, pp. 270–273.

[13] S. Yu, J. Yang, Y. Wang, and Y. Zhang, "Color active contour models based tongue segmentation in traditional Chinese medicine," in *Proc. 1st Int. Conf. Bioinformat. Biomed. Eng. (ICBBE)*, Wuhan, China, 2007, pp. 1065–1068.

[14] L. Zhi, D. Zhang, J.-Q. Yan, Q.-L. Li, and Q.-L. Tang, "Classification of hyperspectral medical tongue images for tongue diagnosis," *Computerized Med. Imag. Graph.*, vol. 31, no. 8, pp. 672–678, 2007.

[15] X. Wang, B. Zhang, Z. Yang, H. Wang, and D. Zhang, "Statistical analysis of tongue images for feature extraction and diagnostics," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 5336–5347, Dec. 2013.

[16] B. Huang, J. Wu, D. Zhang, and N. Li, "Tongue shape classification by geometric features," *Inf. Sci.*, vol. 180, no. 2, pp. 312–324, 2010.

[17] H. Wang, X. Zhang, and Y. Cai, "Research on teeth marks recognition in tongue image," in *Proc. Med. Biometrics Int. Conf.*, Shenzhen, China, 2014, pp. 80–84.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[19] F. Chollet *et al.* (2015). *Keras*. [Online]. Available: https://github.com/fchollet/keras

[20] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2003, pp. 577–584.

[21] J. Amores, "Multiple instance classification: Review, taxonomy and comparative study," *Artif. Intell.*, vol. 201, pp. 81–105, Aug. 2013.

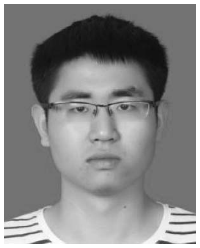
[22] G. Doran and S. Ray, "A theoretical and empirical analysis of support vector machine methods for multiple-instance classification," *Mach. Learn.*, vol. 97, nos. 1–2, pp. 79–102, 2014.

[23] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," in *Proc. Adv. Neural Inf. Process. Syst.*, Denver, CO, USA, 1998, pp. 570–576.

[24] T. Gärtner, P. A. Flach, A. Kowalczyk, and A. J. Smola, "Multi-instance kernels," in *Proc. ICML*, vol. 2. 2002, pp. 179–186.

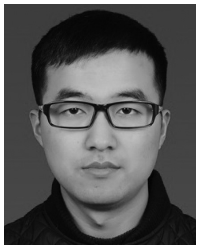
Xiaoqiang Li (M'14) received the Ph.D. degree in computer science from Fudan University, Shanghai, China, in 2004.

He is currently an Associate Professor of computer science with Shanghai University, Shanghai. His current research interests include image processing, pattern recognition, computer vision, and machine learning.



Yin Zhang received the B.S. degree in computer science from Shanghai University, Shanghai, China, where he is currently pursuing the M.S. degree in computer science.

His current research interests include computer vision and machine learning.



Qing Cui received the B.S. degree in computer science from Yantai University, Yantai, China, in 2014. He is currently pursuing the M.S. degree in computer science with Shanghai University, Shanghai, China.

His current research interests include computer vision and machine learning.

Xiaoming Yi received the B.S. degree in computer science from Southwest Petroleum University, Sichuan, China, in 2015 and the M.S. degree in computer science from Shanghai University, Shanghai, China, in 2017.

His current research interests include software engineering and machine learning.

Yi Zhang received the B.S. degree in computer science from Shanghai Business School, Shanghai, China, in 2014. He is currently pursuing the M.S. degree in computer science Shanghai University, Shanghai.

His current research interests include computer vision and machine learning.