Research Article | Vol. 11, No. 8 / 1 August 2020 / Biomedical Optics Express 4695
Biomedical Optics EXPRESS

Two step convolutional neural network for automatic glottis localization and segmentation in stroboscopic videos

VARUN BELAGALI,[1] ACHUTH RAO M V,[2] PEBBILI GOPIKISHORE,[3] RAHUL KRISHNAMURTHY,[4] AND PRASANTA KUMAR GHOSH[5,*]

[1]Computer Science and Engineering, RV College of Engineering, Bangalore 560059, India
[25]Electrical Engineering, Indian Institute of Science, Bangalore 560012, India
[3]All India Institute of Speech and Hearing, Mysuru, 570006, India
[4]Department of Audiology and Speech Language Pathology, Kasturba Medical College, Mangalore, Manipal Academy of Higher Education, Manipal, India
[*]prasantg@iisc.ac.in

Abstract: Precise analysis of the vocal fold vibratory pattern in a stroboscopic video plays a key role in the evaluation of voice disorders. Automatic glottis segmentation is one of the preliminary steps in such analysis. In this work, it is divided into two subproblems namely, glottis localization and glottis segmentation. A two step convolutional neural network (CNN) approach is proposed for the automatic glottis segmentation. Data augmentation is carried out using two techniques : 1) Blind rotation (WB), 2) Rotation with respect to glottis orientation (WO). The dataset used in this study contains stroboscopic videos of 18 subjects with Sulcus vocalis, in which the glottis region is annotated by three speech language pathologists (SLPs). The proposed two step CNN approach achieves an average localization accuracy of 90.08% and a mean dice score of 0.65.

1. Introduction

1.1. Motivation

In speech production, vocal folds play a vital role in modulating airflow from lung through its quasiperiodic vibration [1]. In between the vocal folds, a narrow opening is present which allows air to pass through the trachea called as glottis. The changes in shape or muscular properties of vocal folds lead to changes in the glottis opening [2]. This causes variation in the voice like hoarseness and dysphonia. Some people face a vocal fold condition called Sulcus vocalis (SV). According to cadaver studies, the extensiveness rate of the SV has been found to vary from 0% to 9% [3]. In this condition a groove is formed in the vocal fold, which causes a reduction in the mass of the vocal folds. This leads to an incomplete closure of the glottis during production of voice, this is termed as glottic chink. The glottic chink is typically visualized utilizing endoscopic video by speech language pathologists (SLPs). The severity of the glottic chink in SV is assessed by examining the endoscopic video. It is challenging for the SLP to quantitatively assess the glottic chink from the endoscopic video using naked eyes. This brings the need for automatic detection and segmentation of glottis from these videos. The area quantification of the segmented glottis can be used to calculate the minimal glottis opening in the video, which could assist SLPs in their assessment.

To visualize the glottic chink in SV, a technique called stroboscopy is utilized. Stroboscopy has been a clinical standard routine for visualization of glottis and it is believed to remain as a benchmark for the next few decades, according to T. Nawka and U. Konerding [4]. The description of the typical parts of a videostroboscopic image is given in the work by L. Rudmik [5]. Stroboscopy is an important tool in the field of medical voice analysis, although it has some limitations. These limitations are due to the fact that the clinical parameters acquired from it

#396252 https://doi.org/10.1364/BOE.396252
Journal © 2020 Received 28 Apr 2020; revised 16 Jul 2020; accepted 16 Jul 2020; published 28 Jul 2020

Biomedical Optics EXPRESS

show low inter-observer reliability and are highly specific to subjects. Several factors make the stroboscopy more challenging: 1) There might be incorrect illumination in some of the images. 2) Some images might be taken at a wrong instant where glottis is occluded. 3) The change in orientation of the camera with respect to glottis causes the glottis to appear in different angles increasing the difficulty in segmentation. Illustrative images extracted from stroboscopy videos of different subjects with SV, used in this study are as shown in Fig. 1. From the figure we can see that there is a high variation of glottis shape and size across the subjects. In some subjects like (S_8, S_9), low illumination can be observed. In the case of $(S_5, S_{11}, S_{12}, S_{13})$, the glottis appears in an angle different from the rest of the images due to different orientation of the camera. Further, additional challenges arise due to the supraglottic structures that block the view of the glottis as shown in Fig. 2 during the stroboscopy recording. These are some of the challenges posed by the stroboscopic technique on automatic glottis segmentation.

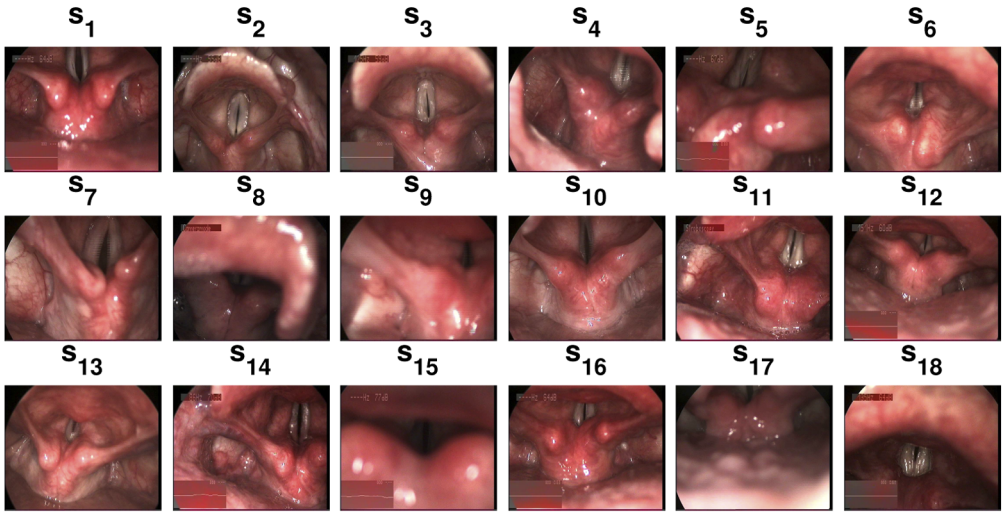

Fig. 1. Sample videostroboscopy images from 18 subjects with SV showing the variation in terms of shape of the glottis, illumination and position of the camera while recording the data.

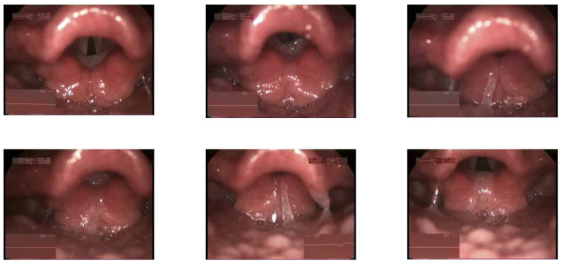

Fig. 2. Frames where the supraglottic structures block the glottis.

1.2. Literature survey

These observations underline the need for development of automatic image analysis methods on stroboscopic video to assist in diagnosis of voice disorders. The first step in this analysis is segmentation of the glottis opening region. The changes in the muscular properties of the glottis

Biomedical Optics EXPRESS

region can be studied by the shape and vibration patterns of the glottal region. The SLPs can be assisted in the severity assessment of glottic chink by automated glottis segmentation, followed by quantification of the area of the glottis opening region. There are only a few methods in the literature for automatic glottis segmentation from stroboscopic videos. Deep neural network (DNN) based approach was proposed and the problem was posed as a pixelwise classification [6]. They used RGB values from a 3×3 image patch surrounding a pixel to form a feature vector, to predict whether the pixel belongs to inside or outside the glottis region. A combination of region growing method and active shape model was proposed for glottis localization and segmentation, respectively, but the active shape model based algorithm might fail to generalize on the unseen data [7]. The use of a heat map for the detection of region of interest was proposed followed by a fully convolutional neural network (FCN) for segmentation [8]. A semi-automatic seeded region-growing algorithm was proposed to segment the glottal area [9]. A comparison was done on various CNN architectures for the glottis segmentation from laryngeal endoscopic videos [10]. A combination of CNN and long short term memory (LSTM) was proposed for the glottis segmentation from laryngeal high speed videos [11]. The LSTM was used to take the advantage of temporal relation across sequential video frames. Unlike glottis segmentation, to the best of our knowledge, there is no work on glottis localization in the literature. In this work, we hypothesize that glottis localization prior to segmentation would improve the glottis segmentation accuracy.

Image segmentation is an important problem in the field of computer vision. Recently, deep learning techniques have provided very good results for image segmentation. Among others, segmentation has been done using a Fully convolutional neural network (FCNs) which does not contain fully connected layers [12]. Examples of the FCNs include SegNet and U-Net [13] [14]. U-Net was particularly designed for bio-medical image segmentation. It has an encoder decoder architecture with skip connections from encoder layers to decoder layers to pass the information from encoder layers to the corresponding decoder layers. Such an architecture results in an effective segmentation. An application of U-Net for Artery-vein segmentation in fundus images was proposed [15]. SegNet also uses encoder decoder architecture for image segmentation but no skip connections are present. A LeNet based FCN was proposed for cell image segmentation [16]. It was shown that the FCN with transfer learning results in better segmentation [17]. When the size of dataset is smaller compared to the number of parameters of FCN, transfer learning can enhance the FCN with effective feature learning.

1.3. Key contributions and results

We divide the problem of automatically segmenting the glottis into two steps: 1) glottis localization, 2) glottis segmentation. The localization step is used to detect the region of the image where the glottis opening is present. This detected region of glottis is used in the segmentation step. The segmentation step predicts the segment of the glottis opening. We use two CNNs referred to as CNN1 and CNN2 for glottis localization and glottis segmentation steps, respectively. The CNN1 is used to draw a bounding box around the glottis region. The CNN2 is used to predict the segment of the glottis opening from the region within the bounding box. The CNN2 gets a clear representation of the glottis as the unnecessary background is removed in the localization step. We term this methodology as a two-step CNN approach. For glottis localization and segmentation, we use a modified version of SegNet as the architecture.

CNNs have a large set of parameters to be learnt from the data used in the training. The effective learning of the CNN requires collection of large set of data. In medical applications, it is, in general, difficult to obtain large annotated database as annotation requires domain experts, and is expensive and time consuming. The dataset in our study contains 921 images, which is smaller compared to the large number of parameters in the CNN. To effectively train the CNN parameters, data augmentation is done to increase the size of the dataset. Image rotation is used

Biomedical Optics EXPRESS

for augmenting the data. We use two types of data augmentation: 1) Blind rotation (WB) based data augmentation, 2) Rotation with respect to glottis orientation (WO) based data augmentation. In WB, the images are rotated clockwise and anticlockwise without the consideration of glottis orientation. On the other hand, in WO, the images are rotated clockwise and anticlockwise with respect to glottis orientation. WO is a knowledge based augmentation. In addition to simulating large dataset, the augmented data makes the model robust to glottis orientation caused by a rotation of the camera. Segmentation results are found to be better with WO images compared to those using WB images.

We perform experiments to compare the two step CNN approach trained with different data augmentation techniques. A statistical test is performed on the results achieved by various data augmentation techniques [18]. We show that the two-step CNN approach with WO based augmentation outperforms all the other methods. This shows the importance of the knowledge based augmentation. The correlation between the dice score [19] achieved by the proposed method and the inter-annotator agreement is found to be high indicating the proposed method to be equivalent to a human annotator. The proposed method performs better both in terms of localization accuracy and segmentation compared to the baseline method [6] used in this work.

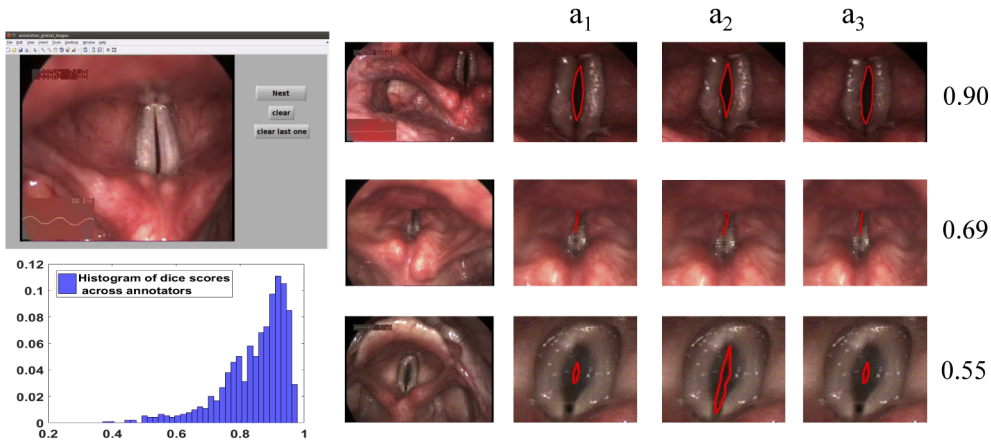

Fig. 3. Screenshot of Matlab based GUI (top left) used for annotation by SLPs. Histogram of the dice score (bottom left) calculated for all pairs of annotators to show the inter-annotator agreement and three exemplary images each annotated by three SLPs ($\mathbf{a_1}, \mathbf{a_2}, \mathbf{a_3}$) and the corresponding mean dice score obtained by averaging the dice scores obtained from all three pairs of annotators, (($\mathbf{a_1}, \mathbf{a_2}$), ($\mathbf{a_2}, \mathbf{a_3}$), ($\mathbf{a_1}, \mathbf{a_3}$)), shown in the right side of each row.

2. Dataset

Stroboscopic videos from 18 subjects with SV have been used in this work. Prior to the stroboscopic video recording, the subject was made to sit on a stool facing the recorder. The Xylocaine solution was sprayed to the participant's oropharyngeal region to eliminate gag reflex. The subject was asked to extend the tongue out and phonate the vowel /i/ (as in word 'bee') for a duration of 4 to 5 seconds. The subject was asked to repeat the phonations until a clear view of the glottis was obtained by the Otolaryngologist. Xion Endostrob E with a scope of 70 degrees was used in the recording of the videos which was operated by an Otolaryngologist.

18 patients, used in this study comprising 12 males and 6 females, are $S_i, i = 1, \ldots, 18$. The average age of subjects is 30.72 years ($\pm$ 5.65 years). The videos were chosen to ensure that there were audible recording events with a complete view of the glottis. All videos were converted into avi format with 25 frames per second. The resolution of each frame is 720×576. From these 18

Biomedical Optics EXPRESS

videos, a set of 921 frames (images) was randomly selected for annotation and was used in our study. The number of images extracted from each of 18 subjects were 102, 40, 27, 17, 97, 33, 47, 34, 10, 76, 28, 64, 39, 124, 73, 50, 19, 41, respectively. The number of phonations present in the 18 videos vary between 3 to 15. The average duration of the recorded video is 44 seconds with the duration of each video ranging from 11 seconds to 84 seconds.

The boundary of the glottis opening region was annotated by three SLPs (a_1, a_2, a_3) in each image. A graphical user interface (GUI) was designed for the annotation of the images by SLP. The GUI was developed using MATLAB. Figure 3 shows three exemplary images and the corresponding annotations done by three SLPs. It can be observed that, for the same image, there is a variation among the boundaries marked by the SLPs, especially when the glottis opening is too small to visually identify the exact boundaries by the experts. In particular, annotations corresponding to subjects S_6, S_8, S_{15} show less consistency. This could be primarily due to a low illumination present in the images. Dice score was used to measure the consistency between any two annotations from three SLPs [19]. The dice score was computed between each pair of annotators for every image followed by averaging across three pairs $((a_1, a_2), (a_2, a_3), (a_3, a_1))$. The histogram of the average dice score is shown in the Fig. 3. In most of the cases (72.31%) the dice score is above 0.8 indicating a high agreement among the annotators. But there are some images with dice scores around 0.5 due to the small glottis opening or low illumination. In such cases the SLPs might not be sure about the glottis boundaries because they are visually hard to detect.

3. Proposed two step CNN model for localization and segmentation

The task of automatically segmenting the glottis is broken into two subproblems, glottis localization and glottis segmentation. In localization step, we identify the region of interest corresponding to the glottis opening in an image by automatically drawing a bounding box around it. The localization step is performed with the intention to: 1) eliminate regions in the image which do not correspond to glottis, 2) zoom the glottis opening region with respect to its background. In the segmentation step the localized glottis is used to predict the boundary of glottis opening. We train two CNN models, one for localization (CNN1) and another for segmentation (CNN2). The advantage of the two step CNN approach is that the CNN2 gets a clear and magnified representation of the glottis opening. This results in a better feature extraction and learning about the glottis shape. If we train a single model for segmentation from the original image, it may not be able to learn glottis specific features effectively as the glottis is very small compared to the rest of the image (this is later discussed in the section 5.4.3 using Fig. 9). Figure 4 shows the block diagram of proposed two step approach. The resizing mentioned in this section uses bicubic interpolation (BI) [20].

3.1. CNN architecture

The CNN architecture used in our approach is identical to that of SegNet [13]. It has an encoder and a decoder followed by a pixelwise classification layer as shown in Fig. 4.

3.1.1. Encoder

The encoder consists of 4 blocks (Fig. 4). There are a total of 10 convolutional layers in the encoder. The first block has 2 convolutional layers (blue), with 64 filters in each having a size of 3×3 with a stride of 1 and a ReLu activation function. This is followed by a max pooling layer (red) with a window size of 2×2 and a stride of 2. The next block has 2 convolutional layers (blue) each having 128 filters with a filter size of 3×3 and a stride of 1 with ReLu activation. This is followed by a max pooling layer (red) with a window size of 2×2 and a stride of 2. The third block consists of 3 convolutional layers (blue) each having 256 filters having filter size of 3×3 with a stride of 1 and a ReLu activation function. This is followed by a max pooling layer

Biomedical Optics EXPRESS

(red) with a window size of 2 × 2 and a stride of 2. Finally, fourth layer has 3 convolutional layers (blue) each having 512 filters of size 3 × 3 with a stride of 1 and a ReLU activation function. This is followed by a max pooling layer (red) with a window size 2 × 2 and a stride of 2. The weights of the encoder are initialized with pretrained VGG weights trained on ImageNet dataset [21,22]. The pretrained weights enhance training of the model because the weights are already tuned to extract certain edge features which could be difficult to learn from the small dataset available for our study.

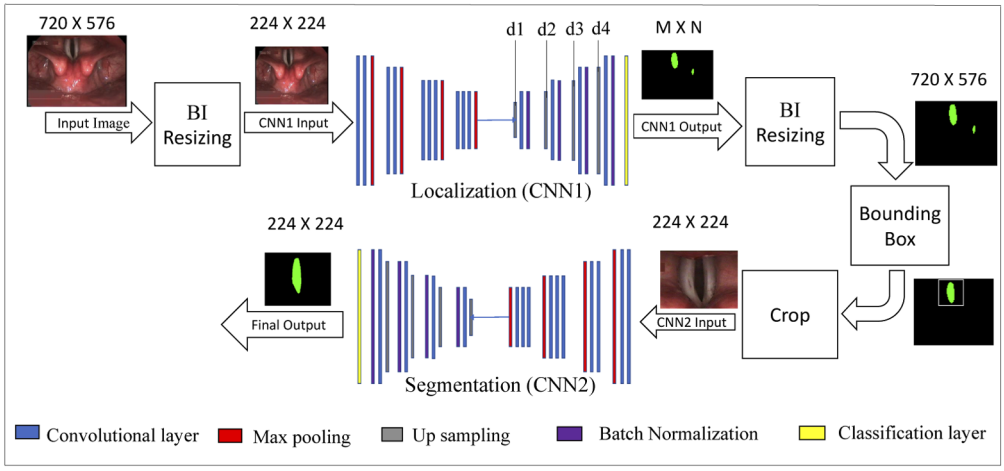

Fig. 4. Block diagram of the proposed glottis segmentation approach where an image is passed through two steps: 1) localization step which uses CNN1 architecture, 2) segmentation step which uses CNN2 architecture. BI denotes bicubic interpolation.

3.1.2. Decoder

The decoder upsamples the output of the encoder in four steps to get the final image segmentation. The decoder has 4 blocks corresponding to the four blocks of the encoder (Fig. 4). The first block has an upsampling layer (grey,d1) with a stride of 2, followed by a convolutional layer (blue) with 512 filters having a size of 3 × 3 and a stride of 1, followed by a batch normalization layer (purple). The second block has an upsampling layer (grey,d2) followed by a convolutional layer (blue) with 256 filters having a size of 3 × 3 and a stride of 1 followed by a batch normalization layer (purple). The third block consists of an upsampling layer (grey,d3) followed by a convolutional layer (blue) with 128 filters having a size of 3 × 3 and a stride of 1 followed by a batch normalization layer (purple). Finally, fourth block contains an upsampling layer (grey,d4) followed by a convolutional layer (blue) with 64 filters having a size of 3 × 3 with stride 1 followed by a batch normalization layer (purple).

In localization step the aim is to find the region of interest where glottis is present. The standard SegNet has 4 encoders blocks and 4 decoder blocks. Varying the number of upsampling layers and their positions in the decoder results in different models with variations in the output resolution of the model. The SegNet considered for our study has a output resolution same as that of the input image. The input image has a resolution of 224 × 224 in our experiment which results in an output resolution of 224 × 224. The upsampling layers used have a window size of 2 × 2, removal of an upsampling layer results in a reduction of output resolution by half. The removal of one upsampling layer results in a resolution of 112 × 112. Similarly removal of two upsampling layers results in a resolution of 56 × 56. As there are a total of four upsampling layers, four different models can be created by removing each of these layers at a time. Similarly,

Biomedical Optics EXPRESS

six models can be formed by removing all possible pairs of upsampling layers. Thus, the removal of one and two upsampling layers results in a combination of 10 models. We compare all these 10 models along with the model having all upsampling layers, i.e., the standard SegNet architecture. We train the two step CNN approach with localization network (CNN1) separately having each of the eleven combinations and the segmentation (CNN2) having the standard SegNet architecture. The 11 combinations of the localization network are denoted as $M_{d1,d2,d3,d4}$, where di is 1 (0) if the upsampling layer di is present (absent) in the deocder of the localization network.

3.1.3. Classification layer

It takes the decoder output as the input and predicts the class to which each pixel belongs. It consists of a convolutional layer with a filter of size 3×3 and stride of 1 with a sigmoid activation function (yellow).

3.2. Glottis localization

We pose the glottis localization as a pixelwise classification problem for which CNN1 model is used. The pixelwise classification is used to predict the pixels which belong to the glottis region. Then the bounding box is drawn around the area comprising the pixels predicted to find the region of interest.

3.2.1. CNN1 based pixel wise classification

In this step, each pixel of the image is classified as belonging to inside or outside the glottis region. For this purpose, CNN1 is used. Each image of size 720×576 is resized to 224×224, because the input size of CNN1 is 224×224. This image is then passed to the CNN1 for pixelwise classification to determine whether a pixel is inside or outside the glottis region. The objective function used for training the CNN1 is a weighted binary cross entropy. The weight ratio used for objective function is 200:1 from inside to outside the glottis region, as the average ratio of the number of pixels inside the region to the number of pixels outside the region is 1:200. This is done to obtain a class balanced objective function. The weights are updated using Adam optimizer [23]. The output of the CNN1 is either 224×224 or 112×112 or 56×56 sigmoid output (Pr) depending on the number of upsampling layers removed in CNN1. (i,j)-th element of Pr is denoted by $Pr_{i,j}$. We determine the pixel wise classification by thresholding. The thresholding method is utilized to form a binary image which is further resized to 720×576. The thresholded binary output is obtained in the following manner:

$$P_{i,j} = \begin{cases} 1 & Pr_{i,j} > max_{i,j}(Pr_{i,j}) - \epsilon \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

The value of ϵ is 0.05. It was observed that the variation of ϵ in the range of 0.01 to 0.1 did not significantly affect the bounding box prediction. The pixels inside the glottis region are labelled as '1' and those outside as '0'.

3.2.2. Drawing the bounding box

The CNN based pixelwise classification independently classifies whether a pixel is inside or outside the glottis region. Due to this independent classification across pixels, there is no constraint that the pixels with label '1' should form a single region. We spatially cluster all pixels with label '1' into regions. The binary image predicted by CNN1 is scanned column by column and the pixel labeled as '1' is assigned a cluster number. The cluster number assigned to the pixel is majority of the cluster numbers assigned to its neighbours. The pixel is assigned a new cluster number, if all its neighbours have not been assigned a cluster number [24]. The region with the highest area (i.e., the highest number of pixels) is selected as the region of interest for glottis

Biomedical Optics EXPRESS

opening. The bounding box is drawn by obtaining the centroid of the region and treating it as the centre of the box (white box in Fig. 4). The input shape of segmentation network (CNN2) is 224 × 224. So, a bounding box of size 224 × 224 is drawn. If the region size exceeds 224 × 224 pixels, a larger bounding box is drawn to accommodate the full region. This resized cropped image from the bounding box is taken as the input to the glottis segmentation.

3.3. Glottis segmentation

The final segmentation is performed on the localized glottis image. The CNN2 is trained for the segmentation. The dataset used in training is generated from the localization step. The bounding box is drawn on both the images and the corresponding annotations. A new dataset comprising cropped images and annotations is generated. During the training phase if the bounding box predicted by the localization step is incorrect on the training sample i.e., if the bounding box partially contains the glottis or it does not contain the glottis, then there is a localization error. Irrespective of this error, the cropped images and annotations are used in training CNN2. This helps in training CNN2 to predict no glottis for the incorrectly localized images that do not contain glottis and to predict partial glottis segment in the images where the bounding box contains glottis partially.

3.3.1. CNN2 based pixel wise classification

Each pixel in the image is classified as inside or outside the glottis region using CNN2. If the image from the result of section 3.2.2 has a size larger than 224 × 224, it is resized to 224 × 224 using BI and passed as an input to the CNN2. The objective function used for training is a weighted binary cross entropy. The weight is in the ratio 2:1 for the pixels inside and outside glottis. The weights are updated using SGD optimizer [25] with a fixed learning rate of 0.01. The CNN2 produces a 224 × 224 sigmoid output. We determine the pixelwise classification using a threshold of 0.5 to form a binary image of size 224 × 224.

3.3.2. Choosing the segment of glottis opening

Usually the predicted segmentation has only one region labelled with '1's. But if there are more than one region, the clustering method mentioned in 3.2.2 is applied. The region with the largest area is selected and marked as the final segment of the glottis opening.

4. Data augmentation

The number of images in the dataset in this work is smaller compared to the large number of parameters in the CNN. Data augmentation is done to increase the number of annotated images simulating an availability of a large dataset. The most popular data augmentation technique for images is rotation [26]. Two types of rotations are used for data augmentation in this work as described below.

4.1. Blind rotation (WB)

In blind rotation, glottis orientation is not taken into account. The images of size 720 × 576 and the corresponding annotations are rotated by same angle to generate the augmented data. The image is rotated clockwise and anticlockwise with angles up to 90 degrees in steps of 3 degrees. This creates a collection of 60 additional images for each image in the original dataset. The regions at the corners generated by rotation are filled with zeros when the corresponding pixels are outside the dimension of the original image. This process of data augmentation creates a dataset comprising 55,260 images.

Biomedical Optics EXPRESS

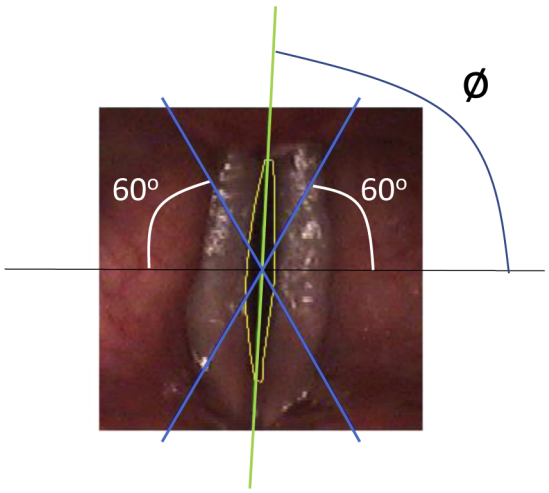

Fig. 5. Orientation estimation involved in WO based augmentation process.

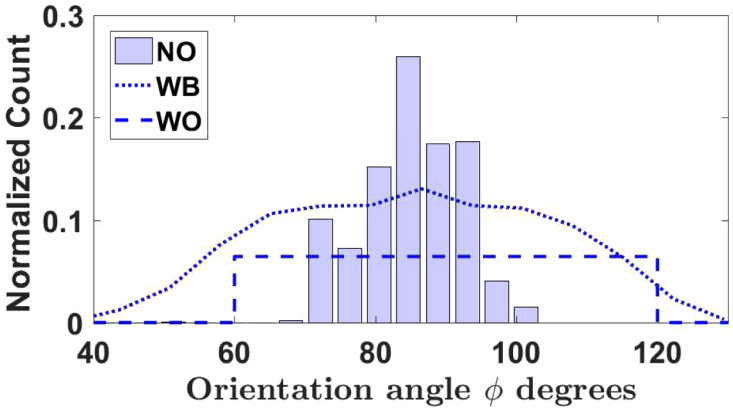

Fig. 6. Histogram of orientation angle (ϕ) between the major axis of glottis and the horizontal axis measured in degrees in NO, WB and WO based augmentation.

4.2. Rotation with respect to glottis orientation (WO)

An ellipse is fit to the manually marked boundary of the glottis opening such that they have identical second-moment [27]. Following this, an angle between the horizontal axis and the major axis of this ellipse is found. This angle ϕ is termed as glottis orientation. Figure 5 shows the process of data augmentation in which orientation of glottis is accounted. The yellow contour represents the annotation boundary marked by an SLP. The green line represents the major axis of the ellipse, fit to the glottis opening boundary. The two blue lines represent limits of rotation. The image is rotated with a step size of one degree such that the resultant ϕ ranges from 60 to 120 degrees (blue lines).

The blue color bar plot in Fig. 6 shows the histogram of the glottis orientation (ϕ) of all the images present in the dataset. This is indicated by NO as no augmentation is performed prior to computing the histogram. This shows that the range of ϕ is between 70 to 110 degrees in the dataset. In blind rotation, the image is rotated irrespective of the glottis orientation. This does not ensure uniform orientation angle post rotation (dotted plot in Fig. 6, indicated by WB) because different frames may have different orientation. Hence, blind rotation results in different

Biomedical Optics EXPRESS

ranges of glottis orientation angle after data augmentation. The glottis opening annotations by the SLPs reveal that the angle of orientation of glottis can be of any value between 60 to 120 degrees. To accommodate small orientational error outside the range between 70 to 110 degrees, a buffer of 10 degrees is considered. The orientation outside this range is not practical in terms of camera orientation error. Each orientation is assumed to be equally likely in the range of 60 to 120 degrees while recording. So an augmentation scheme that results in uniform orientation angles will make the model robust to such orientation error. We rotate the image with respect to glottis orientation in WO based augmentation. The major axis of the glottis is found from the annotated boundary. The image is rotated such that the resultant ϕ after rotation covers the entire range from 60 degrees to 120 degrees. To achieve a large dataset with all possible angles between 60 and 120 degrees, the step of rotation is chosen as 1 degree. Thus, for every original image there are 60 images created in the WO based augmentation. This results in a total of 55,260 images, identical to that obtained by WB based augmentation scheme. Considering all these images, the glottis forms a uniform orientation angle between 60 and 120 degrees (dashed plot in the Fig. 6, indicated by WO). We hypothesize that using WO, the model would be more robust towards the rotation of the camera while recording compared to that using WB.

5. Experiments and results

5.1. Experimental setup

The entire dataset is divided into 4 folds namely fold1 : $\{S_1, S_2, S_3, S_4\}$, fold2 : $\{S_5, S_6, S_7, S_8, S_9\}$, fold3 : $\{S_{10}, S_{11}, S_{12}, S_{13}, S_{14}\}$ and fold4 : $\{S_{15}, S_{16}, S_{17}, S_{18}\}$. The subjects belonging to each fold were chosen such that each fold has approximately same number of frames. In the training phase we use three folds among which one fold is used as the validation set. Test phase comprises the remaining one fold. The folds are used in a round robin fashion for the training and testing. The annotations used in training are obtained from a_1. In the testing phase, annotations from all three annotators are used for evaluation. The pixels belonging to the region inside the glottis opening are denoted as '1' and '0' is used as the label for the pixels outside the glottis opening. The CNN1 used for localization is trained first and then the cropped images and ground truth from this step is used in training the CNN2. The CNN1 and CNN2 are trained separately because the bounding box is a discontinuous function and, hence, poses challenge in joint training of CNN1 and CNN2. The dataset considered for CNN1 contains the original images and the corresponding annotations in the training set. The dataset for CNN2 consists of cropped images and cropped annotations obtained from the training data based on the bounding box predicted by the CNN1. A batch size of 16 was used for training. The CNN models were implemented using keras and theano libraries and trained on Nvidia Titan X GPU with 12 GB of RAM [28,29].

Among the 11 models considered (section 3.1.2), the best two-step CNN model is picked based on the median dice score achieved on the validation data. Further, to improve the best model, the data augmentation is carried out. Three types of data augmentation is done : 1) no data augmentation (NO). 2) WB based data augmentation. 3) WO based data augmentation. The augmented data is only used in the training phase. In the testing phase, the original data without any augmentation is used. Student's *t*-test is performed to compare the results achieved by various methods.

5.2. Baseline scheme

As a baseline we use the method proposed by [6] which uses the 3×3 neighborhood RGB values around a pixel as a feature vector of length 27. A DNN based approach is used to predict the probability of individual pixel belonging to the glottis opening. They pose the problem as a pixelwise classification problem. The DNN contains 3 layers each with 128 units and a single sigmoid output unit. The predicted pixels are clustered together and a post processing method

Biomedical Optics EXPRESS

is applied to pick the glottis region. They achieve a localization percentage of 65.33% and an average dice score of 0.39. We use the dataset and fold structure within the dataset identical to the baseline to evaluate our approach.

5.3. Evaluation metrics

We use three metrics to evaluate the proposed approach: a) Localization accuracy, b) Dice score, c) B_a.

Localization accuracy is measured as the percentage of images present in the test set whose centroid of predicted segment falls inside the annotated region marked by SLPs.

Dice score is used to measure the segmentation accuracy [19]. Dice score is a believed to be the standard metric for single class shape predictions as it measures the shape similarity and overlap between the predicted and ground truth shapes [30]. The formula of dice score is as follows:

$$D(P, G) = \frac{2 \times N(P \cap G)}{N(P) + N(G)} \tag{2}$$

where N represents the number of pixels with label '1' and P and G represent the predicted output labels and the ground truth (annotation) labels. The pixelwise accuracy is not used as a metric, as the ground truth labels for within and outside glottis opening are highly biased. The average ratio of the number of label '1's and label '0's in annotations is 1 : 200. So a prediction with label '0' for all pixels would result in a good pixelwise accuracy but a dice score of zero.

The performance of the CNN1 in two step CNN methods in evaluated by a metric B_a. We define B_a as the percentage of the glottis area covered by the bounding box drawn from the predictions of CNN1.

5.4. Results and discussion

For comparison, the segmentation is also performed by only training CNN1 followed by clustering. This method is referred as to CNN_C. The CNNs in the two step CNN and CNN_C are trained without data augmentation (NO) as well as WB based and WO based augmentation. Each approach is evaluated on three annotators. This results in a total of six approaches namely CNN_C, two step CNN, CNN_C WB, two step CNN WB, CNN_C WO and two step CNN WO.

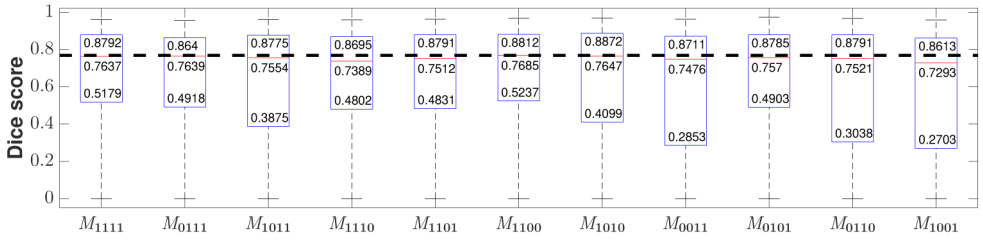

Fig. 7. Boxplot of dice score achieved by the 11 combinations of network architecture on validation data. M_{1100} achieves the highest median dice score indicated by dashed horizontal line. The labels on the boxplot indicate quartiles and medians.

5.4.1. Selection of upsampling layer

Fig. 7 shows the boxplot of the dice scores achieved by the eleven combinations in the two step CNN approach [31]. These models are trained without any data augmentation. The dice scores are evaluated on the validation set across all the folds to select the best model. We use the median dice score (red lines in boxplot) to compare the models. The model M_{1100} has the highest median dice score of 0.7685. We pick this model as the best two step CNN model and perform the

Biomedical Optics EXPRESS

data augmentation experiments on it. The removal of the upsampling layer further degrades the performance of the two step CNN approach because the resolution of the localization step output is too low to represent the glottis region. From now onward, two-step CNN refers to the proposed segmentation scheme where M_{1100} is used in CNN1. On the other hand, CNN_C refers to the scheme where only CNN1 is used followed by clustering as described in section 3.1.2.

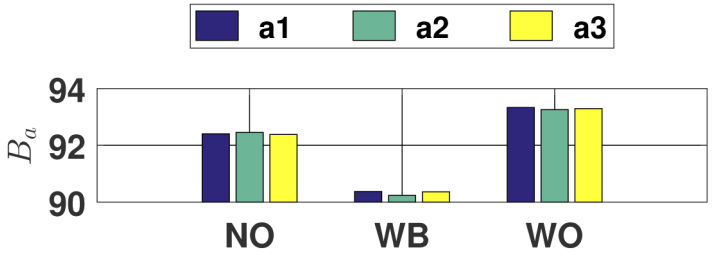

Fig. 8. B_a using CNN1 of the two step CNN approach with three data augmentation schemes, NO, WB and WO evaluated on three SLPs, namely a1, a2, a3. It is clear that the WO data augmentation technique achieves the best B_a.

5.4.2. Performance of CNN1

The performance of the CNN1 in two step CNN methods is measured by B_a. NO is used to indicate data without any augmentation. The percentage of the training images for which the B_a is equal to 100% in NO, WB and WO based augmentation are 95.71%, 87.63% and 93.59%, respectively. Figure 8 shows the B_a calculated using CNN1 of the two step CNN method evaluated on three annotators on the test data. The average B_a achieved by NO, WB and WO based augmentation are 92.41%, 90.32% and 93.29%, respectively. The CNN1 of the two step CNN WO outperforms CNN1 of two step CNN WB and CNN1 of the two step CNN NO in terms of B_a. This shows that WO based augmentation achieves a better localization compared to WB and NO.

5.4.3. Two step CNN vs CNN_C

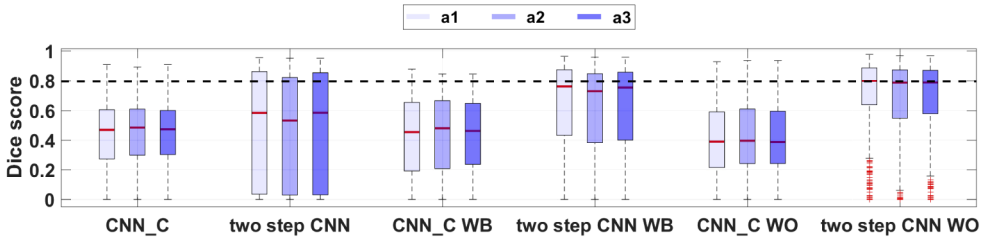

Fig. 9. Boxplot of dice score obtained using CNN_C and two step CNN approaches without augmentation (NO), with WB based augmentation and WO based augmentation. The dashed horizontal line indicates the average of three median dice scores obtained using the two step CNN WO evaluated on three SLPs.

Fig. 9 shows the box plot of dice scores on the test data achieved by six approaches evaluated on three annotators. The horizontal line in each box represents the median of the dice scores from respective evaluations. The average mean and median are defined as the mean and median dice scores averaged across three annotators, respectively. The dice score achieved by the six approaches are as follows : CNN_C median 0.47 (0.44 ± 0.23), two step CNN median 0.57

Biomedical Optics EXPRESS

(0.48 ± 0.37), CNN_C WB median 0.46 (0.43 ± 0.26), two step CNN WB median 0.75 (0.60 ± 0.33), CNN_C WO median 0.39 (0.41 ± 0.25) and two step CNN WO median 0.79 (0.67 ± 0.29). We do a statistical analysis to compare all the above mentioned segmentation methods. The Student's *t*-test is performed on the dice scores obtained from all pairs of methods. Table 1 shows the *p*-value obtained by performing the Student's *t*-test between every pair of methods. The threshold of *p*-value used for identifying the significance is 0.01. It is observed that the dice score from CNN_C is significantly ($p< 10^{-5}$) higher than that from CNN_C WO. The dice score from two step CNN is significantly ($p< 10^{-5}$, 10^{-10}, 10^{-16}) higher than that from CNN_C, CNN_C WB and CNN_C WO. The difference in dice scores between CNN_C WB and CNN_C WO is not significant ($p = 0.1323$). The dice score from two step CNN WB is significantly ($p< 10^{-83}$, 10^{-30}, 10^{-89}, 10^{-109}) higher than that from CNN_C, two step CNN, CNN_C WB and CNN_C WO. The dice score from two step CNN WO is significantly ($p< 10^{-183}$, 10^{-77}, 10^{-186}, 10^{-13}, 10^{-216}) higher than that from CNN_C, two step CNN, CNN_C WB, two step CNN WB and CNN_C WO. The two step CNN WO outperforms all the other methods in terms of both the average median and mean dice scores. All the two step CNN methods outperform their respective CNN_C methods, which shows the importance of adding the CNN2 for segmentation. Therefore the two step CNN methods result in a better segmentation compared to that using a CNN_C. We resize the input image from 720 × 576 to 224 × 224 in CNN_C. Hence, the glottis opening occupies less number of pixels in the input image and the boundary of the glottis opening may not be clearly present. In case of the two step CNN, the 224 × 224 box is drawn on the image, so the number of pixels representing the glottis is same as that of the original image. The bounding box obtained from the localization step increases the percentage of pixels belonging to the glottis opening region compared to considering the entire image. The localization step helps the segmentation step to have a clear region of interest where the boundary of glottis is clearly present. An other advantage of the CNN1 in the proposed two step CNN is that the glottis in an image the CNN2 receives, is always located in the center of the image, making it easier for the CNN2 to segment it. This ensures a good quality segmentation. Overall, the dice score is enhanced due to two step CNN and WO based augmentation method. Therefore the two step CNN WO is selected as the best model and is used for comparison with the baseline.

Table 1. *p*-values of the Student's *t*-test performed on dice scores between all pairs of segmentation methods. A bold entry in a cell indicates that the method in the corresponding column has higher mean dice score than that in the corresponding row. Blue entry in a cell indicates that the methods in the corresponding row and column do not yield significantly different average dice score

	CNN_C	Two-step CNN	CNN_C WB	Two-step CNN WB	CNN_C WO	Two-step CNN WO
CNN_C	*	**$<10^{-7}$**	0.2506	**$<10^{-83}$**	$<10^{-5}$	**$<10^{-183}$**
Two-step CNN	$<10^{-7}$	*	$<10^{-10}$	**$<10^{-30}$**	$<10^{-16}$	**$<10^{-77}$**
CNN_C WB	0.2506	**$<10^{-10}$**	*	**$<10^{-89}$**	0.1323	**$<10^{-186}$**
Two-step CNN WB	$<10^{-83}$	$<10^{-30}$	$<10^{-89}$	*	$<10^{-109}$	**$<10^{-13}$**
CNN_C WO	**$<10^{-5}$**	**$<10^{-16}$**	0.1323	**$<10^{-109}$**	*	**$<10^{-216}$**
Two-step CNN WO	$<10^{-183}$	$<10^{-77}$	$<10^{-186}$	$<10^{-13}$	10^{-216}	*

Fig. 10 shows the bar graph of localization accuracy achieved by CNN_C, two step CNN, CNN_C WB, two step CNN WB, CNN_C WO and two step CNN WO. It is evaluated across all three annotators. Increased bar height for two step CNN compared to CNN_C method suggests that the localization accuracy is more using two step CNN methods compared to CNN_C methods. The increase in both dice score and localization accuracy further underlines the advantage of two step CNN methods. The average localization accuracy across 3 annotators using CNN_C, two step CNN, CNN_C WB, two step CNN WB, CNN_C WO and two step CNN WO are 74.18%,

Biomedical Optics EXPRESS

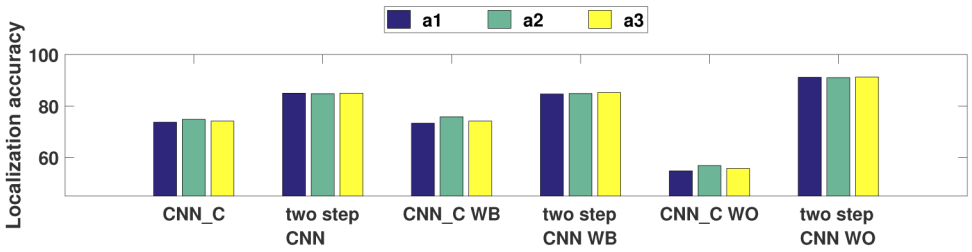

Fig. 10. Bar graph indicating localization accuracy using various approaches evaluated across three annotators.

84.90%, 74.40%, 84.90%, 55.79%, 91.14%, respectively. The two step CNN WO achieves the highest localization accuracy.

5.4.4. Proposed method vs baseline

Table 2. Localization accuracy and corresponding dice score achieved by the baseline and the proposed approach

	Localization accuracy (a)		Dice score (b)	
	Baseline	Proposed	Baseline	Proposed
fold1	89.9,78,79.1	97.75,98.09,98.07	0.56,0.54,0.54	0.66,0.62,0.62
fold2	38.0,69.0,40.0	93.92,92.82,92.82	0.28,0.31,0.28	0.60,0.54,0.59
fold3	72.2,68.6,72.8	80.56,81.67,81.68	0.38,0.42,0.38	0.79,0.78,0.77
fold4	39.9,76.2,41.4	88.18,87.27,88.18	0.32,0.32,0.32	0.65,0.63,0.63
average	**60.0,73.1,63.2**	**90.10,89.96,90.18**	**0.39,0.40,0.38**	**0.68,0.64,0.65**

Table 2a shows the fold-wise localization accuracy calculated across three SLPs using the baseline and the proposed two step CNN WO approach. Table 2b shows corresponding fold-wise dice score calculated across three SLPs using the baseline and the proposed approach. Both localization accuracy and dice scores are better using the proposed two step CNN WO approach compared to the DNN based baseline across all the three SLPs. The p-value obtained from the Student's t-test on localization accuracy and dice score between proposed method and DNN method are 0.0002 and 0.0005 respectively. The reason for the increased dice score could be due to the fact that a CNN has a larger receptive field for each pixel as the network gets deeper, whereas the DNN based approach only takes 8 pixels neighbourhood for prediction. There might be a lot of false positives in DNN method before clustering, which do not belong to a glottis region, rather belong to some dark region in the image where the illumination is similar to that of the glottis. In DNN approach, if such false positive regions are bigger than the glottis, they get selected after clustering which leads to poor localization. The average localization accuracy and dice score of the proposed approach is 90.08% and 0.65 which are better than those using the baseline scheme by 24.64% and 0.26 (absolute), respectively.

Fig. 11 illustrates glottis segmentation on several sample images, one from each of the 18 subjects with SV. Column (a) is labelled with the subject number with the corresponding average dice score (in blue) achieved by the proposed approach on the sample image calculated across 3 SLPs and the corresponding average inter-annotator agreement (in red). Column (b) contains the frames extracted from videos of corresponding subject with bounding box obtained from the localization step. The resolution of a frame is 720×576. Column (c) contains images within the bounding box predicted from the localization step on the original image. The resolution of this image is 224×224. Column (d) contains the glottis opening segment predicted in the

Biomedical Optics EXPRESS

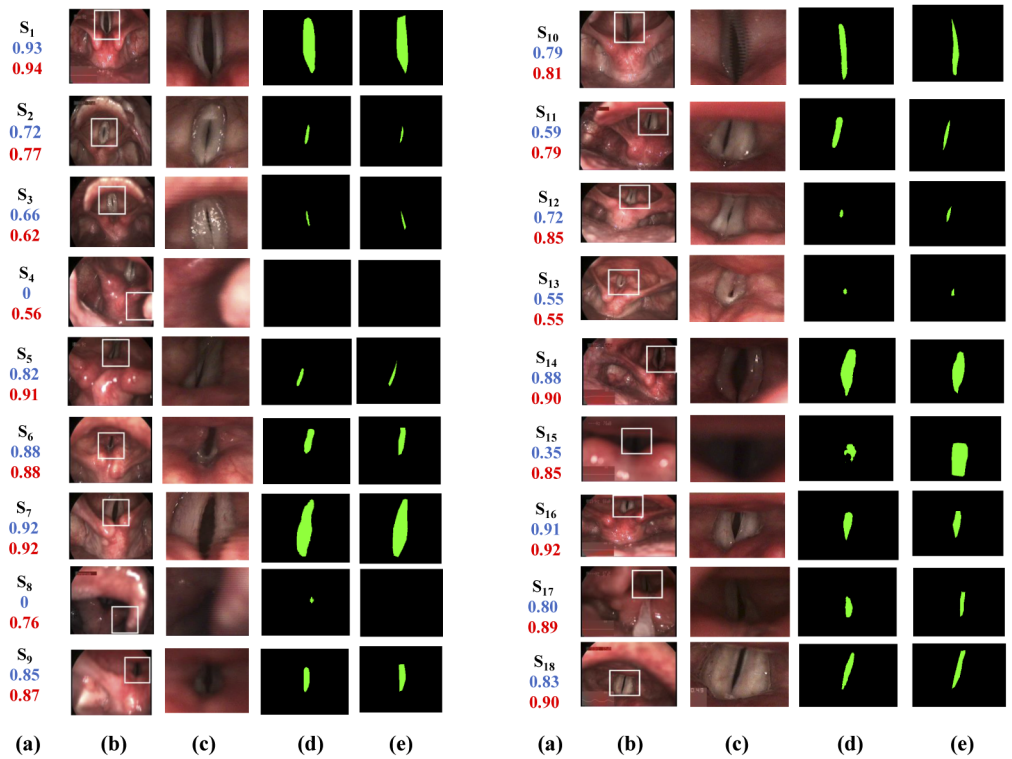

Fig. 11. Illustration of the glottis segmentation on various sample images, one from each of the 18 subjects with SV. Column (a): Subject number with corresponding dice score achieved by the proposed approach (blue) and average inter annotator agreement (red) on the sample image obtained by averaging the dice scores from every pair of annotators. Column (b): 720×576 original image with bounding box obtained in the localization step. Column (c): cropped image obtained from the localization step. Column (d): The predicted segment by the two step CNN approach. Column (e): Groundtruth segment corresponding to the cropped image labeled by annotator a1.

segmentation step. Column (e) contains the groundtruth from a_1 obtained by drawing a bounding box (obtained from the localization step) on the annotation. Four different patterns can be observed from the results: 1) The glottis opening is of large size and a good illumination is present. 2) The glottis opening is very small and a good illumination is present. 3) The boundary of glottis is not clearly visible because of low illumination. 4) The glottis opening is blocked by supraglottic structures. In case 1 both the localization accuracy and the dice score are high. The bounding box drawn over the ground truth covers the whole segment i.e., the B_a is 100%. The dice score is high because the proposed approach is able to detect the exact boundary. A slight deviation from the ground truth decreases the dice score by a small amount. This is because of the large size of the glottis opening. The subjects S_1, S_6, S_7, S_{10} and S_{14} come under case 1. In case 2 the localization accuracy is high, but not the dice score. The reason for less dice score is that the glottis opening region is small, so a small deviation from ground truth would result in a large percentage of mismatch between the predicted and ground truth glottis segment. The subjects S_2, S_3, S_4, S_5, S_9, S_{11}, S_{12}, S_{13}, S_{16}, S_{17} and S_{18} come under case 2. High dice score for some subjects are observed even for case 2 because the edge of glottis is clearly present. The glottis opening in case of subject S_4 is too small to be detected by localization step. The localization step goes wrong in such case which results in zero dice score. In case 3 the localization accuracy is high

Biomedical Optics EXPRESS

because the glottis region is much darker than the surrounding. The subject S_{15} comes under case 3. Because of poor illumination, the glottis opening boundary is not visible. The CNN2 fails to predict the exact glottis which results in a poor dice score. In case 4, the supraglottic structures block the view of the glottis. In case of S_8 the epiglottis completely blocks the glottis. Hence, the bounding box drawn by the localization step does not contain glottis. Therefore segmentation results in a dice score of zero. The outliers present in the box plot in Fig. 9 correspond to the images which belong to case 3 and case 4 as well as condition that arises for subject S_4.

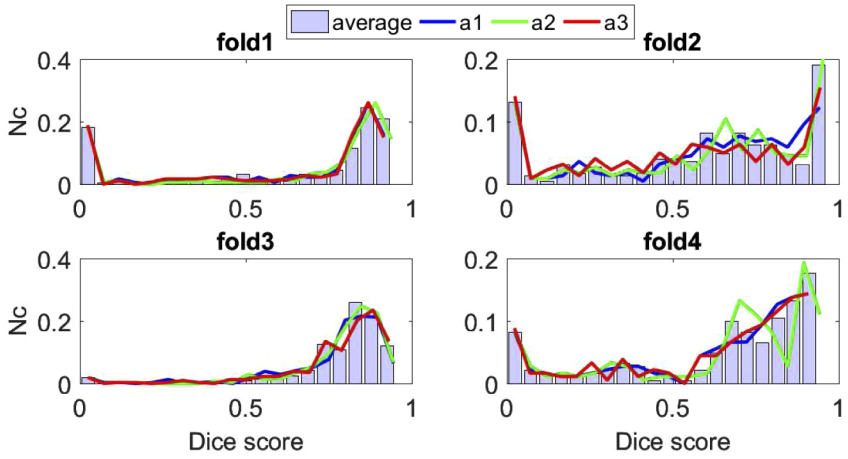

Fig. 12. Foldwise normalized count (Nc) histogram of mean dice scores from every fold in the test set evaluated on three SLPs and three histogram curves corresponding to evaluation on three annotators separately.

Figure 12 shows the histograms of mean dice score between a predicted glottis boundary and the corresponding ground truth annotation across all annotators. The figure also shows the histogram curves corresponding to the dice score evaluated separately on three annotators. There are four histograms corresponding to four folds used in the test set. From the figure, it is clear that all the four histograms have good amount of frames having a dice score greater than 0.8. The frames corresponding to the dice score greater than 0.8 belong to case 1 and case 2, as described in Fig. 11. Most images belonging to case 2 spread across the dice score ranging from 0.4 to 0.9. The frames that fall between 0 to 0.1 in fold1 correspond to S_4 of case 2. Those between 0 to 0.1 in fold2 correspond to case 4. The short histogram bars between 0 and 0.1 in fold3 correspond to S_{12} of case 2 which has a very small glottis that go undetected sometimes. The frames between 0 to 0.1 in fold4 belong to case 3.

Fig. 13 shows the plot of average dice score across three SLPs versus the average dice score of the proposed algorithm evaluated on three SLPs. It represents a correlation between the inter-annotator agreement and the dice score achieved by the proposed algorithm. Only the dice scores of correctly localized images are taken into account. A correctly localized image is an image in which the centroid of the predicted glottis opening segment falls inside the annotated glottis opening. The l_1 line fit [32] to the points in the plot has a slope of 1. It can be inferred that the average dice score using the proposed algorithm follows the trend of inter-annotator agreement with a correlation coefficient of 0.46. The average dice score is low, when inter-annotator agreement among the SLPs is low. Average dice score is high when the inter-annotator agreement is high. This suggests the robustness of the proposed approach.

Fig. 14 shows a scatter plot of average dice score obtained using the proposed method evaluated across three SLPs versus the normalized glottis opening area evaluated on each image. The normalized glottis opening area is the ratio of number of pixels which are present inside the

Biomedical Optics EXPRESS

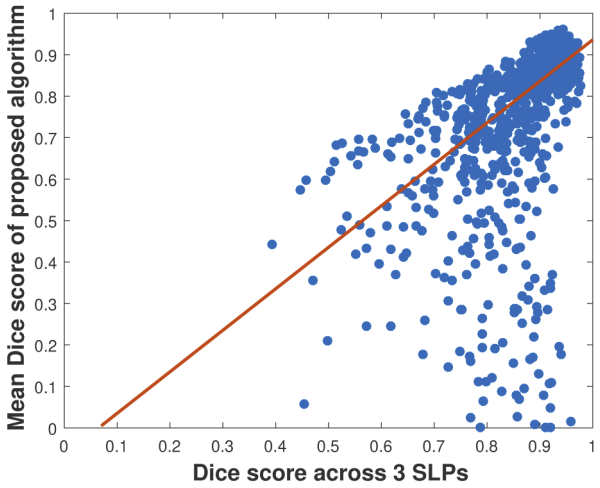

Fig. 13. Mean dice score averaged across three SLPs vs the mean dice score obtained using the proposed algorithm evaluated on three SLPs and the l_1 line fit to plot (red).

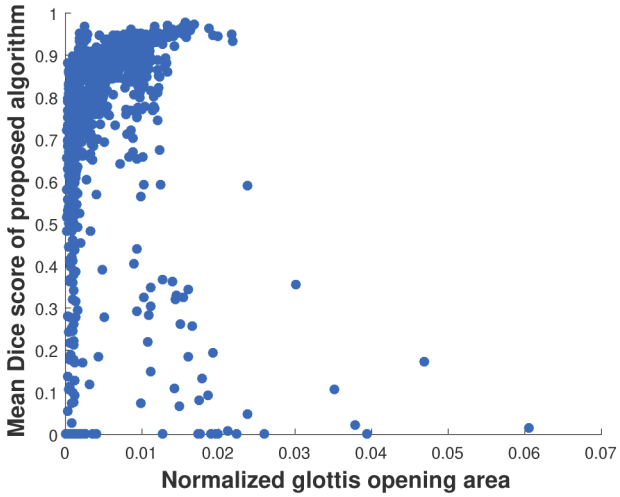

Fig. 14. Mean dice score using the proposed method evaluated across three SLPs vs the normalized glottis opening area with respect to the image size.

Biomedical Optics EXPRESS

glottis region and the total number pixels in the image. The figure shows how the glottis area and the dice score could be related. The dice scores in the range of 0.6 to 1 show an exponential increase in number of images with increase in the normalized glottis opening area. Dice scores tend to be higher for the larger glottis opening. When glottis opening is large, a small number of pixel classification error decreases the dice score by a small amount, since the percentage of misclassification is less due to the large opening. But in case of small glottis opening, a small number of pixel misclassification would result in a moderate percentage of misclassification. The dice score tends to be lower for images where small glottis opening is present. The purpose of glottis segmentation is to segment the small glottis regions accurately. Dice score penalizes the small number of misclassifications more when the glottis opening is relatively smaller. This indicates that dice score is a good evaluation metric for glottis segmentation.

6. Conclusion

We present a two step CNN method for automatic localization and segmentation of glottis. In the localization step, a bounding box is automatically drawn around the glottis region detected. This is passed to a second CNN for final segmentation of localized glottis. A dataset consisting of stroboscopic videos from 18 subjects with SV are annotated by three SLPs which is used to evaluate the proposed approach. The two step CNN WO method achieves an average localization accuracy of 90.08% and a dice score of 0.65 outperforming the baseline scheme by 24.64% and 0.26, respectively. The localization network is experimented with variations in the upsampling layers to find the best localization network. The dice score achieved by the two step CNN WO method shows a dependence on the glottis opening area. As a part of future work we would like to impose the glottis specific shape constraints on the CNN architecture to improve the segmentation results. It will also be interesting to design a smooth localization algorithm rather than posing it as a segmentation problem and then drawing the bounding box. We would also want to experiment with the mask R-CNN which predicts both localization and segmentation. In segmentation step, we would want to implement a recursive pass to the CNN to fine tune the predicted segment. We also want to experiment on the methods for quantifying the minimal glottis opening in the stroboscopic recording from the segmentation results. Although the evaluation of the proposed two step CNN method is done only on stroboscopic video from subjects with SV, the proposed method is not limited to this voice disorder only. As videostroboscopy is the primary method for clinical assessment of various voice disorders, we would like to evaluate the effectiveness of proposed method on videostroscopic recordings from individuals with voice disorders that result in similar form of glottic chink such as the vocal fold paralysis or presbyphonia.

Disclosures

The authors declare that there are no conflicts of interest related to this article.

References

1. I. R. Titze and F. Alipour, *The myoelastic aerodynamic theory of phonation* (National Center for Voice and Speech, 2006).
2. 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 Trans. Biomed. Eng. **62**(3), 795–806 (2015).
3. J. Demeyer, T. Dubuisson, B. Gosselin, and M. Remacle, "Glottis segmentation with a high-speed glottography: a fully automatic method," in *3rd Adv. Voice Funct. Assess. Int. Workshop*, (2009).
4. T. Nawka and U. Konerding, "The interrater reliability of stroboscopy evaluations," J. Voice **26**(6), 812.E1–812.E10 (2012).
5. L. Rudmik, *Evidence-based Clinical Practice in Otolaryngology* (Elsevier Health Sciences, 2018).
6. A. Rao MV, R. Krishnamurthy, P. Gopikishore, V. Priyadharshini, and P. K. Ghosh, "Automatic glottis localization and segmentation in stroboscopic videos using deep neural network," in *Proc. Interspeech 2018*, (2018), pp. 3007–3011.
7. J. J. Cerrolaza, V. Osma-Ruiz, N. Sáenz-Lechón, A. Villanueva, J. M. Gutiérrez-Arriola, J. I. Godino-Llorente, and R. Cabeza, "Fully-automatic glottis segmentation with active shape models," in *MAVEBA*, (2011), pp. 35–38.

Biomedical Optics EXPRESS

8. J. Lin, E. S. Walsted, V. Backer, J. H. Hull, and D. S. Elson, "Quantification and analysis of laryngeal closure from endoscopic videos," IEEE Trans. Biomed. Eng. **66**(4), 1127–1136 (2019).
9. 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," Med. Image Anal. **11**(4), 400–413 (2007).
10. M.-H. Laves, J. Bicker, L. A. Kahrs, and T. Ortmaier, "A dataset of laryngeal endoscopic images with comparative study on convolution neural network-based semantic segmentation," Int. J. CARS **14**(3), 483–492 (2019).
11. M. K. Fehling, F. Grosch, M. E. Schuster, B. Schick, and J. Lohscheller, "Fully automatic segmentation of glottis and vocal folds in endoscopic laryngeal high-speed videos using a deep convolutional lstm network," PLoS One **15**(2), e0227791 (2020).
12. J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2015), pp. 3431–3440.
13. V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," IEEE Trans. Pattern Anal. Mach. Intell. **39**(12), 2481–2495 (2017).
14. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, (Springer, 2015), pp. 234–241.
15. R. Hemelings, B. Elen, I. Stalmans, K. Van Keer, P. De Boever, and M. B. Blaschko, "Artery–vein segmentation in fundus images using a fully convolutional network," Comput. Med. Imag. Grap. **76**, 101636 (2019).
16. F. H. Araújo, R. R. Silva, D. M. Ushizima, M. T. Rezende, C. M. Carneiro, A. G. C. Bianchi, and F. N. Medeiros, "Deep learning for cell image segmentation and ranking," Comput. Med. Imag. Grap. **72**, 13–21 (2019).
17. Z. Jiang, H. Zhang, Y. Wang, and S.-B. Ko, "Retinal blood vessel segmentation using fully convolutional network with transfer learning," Comput. Med. Imag. Grap. **68**, 1–15 (2018).
18. D. Owen, "The power of student's t-test," J. Am. Stat. Assoc. **60**(309), 320–333 (1965).
19. L. R. Dice, "Measures of the amount of ecologic association between species," Ecology **26**(3), 297–302 (1945).
20. D. Abdullah, F. Fajriana, M. Maryana, L. Rosnita, A. P. U. Siahaan, R. Rahim, P. Harliana, H. Harmayani, Z. Ginting, and C. I. Erliana, *et al.*, "Application of interpolation image by using bi-cubic algorithm," in *Journal of Physics: Conference Series*, vol. 1114 (IOP Publishing, 2018), p. 012066.
21. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556 (2014).
22. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein, "Imagenet large scale visual recognition challenge," Int. J. Comput. Vis. **115**(3), 211–252 (2015).
23. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980 (2014).
24. R. M. Haralick and L. G. Shapiro, *Computer and Robot Vision*, vol. 1 (Addison-Wesley Reading, 1992).
25. L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*, (Springer, 2010), pp. 177–186.
26. C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," >J. Big Data **6**(1), 60 (2019).
27. S. J. Ahn, W. Rauh, and H.-J. Warnecke, "Least-squares orthogonal distances fitting of circle, sphere, ellipse, hyperbola, and parabola," Pattern Recognit. **34**(12), 2283–2303 (2001).
28. F. Chollet, *et al.*, "Keras," (2015).
29. T. T. D. Team, R. Al-Rfou, G. Alain, A. Almahairi, C. Angermueller, D. Bahdanau, N. Ballas, F. Bastien, J. Bayer, and A. Belikov, *et al.*, "Theano: A python framework for fast computation of mathematical expressions," arXiv preprint arXiv:1605.02688 (2016).
30. W. R. Crum, O. Camara, and D. L. Hill, "Generalized overlap measures for evaluation and validation in medical image analysis," IEEE Trans. Med. Imaging **25**(11), 1451–1461 (2006).
31. D. F. Williamson, R. A. Parker, and J. S. Kendrick, "The box plot: a simple visual method to interpret data," Ann. Intern. Med. **110**(11), 916–921 (1989).
32. A. Sadovski, "Algorithm as 74: L1-norm fit of a straight line," J. Royal Stat. Soc. Ser. C (Applied Stat.) **23**(2), 244–248 (1974).