Automatic glottis segmentation from laryngeal high-speed videos using 3D active contours

Fabian Schenk[1]
fabian.schenk@student.tugraz.at
Martin Urschler[2]
martin.urschler@cfi.lbg.ac.at
Christoph Aigner[3]
christoph.aigner@student.tugraz.at
Imme Roesner & Philipp Aichinger[4]
philipp.aichinger@meduniwien.ac.at
Horst Bischof[1]
bischof@icg.tugraz.at

[1] Inst. f. Computer Graphics & Vision
Graz University of Technology, Austria
[2] Ludwig Boltzmann Inst. f. Clinical
Forensic Imaging, Graz, Austria
[3] Signal Processing & Speech
Communication Laboratory
Graz University of Technology, Austria
[4] Dept. of Otorhinolaryngology
Div. of Phoniatrics-Logopedics
Medical University Vienna, Austria

Abstract

Laryngeal high-speed videos are a state of the art method to investigate vocal fold vibration but the vast amount of data produced prevents it from being used in clinical applications. Segmentation of the glottal gap is important for excluding irrelevant data from video frames for subsequent analysis. We present a novel, fully automatic segmentation method involving rigid motion compensation, saliency detection and 3D geodesic active contours. By using the whole color information and establishing spatio-temporal volumes, our method deals with problems due to low contrast or multiple opening areas. Efficient computation is achieved by parallelized implementation using modern graphics adapters and NVidia CUDA. A comparison to a semi-automatic seeded region growing method shows that we achieve improved segmentation accuracy.

1 Introduction

In our service oriented society speech is the main form of communication between people and has gained a tremendous economic value. In the US around 60% of the jobs require communication skills [8] and therefore voice disorders are a factor in a country's economy. Diagnosis and classification of these disorders have become important research topics in recent years. In this context laryngeal high-speed videos (LHSVs) have emerged as a very sophisticated tool [3]. An important limitation of this method is the vast amount of video material produced in a single investigation, which makes it nearly impossible to use in clinical applications. Therefore, methods for automatically processing LHSVs with the aim of detecting vocal fold vibration patterns are highly relevant for voice disorder research. Especially the automated segmentation of the glottis area, which is the opening between the vocal

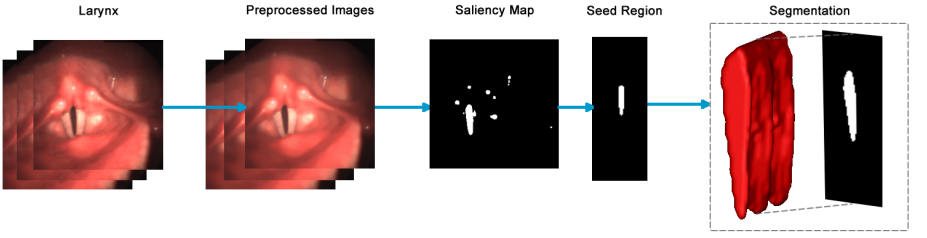

Figure 1: Image processing pipeline for automatic glottis segmentation.

folds (see Fig. 1), is an important preliminary step for later assessment of spatio-temporal plots (i.e., phonovibrograms [7]). Typical obstacles for the segmentation are the drift of the glottis due to breathing and patient movement, fluid artifacts, or brightness changes and contrast inadequacies during acquisition. The clinical standard method [6] is far from ideal in terms of accuracy and user input efficiency. In this work we propose a method to automatically segment the glottis from LHSVs using the full color information, taking motion compensation into account and processing the 3D temporal volume at once. Efficiency is achieved by a parallelized implementation using modern graphics adapters.

1.1 Related work

One of the first LHSV segmentation methods was presented by Lohscheller et al. [6] based on a seeded region growing (SRG) algorithm. The manual seed point and threshold selection on a large number of frames throughout the video is a time consuming task, that crucially depends on the choice of a proper homogeneity criterion. Another drawback of this method is its lack of motion compensation. Demeyer et al. [4] extended this method to get rid of user intervention. After locating the maximal glottal opening in every cycle, they look for dark elliptical regions to estimate glottis center and area. The center point is used as seed in the following SRG algorithm, which continuously adapts its threshold until the segmented region is larger than the previously estimated one. This segmentation result is adapted to the other frames in the cycle by using level sets. Similar to [6] this algorithm has problems due to leaking of the SRG and the parameters of the 2D level set are crucial but hard to select. Level set propagation heavily depends on the SRG initialization. Karakozoglou et al. [5] proposed a way to localize the glottis and find frames with the maximum glottal opening to calculate a bounding box for every cycle. The following image segmentation uses frame by frame 2D active contours. For this method the starting curve is of great importance and they propose two automatic ways for curve initialization on the landmark frames. Despite showing promising results, initital curve calculation per frame is not very robust due to leaking, leading to the need for manual post-correction.

Our contribution is an algorithm that extends these methods by performing a 3D segmentation on the spatio-temporal volumes at once, making use of the full color information. We incorporate a motion compensation step to correct for rigid patient or camera drift in a pre-processing step and tackle the leaking problem by a salient region detection initializing the segmentation. No user intervention or manual refinement is required in our method.

2 Method

We set up an image processing pipeline as depicted in Fig. 1 to tackle the specific problem of glottis segmentation from LHSVs. The typical behavior of the glottis during voice production is a repeated opening and closing process. The frame rate is much higher than the fundamental frequency of the vocal folds, which allows treating the glottography videos as 3D temporal volumes. To get rid of artifacts we apply an edge-preserving denoising filter to the 3D temporal volume, followed by a rigid registration of subsequent frames. The core of our method is the saliency detection using a boolean map approach, which gives us the seed regions located in the interior part of the glottis. The following 3D segmentation uses these seeds to compute the opening and closing glottis volumes.

2.1 Preprocessing

Edge-preserving denoising is used on the frames to get rid of light and fluid artifacts while keeping important edge information intact. Our implementation follows the total variation based 3D denoising paradigm that regularizes the L1 norm of the gradients in the denoising solution. The continuous convex optimization scheme is based on the primal-dual algorithm from Chambolle and Pock [1]. We apply this method to each color channel separately. Preventing smoothing over the important edge information is crucial for later segmentation.

To get rid of global patient or camera movement, the frames have to be registered onto each other. The high frame rate of the recording provides a certain advantage for the image registration because we can assume that there is no movement between two or even more consecutive frames, since camera or patient drifts occur more slowly over the videos. Therefore, to speed up this process we only register every tenth image to our fixed image and apply the transformation to the four frames before and the five after the registered frame. For image registration we use a slightly modified version of the *magnitude difference minimization method* proposed by Delisyki et al. [2] by additionally taking rotation into account. This is essentially an intra-modality registration using the sum of absolute color pixel differences as a similarity measure. Since individual differences between frames are small, it is feasible to solve this registration problem globally by an exhaustive search strategy over the two translation and the rotation parameters in a limited parameter range. Efficient computation of this step is achieved by a parallel NVidia CUDA GPU implementation.

2.2 Saliency detection

The saliency detection is crucial in our image processing pipeline as it provides seed points in the interior of the glottis that initialize the following segmentation algorithm. We face a saliency detection problem, where we have to calculate a saliency map from an RGB-image to represent the interesting regions, defined as areas fully surrounded by background. The glottal opening is a dark area surrounded by tissue, thus fulfilling these requirements. Without the previous denoising step fluid artifacts or tissue anomalies would be detected as interesting regions as they are usually areas on a homogeneous background.

2.2.1 Boolean Map based Saliency

We adapted a method proposed by Zhang and Sclaroff [10], as depicted in Fig. 2. The main idea is to generate boolean maps by thresholding the different color channels. We use the

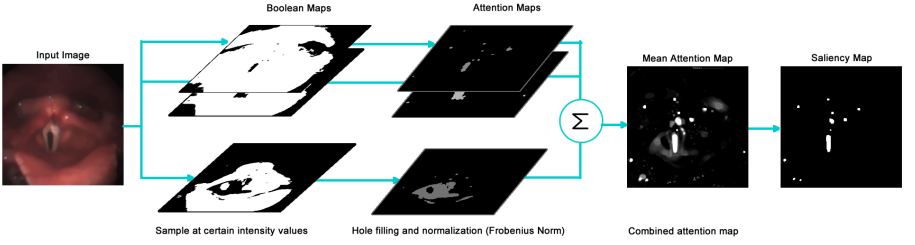

Figure 2: Saliency detection based on boolean maps, adapted from [10].

CIE Lab color space due to its perceptual uniformity with all channels scaled to [0,255]. In order to generate boolean maps we iterate through all the channels and sample them at a certain intensity threshold value θ, which is increased by a fixed step size δ after each iteration. An attention map is calculated by closing the surrounded regions or holes of the boolean map followed by a division by its Frobenius norm to emphasize small concentrated areas. The results are then combined to a mean attention map and a threshold operation filters out small and irrelevant signals. With the use of a connected component analysis the darkest and biggest area is found and used as seed region for the segmentation.

2.3 Image Segmentation

The 3D spatio-temporal volume segmentation problem can be formulated as a Markov Random Field and solved using graph cuts. We formulate this problem as the equivalent continuous formulation using the calculus of variation, which leads to a convex geodesic active contour based on the weighted total variation (GAC-TV). Minimization of this functional has been demonstrated in [9] and leads to a temporally smooth globally optimal minimal surface solution. We work solely on the gradient information of the 3D temporal volume, computed from the RGB values according to $V_{grey} = 0.2126R + 0.7152G + 0.0722B$.

The initialization regions come from the salient region detection step described above and the 3D segmentation iteratively evolves them towards the edges of the temporal volume. This segmentation method in 3D gives an advantage over traditional active contour algorithms like [4, 5], since it allows to fill in topologically disconnected regions in single frames.

3 Experiments and Results

Our algorithm is intended to process LHSVs that typically consist of 4000 frames/s at a resolution of 256×256 pixels covering two seconds of phonation. To quantitatively evaluate our method, on the one hand temporal sub-volumes of consecutive frames and on the other hand randomly chosen single frames from a number of videos were selected. Our experimental setup consists of two 60- and one 30-frame video sequences as well as 25 single frames randomly picked from different videos, involving healthy and disordered subjects. For all frames an experienced computer vision researcher has performed a manual segmentation of the glottis, which was investigated and corrected by an expert in kymography and voice disorders. These segmentations are the ground truth to which we compare our proposed algorithm as well as an implementation of the clinical standard [6]. For quantitative

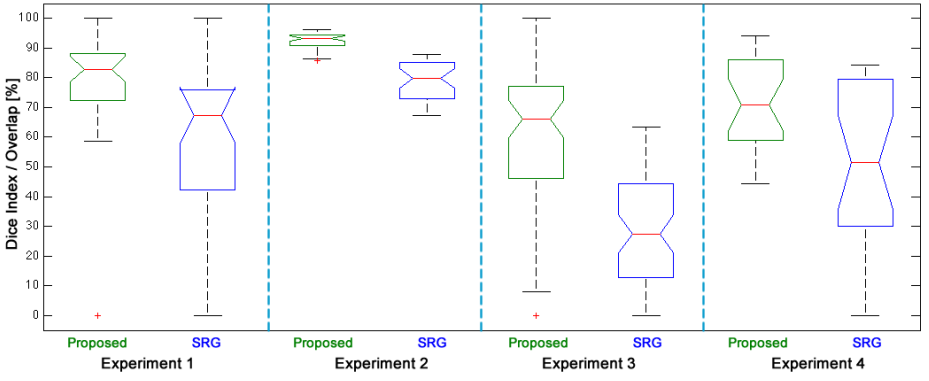

Figure 3: Box-whisker plots showing four different experiments with the proposed method in green and SRG [6] in blue. Exp. 1 and 2 are the two 60-frame videos, Exp. 3 is the 30-frame sequence and Exp. 4 represents the 25 randomly picked frames.

evaluation we compute the Dice coefficient as a segmentation overlap measure for each segmentation approach and the ground truth segmentation. The Dice coefficient (DC) is defined as $DC = \frac{2|A \cap B|}{|A|+|B|}$, with A the ground truth and B a segmentation. It computes a value between 0 and 1, which we present as an overlap percentage. All the calculations were performed on a Linux machine using an Intel Core i7 and an NVidia GeForce GTX 580 graphics adapter.

The segmentation results can be seen in Fig. 3, where the proposed method is far more accurate with median values of 82.9, 93.2, 66.1 and 70.1 % compared to 67.4, 79.8, 27.5 and 51.5 % of the SRG method. Some of the outliers can be explained due to the nature of the dice index, which over-emphasizes relatively small mistakes for small structures (i.e. missing a ground truth segmentation that only consists of a single pixel would result in overlap $DC = 0$). There is also a great advantage when it comes to speed, as the proposed fully automatic method takes only 50 seconds to segment 100 frames compared to 3.5 minutes of the SRG method, which also highly depends on the experience of the user. Figure 4 shows exemplary qualitative segmentation results of the two methods in the three different videos. The low contrast at the bottom of Fig. 4(a,i) poses a leaking problem for the threshold based SRG method (see Fig. 4(c,k)), whereas the proposed method does not have this problem (see Fig. 4(b,j)). Figure 4(j) indicates a problem of the proposed algorithm. Due to the narrow connection between the bottom and top region, the edges in this part of the image are not sufficient to detect, which makes it expensive for the GAC-TV segmentation to connect these two areas. The frames in Fig. 4(e-h) show that on an image with a high contrast difference and a big opening area both methods work well.

4 Conclusion

A fully automatic algorithm for glottis segmentation from LHSVs has been proposed, which combines motion compensation, robustness to video artifacts and a fully 3D segmentation algorithm to efficiently process glottography video material. A comparison with the clinical standard method has revealed a higher segmentation accuracy and at the same time requires no user interaction, thus our expected run-time scales much better to the real-world applica-

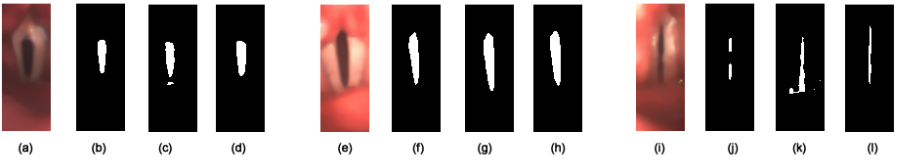

Figure 4: Qualitative results. (a,e,i) original frames, (b,f,j) segmentation results of the proposed method (c,g,k), segmentation results of the SRG method and (d,h,l) ground truth.

tion of segmenting thousands of video frames. Future work will concentrate on evaluating a larger database as well as a comparison to further related algorithms.

References

[1] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal Math Imaging and Vision*, 40(1):120–145, 2011.

[2] D. Deliyski, S. Cieciwa, and T. Zielinski. Fast and robust endoscopic motion estimation in high-speed lyrngoscopy. In *Proc Conf Advances in Quantitative Laryngology, Voice and Speech Research (AQL)*, 2006.

[3] D. Deliyski, P. Petrushev, H. Bonilha, T. Gerlach, B. Martin-Harris, and R. Hillman. Clinical implementation of laryngeal high-speed videoendoscopy: Challenges and evolution. *Folia Phoniatrica et Logopaedica*, 60:33–44, 2008.

[4] J. Demeyer, T. Dubuison, B. Gosselin, and M. Remacle. Glottis segmentation with a high-speed glottography: a fully automatic method. In *3rd Int. Workshop Adv. Voice Funct. Assess.*, 2009.

[5] S. Karakozoglou, N. Henrich, C. d'Alessandro, and Y. Stylianou. Automatic glottal segmentation using local-based active contours and application to glottovibrography. *Speech Communication*, 54(5):641–654, 2012.

[6] J. Lohscheller, H. Toy, F. Rosanowski, U. Eysholdt, and M. Doellinger. Clinically evaluated procedure for the reconstruction of vocal fold vibrations from endoscopic digital high-speed videos. *Medical Image Analysis*, 11(4):400–413, 2007.

[7] J. Lohscheller, U. Eysholdt, H. Toy, and M. Doellinger. Phonovibrography: Mapping high-speed movies of vocal fold vibrations into 2D diagrams for visualizing and analyzing the underlying laryngeal dynamics. *IEEE Trans Med Imag*, 27:300–309, 2008.

[8] R.J. Ruben. Redefining the survival of the fittest: communication disorders in the 21st century. *Laryngoscope*, 110:241–245, 2000.

[9] M. Unger, T. Pock, W. Trobin, D. Cremers, and H. Bischof. TVSeg – interactive total variation based image segmentation. In *Proc British Machine Vision Conf*, 2008.

[10] J. Zhang and S. Sclaroff. Saliency Detection: A Boolean Map Approach. In *Proc Int Conf Computer Vision (ICCV)*, pages 153–160, 2013.