Automated Segmentation of the Vocal Folds in Laryngeal Endoscopy Videos using Deep Convolutional Regression Networks

Ali Hamad[1], Megan Haney[2], Teresa E. Lever[3], and Filiz Bunyak* [1]

[1]Department of Electrical Engineering and Computer Science
[2]Department of Veterinary Pathobiology
[3]Department of Otolaryngology–Head and Neck Surgery, School of Medicine
University of Missouri-Columbia
Columbia, MO, USA

Abstract

Swallowing and breathing are vital, life-sustaining upper airway functions that require precise, reciprocal coordination of the vocal folds (VFs). During swallowing, the VFs must fully close to prevent aspiration of food/liquid into the lungs, whereas during breathing, the VFs must remain open to prevent obstruction of airflow into and out of the lungs. This coordination may become impaired by a variety of neurological conditions and diseases. Clinical evaluation relies on transnasal endoscopy to visualize the VFs within the larynx, and subjective interpretation of VF function by clinicians. However, objective, quantitative, and high-throughput analysis of VF function is important for early diagnosis, monitoring disease progression, treatment monitoring, and treatment discovery. In this paper we propose a fully automated, deep learning based VF segmentation system for the analysis of VF motion behavior captured using flexible endoscopes with low-speed capability. Experimental results on human laryngeal videos showed promising results that were robust to many challenges caused by imaging, anatomical, and behavioral variations. The proposed segmentation and tracking system will be used to compute quantitative outcome measures describing VF motion behavior in order to help clinical practice and scientific discovery.

1. INTRODUCTION

Swallowing and breathing are life-sustaining physiological functions of the upper airway that require precise, reciprocal coordination of the vocal folds (VFs). During swallowing, safe passage of food and/or liquid boluses from the mouth to the stomach without entering the lungs (i.e., aspiration) is ensured by complete closure of the VFs to seal the glottis (opening between the VFs), which momentarily prevents breathing. During breathing, the glottis must remain wide open to prevent obstruction of airflow into and out of the lungs for respiration. These opposing behaviors of the VFs are predominantly controlled by a single cranial nerve (vagus), its laryngeal branches, and multiple laryngeal muscles [1] that are vulnerable to many neurological conditions/diseases such as degenerative (e.g., amyotrophic lateral sclerosis/ALS, Parkinsons, and advanced aging), cerebrovascular (e.g., cerebral ischemia or infarction), congenital (e.g., DiGeorge syndrome and Rett syndrome), and neoplastic processes (e.g., tumors) and their associated surgical and/or medical interventions [2] [3]. A leading cause of death in these conditions/diseases is respiratory failure resulting from aspiration pneumonia, often due to impaired VF function that fails to protect the airway during swallowing [4].

Clinical evaluation of VF function during swallowing and breathing relies on laryngeal endoscopy (laryngoscopy), a medical test in which a flexible fiberoptic camera is passed through the nose into the upper throat in the awake patient. Current practice relies predominantly on subjective interpretation of VF motion during real-time viewing or video playback [5] [6] [7]. While objective, manual analysis of laryngoscopy videos is possible, the process is extremely labor intensive and creates a major bottleneck in scientific discovery and clinical use. In this paper, we explore the use of computer vision and machine learning approaches for automated and quantitative analysis of laryngoscopy videos. This work addresses the criti-

*Corresponding author. This work was funded by the Coulter Translational Partnership Program for development of vocal fold motion-tracking software.

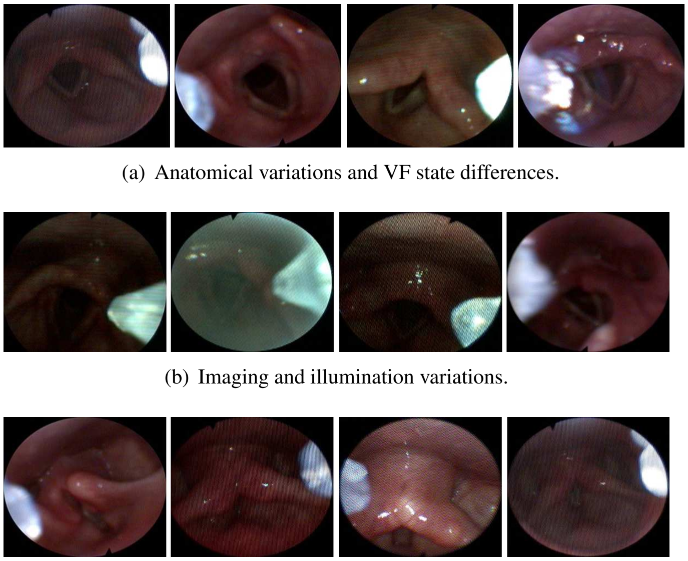
cal need for automated, high-throughput, quantitative analysis of imaging data to increase the diagnostic utility of this medical test.

The first and most critical step in automated analysis of laryngoscopy videos is segmentation of the VFs and glottal region. Measurements characterizing the shape and motion of the VFs can then be computed using the output of this step. Recently, several video analysis pipelines have been developed using rigid endoscopes with high-speed cameras to visualize the VFs. However, the large diameter and inflexible structure of rigid endoscopes require transoral (through the mouth) insertion, which restricts assessment of VF function to vocalization (mainly vowel production) rather than the coordination of swallowing and breathing. For example, in [8], vibrating vocal fold edges during vocalization are segmented in high-speed rigid laryngoscopy videos using a seeded region-growing algorithm and adaptive thresholding. In [9], Zernike moments operator and level set algorithm are used to detect the glottal edges at a subpixel level. In [10], spatio-temporal information is exploited by using rigid motion compensation, saliency detection, and 3D geodesic active contours to segment the glottal region. In [11], flexible thresholding technique is combined with a refining level set method that incorporates prior glottal shape knowledge. However, video data obtained using rigid, high-speed ($>$ 100 frames per second, fps) endoscopes has markedly different characteristics than video data captured using flexible, low-speed (30 fps) endoscopes. In particular, the larger diameter of rigid endoscopes typically results in high resolution images, better illumination, and less degradation of images compared to flexible endoscopes. High-speed video capture further facilitates VF tracking because the high frame rate results in observation of smoother motion that is more predictable.

In this paper, we focus on the more challenging analysis of VF videos obtained using flexible endoscopes. Inspired by the recent successes of deep learning in computer vision in general, and biomedical image analysis in particular [12] [13] [14] [15], we have developed a deep learning system for segmentation of the VFs and glottal region in transnasal flexible laryngoscopy videos. Section 2 provides details of the proposed system and the network training procedures. Section 3 presents system evaluation approaches and experimental results. Section 4 concludes the paper and describes future directions.

2. METHODOLOGY

We have developed a deep learning system for segmentation of the glottal region in laryngoscopy videos using Fully Convolutional Regression Networks (FCRN). The trained network has the ability to automatically segment the glottal region despite large appearance variations in the acquired images due to anatomical differences, imaging factors (i.e., type/model of endoscope, endoscope position, and strength of illumination), state of the VFs (fully open, fully closed, partially open, etc.), gradual or sudden motion of the camera or subject during imaging, and saliva that affects the visibility of the laryngeal region. Some of glottal region appearance variations are illustrated in Figure 1.


(a) Anatomical variations and VF state differences.

(b) Imaging and illumination variations.

(c) Full VF closure and glottal gap occlusion.

Figure 1. Various factors affecting appearance of VFs and glottal region.

2.1. Semi-automated Vocal Fold Tracking

While deep learning methods are powerful, a large amount of labeled training data is needed to train a reliable model. Training data generation for segmentation tasks is typically more labor intensive compared to detection or classification tasks, particularly for biomedical images, because manual segmentation often requires expert knowledge and careful boundary tracing. We have developed VFTrack [16] (Figure 2), an interactive VF motion tracking software for efficient training data generation for glottal region segmentation. VFTrack operates as follows:

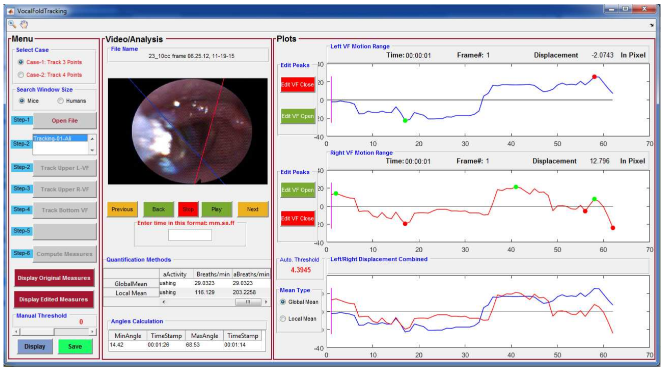
1. On the first video frame, the user selects three points: one on the left (p_L), one on the right (p_R), and one in the middle (p_o) at the junction of the left and right VFs respectively (blue, red, and green points in Figure 3).

2. VFTrack independently tracks each point (p_L, p_R, p_o) in time. When a track is lost, VFTrack prompts the user to re-select the point of interest and restarts the tracking process.

3. Once all three points are tracked for the duration of the video clip, the VFs are modeled in each frame with two lines L_L passing through the points p_L and p$_o$ and

L_R passing through the points p_R and p_o (blue and red lines in Figure 3).

Detailed performance analysis of VFTrack compared to fully manual VF marking is given in [16]. When tested on endoscopy videos of a laryngeal nerve injury mouse model, VFTrack produced a pixel distance error comparable to inter-reviewer distance with drastically reduced processing time (18 minutes shorter compared to manual analysis) [16].



Step-2: Tracking-01-Ali

File Name: 23_10cc frame 06.25.12, 11-19-15

Manual Threshold: 0

	aActivity	Breaths/min	aBreaths/min
GlobalMean	ushing	29.0323	29.0323
Local Mean	ushing	116.129	203.2258

MinAngle	TimeStamp	MaxAngle	TimeStamp
14.42	00:01:26	68.53	00:01:14

Left VF Motion Range: Time: 00:00:01 Frame#: 1 Displacement -2.0743 In Pixel

Right VF Motion Range: Time: 00:00:01 Frame#: 1 Displacement 12.796 In Pixel

Auto. Threshold: 4.3945

Figure 2. VFTrack, our semi-automatic vocal fold tracking and vocal fold motion analysis software.

2.2. Training Data Generation

Once the VFs are interactively tracked, training data is prepared using the following steps:

1. *Training mask generation:* A triangle defined by the points p_L, p_R, p_o is used to produce the glottal region training mask, where pixels inside and outside of the triangle are assigned 1 and 0 values, respectively. Note that this triangle may not fully cover the upper section of the glottal gap.

2. *Normalization, cropping, and resizing:* Input images are normalized by mean subtraction. The endoscope field of view is automatically detected. Training images and masks are first cropped using the bounding box of the endoscope field of view, then resized to 128×128 patches.

3. *Distance transformation:* Euclidean distance transform is applied to the training masks to produce continuous valued training labels that are lower (closer to the glottal region boundaries), higher (towards glottal region center), and zero (everywhere else).

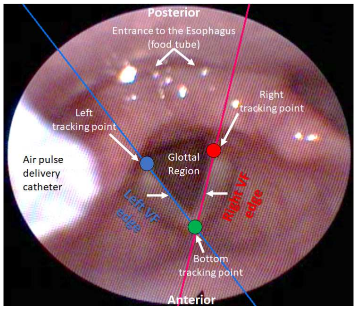


Figure 3. Anatomical structures and interest points selected for VF tracking.

Figure 4 shows a sample training image and corresponding glottal region mask and training labels (distance map).

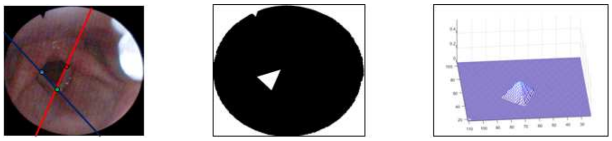


Figure 4. Dataset preparation steps for the fully convolutional regression network (FCRN).

2.3. Network Architecture

We have designed and implemented a fully convolutional regression network (FCRN) that maps its 3-channel RGB input to a 1-channel distance map that acts like a glottal region likelihood map. The network architecture consists of 13 layers as shown in Figure 5 and detailed in Table 1. The proposed network was implemented using Matlab deep learning toolbox [17]. Training of the proposed FCRN architecture was done using the images and corresponding distance maps prepared as described in Section 2.2. FCRN network learns a mapping from the endoscopy image to the distance map of the glottal region, leading to a robust segmentation that can capture both localization and shape information of the glottal region. The number of layers in the network was experimentally determined to optimize the segmentation performance. A not very deep network was proposed to be able to train the network with a limited amount of training data.

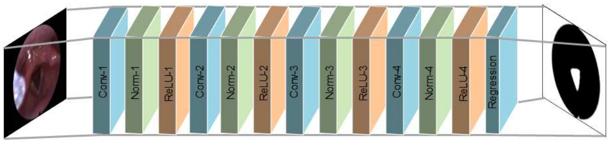


Figure 5. Proposed fully convolutional regression network (FCRN) for glottal region segmentation.

2.4. Glottal Region to Vocal Fold Model Conversion

The ultimate goal of this study is to develop high-throughput image analysis techniques for automated analysis of VF shape and motion patterns. Laryngeal region appearance and motion behavior can be captured through region-based (glottal mask) or boundary-based (VF lines)

Table 1. Layer information for the fully convolutional regression network (FCRN) proposed for glottal region segmentation.

Layer	Layer Type	Parameters
1	Convolution	32 5 × 5 × 3 convolutions; stride [1 1]; padding 2
2	Normalization	Batch normalization; 32 channels
3	Rectified Linear Unit	ReLU
4	Convolution	32 3 × 3 × 32 convolutions; stride [1 1]; padding 1
5	Normalization	Batch normalization; 32 channels
6	RELU	Rectified Linear Unit
7	Convolution	32 3 × 3 × 32 convolutions; stride [1 1]; padding 1
8	Normalization	Batch normalization; 32 channels
9	RELU	Rectified Linear Unit
10	Convolution	32 3 × 3 × 32 convolutions; stride [1 1]; padding 1
11	Normalization	Batch normalization; 32 channels
12	RELU	Rectified Linear Unit
13	Regression	1 3 × 3 × 32 convolutions; stride [1 1]; padding 1

descriptors. These approaches have different advantages and disadvantages. While the glottal region mask better captures complex shape information, modeling the VFs will allow independent analysis of the left and right VF behaviors, enabling capture of motion symmetry and synchrony information (as we used in [16] for quantitative analysis of laryngeal nerve injury). This section describes modeling of the VFs given a glottal region mask. Processing steps are illustrated in Figures 6, 7 and described as follows. First, given the glottal region mask, glottal region boundaries are extracted. Then, boundary points are clustered using K-means clustering to identify left and right VFs and upper boundary of the glottis. Note that during imaging, the endoscope can be tilted left or right. K-nearest neighbor search is used to further improve VF assignments. Finally, two lines are fitted to the left and right boundaries to model the left and right VFs.

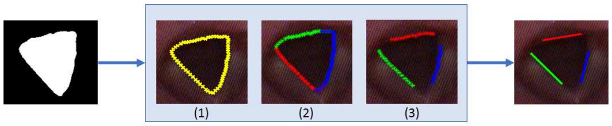


Figure 6. Extraction of VF line models from glottal segmentation mask. (1) Glottal mask boundaries, (2) preliminary clustering of boundary points, (3) refined clustering of boundary points used to extract VF line models.

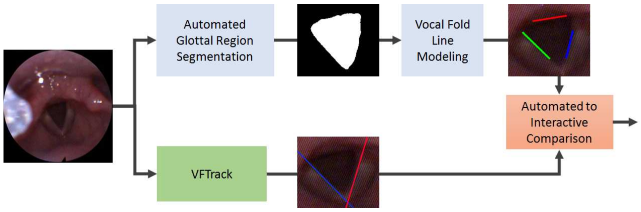

Automated Glottal Region Segmentation
Vocal Fold Line Modeling
Automated to Interactive Comparison
VFTrack

Figure 7. Evaluation of the proposed fully automated glottal region segmentation method using interactive VF tracking results.

3. Experimental Results

The dataset used to train and evaluate the proposed deep learning based glottal region and VF segmentation system has been collected from 20 participants according to the protocol described in [18]. The data collection protocol was approved by the University of Missouri Institutional Review Board. Twenty healthy nonsmoking human subjects (7 men and 13 women) aged 20 to 40 years were recruited and tested. The test procedure entailed transnasal passage of a flexible 3.7-mm outer-diameter endoscope with a 1.5-mm inner-diameter working channel (11302BD2, Karl Storz). The endoscope tip was positioned at a typical level for viewing laryngeal pathology to permit visualization of the bilateral VFs throughout the procedure. In total 58 videos were collected, and 7892 color images of size $480 \times 720 \times 3$ were extracted from these videos. The proposed network was trained with 1661 images and tested with 6231 images. Training labels for these images were obtained using our interactive VF tracking software, VFTrack, as described in Section 2.2 and illustrated in Figure 4. For comparison purposes, we have also built and trained a convolutional encoder-decoder network similar to semantic segmentation network described in [19]. However, this network, as in the case of the proposed regression network, was built with much less layers due to limited availability of training data. Both networks contained 4 convolutional layers, plus one classification or regression layer (also other types of layers such as pooling, RELU etc. not listed here). The encoder-decoder network was trained on the same training dataset, and for the same number of epochs (1000) as the proposed deep convolutional regression network.

3.1. Qualitative Evaluation

Sample results from the proposed system are presented in Figures 8 and 11. Figure 8 shows single frames from different videos and corresponding results (network outputs and produced binary masks) obtained using the proposed deep regression network. The frames are selected to show variations in endoscope position and illumination, opening angles of the VFs, mucosal color, visibility of the VFs, and glare. Satisfactory segmentation results were obtained in each case despite these variations. Figure 11 shows performance of the proposed system on a sequence of frames in two selected videos. The set of frames are selected to show the segmentation behavior through the laryngeal adductor reflex (LAR), which entails brief bilateral closure of the VFs in response to air puff stimulation of the laryngeal mucosa [18]. LAR events are a challenging case for both VF tracking and glottal region segmentation approaches. When the VFs are fully closed, the glottal region becomes invisible, leading to false detections during segmentation. VF tracking typically relies on frame to frame matching of VF transition pattern (tissue to glottal gap). Disappearance of the glottal gap during LAR event disrupts this pattern, leading to tracking failures. Figure 11 demonstrates that the proposed segmentation scheme is even able to handle LAR events.

3.2. Performance Evaluation using VF Line Models

In order to quantitatively evaluate the system performance, we have performed two types of analysis: (1) comparison to interactively tracked VFs on all test frames; (2) comparison to manually segmented glottal regions on a subset of test frames.

We have interactively tracked the VFs in the 6231 test frames using our VFTrack software as described in Section 2.2. Any tracking error was manually fixed using the same software. Using the tracking output, for each frame in the test set, we have produced two lines L_L^{GT} and L_R^{GT} modeling the VFs, to be used as our ground-truth. For the proposed fully automated segmentation system, we have computed VF lines L_L^{Seg} and L_R^{Seg} from the glottal region segmentation masks using the steps described in Section 2.4. Evaluation is done in terms of line to line distance between the ground-truth and the proposed segmentation results at left and right VFs, and in terms of difference of VF opening angles (angle between lines L_L and L_R). For line to line distance, we find the unique vector of two points between the two lines, where this vector is perpendicular to both lines and represents the shortest distance between them. The Mean errors are reported in Table 2. VF opening angle distributions are compared in Figure 9. The differences in VF angle estimation are largely due to line fitting and ground-truth errors. While we choose to represent the VFs with lines for the sake of model simplicity, the VFs are actually deformable curves. Lines fitted at different portions of the curve produce different linear models leading to VF opening angle differences. Sample problem cases are shown in Figure 10.

3.3. Performance Evaluation using Manually Generated Ground-truth Regional Masks

We have manually generated ground-truth segmentation masks for 500 randomly selected images from the test set. Sample segmentation masks produced using the proposed deep regression network and encoder-decoder network used for comparison are shown in Figure 12. Both networks are able to detect the glottal region. Proposed deep regression network produces more accurate masks of the glottal region, compared to the encoder-decoder segmentation network. This is predominantly due to use of distance transform that captures shape information better than binary mask. For quantitative evaluation, ground-truth and segmentation masks are compared in terms of accuracy, intersection over union (IOU), rand index (RI), dice similarity, and Hausdorff distance measures [17] [20] [21] described in Table 3. Region-based segmentation evaluation measures for the randomly selected 500 images are reported in Table 4. The proposed deep regression network outperforms comparable encoder-decoder segmentation network in all measures. We have also computed VF lines for the ground-truth segmentation masks using the steps described in Section 2.4. VF line comparison measures as described in Section 3.2 were made between proposed segmentation and ground-truth masks for the selected 500 images (Table 5).

Table 2. Segmentation performance evaluation using VF line model.

Measure	**Value**
Left VF distance $Dist_{Line}(L_L^{Seg}, L_L^{GT})$	8.6 pixels
Right VF distance $Dist_{Line}(L_R^{Seg}, L_R^{GT})$	11.9 pixels
Difference in VF opening angles $Dist_{Angle}(Angle^{Seg}, Angle^{GT})$	17°

Table 3. Segmentation mask evaluation measures. TP, FP, FN, TN, I_S, I_{GT} denote true positives, false positives, false negatives, true negatives, segmentation mask, ground-truth mask respectively.

Evaluation measure	**Equation**
Accuracy	$\frac{TP}{TP+FN}$
Intersection over Union (IOU)	$\frac{TP}{TP+FP+FN}$
Rand Index (RI)	$\frac{TP+TN}{TP+TN+FP+FN}$
Dice similarity	$\frac{2*TP}{2*TP+FP+FN}$
Hausdorff distance	$h(I_S, I_{GT}) = \max_{i_s \in I_S} \{ \min_{i_{gt} \in I_{GT}} \{d(I_S, I_{GT})\}\}$

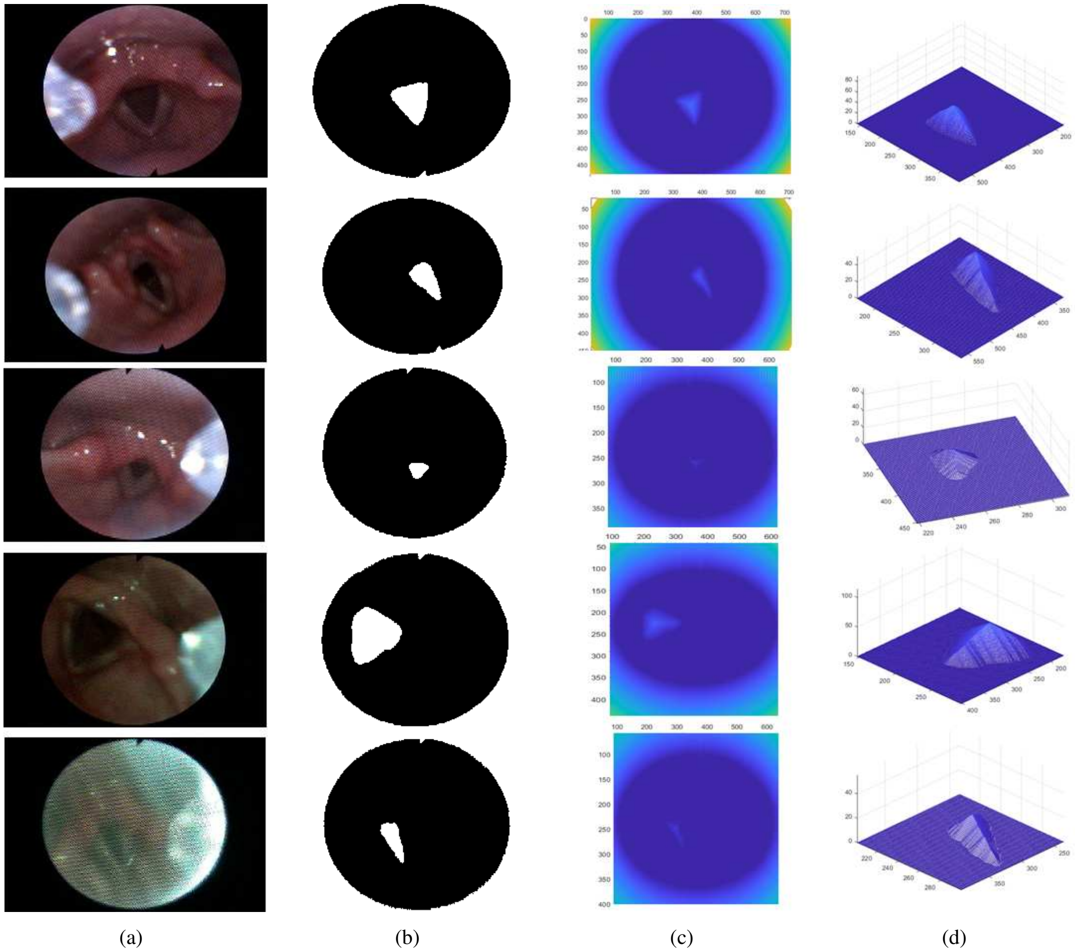


Figure 8. Sample glottal region segmentation results. (a) Original image, (b) segmentation output, (c) segmentation output visualized in 3D (top view), (d) segmentation output visualized in 3D (side view enlarged)

Table 4. Performance evaluation using ground-truth regional masks. Regional evaluation measures.

Measure	**Proposed**	**Encoder-Decoder**
Accuracy	0.9469	0.7358
Intersection over union (IOU)	0.8807	0.7137
Rand index (RI)	0.9936	0.9481
Dice Similarity	0.8576	0.6897
Hausdorff distance	4.26	7.42

Table 5. Performance evaluation using ground-truth regional masks. VF line model comparison measures.

Measure	**Value**
Left VF distance $Dist_{Line}(L_L^{Seg}, L_L^{GT})$	3.3 pixels
Right VF distance $Dist_{Line}(L_R^{Seg}, L_R^{GT})$	3.6 pixels
Difference in VF opening angles $Dist_{Angle}(Angle^{Seg}, Angle^{GT})$	10°

4. Conclusion and Future Work

We have presented a deep fully convolutional regression network for segmentation of the glottal region in human la-

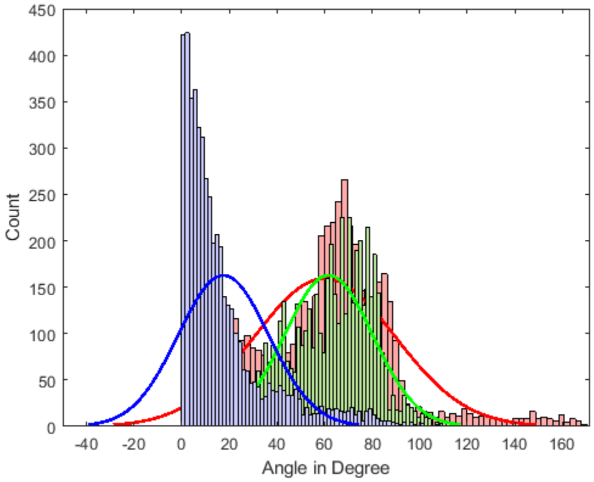


Figure 9. VF opening angle distributions. Red and green plots denote VF opening angles computed using proposed deep regression segmentation method and VFTrack software. Blue plot denotes distribution of angular differences between the two methods.

ryngeal endoscopy videos. The proposed network learns to map the three channel RGB image into a distance map that acts like a glottal likelihood function that captures the location and precise shape of the glottal region between the VFs. We also describe methods to convert the glottal mask to the VF line model and vice versa. Experimental results on healthy human subjects showed promising performance despite large variations between the images. This network constitutes the first step towards a fully automated, high-throughput, quantitative VF motion behavior analysis pipeline using flexible, low-speed endoscopes. Our goal is to use this processing pipeline of healthy adults to establish normative values of VF motion during a variety of upper airway functions, including swallowing, breathing, and the LAR. Our ultimate goal is to use this system for objective and quantitative analysis of disease progression and treatment outcomes in clinical settings.

References

[1] K. Matsuo and J. B. Palmer, "Anatomy and physiology of feeding and swallowing: normal and abnormal," *Physical medicine and rehabilitation clinics of North America*, vol. 19, no. 4, pp. 691–707, 2008.

[2] H.-C. Chen, Y.-M. Jen, C.-H. Wang, J.-C. Lee, and Y.-S. Lin, "Etiology of vocal cord paralysis," *ORL*, vol. 69, no. 3, pp. 167–171, 2007.

[3] D. Myssiorek, "Recurrent laryngeal nerve paralysis: anatomy and etiology." *Otolaryngologic clinics of North America*, vol. 37, no. 1, pp. 25–44, 2004.

[4] J. H. Ta, Y. F. Liu, and P. Krishna, "Medicolegal aspects of iatrogenic dysphonia and recurrent laryngeal nerve injury," *Otolaryngology–Head and Neck Surgery*, vol. 154, no. 1, pp. 80–86, 2016.

[5] L. F. Giraldo-Cadavid, L. R. Leal-Leaño, G. A. Leon-Basantes, A. R. Bastidas, R. Garcia, S. Ovalle, and J. E. Abondano-Garavito, "Accuracy of endoscopic and videofluoroscopic evaluations of swallowing for oropharyngeal dysphagia," *The Laryngoscope*, vol. 127, no. 9, pp. 2002–2010, 2017.

[6] S. E. Langmore, "Endoscopic evaluation of oral and pharyngeal phases of swallowing," *GI Motility online*, 2006.

[7] H. Tohara, A. Nakane, S. Murata, S. Mikushi, Y. Ouchi, Y. Wakasugi, M. Takashima, Y. Chiba, and H. Uematsu, "Inter- and intra-rater reliability in fibroptic endoscopic evaluation of swallowing," *Journal of Oral Rehabilitation*, vol. 37, no. 12, pp. 884–891, 2010.

[8] J. Lohscheller, H. Toy, F. Rosanowski, U. Eysholdt, and M. Döllinger, "Clinically evaluated procedure for the reconstruction of vocal fold vibrations from endoscopic digital high-speed videos," *Medical image analysis*, vol. 11, no. 4, pp. 400–413, 2007.

[9] X. Qin, S. Wang, and M. Wan, "Improving reliability and accuracy of vibration parameters of vocal folds based on high-speed video and electroglottography," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 6, pp. 1744–1754, 2009.

[10] F. Schenk, M. Urschler, C. Aigner, I. Roesner, P. Aichinger, and H. Bischof, "Automatic glottis segmentation from laryngeal high-speed videos using 3d active contours." in *MIUA*, 2014, pp. 111–116.

[11] O. Gloger, B. Lehnert, A. Schrade, and H. Völzke, "Fully automated glottis segmentation in endoscopic videos using local color and shape features of glottal regions," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 3, pp. 795–806, 2015.

[12] Z. Yu, E.-L. Tan, D. Ni, J. Qin, S. Chen, S. Li, B. Lei, and T. Wang, "A deep convolutional neural network-based framework for automatic fetal facial standard plane recognition," *IEEE journal of biomedical and health informatics*, vol. 22, no. 3, pp. 874–885, 2018.

[13] H. Jiang, H. Ma, W. Qian, M. Gao, and Y. Li, "An automatic detection system of lung nodule based on multigroup patch-based deep learning network," *IEEE journal of biomedical and health informatics*, vol. 22, no. 4, pp. 1227–1237, 2018.

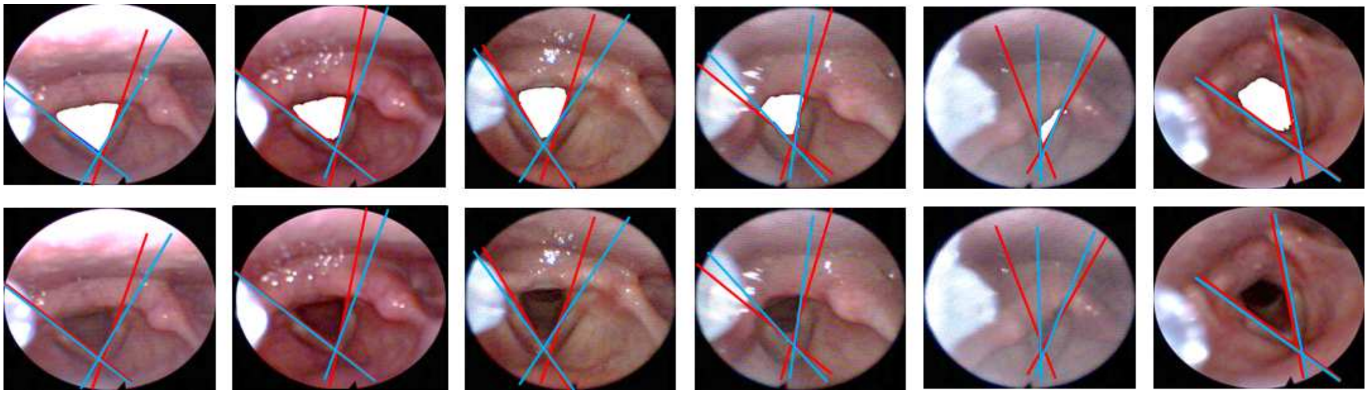
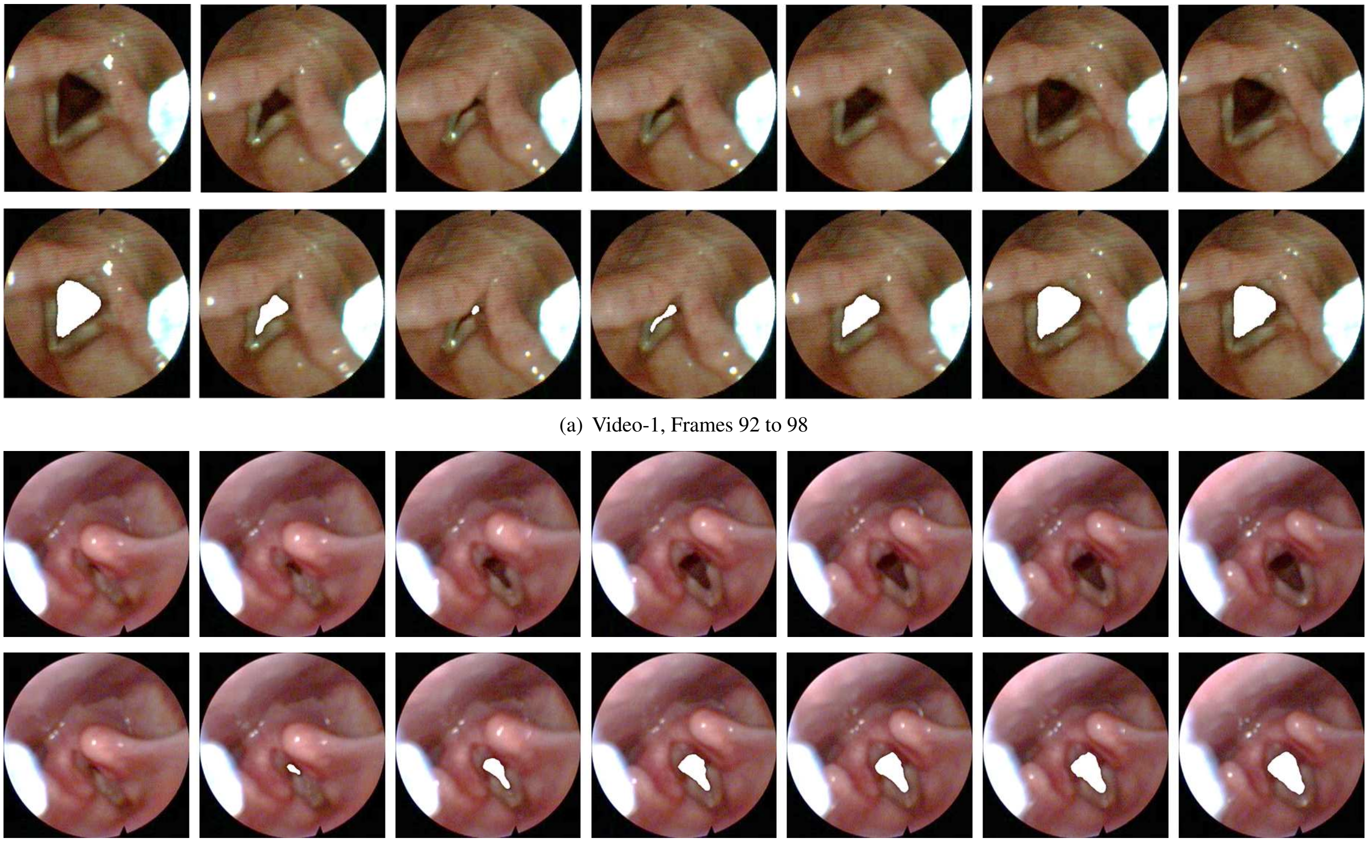


Figure 10. VF line models derived from deep regression network mask (red) versus VFTrack output (blue). Opening angle differences between two models are 16°, 12°, 21°, 22°, 20°, and 4°(left to right).



(a) Video-1, Frames 92 to 98

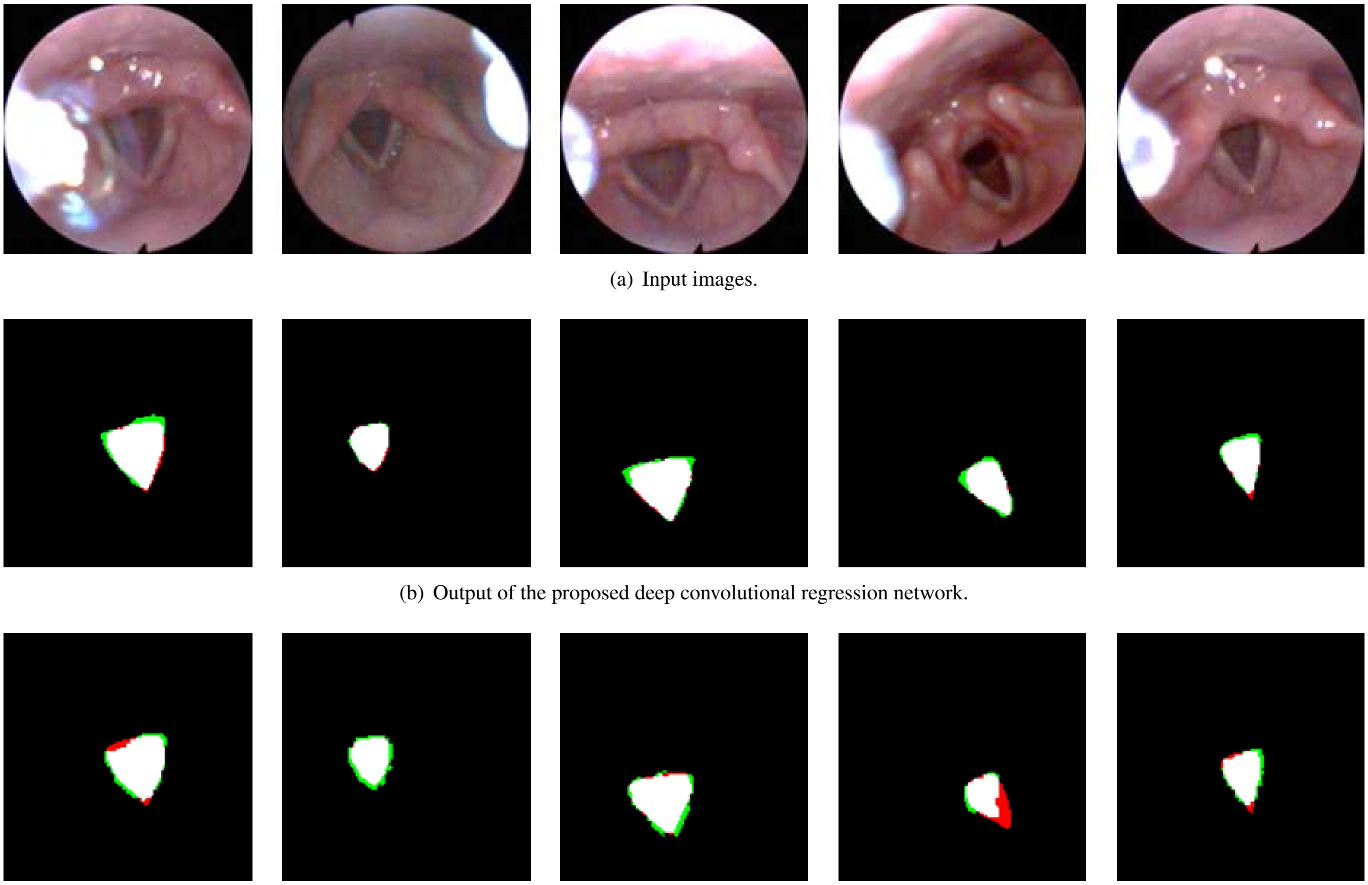
(b) Video-2, Frames 43 to 49

Figure 11. Sequence of frames for Video-1 (frame 92 to frame 98) and Video-2 (frame 43 to frame 49). Raw input images and segmentation masks obtained using proposed deep convolutional regression network overlaid on input images.

[14] D. Nie, R. Trullo, J. Lian, L. Wang, C. Petitjean, S. Ruan, Q. Wang, and D. Shen, "Medical image synthesis with deep convolutional adversarial networks," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 12, pp. 2720–2730, 2018.

[15] J. Schlemper, J. Caballero, J. V. Hajnal, A. N. Price, and D. Rueckert, "A deep cascade of convolutional neural networks for dynamic mr image reconstruction," *IEEE transactions on Medical Imaging*, vol. 37, no. 2, pp. 491–503, 2018.

[16] M. M. Haney, A. Hamad, E. Leary, F. Bunyak, and T. E. Lever, "Automated quantification of vocal fold motion in a recurrent laryngeal nerve injury mouse model," *The Laryngoscope*, 2018.



(a) Input images.

(b) Output of the proposed deep convolutional regression network.

(c) Output of the semantic segmentation network.

Figure 12. Comparison of the segmentation outputs. (a) Network input images (b) output of the proposed deep convolutional regression network (c) output of the semantic segmentation network. White, green, and red regions correspond to true positive (TP), false positive (FP), and false negative (FN) regions.

[17] MATLAB, *Deep Learning Toolbox 2017b*. The MathWorks Inc., Natick, Massachusetts, United States, 2017.

[18] L. A. Shock, B. C. Gallemore, C. J. Hinkel, M. M. Szewczyk, B. L. Hopewell, M. J. Allen, L. A. Thombs, and T. E. Lever, "Improving the utility of laryngeal adductor reflex testing: A translational tale of mice and men," *Otolaryngology–Head and Neck Surgery*, vol. 153, no. 1, pp. 94–101, 2015.

[19] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[20] J. Pont-Tuset and F. Marques, "Measures and meta-measures for the supervised evaluation of image segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2131–2138.

[21] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical association*, vol. 66, no. 336, pp. 846–850, 1971.