DEEPTONGUE: TONGUE SEGMENTATION VIA RESNET

Bingqian Lin, Yanyun Qu, Junwei Xie, Cuihua Li*

School of Information Science and Engineering, Xiamen University, China

ABSTRACT

Accurate tongue image segmentation is helpful to acquire correct automatic tongue diagnosis result. However, traditional methods cannot bring satisfying results in most cases. This paper proposes an end-to-end trainable tongue image segmentation method using deep convolutional neural network based on ResNet. The proposed method, named DeepTongue, segments tongue by using a forward network without preprocessing. The proposed method has no restrictions of the illumination and size of tongue images. Experimental results show that the proposed DeepTongue improves the segmentation accuracy by a noticeable margin. In addition, DeepTongue is much faster than the existing tongue image segmentation methods.

***Index Terms*—** Automatic tongue image segmentation, DeepTongue, deep convolutional neural network, DeepMask, ResNet

1. INTRODUCTION

Tongue diagnosis plays an important role in Traditional Chinese Medicine (TCM). The tongue characterization is an auxiliary tool of tongue diagnosis, which can lead to more accurate diagnosis result. The tongue characterization mainly consists of the following four steps: 1) tongue image collection, 2) automatic tongue segmentation, 3) tongue color correction, 4) disease diagnosis. Among the above four steps, the automatic tongue image segmentation is a key procedure. The accuracy of tongue segmentation can directly affect the tongue diagnosis result.

The existing tongue segmentation methods can roughly be divided into four categories: thresholding tongue segmentation methods [1, 2], edge detection tongue segmentation methods [3, 4], the graph theory based methods [5, 6] and active contour model based methods [7, 8]. These methods can produce fine segmentation results to a certain extent. However, their disadvantages are four folds: 1) They are sensitive to the illumination change and cluttered background. 2) They cannot segment tongue from lip accurately due to the color similarity between them. 3) They usually need some preprocessing such as the detection of tongue body before conducting the segmentation. 4) They are generally slow in running time.

Recently, deep learning has made a lot of breakthroughs in the field of computer vision. Due to the outstanding ability of feature learning and representation, the deep convolutional neural network (CNN) has achieved great success in object detection [9, 10], semantic segmentation [11, 12], image recognition [13, 14] and some other fields in computer vision. Even so, there are few methods using deep convolutional neural network for the automatic tongue segmentation because of the difficulty of collecting and labeling the tongue image datasets. The latest methods [15, 16] applying the deep convolutional neural network for tongue segmentation outperform some traditional tongue segmentation methods. However, they also need some additional preprocessing such as the brightness discrimination and the image enhancement which make the whole process complex.

In this paper, we propose an end-to-end tongue segmentation method, named DeepTongue, which segments tongue with high accuracy based on the deep convolutional neural network. For a tongue image, DeepTongue can simultaneously perform tongue detection and segmentation. Furthermore, the DeepTongue models with ResNet based on different number of layers [13] are implemented to achieve better segmentation performance. Unlike the existing methods mentioned above, DeepTongue does not require additional preprocessing such as size normalization, illumination discrimination and tongue detection. The experiment results demonstrate that the DeepTongue proposed for tongue segmentation is not only feasible but also effective.

The main contributions of this paper are as follows. i) The proposed DeepTongue is an end-to-end segmentation method involving no pre-processing, unlike the previous tongue segmentation methods. ii) The proposed method is adaptive to tongue images with different illumination conditions, image sizes and tongue positions. iii) DeepTongue models with ResNet based on different number of layers achieve promising segmentation results with very fast segmentation speed.

2. TONGUE SEGMENTATION WITH DEEPTONGUE

The proposed DeepTongue framework is illustrated in Fig .1. In the training stage, each tongue image in the training set is sampled to generate image patches. Then the DeepTongue model is trained on these image patches. When a tongue image is queried, a sliding window scheme is implemented to obtain the image patches with different scales and locations. Then the DeepTongue model is used to perform tongue detection and segmentation simultaneously for each image patch. Finally the image patch which is most likely to contain the tongue body is selected to complete the tongue segmentation.

2.1. Architecture of DeepMask

The architecture of DeepTongue is inspired by DeepMask proposed by Pinheiro et al. [17]. The DeepMask architecture is shown in Fig. 2. It consists of the shared feature extraction layers and two branches, which are called the segmentation branch and the scoring branch. The shared part is the VGG-A architecture [14] containing eight 3×3 convolutional layers and five 2×2 max-pooling layers. In order to reserve more spatial information in the convolutional layer, all the final fully connected layers and the last max-pooling layer of the VGG-A model are removed.

The segmentation branch consists of a single 1×1 convolutional layer followed by a classification layer. The classification layer is composed of $h \times w$ pixel classifiers and each pixel classifier is used for indicating whether a given pixel belongs to the object in the center of the patch. The classification layer is decomposed into two

*(Corresponding author: Yanyun Qu)

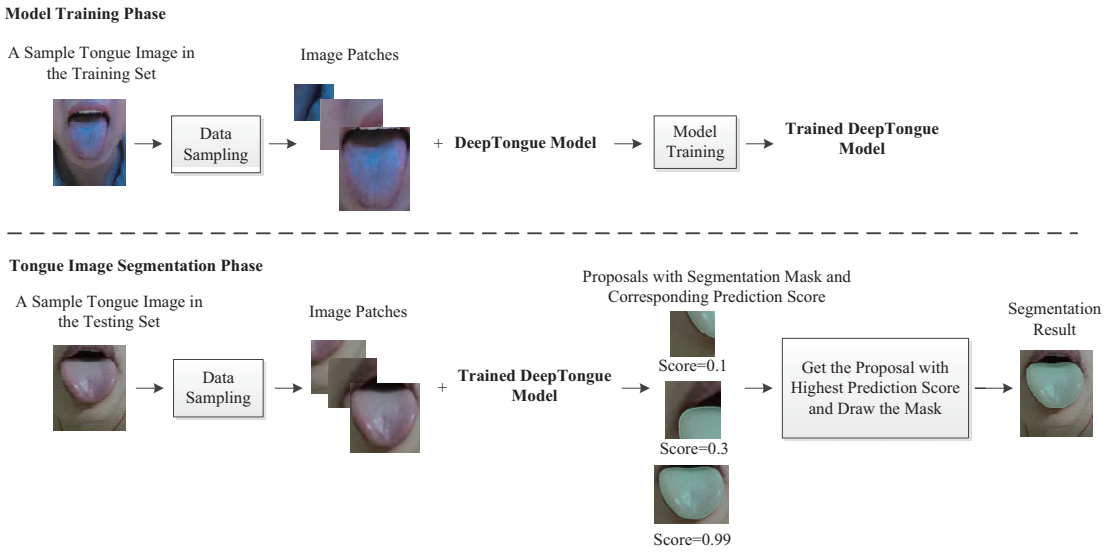

Fig. 1. The proposed DeepTongue framework.

linear layers. Considering the running time, the output of the classification layer is set to be $h^\circ \times w^\circ$ with $h^\circ < h$ and $w^\circ < w$. Finally the output is upsampled to $h \times w$ by bilinear interpolation to generate the segmentation mask with the same dimension of the input image patch. The scoring branch consists of a 2×2 max-pooling layer followed by two fully connected layers. The final output of the scoring branch is a prediction score indicating whether an image patch is fully centered by an object.

2.2. DeepTongue models with ResNet based on different model depths

Pinheiro et al. further optimized the architecture of the original DeepMask in [18]. They called the shared layers as 'trunk' architecture and two branches as 'head' architecture respectively. The 'trunk' architecture was replaced with ResNet and achieved excellent segmentation performance experimentally. For the 'head' architecture of the network, they designed several new architectures which could share more computation than the original one.

The ResNet models [13] can achieve better classification accuracy and lower training error with deep model depth compared to many previous models. In this paper, we design several DeepTongue models with ResNet based on different model depths. Specifically, we use the 18-layer ResNet model, the 34-layer ResNet model, the 50-layer ResNet model and the 101-layer ResNet model respectively as the 'trunk' architecture in the DeepTongue model. The main differences among these ResNet models are the number of layers and the architecture of the residual blocks [13]. For the 'head' architecture of DeepTongue model, we choose Head C in [18] uniformly due to its fast speed and simplicity. Fig .3 shows the architecture of DeepTongue based on 50-layer ResNet. The 'trunk' architecture of the network is the 50-layer ResNet with the additional modification of removing the last fully connected layer, average-pooling layer and convolutional layer $conv5\_x$. For the DeepTongue models based on 18-layer ResNet, 34-layer ResNet and 101-layer ResNet, we implement the same modification on the 'trunk' architecture.

2.3. Implementation details of DeepTongue

The proposed method DeepTongue consists of two stages: the model training stage and the tongue image segmentation stage. The training details of the DeepTongue model and the process of tongue segmentation are similar to [17].

Model training. A training sample k in the training set is a triplet containing x_k , y_k and m_k. x_k denotes the RGB input patch, y_k denotes the label of whether the patch contains a tongue body ($y_k \epsilon \{\pm 1\}$), and m_k denotes the binary segmentation mask of the patch. m_k^{ij} is the mask value of the pixel at location (i, j) in the input patch ($m_k^{ij} \epsilon \{\pm 1\}$).

In the model training phase, the training samples are divided into positive samples and negative samples. The positive samples are image patches fully centered by a tongue body (having some tolerance). The scoring network is trained with an equal number of positive and negative samples while the segmentation network is trained with positive samples only. The loss function of DeepTongue contains the segmentation term and the scoring term and is formulated as

$$Loss = \sum_k \Big(\frac{1+y_k}{2w^\circ h^\circ} \sum_{ij} log(1 + e^{-m_k^{ij} f_{segm}^{ij}(x_k)}) + \lambda log(1 + e^{-y_k f_{score}(x_k)})\Big) \quad (1)$$

where $f_{segm}^{ij}(x_k)$ is the predicted mask value at location (i, j) and $f_{score}(x_k)$ is the prediction score indicating whether the patch contains a tongue body. w° and h° correspond to the width and height respectively of the segmentation mask generating from the classification layer of the segmentation network. λ is the balance factor of the segmentation term and the scoring term.

The DeepTongue model simultaneously generates the segmentation mask and the prediction score for each training sample in the forward pass. Then the loss function of the network is calculated. The segmentation branch and the scoring branch are backward propagated alternatively and the parameters of the network are updated

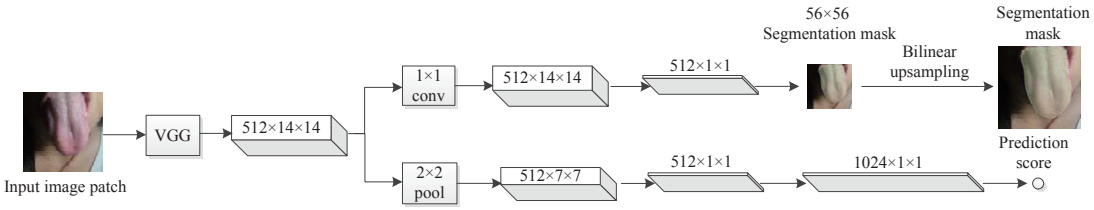

Fig. 2. The architecture of DeepMask.

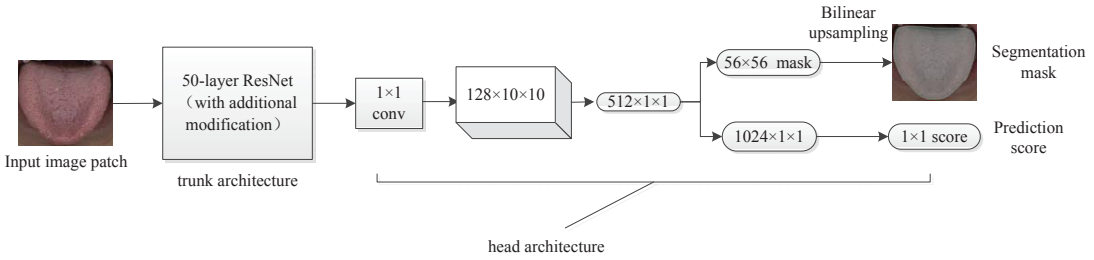

Fig. 3. The architecture of DeepTongue based on 50-layer ResNet.

by using the stochastic gradient descent. The choice of training parameters and network initialization scheme are the same as those in [17].

Tongue image segmentation. A tongue image in the testing set is sampled to generate a set of image patches with different locations (with a stride of 16 pixels) and scales (scales from 2^{-2} to 2^{1} with a step of $2^{1/2}$) at first. This procedure ensures that the tongue body in the image will not be missed. Then the DeepTongue model generates a segmentation object mask and assigns a prediction object score for each image patch. As the tongue segmentation is single-object segmentation, we only use the proposal with the highest prediction score to complete the segmentation.

3. EXPERIMENTAL RESULTS

3.1. Tongue image datasets

We conduct the experiments in Torch7 on the AMAX GPU with 48 G internal storage and NVIDIA GTX 1080 graphics card. The tongue image training set in this paper is composed of 2344 tongue images captured by cell phones. These tongue images have different sizes and illumination conditions with variance of tongue shapes and locations.

So far, there are few public standard tongue image datasets for evaluating the tongue image segmentation performance. Most tongue image segmentation algorithms evaluate the segmentation performance on the tongue image datasets with uniform illumination and small changes of tongue body in location. To evaluate the proposed method more objectively, we use two kinds of tongue image testing sets called TestSet1 and TestSet2 respectively. TestSet1 is made up of 1001 tongue images captured by cell phones. The size and illumination condition of each tongue image are different in TestSet1. TestSet2 is the tongue image dataset published in the Internet by BioHit [19], which is composed of 300 tongue images. The size and illumination condition are uniform in TestSet2. Moreover, the change of tongue body in location is large in Testset1 while small in Testset2. We perform segmentation artificially to the tongue images in both training and testing sets to obtain the ground truth. The training and testing examples with the corresponding artificial binary segmentation masks are shown in Fig .4.

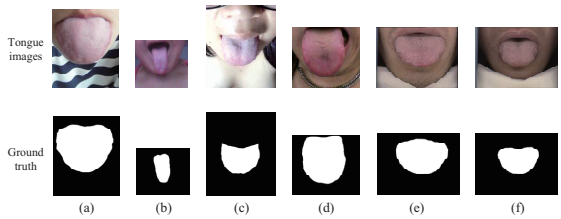

Fig. 4. The tongue images with the corresponding artificial segmentation masks. (a)-(b) are the training examples in the training set, (c)-(d) are the testing examples in TestSet1, (e)-(f) are the testing examples in TestSet2.

3.2. The evaluation criteria

We adopt four criteria to evaluate the segmentation performance of single tongue image: the pixel accuracy (PA), the mean pixel accuracy (MPA), the mean intersection over Union (MIoU) and the segmentation time. The formulations of PA, MPA and MIoU are shown as follows:

$$PA = \frac{\sum_i n_{ii}}{\sum_i t_i} \tag{2}$$

$$MPA = \left(\frac{1}{n_{cl}}\right) \sum_i \frac{n_{ii}}{t_i} \tag{3}$$

$$MIoU = \left(\frac{1}{n_{cl}}\right) \sum_i \frac{n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}} \tag{4}$$

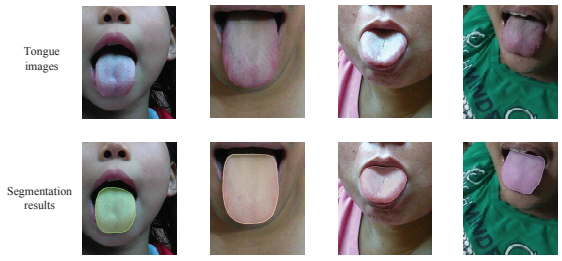

Fig. 5. The visual segmentation results of DeepTongue on TestSet1.

Table 1. The performance results of different methods on TestSet1 and TestSet2.

method	dataset	PA	MPA	MIoU	time
GrabCut	TestSet1	93.11%	87.58%	82.12%	0.54s
	TestSet2	81.21%	59.20%	49.56%	1.10s
ResNet18-D	TestSet1	97.92%	96.39%	94.57%	0.30s
	TestSet2	98.11%	96.19%	94.49%	0.36s
ResNet34-D	TestSet1	97.84%	96.17%	94.31%	0.30s
	TestSet2	98.08%	96.11%	94.38%	0.36s
ResNet50-D	TestSet1	97.96%	96.55%	94.74%	0.45s
	TestSet2	98.13%	96.34%	94.58%	0.57s
ResNet101-D	TestSet1	97.89%	96.45%	94.57%	0.77s
	TestSet2	98.04%	96.22%	94.30%	0.89s

where n_{cl} is the number of the pixel classes, n_{ij} is the number of pixels of class i predicted to belong to class j, and $t_i = \sum_j n_{ij}$ is the total number of pixels of class i. We calculate the average values of all these four evaluation criteria on TestSet1 and TestSet2 respectively.

3.3. Results and analysis

In this section, we first show the visual segmentation results of DeepTongue on Testset1 where tongue images have different illumination conditions, image sizes and tongue positions. The visual segmentation results are shown in Fig .5. Then we compare the segmentation results of our method and the traditional method GrabCut [20] on both two test sets. The performance comparison is shown in Table 1 and the comparison of visual effect is shown in Fig .6. As shown in Fig .5, our method DeepTongue is robust to illumination changes, image sizes and tongue positions of tongue images. The performance results in Table 1 show that the DeepTongue models with ResNet based on different number of layers are all superior to the traditional method GrabCut in terms of PA, MPA and MIoU on both TestSet1 and TestSet2. The DeepTongue model based on 50-layer ResNet achieves the best results among them. Moreover, the segmentation speeds of DeepTongue models based on 18-layer ResNet, 34-layer ResNet and 50-layer ResNet are all faster than GrabCut. From Fig .6 we can see that the visual segmentation results of DeepTongue are much better than the traditional method GrabCut on both TestSet1 and TestSet2. GrabCut performs wrong segmentation on both two testing sets in some cases.

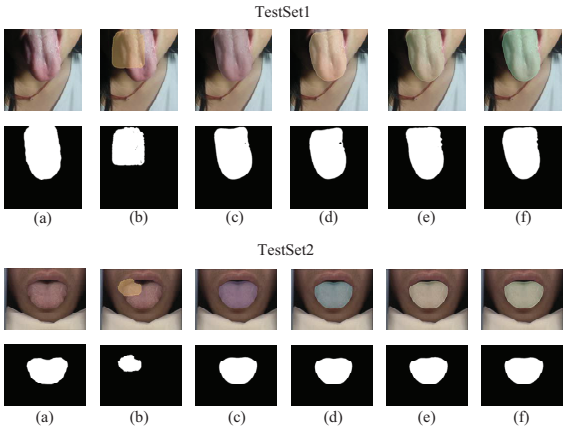

Fig. 6. The visual segmentation results of different methods on TestSet1 and TestSet2. (a) represents the original image, (b) represents the GrabCut, (c)-(f) represent the DeepTongue models based on 18-layer ResNet, 34-layer ResNet, 50-layer ResNet and 101-layer ResNet respectively.

4. CONCLUSIONS

In this paper, we propose an end-to-end tongue image segmentation method named DeepTongue. Unlike the traditional methods which extract the image feature manually, DeepTongue based on the deep convolutional neural network can automatically extract the high-level image feature to perform segmentation well. It is robust to the changes of illumination condition, image size and tongue position. Moreover, there is no need to do any pre-processing in DeepTongue. The experiment results show that the DeepTongue models based on ResNet of different model depths are competitive to the existing tongue image segmentation methods in segmentation accuracy and speed. In the future, we plan to collect more tongue image datasets and design better network architectures to further improve the tongue image segmentation performance.

5. ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China under Grant 61373077.

6. REFERENCES

[1] L. Zhang and J. Qin, "Tongue-image segmentation based on gray projection and threshold-adaptive method," *Journal of Clinical Rehabilitative Tissue Engineering Research*, vol. 14, no. 9, pp. 1638–1641, 2010.

[2] L. I. Dan-Xia and Yu Ke Wei, "Tongue image segmentation method based on adaptive thresholds," *Computer Technology & Development*, 2011.

[3] F. U. Zhi cheng, L. I. Xiao qiang, Shanghai University, and Shanghai, "Tongue image segmentation based on snake model and radial edge detection," *Journal of Image & Graphics*, 2009.

[4] Q. L. Li, Y. Q. Xue, J. Y. Wang, and X. Q. Yue, "Automated tongue segmentation algorithm based on hyperspectral image,"

Journal of Infrared & Millimeter Waves, vol. 26, no. 1, pp. 77–80, 2007.

[5] Yu Ke Wei, Peng Fan, and Gui Zeng, "Application of improved grabcut method in tongue diagnosis system," *Transducer & Microsystem Technologies*, 2014.

[6] Shanchao Chen and F. U. Hongguang, "Application of improved graph theory image segmentation algorithm in tongue image segmentation," *Computer Engineering & Applications*, vol. 48, no. 5, pp. 201–203, 2012.

[7] Jingwei Guo, Yikang Yang, Qingwei Wu, Jionglong Su, and Fei Ma, "Adaptive active contour model based automatic tongue image segmentation," in *International Congress on Image and Signal Processing, Biomedical Engineering and Informatics*, 2017, pp. 1386–1390.

[8] Miao Jing Shi, Guo Zheng Li, and Fu Feng Li, "C 2 g 2 fsnake: automatic tongue image segmentation utilizing prior knowledge," *Science China*, vol. 56, no. 9, pp. 1–14, 2013.

[9] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," in *International Conference on Neural Information Processing Systems*, 2015, pp. 91–99.

[10] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng Yang Fu, and Alexander C. Berg, "Ssd: Single shot multibox detector," in *European Conference on Computer Vision*, 2016, pp. 21–37.

[11] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[12] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han, "Learning deconvolution network for semantic segmentation," *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1520–1528, 2015.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[14] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Science*, 2014.

[15] Jiang Li, Baochuan Xu, Xiaojuan Ban, Ping Tai, and Boyuan Ma, "A tongue image segmentation method based on enhanced hsv convolutional neural network," in *Cooperative Design, Visualization, and Engineering: 14th International Conference, CDVE 2017, Mallorca, Spain, September 17-20, 2017, Proceedings*, 2017, pp. 252–260.

[16] Panling Qu, Hui Zhang, Li Zhuo, Jing Zhang, and Guoying Chen, "Automatic tongue image segmentation for traditional chinese medicine using deep neural network," in *Intelligent Computing Theories and Application: 13th International Conference, ICIC 2017, Liverpool, UK, August 7-10, 2017, Proceedings, Part I*, 2017, pp. 247–259.

[17] Pedro O. Pinheiro, Ronan Collobert, and Piotr Dollr, "Learning to segment object candidates," in *NIPS*, 2015.

[18] Pedro O. Pinheiro, Tsung-Yi Lin, Ronan Collobert, and Piotr Dollr, "Learning to refine object segments," in *ECCV*, 2016.

[19] BioHit, "Tongueimagedataset," `https://github.com/BioHit/TongeImageDataset`, 2014.

[20] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake, ""grabcut": interactive foreground extraction using iterated graph cuts," in *ACM SIGGRAPH*, 2004, pp. 309–314.