applied sciences

MDPI

Article

TongueNet: A Precise and Fast Tongue Segmentation System Using U-Net with a Morphological Processing Layer

Jianhang Zhou †, Qi Zhang †, Bob Zhang *,† and Xiaojiao Chen †

PAMI Research Group, Department of Computer and Information Science, Faculty of Science and Technology, University of Macau, Taipa, Macau 999078, China

* Correspondence: bobzhang@um.edu.mo; Tel.: +853-8822-4425

† Current address: University of Macau, Avenida da Universidade, Taipa, Macau 999078, China.

Received: 6 July 2019; Accepted: 29 July 2019; Published: 1 August 2019

Abstract: Automated tongue segmentation is a critical component of tongue diagnosis, especially in Traditional Chinese Medicine (TCM), where it has been practiced for thousands of years and is generally considered pain-free and non-invasive. Therefore, a more precise, fast, and robust tongue segmentation system to automatically segment tongue images from its raw format is necessary. Previous algorithms segmented the tongue in different ways, where the results are either inaccurate or time-consuming. Furthermore, none of them developed a dedicated, automatic segmentation system. In this paper, we proposed TongueNet, which is a precise and fast automatic tongue segmentation system. U-net is utilized as the segmentation backbone applying a small-scale image dataset. Besides this, a morphological layer is proposed in the latter stages of the architecture. The proposed system when applied to a tongue image dataset with 1000 images, achieved the highest Pixel Accuracy of 98.45% and consumed 0.267 s per picture on average, which outperformed conventional state-of-the-art tongue segmentation methods in both accuracy and speed. Extensive qualitative and quantitative experiments showed the robustness of the proposed system concerning different positions, poses, and shapes. The results indicate a promising step in achieving a fully automated tongue diagnosis system.

Keywords: image segmentation; tongue image analysis; deep learning; convolutional neural network; morphological image processing

1. Introduction

The human tongue is a large and soft piece of flesh found in the mouth and primarily used for tasting and speaking [1]. Besides its essential functions in the human digestive system, a tongue can also act as a key region of interest in disease diagnosis using traditional medicines such as Traditional Chinese Medicine (TCM). Traditional Chinese Tongue Diagnosis (TCTD) [2–7] performs pain-free and non-invasive disease detection on our human bodies by analyzing the different attributes of the tongue (e.g., color, shape and texture). This has been practiced for thousands of years. Due to its convenience and pain-less procedure, Chinese Tongue Diagnosis is widely used and has become a valuable reference in disease inference. To perform TCTD automatically, a computer-aided tongue diagnosis system was proposed [8] which contains tongue segmentation as one of the key procedures. Since any disease diagnosis system makes decisions according to different features from the feature extraction stage [9–12], the quality of its output is directly related to the segmented tongue image from the source image. Therefore, it is critical to perform precise and fast tongue segmentation.

Until now, various existing automatic tongue segmentation techniques have been proposed as part of a more comprehensive system. For example, Bi-Elliptical Deformable Contour (BEDC) [13]

combines model-based techniques [14] and active contour models [15]. This method achieved relatively promising segmentation results while its segmentation quality is heavily dependent on some prior knowledge and is sensitive to the position and initial curves generated from the tongue. To overcome this, Ning et al. replaced the model-based techniques with a region merging strategy [16] to obtain coarse segmentation results [17]. They used ACM (Active Contour Models) as the post-processing step achieving a better segmentation performance compared with BEDC. However, the authors in [16] also depend heavily on some prior knowledge as the position information of the initial marker should be manually defined. Inspired by the effectiveness of the region merging strategy, Wu et al., designed a combination approach using region-based and edge-based [18] that can further remove the influences of the surrounding artifacts in the tongue and therefore boosts the segmentation results and robustness. That being said, this method still shows weak robustness in many different circumstances. For instance, the segmentation results are poor when tongue poses are irregular (refer to Figure 1e) and when the tongue is closely surrounded by some lips (see Figure 1h). Moreover, to the best of our knowledge, none of the previous works [13,16,18–20] have established a dedicated and complete tongue segmentation system. Rather, they focus on tongue segmentation as part of a subsystem in TCTD.

The challenges of tongue segmentation are that the characteristics of a tongue are different, such that it is different to mine common attributes using conventional geometric and iterative image processing methods. Here, we summarized the tongue images into eight circumstances so that each image is able to be categorized properly for further analysis. The eight circumstances displayed in Figure 1 are:

(a) Tongue with an apparent gap in the mouth.
(b) Tongue with abnormal color.
(c) Tongue with abnormal texture.
(d) Tongue with teeth showing.
(e) Tongue with irregular poses.
(f) Tongue not completely protruding.
(g) Tongue with teeth imprints on the edges.
(h) Tongue closely surrounded by the lips.

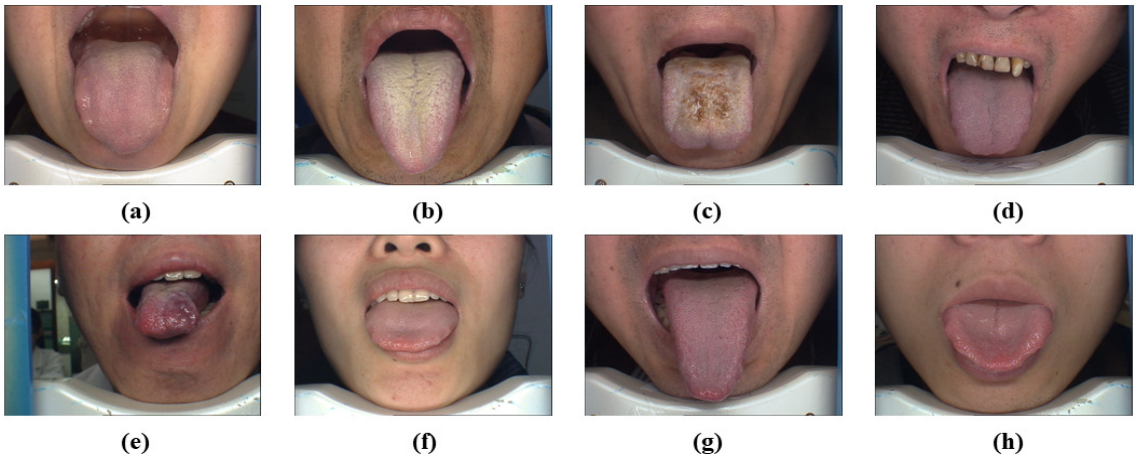

Figure 1. Samples of tongue in images with different circumstances. (**a**) sample of tongue with an apparent gap in the mouth; (**b**) sample of tongue with abnormal color; (**c**) sample of tongue with abnormal texture; (**d**) sample of tongue with teeth showing; (**e**) sample of tongue with irregular poses; (**f**) sample of tongue not completely protruding; (**g**) sample of tongue with teeth imprints on the edges; (**h**) sample of tongue closely surrounded by the lips.

It can be observed from this figure that more than one circumstance appears in a single tongue image.

To handle these challenges, in this paper, we propose a new tongue segmentation system with a specific segmentation model and a set of procedures (named TongueNet). To dig out common attributes of tongues in different images, we applied deep learning techniques [21]. In the image segmentation area, deep learning is widely used. For instance, FCN (Fully Covolutional Nerural Network) [22] , U-net [23], and Segnet [24] are popular deep learning models. In our proposed system (shown in Figure 2), there are four procedures including image acquisition, image grey-scaling, segmentation model prediction, and segmented image extraction. In the image acquisition procedure, a specially designed image capture device is applied to capture the tongue image of an individual in a stable environment. To enhance the segmentation performance as well as its efficiency, the image grey-scaling procedure is designed to obtain a grey-scale image with a dominant color channel. In the segmentation model, U-net with a morphological processing layer is proposed to overcome the limitation of the size of the dataset and characteristics of the tongue images. Finally, the segmented image extraction procedure is utilized to automatically segment the tongue from the raw image and takes it as the final result. Generally speaking, the following main contributions in this paper are:

1. A dedicated automatic tongue segmentation system is proposed.
2. A deep architecture: U-net with a morphological processing layer is applied.
3. The proposed tongue segmentation system is more precise and much faster than other state-of-the-art tongue segmentation methods.

This paper is organized as follows: In Section 2, the methodology of the proposed method is presented with theoretical interpretation and algorithm design. In Section 3, the experiments are to prove the effectiveness and efficiency of the proposed system. In Section 4, details regarding the design of TongueNet and its experimental results are fully discussed. In Section 5, we reach a conclusion.

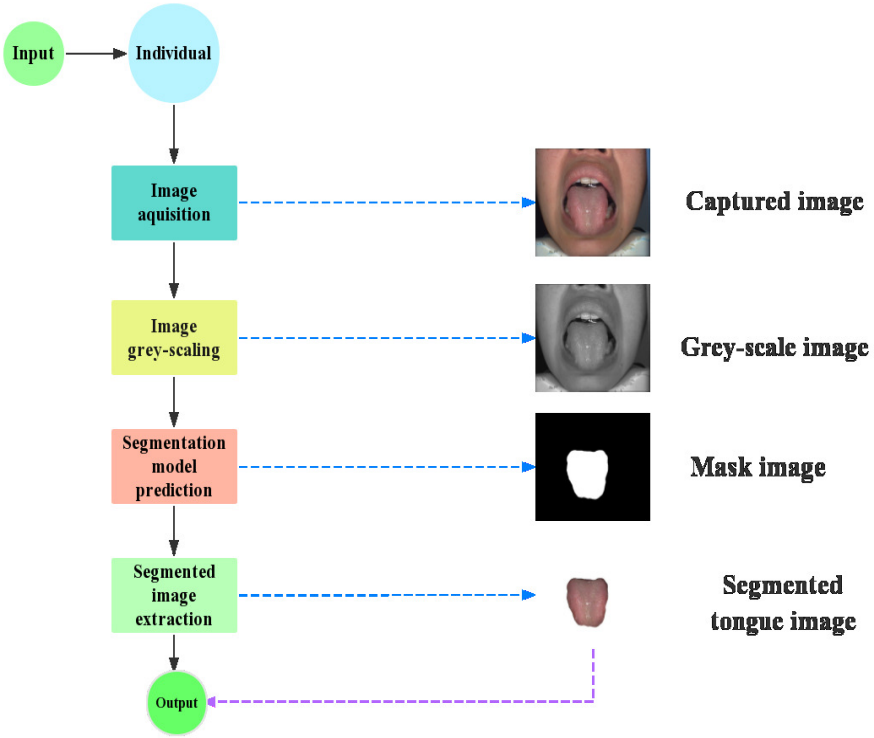

Figure 2. The pipeline of TongueNet. An individual's tongue image is first taken as the input. Figures on the right of each dotted line are the output of each procedure. In the image grey-scaling stage, the captured image is converted to grey-scale. Next is the segmentation model prediction stage, where the mask image is predicted. Afterwards, in the segmented image extraction stage, a tongue is segmented from its raw image. The final output is a segmented tongue image.

2. Methodology

In this section, we first introduce the basic idea of TongueNet. Next, we present the deep convolutional neural network architecture applied to the Segmentation Model Predication step in detail. Afterwards, a newly proposed morphological layer is described.

2.1. Overview of TongueNet

As mentioned above, the pipeline of TongueNet is displayed in Figure 2. The input of the system is an individual's tongue image captured (using a uniquely designed imaging apparatus [8]) using a special device, while the output of the system is a well-segmented tongue image.

After a raw image showing the lower half of an individual's face with their tongue protruding is captured in the first step. The next step takes the captured image and converts it to a grey-scale image after selecting one channel of the raw image to enhance the tongue domain [13]. The dominant color analysis of one sample in the dataset is shown in Figure 3. In Figure 3, (a) is a sample selected from the tongue image dataset, (b), (c) and (d) are the bivariate histograms to describe the different RGB values coming from the channels of each pixel. The RGB model is a color model which takes red, blue and green as the primary colors. In the histograms of Figure 3b–d, the color ranges from deep blue to yellow representing the frequency of pixels in the image. The color bar beside each histogram displays the relationship between the colors and the pixel frequencies. As can be seen from the yellow regions (high-frequency regions) of Figure 3b,c, most pixels have a higher red channel value with a lower blue and green channel value. Meanwhile, (d) shows that the majority of pixels (in the yellow region) have almost the same number of blue and green values. In the end, this indicates that the red color is the dominant color of tongue images. Therefore, for all samples, the red channel is selected since it is the dominant color. Next, the third step is mask prediction. In this step, a grey-scale image from the previous step is fed into the segmentation model. Afterwards, the model will output a well-refined mask image, only containing binary information to indicate which pixels belong to the tongue. Followed by this, the fourth step is tongue image extraction. In this step, the completed tongue image is extracted from the original raw image according to the mask image. The extracted result is regarded as the final output of the system.

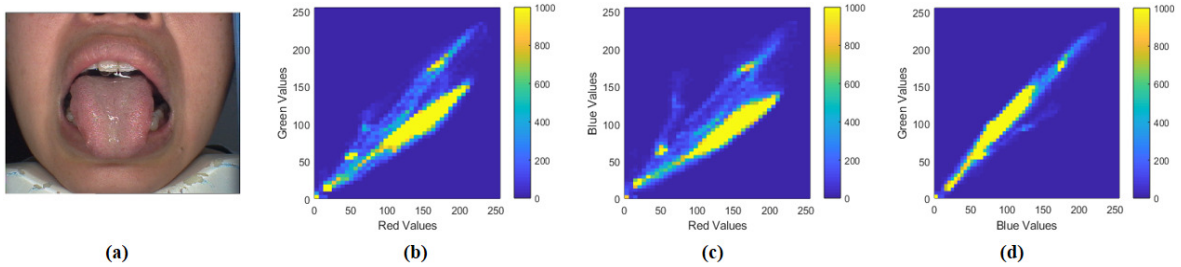

Figure 3. An example of color analysis to determine the dominant a color. The color bar beside the bivariate histograms show the relationship between the colors in the histogram and the pixel frequencies. (**a**) a sample from the dataset; (**b**) bivariate histogram of red vs. green; (**c**) bivariate histogram of red vs. blue, and (**d**) bivariate histogram of blue vs. green.

2.2. Network Architecture

Due to its effectiveness and efficiency in medical image segmentation, U-net [23] is applied in this system to perform tongue mask image prediction. There are three reasons for applying U-net in tongue image segmentation. First, U-net is sturdy and robust in small-scale datasets. Second, U-net maintains a good balance between the effectiveness and the efficiency in segmentation, which ensures that the overall system performs well. Thirdly, U-net does well in both segmentation and localization [25].

The structure and settings of the U-net architecture applied to tongue segmentation are shown in Figure 4. In this figure, the size of the input is 576 $\times$ 768 and the output image is a binary image of the same size. The green lines between two blocks are the contracting paths which combine

shallow features (left) with deep features at the same level (right). At the last stage of the network, a morphological processing layer is appended, to refine the feature map of the previous layer. We will introduce the principles and details of this morphological processing layer in Section 2.3.

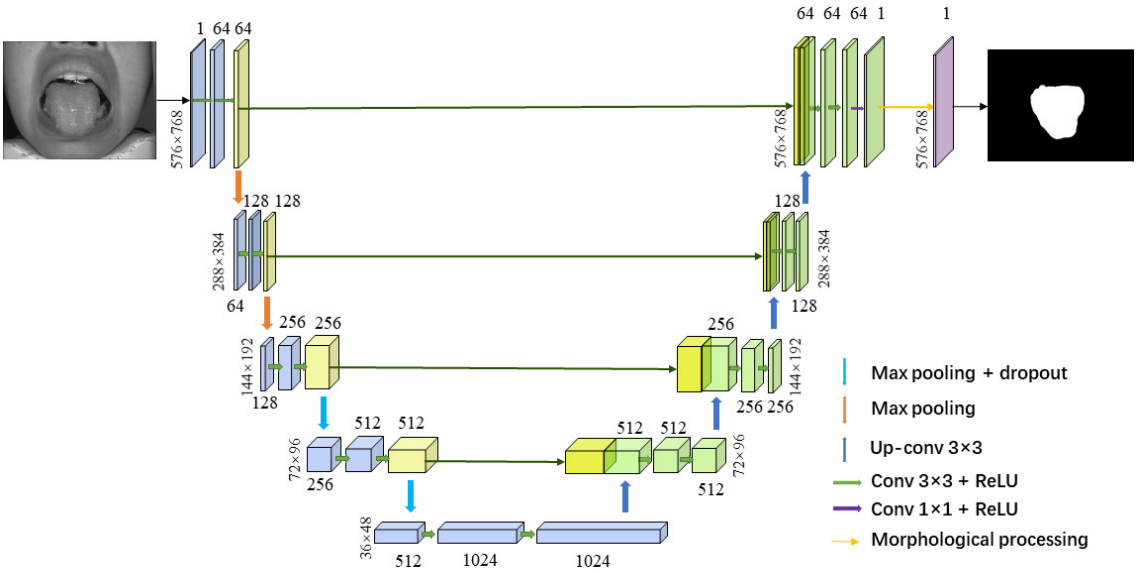

Figure 4. The U-net architecture applied in tongue segmentation. Boxes in yellow are the feature maps which are combined with the corresponding feature maps after up-sampling (green boxes). The morphological processing layer is placed at the end of network.

The parameter settings of each layer in the architecture are shown in Table 1, where the size of the training set is composed of 800 images in total (refer to Section 3.1). To achieve a model with higher accuracy and less training time, the learning rate is fixed to $\alpha = 10^{-4}$. Figure 5 shows loss and accuracy plots on the training set and test set using different learning rates ($\alpha = 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}$). From this figure, we can reach two conclusions as follows:

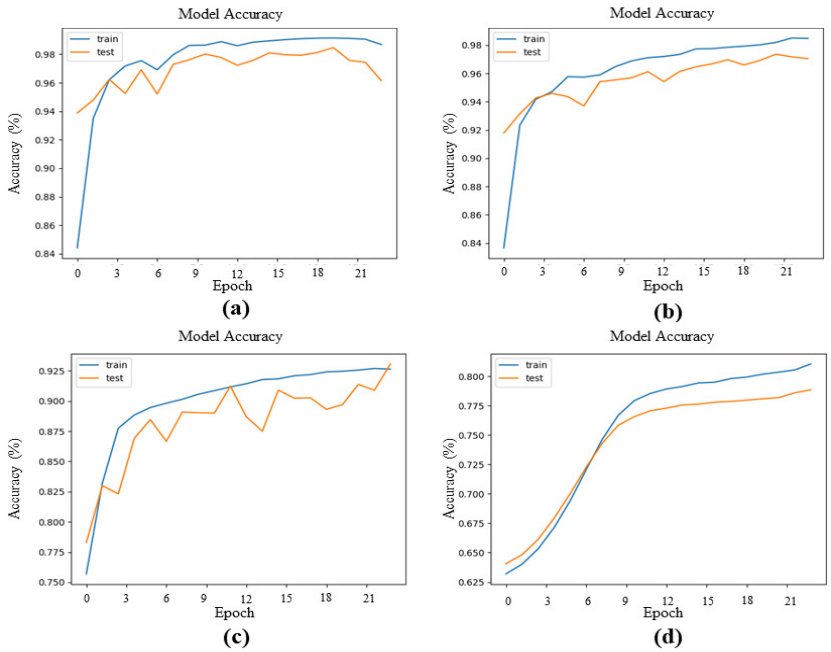

Figure 5. Accuracy on the training and test sets using different learning rates (α). (**a**) accuracy with $\alpha = 10^{-4}$; (**b**) accuracy with $\alpha = 10^{-5}$; (**c**) accuracy with $\alpha = 10^{-6}$; (**d**) accuracy with $\alpha = 10^{-7}$.

1. The model achieves its best performance when the learning rate is $\alpha = 10^{-4}$ (Figure 5a) compared with $\alpha = 10^{-5}$ (Figure 5b), $\alpha = 10^{-6}$ (Figure 5c), and $\alpha = 10^{-7}$ (Figure 5d).
2. The model with $\alpha = 10^{-4}$ (Figure 5a) converges faster than others (Figure 5b–d).

Therefore, we select the learning rate $\alpha = 10^{-4}$ when training the model.

Table 1. Parameter settings of each layer.

Layer	Type	Kernel Size	Number of Kernels	Input Dimensions	Activation Function
1	Convolution2D	3 × 3	64	576 × 768	ReLU
2	Convolution2D	3 × 3	64	576 × 768	ReLU
3	Maxpooling	2 × 2	-	576 × 768	-
4	Convolution2D	3 × 3	128	288 × 384	ReLU
5	Convolution2D	3 × 3	128	288 × 384	ReLU
6	Maxpooling	2 × 2	-	288 × 384	-
7	Convolution2D	3 × 3	256	144 × 192	ReLU
8	Convolution2D	3 × 3	256	144 × 192	ReLU
9	Maxpooling	2 × 2	-	144 × 392	-
10	Dropout	2 × 2	-	-	-
11	Convolution2D	3 × 3	512	72 × 96	ReLU
12	Convolution2D	3 × 3	512	72 × 96	ReLU
13	Maxpooling	2 × 2	-	72 × 96	-
14	Dropout	2 × 2	-	-	-
15	Convolution2D	3 × 3	1024	36 × 48	ReLU
16	Convolution2D	3 × 3	1024	36 × 48	ReLU
17	Up-convolution2D	2 × 2	512	36 × 48	-
18	Convolution2D	3 × 3	512	72 × 96	ReLU
19	Convolution2D	3 × 3	512	72 × 96	ReLU
20	Up-convolution2D	2 × 2	256	72 × 96	-
21	Convolution2D	3 × 3	256	144 × 192	ReLU
22	Convolution2D	3 × 3	256	144 × 192	ReLU
23	Up-convolution2D	2 × 2	128	144 × 192	-
24	Convolution2D	3 × 3	128	288 × 384	ReLU
25	Convolution2D	3 × 3	128	288 × 384	ReLU
26	Up-convolution2D	2 × 2	64	288 ×	-
27	Convolution2D	3 × 3	64	576 × 768	ReLU
28	Convolution2D	3 × 3	64	576 × 768	ReLU
29	Convolution2D	1 × 1	1	576 × 768	ReLU
30	Morphological processing	-	1	576 × 768	-

2.3. The Morphological Processing Layer

Although each tongue in the image is a relatively large and complete object, there are still inevitable noises around or inside the segmented regions. Furthermore, information like the shape and texture of a tongue provided by the training data is complicated, which to some extent influences the prediction of the pixels, especially for ones on the contours or between the border of the lips and tongue. Therefore, it is difficult to obtain a smooth and noise-free mask in one step. Taking the difficulties as well as the efficiency into consideration, we propose a morphological processing layer (MPL) to refine the coarse mask image produced by the network using specifically designed filters.

We formulate the refinement into a problem that can be described in Figure 6a and define the tongue region in Definition, where a binary mask M is generated from a grey-scale mask image G filtered by threshold τ:

$$M(x,y) = \begin{cases} 1, & G(x,y) \geqslant \tau, \\ 0, & G(x,y) < \tau. \end{cases} \tag{1}$$

where the threshold τ is calculated according to Otsu's method [26].

Definition 1. *A connected component C is a set of pixels whose foreground pixel p has a connectivity of 8-connected neighborhoods with an arbitrary pixel q in C.*

Definition 2. *The segmented tongue region T is the largest connected component in the binary mask image M.*

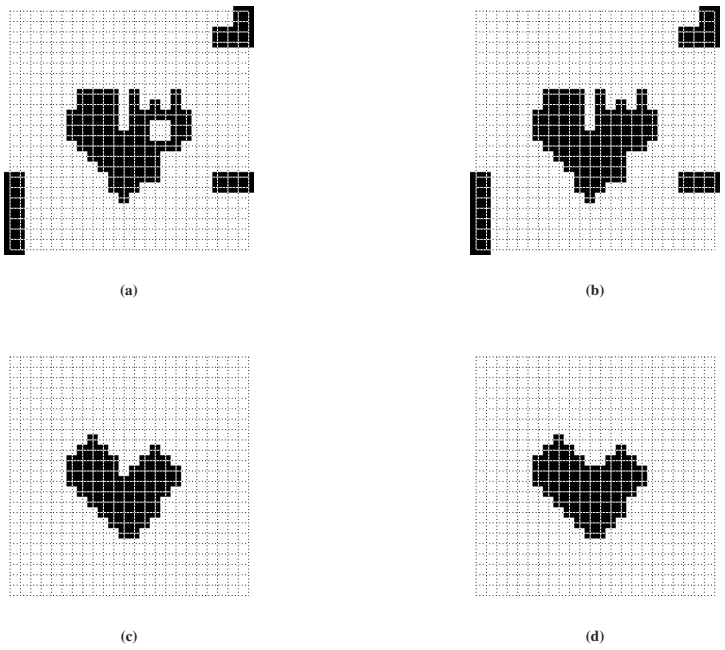

Figure 6. The refinement procedure. (**a**) the mask image with the tongue mask approximately located at the center. As can be observed, there are noises surrounding the tongue and the tongue is coarse with a small 'hole' inside; (**b**) the mask image after using the morphological reconstruction method (Algorithm 1); (**c**) the mask image after using the open operation; (**d**) the mask image after using the close image.

According to Definition 2, a tongue region and is located at the center and is the largest connected component in Figure 6a. Apparently, there is noise at or close to the corners. Moreover, the main body of the mask representing the tongue is not smooth. In addition, some inevitable small 'holes' exist inside the mask. To address these problems, we propose Algorithm 2.

Generally speaking, there are four procedures in the morphological processing layer as shown in Figure 7. The input to the morphological processing layer is a binary image generated from the U-net architecture. In the first procedure, binarization is performed on a coarse result to find all connected components. The second procedure is 'hole' filling, where there are three sub-procedures in this step consisting of reverse image, obtain edges, and morphological reconstruction. After that, the coarse result is overlapped with the result of 'hole' filling. The third procedure is the open operation, which clears all noise outside the tongue region. The fourth procedure is the close

operation, which refines the boundary of a tongue region. Finally, the refined result is output from the morphological processing layer.

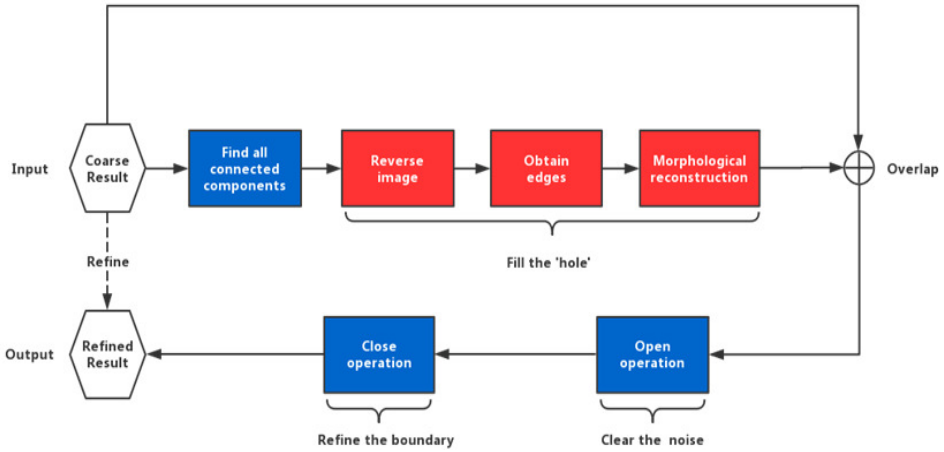

Figure 7. The pipeline of the morphological processing layer. The input (Coarse Result) is a binary image which is generated from U-net. First, all connected components of the Coarse Result are computed. Then, all 'hole' filling procedures (Reverse image, Obtain edges, Morphological reconstruction) are performed. Next, overlap the Coarse Result with the result of 'hole' filling. After that, the Open operation is performed to clear noise. In the next step, the Close operation is performed to refine the boundary of the result. Finally, the Refined Result is the output.

First, for the small 'holes' inside the tongue mask, morphological reconstruction [27] is utilized. In morphological reconstruction, there are two images, one is the maker image, while the other is the template image. Morphological reconstruction is a means of applying morphological transformations to keep the key region defined in the template image via the marker image. Algorithm 1 shows the details of the morphological reconstruction algorithm.

In the first step, we calculate the reverse image R of a given image I to switch hole region to foreground and non-hole region to background:

$$R = 1 - I. \tag{2}$$

Next, we perform edge detection on R, where each pixel belongs to an edge consisting of a point set E. Then, pixels that do not belong to edges are set to zero and we obtain image L:

$$L(x,y) = \begin{cases} 1, & L(x,y) \notin E, \\ 0, & otherwise. \end{cases} \tag{3}$$

Afterwards, L is a marker image for morphological reconstruction in Algorithm 1. Meanwhile, we take R as the template image in Algorithm 1. Based on R and L, we perform morphological reconstruction described in Algorithm 1 on image R to generate image H with the holes filled:

$$H = MR(R, L, c, t), \tag{4}$$

where c is the size of structural element, and t is the tolerance to control the iterative reconstruction operation. The effect of hole filling is shown in Figure 6b. Finally, we overlap the image I and image H:

$$F = I + H. \tag{5}$$

Second, to remove the external noise from the tongue image region, we perform the open operation:

$$M \circ f = \bigcup \{ f_d | f_d \subseteq M \}, \tag{6}$$

where M is the mask image, f is the structural element, and f_d represents the origin point in filter f. The effect of the open operation is shown as Figure 6c. We can find that the noise around the tongue image region is removed and the tongue image region becomes smoother than ever.

Third, to eliminate the internal noise and irregular edge, the close operation is performed:

$$M \circ f = \bigcup \{d | f_d \subseteq W\}, \tag{7}$$

where M is a mask image, f is a structural element, and f_d represents the origin point in filter f. $W = \{d | f_d \cap M \neq \phi\}$ represents the results of M dilated by filter f. The effect of the close operation is shown as Figure 6d. Here, it can be observed that the internal noise is removed.

Algorithm 1: Morphological reconstruction

Input: Template T, Marker L, Size of the structural element r, Tolerance t

Output: Binary Image I

1 Initialize structural element $K = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}_{r \times r}$ and image $I = \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix}_{m \times n}$

2 **do**

3 $\quad I_{p+1} = (I_p \oplus K) \cap T$

4 **while** $\sum_{i+1}^{n} \sum_{j+1}^{m} |I(i,j) - T(i,j)| \geqslant t$

Algorithm 2: Morphological processing layer

Input: Mask Image M, Threshold γ, Open operation Filter size ρ, Close operation filter size σ

Output: Binary Image G

1 Initialize structural elements: $s_1 = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}_{\rho \times \rho}$, $s_2 = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}_{\sigma \times \sigma}$.

2 Perform global binarization on M using Equation (1) to obtain connected components.

3 Generate binary image R according to Equation (2).

4 Perform edge detection using the prewitt operator, set S as the set of pixels from the edge.

5 Reserve pixels of S in R according to Equation (3).

6 Generating hole-filled image H by using using Equation (4).

7 Overlapping binary mask image I and hole-filled image H according to Equation (5).

8 Perform the open operation using s_1 according to Equation (6): $O = open(F, s_1)$.

9 Perform the close operation using s_2 according to Equation (7): $G = close(O, s_2)$.

10 Return G.

3. Experimental Results

In this section, we perform segmentation experiments using our proposed system to demonstrate its effectiveness and efficiency. First, the dataset used in the experiments is described followed by a discussion of the experimental setup. Afterwards, qualitative evaluation is performed to directly show the results on different groups of tongue images to evaluate the quality of the segmentation results in an intuitive way. Next, quantitative evaluation is carried out using five different metrics to evaluate the performance of the proposed system in different aspects to prove the superiority of proposed method in a numerical way. Later on, TongueNet is compared with Floodfill [28] and Snake [15] as well as a state-of-the-art method entitled Region-based and Edge-based fusion approach (REF) [18]. Finally, an efficiency comparison is carried out to compute the computation time.

3.1. Dataset Description

The tongue image dataset contains 1000 samples captured by a device described in [8]. The images were stored in Bitmap (BMP) format and divided into various classes of diseases. Each image has RGB channels in the depth of 24 bits. The resolution of each image is 576 × 768. Some samples of the tongue image are displayed in Figure 8. An overview of the dataset is summarized in Table 2. For the background of each image (in Figure 8), there are a variety of non-tongue artifacts: chin rest of the capture device, cheeks, nose, lips, along with various teeth. For the foreground of each image, the tongues are presented in different positions and poses, textures, shapes, and colors. More information on the dataset can be found in [10].

Figure 8. Samples of tongue image dataset.

Table 2. Overview information of the tongue image dataset.

Size	Resolution	Format	Bit Depth	Color Model
1000	576 × 768	Bitmap	24	RGB

There are two steps in the training preparation. First, the dataset is split into two parts: a training set and a test set. The number of images in the training set and test set is 800, while the number of images in the test set is 200. Afterwards, each mask of the training set is generated manually using the labelme [29] annotation tool and then converted into a 576 × 768 bitmap file with two bits of bit depth. The mask image is a binary image, whose pixels in the foreground are 255 (white), while the pixels in the background are 0 (black). An overview of the information of the mask set is summarized in Table 3. The mask images corresponding to the samples in Figure 8 are shown in Figure 9. In Figure 9, the region in white is the foreground (255), which covers the area of the tongue and draws the contours precisely. The region in black is the background (0).

Table 3. Overview information of a the mask set corresponding to the tongue images in Figure 8.

Size	Resolution	Format	Bit Depth	Color Model
800	576 × 768	Bitmap	2	Binary (255-foreground, 0-background)

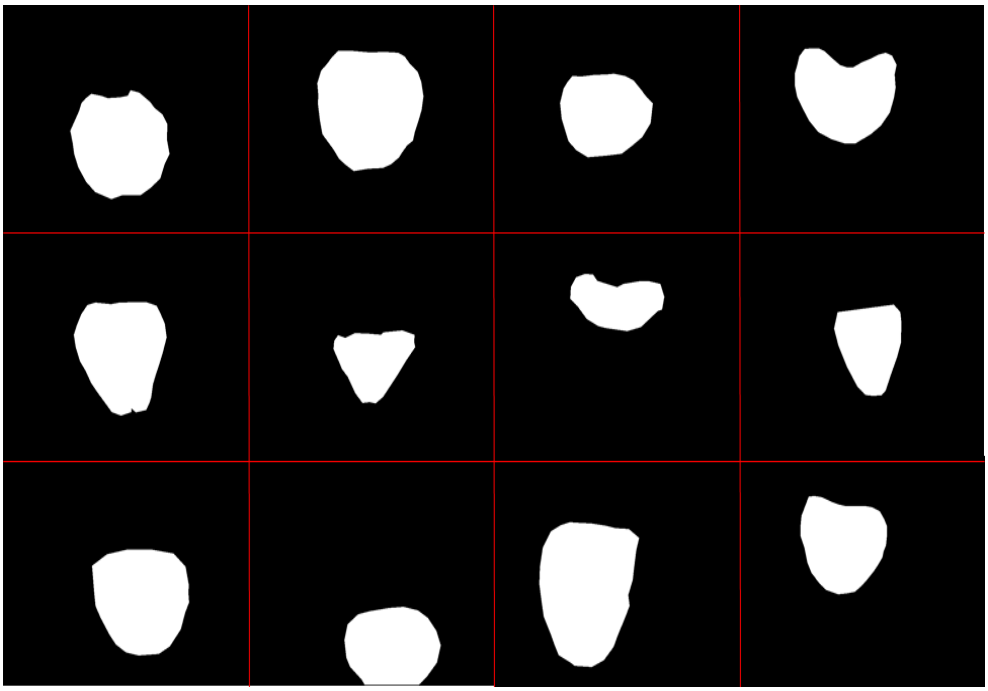

Figure 9. Samples from the mask set corresponding to the tongue images in Figure 8.

3.2. Experimental Setup

The experiments were performed on the 200 test samples from the tongue dataset (refer to Section 3.1 and Table 2). All experiments were conducted on a PC with a 3.40 GHz Intel Core i7-2600 CPU and 12 GB of RAM. The network architecture is built on a tensorflow 1.13 platform and using a NVIDIA GEFORCE GTX 1070 Ti graphics card. For the parameters used in training, we set the learning rate $\alpha = 10^{-4}$, batch size $b = 20$, number of epoch $N = 100$, and applied the Adam optimizer [30]. In the morphological layer, we set the open operation filter size $\rho = 20$, the close operation filter size $\sigma = 5$, the filter size in morphological reconstruction $r = 3$, and the tolerance $t = 0.001$.

3.3. Qualitative Evaluation

The qualitative evaluation was performed on various groups of tongue images with different circumstances to assess the robustness and effectiveness of the proposed system. As mentioned in Section 1, tongue images can be divided into the following groups: (1) Tongue with an apparent gap in the mouth, (2) Tongue with abnormal color, (3) Tongue with abnormal texture, (4) Tongue with teeth showing, (5) Tongue with irregular poses, (6) Tongue not completely protruding, (7) Tongue with teeth imprints on the edges, and (8) Tongue closely surrounded by lips. Figure 10 shows the segmentation results of the above groups on some samples from the dataset. The highlighted regions in each sub-figure are the segmentation results. As can be seen in each group, tongue regions are segmented precisely by TongueNet under different circumstances, which demonstrates its robustness. In particular, the edge pixels are predicted well, indicating that TongueNet is capable of handling various changes on the tongues. Furthermore, the segmented tongue regions are well separated from its surroundings, such as the lips and teeth. Overall, the proposed TongueNet system achieves a promising segmentation performance.

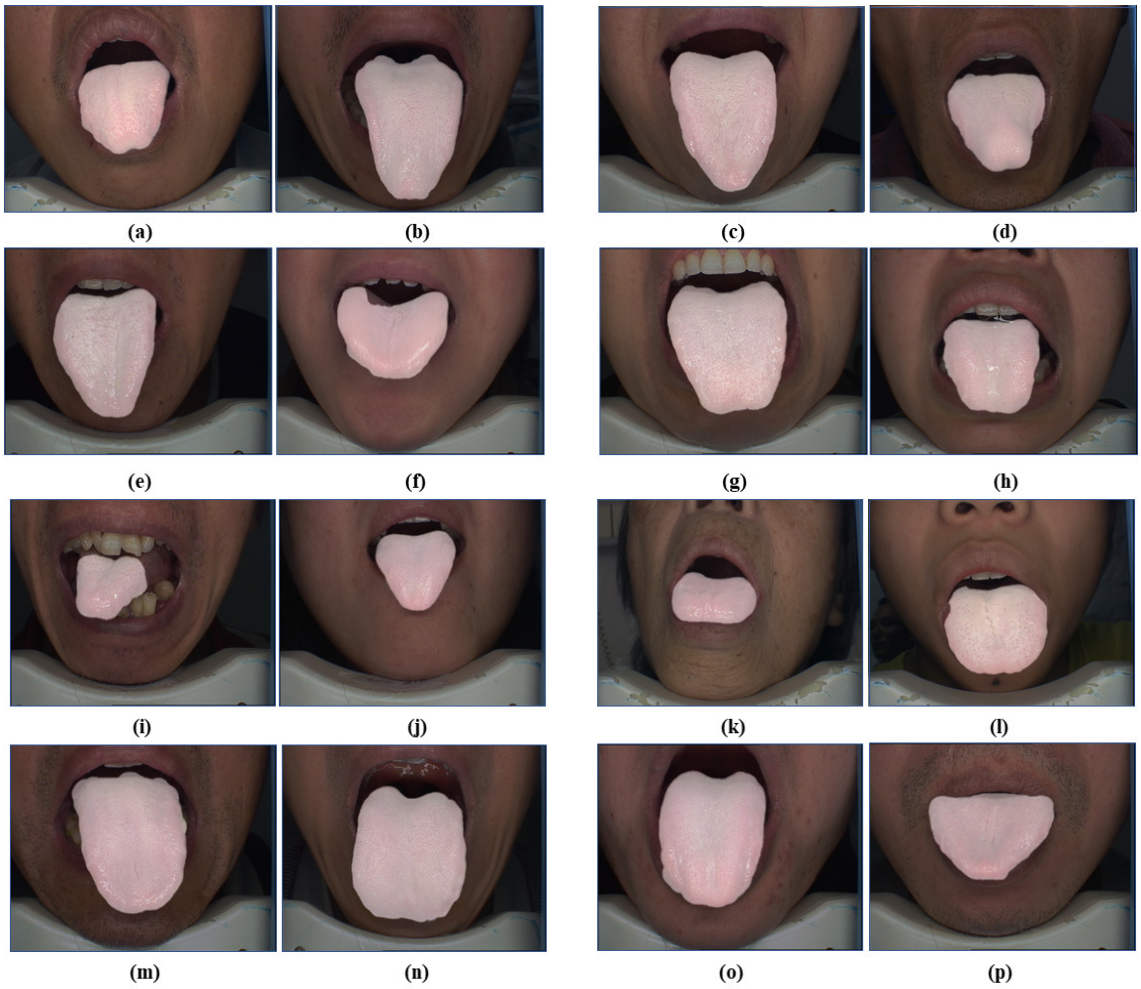

Figure 10. Segmentation results of TongueNet in qualitative evaluation (highlighted). (**a**) and (**b**) results of a tongue with an apparent gap in the mouth; (**c**) and (**d**) results of a tongue with abnormal color; (**e**) and (**f**) results of a tongue with abnormal texture; (**g**) and (**h**) results of a tongue with teeth showing; (**i**) and (**j**) results of a tongue with irregular poses; (**k**) and (**l**) results of a tongue not completely protruding; (**m**) and (**n**) results of a tongue with teeth imprints on the edges; (**o**) and (**p**) results of a tongue closely surrounded by lips.

Figure 11 displays eight pairs of comparison groups with the with the proposed TongueNet (lowercase) and the results of REF [18] (uppercase). Noticeably, we made comparisons with REF as it is the most representative and currently the best tongue-oriented segmentation method among various state-of-the-art methods. From Figure 11a–e,g,A–E,G, it can be observed that TongueNet exactly segmented the tongue without teeth and lips, while REF failed to produce a clean segmentation. This clearly shows the superiority and robustness of the proposed method when the surroundings are complicated. If the color, shape, and texture of the tongue are abnormal, TongueNet can also perform a smooth segmentation (refer to Figure 11f,F,h,H).

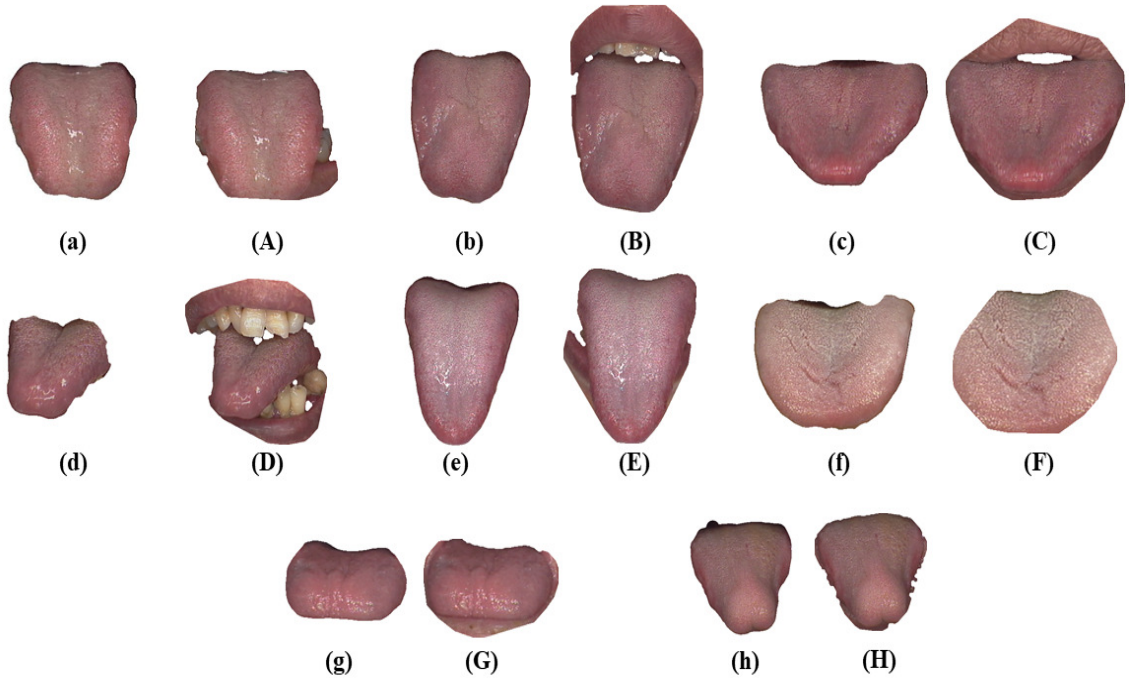

Figure 11. Comparison of a segmented tongue image from TongueNet (lowercase) and REF (uppercase).

3.4. Quantitative Evaluation

In this sub-section, quantitative evaluation is performed to evaluate the effectiveness of TongueNet a numerical way. To assess the performance in different aspects, five metrics are utilized, which are Intersection over Union (IOU) [31] , Pixel Accuracy (PA) [25], Precision [32], Recall [32], and F-score [33]. We first give an introduction to these five metrics. Then, we compare TongueNet and REF under the metric of IOU, PA, F-score, Precision and Recall on a subset of 20 test samples covering all circumstance summarized in Section 1 (refer to Figure 1). Then, we provide comparisons between TongueNet and other state-of-the-art methods by calculating the average value of all test samples. Afterwards, an efficiency comparison between REF and TongueNet is carried out.

3.4.1. Metrics

For pixels p_{ij} in a binary image, $p_{ij} = 1$ means p_{ij} belongs to the positive class, and $p_{i,j} = 0$ means p_{ij} belongs to a negative class. Given a ground truth image G and an image, we predict P. The IOU metric can be defined as follows:

$$IOU = \frac{|GR \cap PR|}{|GR| + |PR| - |GR \cap PR|}, \tag{8}$$

where GR and PR represent the set of positive pixels in image G and P, $|GR|$ and $|PR|$ represent the number of pixels in GR and PR. The IOU measures the similarity between a ground truth image and a prediction image by calculating the number of pixels in the intersection set divided by the number of pixels in union set excluding the intersection region.

We denote TP, FP, TN, FN as true positive pixels, false positive pixels, number of true negative pixels and false negative pixels, respectively. The Pixel Accuracy (PA) is defined as follows:

$$PA = \frac{TP}{|P|}. \tag{9}$$

The PA measures the percentage of true positive pixels from all pixels in an image, providing an overall assessment of prediction.

Besides the PA, the Precision metric is utilized to measure the accuracy of positive pixels predicted in P, which is defined as follows:

$$Precision = \frac{TP}{|TP + FP|}. \tag{10}$$

The Recall metric is utilized to measure the degree to which all the positive pixels are predicted correctly:

$$Recall = \frac{TP}{|TP + TN|}. \tag{11}$$

Based on Precision and Recall, the F-score is calculated, to measure the performance comprehensively:

$$F - score = 2 \cdot \frac{precision \cdot recall}{precision \cdot recall} = \frac{2TP}{2TP + FN + FP}. \tag{12}$$

3.4.2. IOU Results on 20 Test Samples

The IOU results of TongueNet and REF are displayed in Figure 12. The proposed TongueNet achieved an IOU of over 90% on almost all samples with the highest value reaching 96.38% (the 10th sample). This indicates that most of the pixels in the tongue region were predicted precisely. Furthermore, TongueNet outperforms REF on all samples with the most significant difference being 33.7% (the 14th sample). To present an overall perspective of IOU, the mean IOU (MIOU) is 93.11% for TongueNet, while the MIOU of the REF method is 83.1%. In general, more pixels in the tongue region will be predicted correctly by TongueNet than REF.

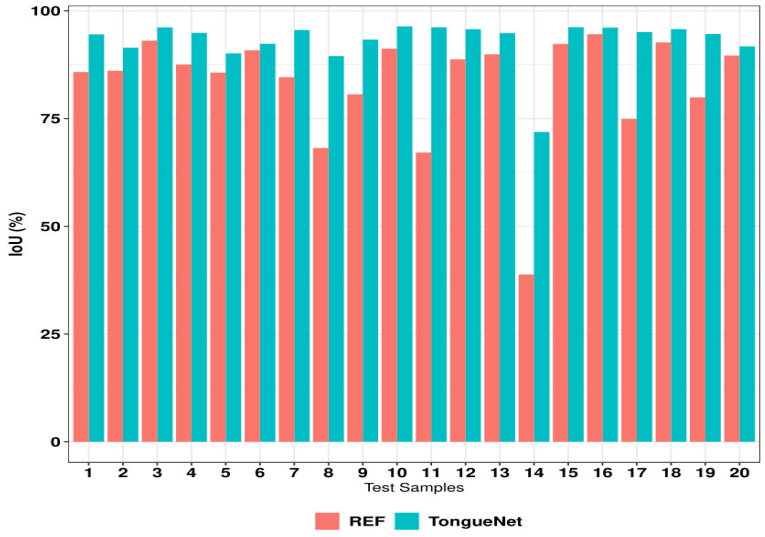

Figure 12. IOU comparison of TongueNet (green) and REF (red).

3.4.3. PA Results on 20 Test Samples

The PA comparison of TongueNet and REF on 20 test images is illustrated in Figure 13. The proposed TongueNet obtained over 97% for all test samples, with the highest being 99.33% (the 9th sample). This means TongueNet can correctly predict on average over 97% of the pixels from the images in the test set, indicating that TongueNet holds a high prediction accuracy. Similar to the IOU results (above), TongueNet also outperformed REF on all test samples, where the mean PA (MPA) of TongueNet is 98.44%, while the MPA of REF is 96.8%. Thus, it proves that TongueNet has a stronger ability at pixel prediction.

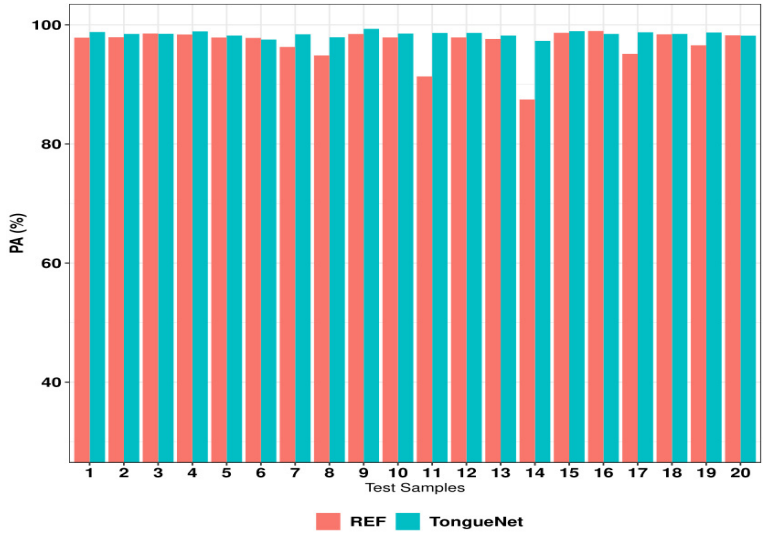

Figure 13. PA comparison of TongueNet (green) and REF (red).

3.4.4. F-score, Precision, and Recall Results on 20 Test Samples

The F-score (left), Precision (middle), and Recall (right) between the proposed method and REF for 20 test images are shown in Figure 14. From the radar charts in this figure, it can be observed that TongueNet outperforms REF in Precision and Recall on almost all 20 test samples. This reflects the high reliability as well as the comprehensiveness of tongue region prediction using the proposed method. Although some Recall values of TongueNet in Figure 14 (right) are lower than REF (the 11th sample and the 14th sample), the F-score of each test sample is still higher, indicating the general superiority of TongueNet over REF.

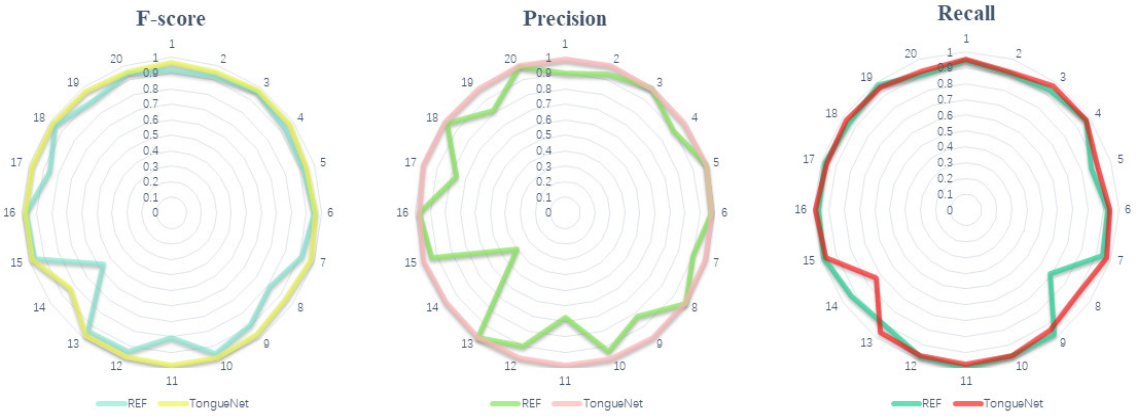

Figure 14. F-score, precision and recall comparison of TongueNet and REF using radar charts. (**left**) F-score; (**center**) precision; (**right**) recall.

3.4.5. Overall Comparison with Other Segmentation Methods

A comparison with other methods (REF [18], Snake [15], Flood fill [28]) using all 200 test samples is carried out to ensure adaequate experimentation in proving the effectiveness of TongueNet in a quantitative way (refer to Table 4). The mean IOU (MIOU), mean PA (MPA), mean F-score (MFscore), mean Precision (MPrecision), and mean Recall (MRecall) are utilized. The mean value is the average of all test samples:

$$Mmetric = \frac{1}{n}\sum_{i=1}^{n} metric(i), \tag{13}$$

where $metric(i)$ represents the ith value of a specific metric and n represents the number of test samples. The highest value of each metric is marked in bold font.

Table 4. Comparison of TongueNet with other segmentation methods (the significance level $\beta = 0.05$)

Methods	MIOU (%)	MPA (%)	MF-score (%)	MPrecision (%)	MRecall (%)
Snake	69.30 ± 0.05	93.51 ± 0.02	81.44 ± 0.08	83.68 ± 0.05	83.30 ± 0.03
Flood fill	44.79 ± 0.04	90.95 ± 0.01	59.11 ± 0.06	89.59 ± 0.04	44.90 ± 0.04
REF	83.10 ± 0.06	96.81 ± 0.01	92.13 ± 0.06	88.94 ± 0.03	93.45 ± 0.04
TongueNet	$\mathbf{93.11 \pm 0.02}$	$\mathbf{98.45 \pm 0.01}$	$\mathbf{95.38 \pm 0.03}$	$\mathbf{98.72 \pm 0.02}$	$\mathbf{94.26 \pm 0.01}$
p-value	**0.002**	**0.002**	**0.010**	**0.004**	**0.015**

From Table 4, it can be seen that the proposed TongueNet outperforms other methods, especially in MIOU, which is over 10% higher than others on average, implying a higher accuracy in tongue region prediction. In terms of MF-score, the gap between TongueNet and other comparison methods range from 3.25% to 36.27%, indicating the superiority of TongueNet (from a comprehensive perspective way). As for MPA, MPrecision, and MRecall, the improvements achieved by the proposed method varied from 1.63–7.50%, 9.13–15.04%, and 0.81–49.36%, respectively. All three of these metrics proved the superiority of TongueNet not only in reliability and completeness, but also in overall pixel prediction. In the statistical significance test, we set the significance level to be $\beta = 0.05$, where the p-values for different metrics have been calculated to show the significance difference of TongueNet and REF.

3.5. Efficiency Comparison

The computation time is calculated to evaluate the efficiency of TongueNet. The comparison of TongueNet and REF is displayed in Table 5. To reflect the computation time, we calculated the average time consumed per image after applying TongueNet and REF. The results show that TongueNet requires less than 0.3 s compared to REF, which used 0.302s. The training time of TongueNet is 2435 s.

Table 5. Computation time comparison between TongueNet and REF.

Methods	Computational Time (s)
REF	0.302
TongueNet	**0.267**

4. Discussion

To fully analyze and discuss the details in the methodology and experimental results presented in the previous sections, we can reach the following inferences:

1. The comparisons above (in the experimental results) show the effectiveness as well as the efficiency of TongueNet in tongue segmentation with the superiority of robustness in many complicated circumstances, which makes it a promising tool when integrated with TCTD. The proposed TongueNet achieves a better segmentation performance compared with other state-of-the-art methods on both qualitative and quantitative evaluations (refer to Figure 11 and Table 4). Furthermore, TongueNet uses less computation time compared with REF (see Table 5), indicating a much faster processing speed.
2. Although TongueNet performs better at segmentation in general, the predictions of the pixels still need improvement as it is not complete in some occasion in terms of quantitative evaluation. From Table 4, even if the mean F-score of TongueNet (95.38%) is higher than other methods, the Recall values of some samples (from the 200 test images) are lower than REF (e.g., the 11th sample: TongueNet-97.94%, REF-99.77% Figure 14). This indicates indicating that the prediction of pixels cannot entirely cover all tongue regions. The cause of this is due to a higher loss of the

boundary in the tongue. Therefore, more focus will be placed at the boundary loss as part of our future work.

3. The parameters in the morphological layer determine the effectiveness of the refinement as well as the efficiency of the whole system. Since there are three filters (morphological reconstruction, open operation and close operation) (refer to Section 2.3) utilized to perform morphological processing for different purposes, the system sensitivity towards parameters is high. There exists a trade-off between the effectiveness and efficiency. To overcome the high sensitivity and trade-off, we fine-tune the system using different parameters and combinations, including the filter kernel size, type of operator, tolerance of morphological reconstruction, and the global threshold to generate the binary mask.

5. Conclusions

In this paper, we proposed a precise and fast tongue segmentation system named TongueNet. The system captures an individual's tongue image and uses it as input before outputting a segmented tongue. Within the system, there are four procedures consisting of tongue image acquisition, image grey-scaling, segmentation model prediction, and segmented image extraction. To train a powerful prediction model with less data, a fine-tuned U-net is selected as the backbone. Moreover, a newly proposed morphological processing layer is proposed to perform segmentation refinement. The experimental results show that TongueNet achieved the highest mean IOU—93.11%, mean PA—98.45%, mean F-score—95.38%, mean Precision—98.72%, and mean Recall—94.26% compared with other state-of-the-art tongue segmentation methods. Furthermore, it can accurately segment tongue images where teeth and lips are present, as well as cases with irregular tongue shapes, colors, and textures. With a computation time of 0.267 seconds per image, TongueNet has the potential to be fully integrated as part of any TCTD system.

Author Contributions: For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used "conceptualization, J.Z., Q.Z. and B.Z.; methodology, J.Z. and Q.Z.; software, J.Z. and Q.Z.; validation, J.Z.; formal analysis, J.Z.; investigation, J.Z.; resources, B.Z.; data curation, B.Z.; writing—original draft preparation, J.Z.; writing—review and editing, J.Z., Q.Z., X.C. and B.Z.; visualization, J.Z.; supervision, B.Z.; project administration, J.Z.; funding acquisition, B.Z.

Funding: This research was funded by the University of Macau Grant No. MYRG2018-00053-FST.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

TCM	Traditional Chinese Medicine
TCTD	Traditional Chinese Tongue Diagnosis
BEDC	Bi-Elliptical Deformable Contour
FCN	Fully Conolutional Neural Network
BMP	Bitmap
RGB	RGB color model
REF	Region-based and Edge-based fusion approach
IOU	Intersection over Union
PA	Pixel Accuracy
TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative
MIOU	Mean IOU
MPA	Mean PA

MPrecision	Mean Precision
MRecall	Mean Recall
MFscore	Mean F-score

References

1. Press, C.U. *Cambridge Academic Content Dictionary*; Cambridge University Press: Cambridge, UK, 2017.
2. Maciocia, G. *Tongue Diagnosis in Chinese Medicine*; Eastland: Seattle, WA, USA, 1995.
3. Kirschbaum, B. *Atlas of Chinese Tongue Diagnosis*; Eastland: Seattle, WA, USA, 2000.
4. Zhao, Q.; Zhang, D.; Zhang, B. Digital tongue image analysis in medical applications using a new tongue ColorChecker. In Proceedings of the 2016 2nd IEEE International Conference on Computer and Communications (ICCC), Chengdu, China, 14–17 October 2016; pp. 803–807.
5. Zhang, H.; Zhang, B. Disease detection using tongue geometry features with sparse representation classifier. In Proceedings of the 2014 International Conference on Medical Biometrics, Shenzhen, China, 30 May–1 June 2014; pp. 102–107.
6. Zhang, B.; Nie, W.; Zhao, S. A novel Color Rendition Chart for digital tongue image calibration. *Color Res. Appl.* **2018**, *43*, 749–759. [CrossRef]
7. David, Z.; Hongzhi, Z.; Bob, Z. *Tongue Image Analysis*; Springer: Berlin, Germany, 2017.
8. Zhang, H.; Wang, K.; Zhang, D.; Pang, B.; Huang, B. Computer aided tongue diagnosis system. In Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, Shanghai, China, 17–18 January 2006; pp. 6754–6757.
9. Zhang, B.; Zhang, H. Significant geometry features in tongue image analysis. *Evid.-Based Complement. Altern. Med.* **2015**, *2015*, 897580. [CrossRef] [PubMed]
10. Wang, X.; Zhang, B.; Yang, Z.; Wang, H.; Zhang, D. Statistical analysis of tongue images for feature extraction and diagnostics. *IEEE Trans. Image Process.* **2013**, *22*, 5336–5347. [CrossRef] [PubMed]
11. Zhang, B.; Wang, X.; You, J.; Zhang, D. Tongue Color Analysis for Medical Application. *Evid.-Based Complement. Altern. Med.* **2013**, *2015*, 264742. [CrossRef] [PubMed]
12. Zhang, B.; Kumar, B.V.; Zhang, D. Detecting Diabetes Mellitus and Nonproliferative Diabetic Retinopathy Using Tongue Color, Texture, and Geometry Features. *IEEE Trans. Biomed. Eng.* **2014**, *61*, 491–501. [CrossRef] [PubMed]
13. Pang, B.; Zhang, D.; Wang, K. The bi-elliptical deformable contour and its application to automated tongue segmentation in Chinese medicine. *IEEE Trans. Med. Imaging* **2005**, *24*, 946–956. [CrossRef] [PubMed]
14. McInerney, T.; Terzopoulos, D. Deformable models in medical image analysis: a survey. *Med. Image Anal.* **1996**, *1*, 91–108. [CrossRef]
15. Kass, M.; Witkin, A.; Terzopoulos, D. Snakes: Active contour models. *Int. J. Comput. Vis.* **1988**, *1*, 321–331. [CrossRef]
16. Ning, J.; Zhang, L.; Zhang, D.; Wu, C. Interactive image segmentation by maximal similarity based region merging. *Pattern Recognit.* **2010**, *43*, 445–456. [CrossRef]
17. Ning, J.; Zhang, D.; Wu, C.; Yue, F. Automatic tongue image segmentation based on gradient vector flow and region merging. *Neural Comput. Appl.* **2012**, *21*, 1819–1826. [CrossRef]
18. Wu, K.; Zhang, D. Robust tongue segmentation by fusing region-based and edge-based approaches. *Expert Syst. Appl.* **2015**, *42*, 8027–8038. [CrossRef]
19. Liu, Z.; Yan, J.Q.; Zhang, D.; Li, Q.L. Automated tongue segmentation in hyperspectral images for medicine. *Appl. Opt.* **2007**, *46*, 8328–8334. [CrossRef] [PubMed]
20. Zhang, D.; Zhang, H.; Zhang, B. A Snake-Based Approach to Automated Tongue Image Segmentation. In *Tongue Image Analysis*; Springer: Berlin, Germany, 2017; pp. 71–88.
21. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436. [CrossRef] [PubMed]
22. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 431–3440.
23. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Berlin, Germany, 2015; pp. 234–241.

24. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]
25. Garcia-Garcia, A.; Orts-Escolano, S.; Oprea, S.; Villena-Martinez, V.; Martinez-Gonzalez, P.; Garcia-Rodriguez, J. A survey on deep learning techniques for image and video semantic segmentation. *Appl. Soft Comput.* **2018**, *70*, 41–65. [CrossRef]
26. Otsu, N. A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man, Cybern.* **1979**, *9*, 62–66. [CrossRef]
27. Gonzalez, R.C.; Woods, R.E. *Digital Image Processing*, 3rd ed.; Prentice-Hall, Inc.: NJ, USA, 2006.
28. Torbert, S. *Applied Computer Science*; Springer: Berlin, Germany, 2012.
29. Wada, K. labelme: Image Polygonal Annotation with Python. 2016. Available online: https://github.com/wkentaro/labelme (accessed on 31 July 2019).
30. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
31. Kosub, S. A note on the triangle inequality for the jaccard distance. *Pattern Recognit. Lett.* **2019**, *120*, 36–38. [CrossRef]
32. Powers, D.M. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *J. Mach. Learn. Technol.* **2011**, *2*, 37–63.
33. Sasaki, Y. The truth of the F-measure. *Teach Tutor Mater* **2007**, *1*, 1–5.