*i*Journals: International Journal of Software & Hardware Research in Engineering (IJSHRE)
ISSN-2347-4890
Volume 10 Issue 2 February 2022

Comparative Analysis of Tongue Segmentation Techniques

Authors: Vibha Bhatnagar[1]; Gauri Gupta[2]; Veerendra Patel[3]

Affiliation: Department of Biomedical Engineering, Shri G. S. Institute of Tech. & Science, 23 Sir M. Visvesvaraya Marg, Indore, Madhya Pradesh 452003, India. R.G.P.V University

E-mail: bhatnagarvibha09@gmail.com[1], gauri2180@gmail.com [2], veerendrapatel50@gmail.com [3]

ABSTRACT

Tongue analysis for disease diagnosis is an integral part of traditional medicine in Asian countries. To overcome the subjective diagnosis based on expert clinician's knowledge, automation of the entire analysis for quantifying the features for diagnosis is required. For achieving successful quantization accurate extraction of the tongue body from complete image is first step towards computer aided analysis system. In this paper different traditional methods are employed for emphasizing the usefulness of deep learning techniques which prove to be more accurate less time consuming and robust.

Keywords: — Residual U-Net, U-Net, tongue segmentation, computer aided analysis system feature extraction.

1. INTRODUCTION

Challenges associated with tongue body segmentation are interference of pathological details with the tongue body, different size and shapes, blurred edges and similar intensity values of surrounding area. Traditional Methods for segmentation roughly falling under basic classes such as thresholding, edge detection, graph theory, level set, active contours gave fairly good accuracy, but had certain limitations associated along-with them. Any image processing task if performed on similar resolution images captured by high resolution device no doubt will be much easier than images captured in different illuminations with different smartphone cameras. Tongue analysis for disease diagnosis is being practiced in Ayurveda along with other Asian traditional medicine practices. No standard dataset covering all diseases that could be predicted with respect to tongue features is available. Over the period of two decades researchers have put in tremendous efforts to predict either some specific tongue feature or disease.

Segmentation using conventional methods fail to give satisfactory results, whereas machine learning methods are successful in giving fairly accurate results. Deep Convolution Neural Networks present an outstanding ability of feature learning and representation, with viable solutions for automatic, generalizable and efficient semantic image segmentation. There are some key challenges associated with medical image segmentation such as unavailability of large number of annotated and labelled dataset for training the model, one of the basic requirements of deep learning models to give accurate results, lack of standard segmentation protocol and huge variations of images among patients. With the existing challenges in this domain, a robust and efficient technique for segmenting the body organ or anomaly under investigation was of utmost priority for automation in biomedical applications. In this paper we have experimented with traditional segmentation algorithms like Watershed, Canny and Snake method and finally used Deep learning model U-Net with variation in the Neural block incorporating ResNet (Res U Net) for comparative analysis.

Olaf Ronnerberg, et.al. [1] designed a network whose architecture visually matched alphabet 'U', hence named it U-NET. Network architecture consisted of a contracting path for contextual features followed by a symmetric expanding path for precise localization. U-Net performed well with small, augmented dataset. In this paper U- Net model is used to train tongue image dataset and results are compared with the results obtained with U-NET with Different encoder blocks. Qualitative and Quantitative comparison are presented. Section 2 contains related work in this field, Section 3 gives Methodology, Section 4 -Results, Section 5- Discussion, Section 6- Conclusion.

2. RELATED WORK

Deep learning techniques have shown impressive performance in terms of speed of computation,

complexity and segmentation results with respect to the traditional methods. Jiang Li, et.al. [2] segmented the tongue region from the image by enhanced HSV colour model Convolution Neural Network. To obtain clear edges RGB image after being converted to HSV model was passed through Contract Limited Adaptive Histogram Equalization (CLAHE). Their model tested on a dataset of 264 images taken by digital camera and results compared with Snakes Model. Results of comparison showed better performance of their model enhanced HSV -CNN over Snakes, especially processing time drastically decreased to 0.0275 sec from time of 3.1355 sec for Snakes model. Yushan Xue, et.al. [3] used fully convolution networks (FCN) for tongue body segmentation of images of size 379 x 489 captured by customized equipment. Performance comparison of their model of FCN-8s was evaluated with that of Deep Lab V3, Deep Lab V3 Plus Learning Based Matting (LBM) resulting in highest Mean Intersection of Union (mIOU) of 93.7% being achieved with Deep Lab V3 against a value of 90.48% with FCN-8s. Though Fully convolution model was fastest amongst the three, they concluded Deep Lab V3 to have better performance on quantitative analysis.

Bingqian Lin, et.al. [4] proposed a model that did not have tight constraints regarding the image quality regarding its size and illumination during acquisition, in other words did not require any prior pre-processing. The Deep CNN put forward by them was based on RES- Net (Residual Neural Network) with different number of layers. Primarily they tested their model on 2344 images captured by cell phone, with different number of layered Res-Net models and compared to traditional Grab -Cut model. From all the different number layer architecture considered, 50 layered Res-Net model showed superior performance. Wei Yuan, et.al. [5] quoted to have developed light weight with better accuracy model over their counterpart higher performance algorithms. A labelled annotation method was also developed along with for reducing manual work for annotation on dataset. Model was tested on 5616 images taken by digital camera. Researchers claim that their model could contribute extensively to automatic remote applications and performance can be further enhanced by exploring use of sample mining and hyper parameter optimization.

XinLei Li, et.al. [6] designed a light weight Encoder -Decoder Architecture, tested on dataset of 5,600 images (FDU/SHUTCM) generated by them and also on publicly available dataset BIOHIT (12 images) and PolyU/HIT (300 images). Segmentation accuracy achieved for their dataset is quoted to be 99.15%. Model architecture consisted of TIFE (Tongue Image Feature Extraction) to extract features with larger receptive fields without loss of spatial resolution, whereas a Context Model is used to increase the performance by aggregating multiscale contextual information. Decoder is designed as simple yet efficient feature up sampling module fusing different depth features and refining segmentation results along tongue boundary. Misclassification error due to class imbalance is also taken care of in loss module.

Xiadong Huang, et.al. [7] developed an enhanced fully convolution network with encoder-decoder structure. Encoder consisted of Deep Residual Network for dense feature maps followed behind by Receptive Field Block to capture sufficient global contextual information. Decoder module incorporated feature Pyramid Network to fuse multiscale feature maps to acquire sufficient positional information to recover the contour of tongue body. Results delivered sensitivity of 98.97% with average Dice Similarity Coefficient of 97.26% on a dataset comprising of 700 images. Qichao Tang, et.al. [8] compared performance of deep learning-based model Mobile Net V2 in combination with Single Shot Multibox Detector (SSD) with conventional method Haar like feature-based detection algorithm. When applied on a dataset of 798 images deep learning based model showed better sensitivity.

3. METHODOLOGY

Tongue area extraction from the raw image is a challenge as surrounding skin, lips and teeth need to be eliminated for accurate analysis of features for disease diagnosis. Conventional method using Canny edge detector, Watershed transform, Snake active contour is used to extract the region of interest. Results achieved are not satisfactory plus since each image is different with respect to resolution and quality, a single parameter setting is not justified for every image. This motivated to use Machine Learning models for better performance. U-Net architecture is the base line model used, experimented with pretrained ResNet as Neural blocks (ResU-Net) model. Results are compared for performance analysis

3.1 Overview of Edge Detection Techniques and Model Architecture

3.1.1 Canny Edge Detector

In this method optimization of a functional is achieved by finding a function is, generally it is sum of four exponential terms. It can be approximated by Gaussian first derivative. Processing steps involved requires firstly to smooth out the image and remove noise using a gaussian filter. Then intensity gradients of the image are derived. Next thresholding and double thresholding applied to determine potential edges. Finally, edges are tracked by hysteresis function and unwanted edges removed

3.1.2 Snake Edge Detector

It is an active contour model which works on energy minimizing, deformable planes influenced by constraint and image forces that pull it towards object contours whereas internal forces resist deformation.

Snakes – active deformable mods. Snakes requires knowledge of the desired contour shape beforehand hence not able to solve the entire problem of finding contours in images. They depend on interaction with a user, interaction with some higher-level image understanding process, or information from image data adjacent in time or space.

3.1.3 Watershed Algorithm

It is a classical image segmentation technique based on the concept of watershed transformation. In this method the similarity with adjacent pixels of the image are considered as an important reference to connect pixels with similar spatial positions and intensity values. It performs well in case of segmentation of overlapping or touching objects, exceling in irregular shape objects, gradient based segmentation requirements.

3.1.4 U-Net Architecture

U-NET Architecture developed by Olaf Ranneberger, et.al. [1] a deep neural network architecture consisting of a contracting path for context capturing and a symmetric expanding path that enables precise localization. U-NET architecture mainly consists of two parts analysis (encoder) module and synthesis(decoder) module. Training strategy of the proposed architecture relied on data augmentation of annotated samples in efficient manner. That is, it showed good performance even in case of small data set as opposed to the basic requirement of large dataset for effective training of deep networks. One important characteristic of their model is that while up-sampling there are large number of connected feature channels that enable propagation of contextual information to higher resolution layers. Due to symmetric expansion as in Fig.1 and contracting path of the network it takes a U-shaped architecture.

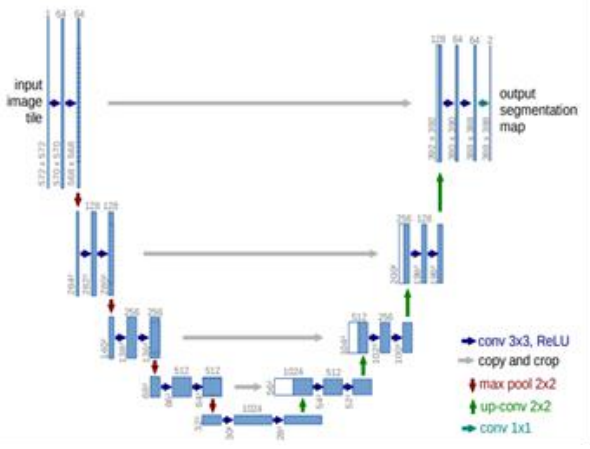


Fig 1: Unet Architecture

3.2 Experimental Methods

This section elaborates upon the experiment performed using conventional and deep learning methods to demonstrate their efficiency and effectiveness. First, the dataset used in the experiment is described followed by some insight into the experimental hardware. Next a qualitative evaluation of different model in an intuitive manner on different groups of tongue images is presented. Finally, metrics considered to evaluate the performance of the applied architecture for quantitative analysis are elaborated

3.2.1 Dataset Description

The tongue image dataset used for the experiment consists of 300 images captured by different Mobile phone in varied environment and locations. For deep learning model training use of Labelme Software to annotate the image and generate a binary mask is done. Data augmentation of images and mask by centre cropping, horizontal flip and grid distortion was done to finally get augmented dataset of 636 images in total. The models also tested on publicly available HIT dataset with 300 images of size 768 x576 bitmap images. First, we performed data augmentation on the available 300 images and mask by centre cropping, horizontal flip and grid distortion to generate augmented dataset. Dataset hereby is split into following sets training set consisting of 1188 images and validation and testing dataset 118 images each

3.2.2 Experimental Setup

The experiments were performed with hp Pavilion laptop with a 1.60 GHz intel i5 8 th generation processor and 8 GB of Ram. Training of the model was done on Google Colab Python 3 Google Compute Engine backend GPU. Parameters set for training are learning rate set to $\alpha = 10\text{-}4$, number of epochs N=300, Early stopping and Reduce LR on Plateau is also used, and applied ADAM Optimizer. Binary cross entropy loss was considered

3.2.3 Qualitative /Quantitative Evaluation

To evaluate the robustness and effectiveness of the applied architecture various groups of tongue images were considered. Tongue images can be divided into various groups such as tongue not completely protruding, tongue with apparent gap in the mouth, tongue with teeth showing and tongue closely surrounded by lips and almost same intensity face area. Fig.2. shows the results of conventional methods on some samples falling in above mentioned groups. These techniques detect the tongue edges, as clearly visible that none of the method could accurately mark precisely the tongue edge

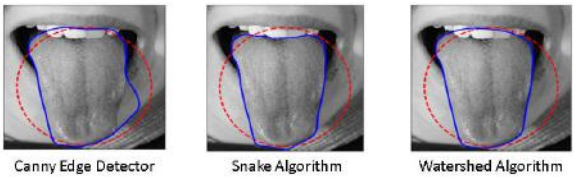


Fig 2: Sample images of edge detection for tongue area extraction using conventional algorithms

3.2.4 Quantitative Evaluation Metrics

Evaluation Metrics considered are IOU, Precision, Recall, and Dice Coefficient. A brief introduction to the metrics considered is given in brief. Then comparison of results achieved by Double U- Net with U-Net and Res U-Net is done based on above metrics as well as qualitative results in the subsequent section.

IOU -Intersection -Over- Union also known as Jaccard Index is the most common commonly used straightforward and extremely effective metric used in semantic segmentation. IOU is the area of overlap between the predicted segmentation 'B' and the ground truth 'A 'divided by the area of union between the predicted segmentation and the ground truth

$$IOU=(|A\cap B|)/(|A\cup B|)$$

The Purity of our positive detections relative to the ground truth describes Precision

$$Precision=TP/|TP+FP|$$

Where TP = True Positive

FP = False Positive

FN = False Negative

TN = True Negative

The completeness of our positive predictions relative to the ground truth is effectively described by Recall. Basically, it gives an idea of all of the annotated masks in our ground truth, how many were captured as positive predictions.

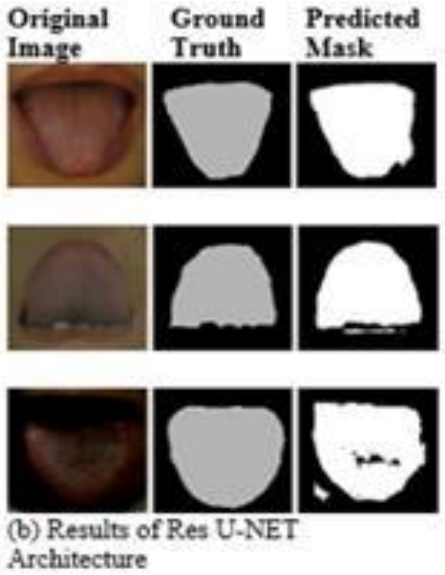
$$Recall=TP/|TP+TN|$$

The Dice coefficient is very similar to the IOU. They are positively correlated. Dice Score serves as means of validating an algorithm by calculating how similar the objects are. It is not only a measure of how many positives you find, but it also penalizes for the false positives that the method finds, similar to precision

$$Dice\ Coefficient=(2\times TP)/(2TP+FP+FN)$$

4. DISCUSSIONS

Results indicate that with conventional techniques segmentation becomes a tedious task specifically in case of smart phone images. Deep Architecture shows robust performance on all the datasets, specifically in case of Dataverse as well as mobile captured dataset where images are not captured in standard conditions. Qualitative results show that U-Net is capable of producing better segmentation mask even for the challenging images, thus suggesting robustness of the model. Limitation of U-Net is that it uses more parameters to learn which leads to increase in training time as compared to other two models.



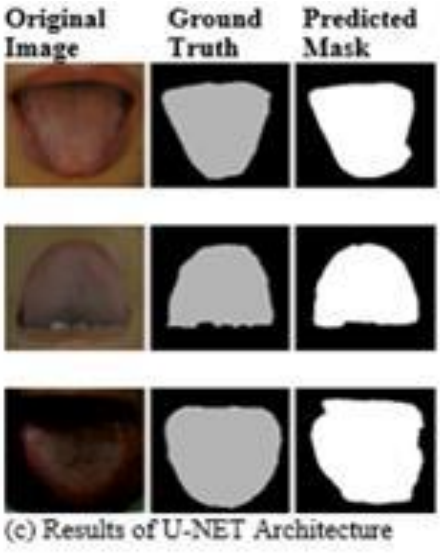


Fig 3: Segmentation results of some samples of various image groups for Double U-Net3)

architectures along with pretrained networks for feature extraction. Future work aims at feature extraction from the segmented tongue region for disease diagnosis.

6. ACKNOWLEDGMENTS

I acknowledge the support of Shri. Govindram Seksaria Institute of Technology & Science, for providing support with TEQIP Project and advanced technical facilities of CIDI (Centre for Innovation and Design Incubation.

7. REFERENCES

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation", Computer Vision and Pattern Recognition (cs.CV), Cornell University, https://arxiv.org/abs/1505.04597v1.

[2] .Jiang Li, Baochuan Xu, Xiaojuan Ban, Ping Tai, and Boyuan Ma "A Tongue Image Segmentation Method Based on Enhanced HSV Convolution Neural Network", © Springer International Publishing AG 2017, Y. Luo (Ed.): CDVE 2017, LNCS 10451, pp. 252–260, 2017, DOI: 10.1007/978-3-319-66805-5_32.

[3] Yushan Xue, Xiaoqiang Li, Pin Wu1, Jide Li, Lu Wang,and Weiqin Tong," Automated Tongue Segmentation in Chinese Medicine Based on Deep Learning", © Springer Nature Switzerland AG 2018, ICONIP 2018, LNCS 11307, pp. 542–553, 2018, doi.:10.1007/978-3-030-04239-4_49.

[4] Bingqian Lin, Yanyun Qu_, Junwei Xie, Cuihua Li," DEEPTONGUE: TONGUE SEGMENTATION VIA RESNET", 978-1-5386-4658-8/18/$31.00 ©2018 IEEE, ICASSP 2018.

[5] Wei Yuan, Changsong Liu, "Cascaded CNN for Real-time Tongue Segmentation Based on Key Points Localization", 2019 the 4th IEEE International Conference on Big Data Analytics,doi: 978-1-7281-1282-4/19/$31.00 ©2019 IEEE.

[6] Xinlei Li, Dawei Yang, Yan Wang, Shuai Yang, Lizhe Qi, Fufeng Li, Zhongxue Gan, Wenqiang Zhang," Automatic Tongue Image Segmentation For Real-Time Remote Diagnosis", 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM),doi: 978-1-7281-1867-3/19/$31.00 ©2019 IEEE.

[7] Xiaodong Huang, Hui Zhang, Li Zhuo , Xiaoguang Li, and Jing Zhang, "TISNet-Enhanced Fully Convolutional Network with Encoder-Decoder Structure for Tongue Image Segmentation in Traditional Chinese Medicine", Computational and Mathematical Methods in Medicine,Volume 2020, Article ID 6029258, doi.:10.1155/2020/6029258.

[8] Qichao Tang, Tingxiao Yang, Yuichiro Yoshimura,Takao Namiki, Toshiya Nakaguchi, "Learning-based tongue detection for automatic tongue color diagnosis system", Artificial Life and Robotics (2020) 25:363–369,doi:10.1007/s10015-020-00623-5.

[9] Zhengxin Zhangy, Qingjie Liuy_, Yunhong Wang, "Road Extraction by Deep Residual U-Net"