Techniques of Glottis Segmentation from Stroboscopic Videos

Varun Belagali[1], Shanta Rangaswamy[2]

[1] Student, Department of Computer Science and Engineering, R V College of Engineering Bengaluru
[2]Associate Professor, Department of Computer Science and Engineering, R V College of Engineering Bengaluru

---***---

Abstract - *The voice patients are evaluated based on the vocal folds vibration. The segmentation of the glottis forms a key part in such evaluation. The area of the glottis opening can be calculated based on the segmentation. Such quantization of the glottis opening helps the doctors in evaluating the voice patients. The stroboscopic videos are recorded by the clinical routines. Such videos are fed to the glottis segmentation algorithm for area quantization. In this paper, a review on recent techniques of glottis segmentation is presented.*

***Key Words*: glottis segmentation, DNN, CNN, ASM, LSTM**

1. INTRODUCTION

There is a narrow opening present between the vocal folds through which the air passes. This narrow opening is termed as glottis. The changes in shape of glottis opening is caused by the variations in the vibration pattern of the vocal folds around it. Some people face a vocal fold condition called Sulcus vocalis (SV). In SV, there is formation of groove which results in reduction of vocal folds mass. During the production of voice this may lead to incomplete glottis closure, which is termed as glottis chink. The glottic chink is observed by the recording of endoscopic videos of the vocal folds recorded from the patients. The technique of recording the endoscopic videos with specific viewing properties is termed as stroboscopy. The severity of the glottic chink in SV is analyzed by the doctors using stroboscopic videos. Figure 1 shows some of the sample frames extracted from stroboscopic videos. The dark region between the white vocal folds is defined as glottis opening. The figure also shows the marking of the glottis opening boundary. The aim of the glottis segmentation is to predict this region bounded by the red contour.

Image segmentation has shown a great impact on the field of computer vision. Image segmentation is defined as the process of the assigning a label corresponding to the object that each pixel belongs to. In case of the glottis segmentation, there are two objects, glottis opening and background. There are various techniques of image segmentation. Recently deep learning has gained a high importance in image segmentation problems. In most of these algorithms training data is required where the objects are marked by humans. Such a labelled collection of data in case of medical imaging is hard because it requires experts of the domain to label the data. Unsupervised methods are most effective in such cases because they do not require any labelled data. Designing of such algorithms is hard, as we have to constraint the algorithm to predict glottis without given the information of the glottis part. Some of the unsupervised algorithms include watershed based methods and Fourier descriptors based methods.

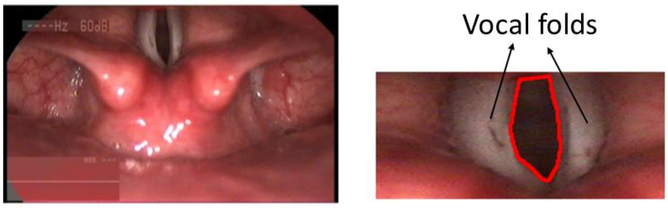

Fig - 1: Sample frame of glottis extracted from stroboscopic video recording [10].

The main challenge of the glottis segmentation is that there is a high variation among the glottis shapes and appearance between the patients. The illumination present during the recording may change with time. This makes algorithm hard to generalize on the various subjects and illumination conditions. Further, challenges are added due the amount available labelled data of the glottis that is less.

2. METHODS OF GLOTTIS SEGMENTATION

The glottis is segmented from each frame of the video. It can be considered as an image segmentation problem on the individual frames of the video. There are various techniques for image segmentation: 1) Manual segmentation, 2) Semi-automatic segmentation, 3) Fully automatic segmentation.

2.1 Manual segmentation

In manual glottis segmentation, the doctors are required to mark the glottis opening. The doctor views the video frame by frame and marks the boundary of the glottis opening. It is a time consuming and a tedious process for the doctor to manually mark the glottis region, because each video is of an average duration of 40 seconds with 25 frames per second. An automatic glottis segmentation system is required for assistance of doctor in evaluating the patients.

2.2 Semi-automatic segmentation

In semi-automatic image segmentation, the images are segmented based on the user intervention, but to do depend on the user completely. It is a combination of manually and automatic system. A semi-automatic glottis segmentation method was proposed which uses seed points [1]. In this method the user is expected to provide the initial seed points my manual marking of the points inside the glottis region. After the annotation of seed points, a region growing algorithm is used of segment the glottis region from the video frame. The limitation of such a method is that the segment predicted is highly dependent on the seed points chosen by the user.

2.3 Fully Automatic segmentation

The fully automatic segmentation algorithms do not require any user intervention. Such an algorithm can help doctor to a great extent. Doctor can record the video of the patients and just feed it into the algorithm to get estimations of the glottis opening area. There are various automatic glottis segmentation techniques proposed in recent studies.

An automatic glottis segmentation algorithm that uses shape and local color features of glottal regions was proposed [2]. Fourier descriptors were used to represent glottal regions. These descriptors were used eliminate region where there is no presence of glottis. Bayesian probability based on local color feature is used in region level set algorithm to find the segment of glottis. As a part of post processing, Principal Component Analysis is done to remove the errors in the prediction. The algorithm achieves an average dice score of 0.85. The limitation of this method is the collection of all possible glottis features for fourier descriptors.

An Active shape model (ASM) has been proposed for glottis segmentation [3]. ASM use the point probability distribution to predict the segment of glottis. In this algorithm, there are three steps: 1) Region growing based algorithm, 2) Morphological processing and 3) ASM Model. The results of the first two steps are used as an initialization for the third step. The method achieved an segmented error of 5.2 pixels on an average. The disadvantage of using ASM is that it might not generalize well on the unseen data that is different when compared to training data.

Glottis segmentation using snakes algorithm (ASM) has been proposed [4]. A median and thresholding filter is applied as a preprocessing stage. The thresholding values are obtained from image histograms. The initialization of the contour of glottis opening is done based on the Gradient Vector Flow value. After this, snakes algorithm is applied to find the final glottis region. The contour is updated based on the energy function of the current contour features. The algorithm converges to point where the energy function reaches minima point. Pratt coefficient [5] is used as the evaluation metric for the results. The results show that the Pratt coefficient is greater than 0.5 for all test images.

Another ASM based algorithm has been proposed which uses 3D Active contours [6]. The features of a model use the color distribution, spatial features and temporal features. The algorithm is able to handle the variation of illumination conditions during recording. Dice score is used as the evaluation metric. An average median dice score of 0.76 is achieved.

An watershed method based algorithm was proposed for glottis segmentation [7]. The image is converted to grey scale using YIQ model. A threshold is applied on the gradient of this image. Then watershed algorithm is used for segmentation. To avoid over segmentation, region merging is used later. JND cost function based merging is used to merge the neighboring regions. This is done as a part of post processing. One limitation of this method is that the result is dependent on the choice of the threshold used. A slight change in threshold results in drastic change in segmentation. The results prove that the glottis region is detected in all the validation set subjects.

2.4 Deep learning based fully automatic segmentation

Deep learning methods have shown a great advancement in image segmentation. Convolutional neural networks(CNNs) are the most effective on images. The need of the deep learning method is the collection of large amount of annotated data for training. The collection of large is a tedious task in the case of stroboscopic videos. In general, data augmentation is done overcome this issue. Among CNNs, Segnet [8] and Unet [9] are the most popular architectures for image segmentation. Unet is generally suitable for medical images. Segnet is formed of encoder decoder structure. Encoder is used to down sample the images and decoder is used construct the segmentation result. Unet is similar to Segnet, except that Unet contains skip connections from encoding layer to decoding layer. In training phase of the CNN, transfer learning is used to have a better initialization of weights to predict edges. Example of transfer learning is the initialization of weights by the VGG weights that are trained on Image net. Such initialization helps in better feature learning at initial stages. Some recent work have also have suggested the use deep neural networks (DNN). Such a network contains fully connected layers.

A DNN based method has been proposed for glottis segmentation [10]. In this method an image patch of 3 X 3 is taken around each pixel to form an input to the DNN. The DNN predicts whether the pixel belongs to the glottal region. A table method is applied as post processing to form the segments using neighborhood predicted information. Further a ellipse is fit to the predicted region to form the final segmentation output. The DNN contains 3 hidden layers with 128 neurons and a single output sigmoid layer. Dice score is used as measure of evaluation. The DNN method results in an mean dice score of 0.74 on the test data. The DNN performs poor on the images which have less illumination. This is because the DNN only uses 3 X 3 patch to predict glottal pixels. Global context of pixel is not taken into consideration.

A CNN based method has been proposed for glottis segmentation [11]. In this method the first a Region of Interest (ROI) is found out. Then segmentation is done on the ROI to find the exact glottis segment. Segnet is used as CNN architecture. The encoder part of the Segnet is used locate the ROI. A bounding box is drawn around ROI detected. Then it is passed through a Segnet for the segmentation of the image. Various geometric shapes of the objects present around the glottis are taken into account. Singular spectrum analysis is used to as part of post processing. It is used to quantify the area of the glottal region detected. Singular spectrum analysis is a signal processing based method.

A combination of CNN and Long Short Term Memory (LSTM) has been proposed for glottis segmentation [12]. In this study a total of 18 CNN architecture were experimented to find the best suitable CNN. In all the previous studies mentioned, segmentation is done on individual frame of the video, the temporal relation between the frames is not taken into consideration. There exists a temporal relation between the sequential frames of the video. There is no much change in position and appearance of the glottis from current frame to next frame. Such temporal information can used to predict more accurate segments. In [12], LSTM is used to take the advantage of the temporal information. Dice score is used as the evaluation metric. The mentioned combination of CNN and LSTM results in an mean dice score of 0.85 on the test data.

3. EVALUATION METRIC

There are various evaluation metrics used for glottis image segmentation. The most popular among them is the Dice score [13]. The main feature of this metric is that in only takes the glottis segment prediction into account. This is important because the glottis forms a small part of the video frame. In such case a metric which focus more on the glottis segment is favorable.

$$\text{Dice} = 2 \times (A \cap B) / (A+B) \qquad (1)$$

Another metric used is pixelwise accuracy. It is defined as the percentage of the total number of pixels that are predicted correctly i.e., the actual pixel belonging to glottis is predicted as glottis and actual background is predicted as background pixel. If there are less number of glottis pixels then more weight is given to the background pixels. This is the limitation of using the metric.

4. CONCLUSION

The recent studies show the use of various fully automatic techniques like ASM, CNN, DNN and LSTM based algorithms for glottis segmentation. There is also a study based on semi-automatic segmentation using region growing. The main requirement of the problem is a fully automatic algorithm. Most of the recent studies propose fully automatic algorithms. Such methods can help the doctors in assessing the patients with vocal disorders by quantization of the glottal region.

REFERENCES

[1] Lohscheller, Jörg, et al. "Clinically evaluated procedure for the reconstruction of vocal fold vibrations from endoscopic digital high-speed videos." Medical image analysis 11.4 (2007): 400-413M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[2] Gloger, Oliver, et al. "Fully automated glottis segmentation in endoscopic videos using local color and shape features of glottal regions." IEEE Transactions on Biomedical Engineering 62.3 (2014): 795-806.

[3] Cerrolaza, Juan J., et al. "Fully-automatic glottis segmentation with active shape models." MAVEBA. 2011.

[4] Miranda, G. Andrade, et al. "A new approach for the glottis segmentation using snakes." BIOSIGNALS 2013. 6th International Conference on Bio-Inspired Systems and Signal Processing. 2013.

[5] McCarl, Bruce A. "Interpretations and Transformations of Scale for the Pratt-Arrow Absolute Risk Aversion Coefficient: Implications for Generalized Stochastic Dominance: Comment." Western Journal of Agricultural Economics 12.2 (1987): 228-230.

[6] Schenk, Fabian, et al. "Automatic high-speed video glottis segmentation using salient regions and 3D geodesic active contours." Annals of the British Machine Vision Association (2015).

[7] Osma-Ruiz, Víctor, et al. "Segmentation of the glottal space from laryngeal images using the watershed transform." Computerized Medical Imaging and Graphics 32.3 (2008): 193-201.

[8] Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." IEEE transactions on pattern analysis and machine intelligence 39.12 (2017): 2481-2495.

[9] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015.

[10] Rao, MV Achuth, et al. "Automatic Glottis Localization and Segmentation in Stroboscopic Videos Using Deep Neural Network." Interspeech. 2018.

[11] Lin, Jianyu, et al. "Quantification and analysis of laryngeal closure from endoscopic videos." IEEE Transactions on Biomedical Engineering 66.4 (2018): 1127-1136.

[12] Fehling, Mona Kirstin, et al. "Fully automatic segmentation of glottis and vocal folds in endoscopic laryngeal high-speed videos using a deep Convolutional LSTM Network." Plos one 15.2 (2020): e0227791.

[13] Dice, Lee R. "Measures of the amount of ecologic association between species." Ecology 26.3 (1945): 297-302.