Annals of the BMVA Vol. 2015, No. 1, pp 1–15 (2015)

# Automatic high-speed video glottis segmentation using salient regions and 3D geodesic active contours

Fabian Schenk[1], Philipp Aichinger[2,3], Imme Roesner[3], Martin Urschler[4,5]

[1] Institute for Computer Graphics and Vision, Graz University of Technology, Austria
[2] Signal Processing & Speech Communication Lab, Graz University of Technology
[3] Department of Otorhinolaryngology, Division of Phoniatrics-Logopedics, Medical University of Vienna, Austria
[4] Ludwig Boltzmann Institute for Clinical Forensic Imaging, BioTechMed, Graz
[5] Department of Legal Medicine, Medical University of Graz, Austria

**Abstract**

Verbal communication plays an important role in our economy. Laryngeal high-speed videos have emerged as a state of the art method to investigate vocal fold vibrations in the context of voice disorders affecting verbal communication. Segmentation of the glottis from these videos is required to analyze vocal fold vibrations. The vast amount of data produced makes manual segmentation impossible in every day clinical applications. Therefore, computer-aided, automatic segmentation is essential for the use of high-speed videos. In this work a novel, fully automatic glottis segmentation method involving motion compensation, salient region detection and 3D Geodesic Active Contour segmentation is presented. By using color information and establishing spatio-temporal volumes, the method overcomes reported problems in related work regarding low contrast and multiple glottal areas. Efficient computation is achieved by parallelized implementation using graphics adapters and NVidia CUDA. A comparison to the seeded region growing based clinical standard shows the benefits of the proposed method in terms of higher segmentation accuracy on manually annotated evaluation data.

## 1 Introduction

During the last fifty years our economy changed from a manual labor to a service oriented one. Speech, as the main form of communication between people, has gained tremendous economic value. With e.g. around 60 % of the jobs in the US requiring communication skills and prevalence of 9.5 % of speech and voice disorders in 0 - 19 years old subjects in developed nations, they have become a crucial part in any country's economy [Ruben, 2000].

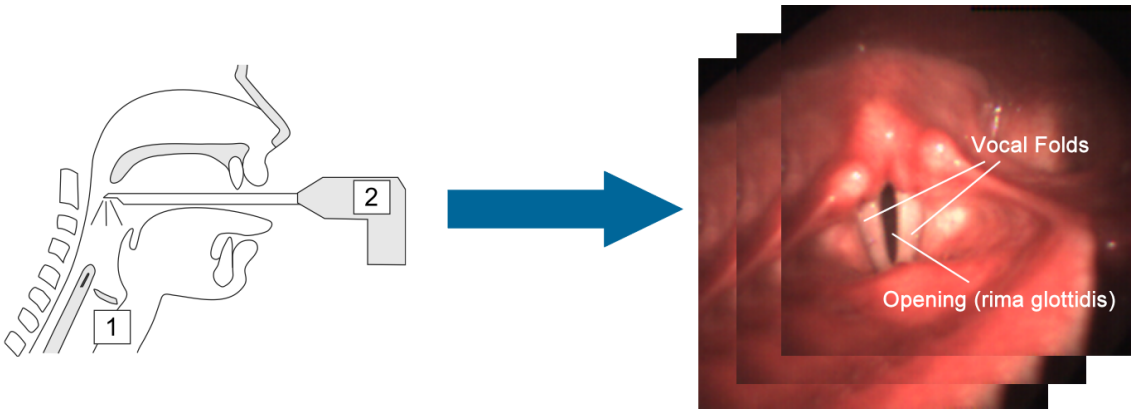


Figure 1: Recording setup for laryngeal high-speed videos (taken from [Lohscheller et al., 2007]). The vocal folds (1) are filmed with a rigid telescopic high-speed camera (2).

Thus, diagnosis and classification of these disorders have become important research topics in recent years.

Voice is generated by the glottis, which is the gap between the vocal folds (see Fig. 1). The primary voice signal is generated within the larynx by the two opposed vocal folds, set into vibrations by streaming air provided by the lungs [Titze, 1994]. The frequency and intensity of the voice signal varies within a wide range due to modification of muscle tension, vocal fold length and lung pressure [Titze, 1976]. Generally, a healthy voice requires symmetric and regular vocal fold oscillations [Doellinger et al., 2003, Hertegard et al., 2003]. Perturbations of the acoustic speech signal can be caused by irregular and asymmetric vibrations, which may result in hoarseness [Hoppe et al., 2001]. [Eysholdt et al., 2003] define hoarseness as the unspecified symptom of a diseased larynx, which originates from irregular vocal fold vibrations. The International Classification of Disease and Related Health Problems of WHO classifies voice disorders according to etiological aspects. Depending on whether organic abnormalities exist or not, they are then further divided into organic and functional dysphonia. Thus, to investigate voice disorders, both anatomy and vocal fold vibration patterns are analyzed.

Due to the frequency of vocal fold vibrations (e.g., 250 Hz in Fig. 2), a laryngeal high-speed video (LHSV) camera with a high frame rate is used for recording vocal fold videos (see Fig. 1). Figure 2 shows a video recording depicting a typical, complete vocal fold oscillation cycle with a duration of 16 frames, including the open and closed phases. The main limitation of LHSVs is the vast amount of video material produced in a single investigation, which makes manual assessment very time-consuming and thus impossible to use for everyday clinical application. Therefore, methods for automated processing of LHSVs are needed.

The automated segmentation of the glottal area is an important preliminary step for later visual or quantitative evaluation of spatio-temporal plots [Lohscheller et al., 2008] of the glottal opening (i.e., phonovibrograms). Typical obstacles for this segmentation problem are a slow spatial drift of the glottis due to patient and camera movements, fluid artifacts on the camera, limited spatial resolution, brightness changes, and contrast inadequacies during acquisition. Those problems arise from difficulties in image acquisition, i.e., the appropriate positioning of the camera is difficult and its insertion may provoke the patient to gag.

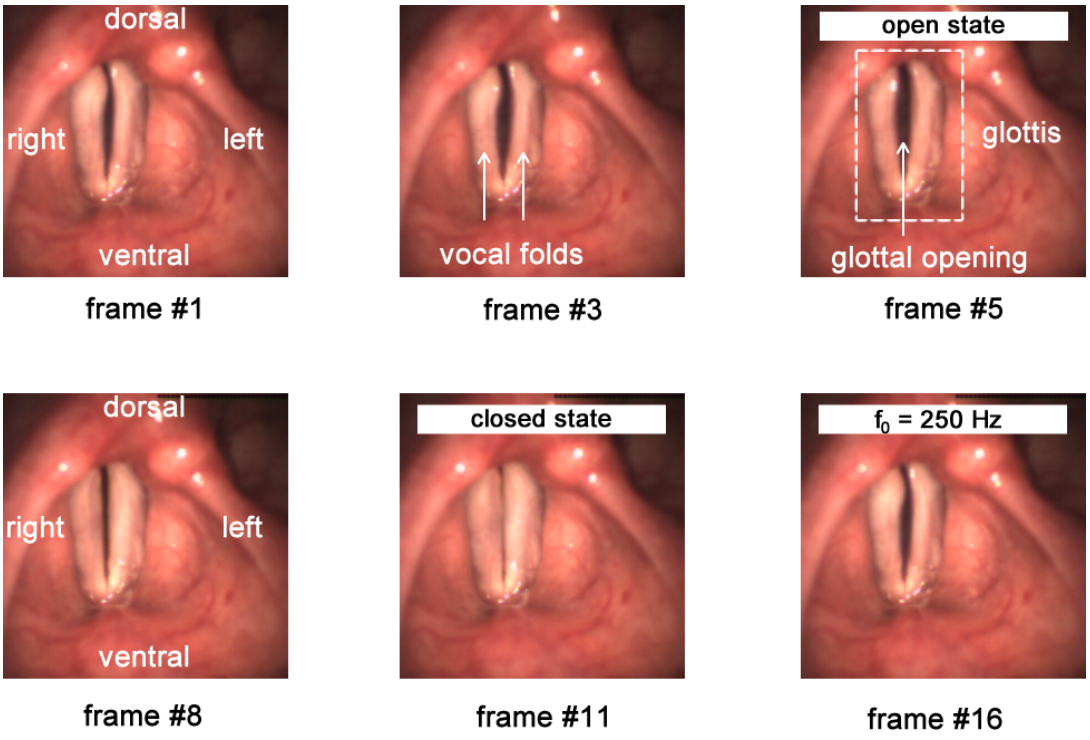


Figure 2: Several frames of a laryngeal high-speed video sequence showing a complete cycle of glottis movement including the open and closed states.

State of the art methods for glottis segmentation have limitations when it comes to accuracy and efficiency. Processing speed is also restricted due to required user input and often necessary manual corrections that complicate the detection process even further. Additionally, these methods neglect the available color information and lack a proper motion compensation step.

## 1.1 Contribution

Extending previous work presented in [Schenk et al., 2014], a fully automatic, computationally efficient glottis segmentation method is developed, starting with a preprocessing step including contrast stretching, motion compensation to correct patient or camera drifts and edge-preserving denoising to get well-defined edges for later segmentation. A novel salient region detection method first detects a region of interest (ROI) and then uses this information to calculate seed regions to initialize the segmentation, also taking full color information of the input videos into account. For image segmentation, videos are treated as spatio-temporal volumes, to which a 3D Geodesic Active Contour segmentation is applied. To overcome slow computation time of the large amount of video data, thus potentially preventing the use in clinical practice, the presented method is partly implemented in CUDA (Compute Unified Device Architecture), a parallel programming platform by NVidia, which uses the immense capabilities of modern graphics adapters to greatly increase the speed of 2D and 3D image processing algorithms.

## 2 Related Work

With the invention of LHSVs, researchers have received a very powerful tool to observe vocal fold vibrations. The huge amount of data produced in LHSVs has led to the development of computer-aided tools to study the video material. Especially the segmentation of the glottis has received a lot of attention in the last decade.

An early method based on seeded region growing (SRG) was developed by [Lohscheller et al., 2007]. Initial seed points and thresholds have to be manually adjusted by the user on a large number of frames throughout the video, which is a tedious and time consuming task. A general obstacle is the crucial choice of an appropriate homogeneity criterion, which can be very difficult. This is due to intensity and brightness variations and unclear transitions between glottal opening and surrounding tissue. A constant set of seed points is used to re-detect the glottis after every full cycle. Even though this method has emerged as a clinical standard, it has a few drawbacks like the need for extensive user intervention and refinement of the final segmentation, or the lack of a proper motion compensation step to account for patient/camera movement.

[Demeyer et al., 2009] avoid user-interaction by an initial analysis step of the video, which detects key frames with maximal glottal opening and estimates the vocal fold oscillation frequency. To localize the key frames, dark, elliptical regions bordered by the brighter vocal folds are detected. Center and size of these regions determine the glottis and an SRG step with an automatically obtained threshold is applied. This segmentation result is then propagated forward and backward along the temporal dimension throughout the image cycle using per image level sets [Sethian, 1999], which evolve the seed regions towards the glottis edges. The detection of the glottal opening is not sufficiently robust. Similar to [Lohscheller et al., 2007] this algorithm is prone to leaking problems and motion is supposed to be compensated through result propagation. Additionally, the 2D level set algorithm requires difficult parameter selection and heavily depends on the SRG initialization.

[Karakozoglou et al., 2012] first look for frames with maximal glottal opening similar to [Demeyer et al., 2009], which are called landmark frames (LF). In these LF, large, nearly vertically oriented areas are identified and an edge detection filter is applied. The vertically oriented object with the largest area in each cycle is found via connected component analysis. A bounding box is then computed for every frame to reduce further calculation time and memory requirements. Glottal drift and camera movement is compensated by keeping the bounding box steady throughout every cycle. An active contour segmentation as proposed by [Chan and Vese, 2001] is used for segmentation, using information obtained from the LF like shape and area of the object of interest. Segmentation always starts from the LF, where either an automatically computed threshold or an elliptic mask for curve initialization is used. After this process, the segmentation results are propagated using the mask of the previous or next frame as initial mask. Despite promising experimental results and full automatism, this method has certain limitations in difficult situations. The calculation of the initial curve for the LF can be problematic due to leaking or in cases, where the glottis does not have an elliptic shape. Further, topological changes (e.g., two glottal openings in pathologic cases) are hard to segment correctly with 2D active contours, suggesting the use of a 3D approach.

A very recent approach proposed by [Koç and Çiloğlu, 2014] is based on image histogram modeling and thresholding. It assumes that the area of the glottal opening changes during the opening and closing cycles, which can be detected by looking at the histogram changes

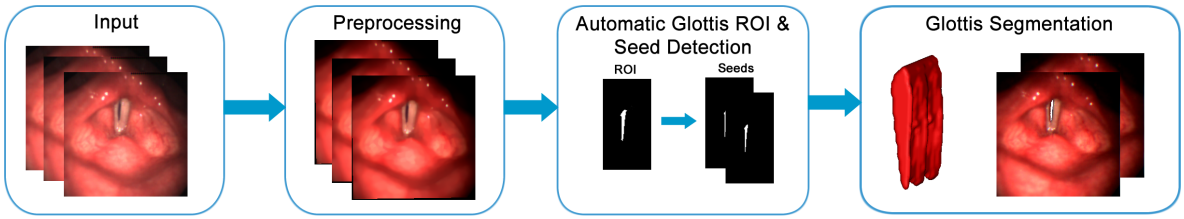


Figure 3: The proposed method can be divided into three major steps: preprocessing, automatic region of interest and seed region detection, and glottis segmentation.

solely in an ROI around the glottis. Based on the vocal folds being the most active structures within the sequence, accurate determination of this ROI is crucial. By calculating the Total Variation (TV) norm over a time sequence, the largest TV values correspond to regions of maximal movement, which coincide with the glottal area. An automatic threshold computation using a Gaussian Mixture Model and a reflectance histogram, which is required to deal with non-uniform illumination, leads to a glottis segmentation. This method neglects the global drift and therefore does not provide any type of motion compensation. This and the assumption that the vocal folds are always the most active regions of the image imply that the detection of the ROI is not very robust, since TV calculation leads to wrong ROI detections if no vibration or drift over a longer period of time exists.

## 3 Method

In related work, no automatic glottis segmentation methods have yet been presented, that make use of the color video information, include a motion compensation step into the algorithm, and rely on 3D segmentation of spatio-temporal volumes instead of stacking together 2D segmentations independently from temporal context. To overcome these limitations, an image processing pipeline is presented as shown in Fig. 3. The presented method consists of three major steps, which are preprocessing, automatic region of interest and seed region detection, and glottis segmentation. In the preprocessing step, problems like non-homogeneous background, illumination artifacts and global drift are dealt with. Then a glottis ROI is automatically computed and seed regions for later segmentation are calculated. This is the core of the presented method, where a salient region detection algorithm localizes the glottis. The generated seed regions are finally used as initialization for a 3D Geodesic Active Contour segmentation algorithm.

### 3.1 Preprocessing

In preprocessing, first image contrast is enhanced by a color contrast stretching operation. A typical LHSV contains noise as well as small interfering structures like illumination artifacts and blood vessels. This greatly affects segmentation and ROI detection, therefore an edge-preserving denoising step is introduced. It is performed according to the model of Rudin, Osher and Fatemi (ROF) [Rudin et al., 1992]:

$$\min_u \left\{ E_{ROF} = \int_\Omega |\nabla u| dx + \frac{\lambda}{2} \int_\Omega (u-f)^2 dx \right\}. \quad (1)$$

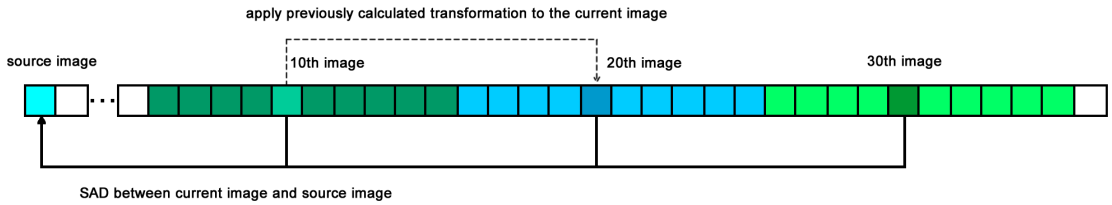


Figure 4: Every 10th frame is registered and used to transform the images before and after.

In this model, denoising is formulated in a variational framework, where the first part represents the regularization term and the second part the data fidelity term. The solution to this optimization problem is a convex, continuous function $u$. Data fidelity accounts for differences of the denoised image $u$ to the given, noisy image $f$, penalizing deviations with a quadratic norm. Regularization accounts for gradients in the reconstructed image $u$, by penalizing the L1 norm of the gradients. A weighting term $\lambda$ serves as a trade-off between data fidelity and regularization. To solve this convex optimization problem, a primal-dual optimization algorithm presented by [Chambolle and Pock, 2011] is applied to the 3D spatio-temporal volume. The last step of preprocessing is image registration, where global patient and camera movements are compensated. For registration, the sum of absolute differences between color images is introduced as the similarity measure in a rigid registration scheme applied to consecutive video frames [Hill et al., 2001, Deliyski et al., 2006]. The three unknown 2D motion parameters from translation and rotation are registered by an exhaustive search strategy, where 9 rotation angles between $[-\frac{\pi}{240}, \frac{\pi}{240}]$ and 5 translation steps between [-2 px, 2 px] in each image dimension specify the search space. The high frame rate of the video recordings allows further speed-up of registration, because no fast global motion between consecutive frames exists. Therefore, only every tenth image is registered to a fixed image, and the resulting transformation is applied to four frames before and five after this image (see Fig. 4).

### 3.2 Automatic Glottis ROI Detection

Automatic glottis detection makes use of the concept of salient regions to detect the ROI and the bounding box around the glottis, as well as to find the seed regions for the final segmentation step. The basic idea of salient region detection [Borji and Itti, 2013, Riche et al., 2013] is to find salient areas in an image and to show their likelihood of being important as intensity values. Therefore, an area surrounded by background and not connected to the image borders is defined as a salient region. The glottal opening fulfills these requirements, since it is a dark area surrounded by tissue and usually located near the center of an image.

The salient region detection is inspired by Boolean Map based saliency [Zhang and Sclaroff, 2013], which relies on topological structural information attracting visual attention [Wolfe and Horowitz, 2004]. A Boolean Map is a spatial representation that partitions a visual scene into two distinct, complementary regions, i.e., fore- and background [Huang and Pashler, 2007]. It is generated from selected feature channels to highlight regions attracting visual attention. The influence of such regions can be represented by an attention map $A(B)$. Figure 5 shows how a set of Boolean Maps $B = B_1, B_2, ..., B_n$ is generated from image $I$ by thresholding its feature map $\phi(I)$ at a value $\theta$:

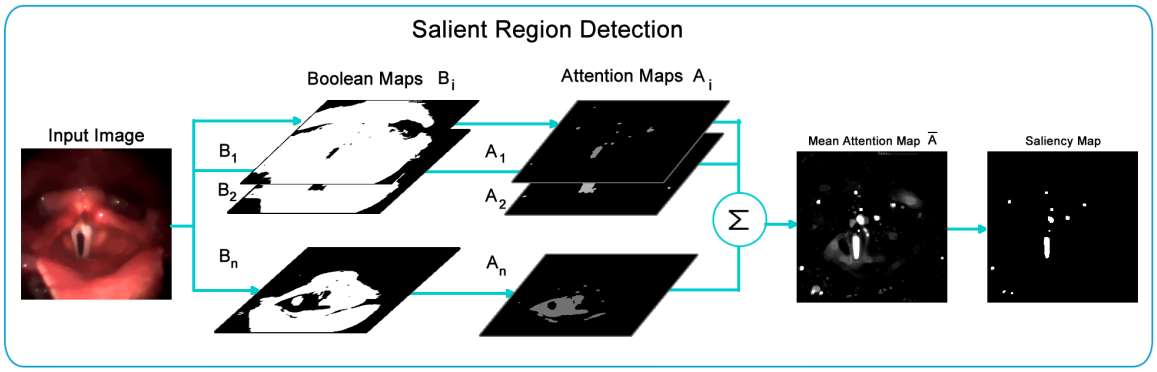


Figure 5: Saliency detection based on Boolean Maps (adapted from [Zhang and Sclaroff, 2013]).

$$B_i = \textbf{THRESH}(\phi(I), \theta),$$
$$\phi \sim p_\phi, \theta \sim p_\theta.$$

The threshold function $\textbf{THRESH}(\phi(I), \theta)$ assigns 0 to every pixel smaller than $\theta$ and 1 otherwise. After each iteration, $\theta$ is increased by a step size $\delta_S$, which can be selected by the user. The feature channels $\phi(I)$ are assumed to be in the range between 0 and 255 and in our case the color information of each frame is used. $\phi \sim p_\phi$ and $\theta \sim p_\theta$ represent the prior distribution of $\theta$ and $\phi$. Salient regions should have a higher chance to be separated from the background when generating a Boolean Map, which can be achieved by a uniform distribution of the threshold $\theta$. The *CIE Lab* color space is used due to its perceptual uniformity with all channels stretched to the range of [0,255]. $L$ is lightness, $a$ the color space position between red and green and $b$ the position between yellow and blue.

The final saliency map can vary depending on step size $\delta_S$ due to image content. A higher $\delta_S$ means faster computation because less Boolean Maps have to be calculated. From the set of Boolean Maps $B = B_1, B_2, ..., B_n$ attention maps $A_i$ are computed based on the Gestalt principle for figure-ground segregation, which states that surrounded regions are more likely to be perceived as figures [Palmer, 1999]. Surrounded regions in a Boolean Map represent salient areas and can be determined by simply filling all the regions connected to the borders. This operation is performed using a region growing method with the image borders as seeds. The resulting attention maps are normalized using the Frobenius norm of the attention map to give small, concentrated areas more emphasis. The normalized attention maps $A_i$ are summed up to a mean attention map $\bar{A}$ and the final saliency map is calculated using a threshold operation with $\frac{1}{\delta_S}$ to filter out weak signals.

For ROI detection the approach shown in Fig. 6 is used. Similar to [Karakozoglou et al., 2012, Demeyer et al., 2009] the frame with lowest intensities is found, which is the one with the largest glottal opening (landmark frame) and then a saliency map is calculated. Because all the salient areas are candidates for the ROI, a connected component analysis is performed and then each component is multiplied with a gray-scale version of the image. To find the correct glottal opening a ranking value, which takes the area and the sum of intensities into account, has to be calculated for each component. The highest ranked one is then selected as ROI for the whole video and used for computing the surrounding, fixed size bounding box.

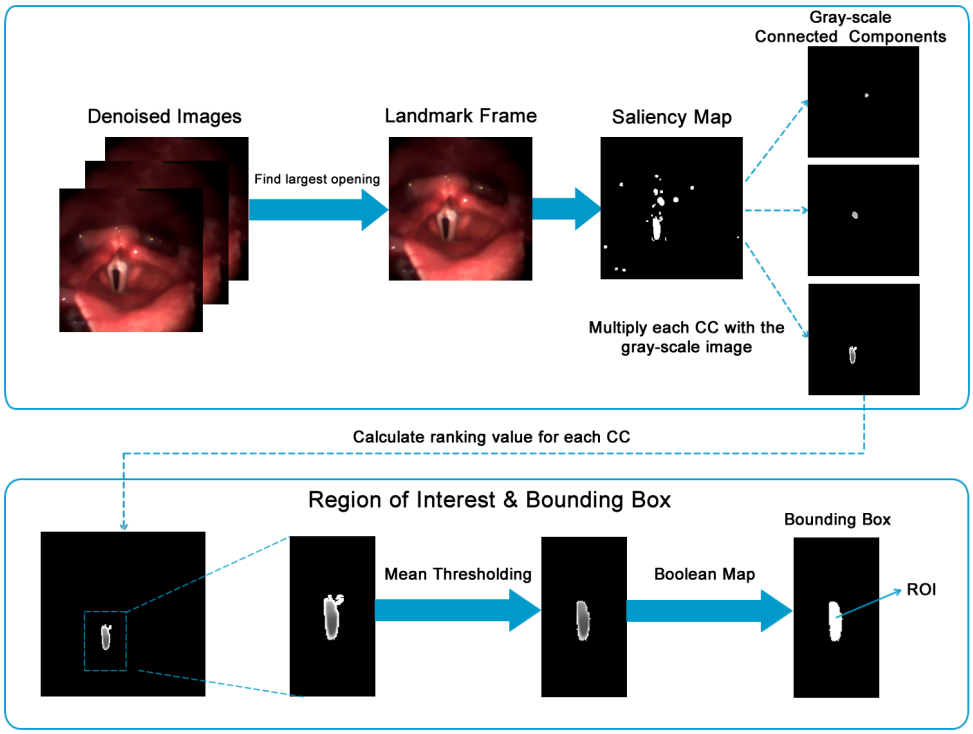


Figure 6: Processing pipeline to find the region of interest and bounding box for the largest glottal opening. Images were contrast-enhanced for visualization purposes.

### 3.3 Glottis Segmentation

The size of the glottal opening is a very important information for speech and voice disorder research and its segmentation is the main purpose of this work. All the previous steps were required to prepare the LHSVs for the segmentation step. As shown in Fig. 7, seed regions and edge information from the preprocessed images restricted to the bounding box from the previous step are used to initialize the segmentation. These seed regions then evolve towards the edges of the glottis to yield the final segmentation. In this work, a novel way of glottis detection utilizing a salient region based algorithm initializes the segmentation step. Seed region detection is applied to each image as opposed to the once per video performed ROI detection of Section 3.2, but otherwise works very similar. After calculating the mean attention map, a threshold filter removes weak salient regions. Then only information inside the previously calculated image bounding box is extracted. By multiplying with the ROI all unwanted salient regions in the image are removed, thus avoiding the time consuming connected component analysis. After a thresholding operation with the mean intensity gray value of the previous seed region and morphological erosion, a seed region inside the glottis is determined.

#### 3.3.1 3D Geodesic Active Contour Segmentation

For final glottis segmentation, the spatio-temporal volume is processed by a Geodesic Active Contour (GAC) model [Caselles et al., 1997]. This model, which is traditionally solved using the level set method [Sethian, 1999] prone to local minimia, is adapted to 3D and formulated

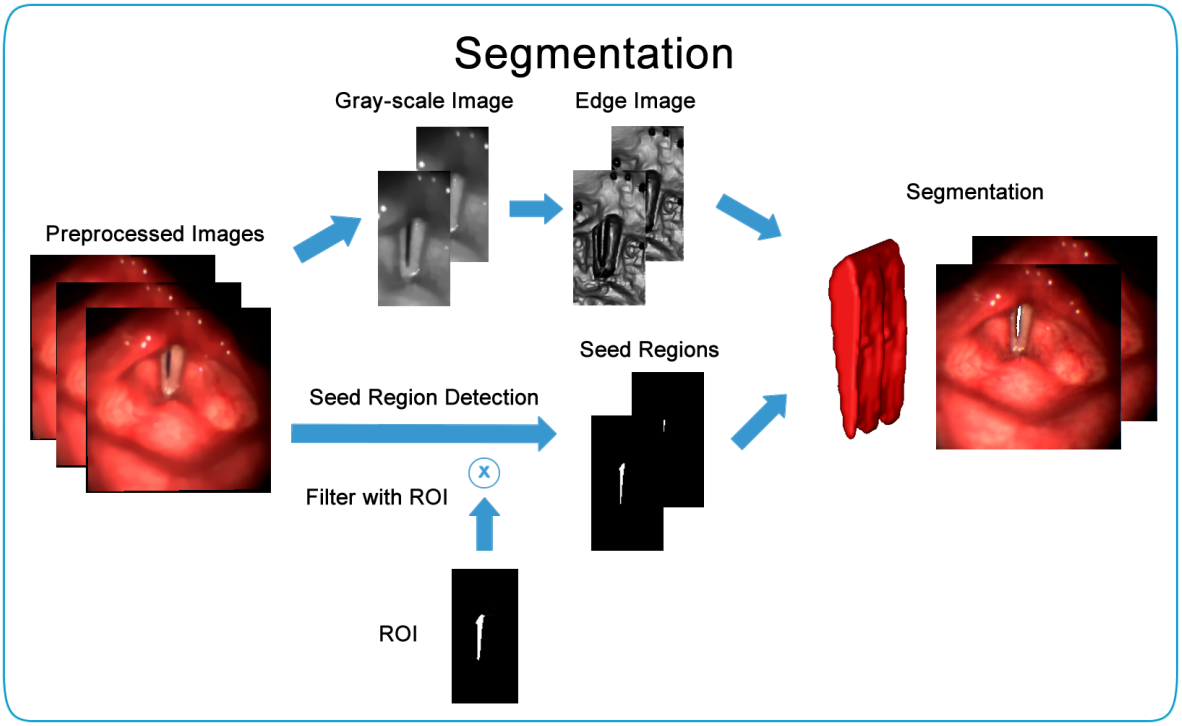


Figure 7: Image segmentation: After extracting a bounding box and computing edge information, seed regions from the salient region detection determine segmentation input.

as a continuous, variational energy minimization problem according to

$$E(C) = \alpha \int_0^1 |C'(q)|^2 dq - \lambda \int_0^1 |\nabla I(C(q))| dq, \tag{2}$$

where a 3D surface $C(q) : [0,1] \rightarrow \mathbf{R^3}$ is sought, that minimizes this energy functional, balancing an internal force penalizing deviations from a smooth surface and a data term penalizing a force derived from $\nabla I$. This data term is an external energy force, that attracts the segmentation surface towards the image gradients. By defining a monotonously decreasing edge indicator function $g : [0, +\infty) \rightarrow \mathbb{R}^+$, a generalization of Eq. 2 can be achieved, which leads to the GAC energy combining internal and external energy in a single term

$$\min_C \left\{ E_{GAC} := \int_0^{L(C)} g(|\nabla I(C(q))|) ds \right\}, \tag{3}$$

where the integration over the infinitesimal area elements $ds$ along $L(C)$ is weighted with $g$, derived from the image information, thus representing computation of a geodesic in a Riemannian space. Extending the model of [Chan and Vese, 2001], it has been shown by [Bresson et al., 2007], that the GAC energy of Eq. 3 can be formulated using the weighted Total Variation, which is defined as

$$TV_{g(u)} = \int_\Omega g(x) |\nabla u| dx. \tag{4}$$

For $u$ being a characteristic function $1_C$, [Bresson et al., 2007] have shown that $E_{GAC}$ is equivalent to $TV_{g(u)}$. The characteristic function $1_C$ is a closed set in the image domain $\Omega$ and C rep-

resents its boundary. By allowing $u$ to vary continuously between [0,1], Eq. 4 becomes a convex functional, enabling computation of a global optimum in this relaxation to Eq. 3. A key benefit of such a convex formulation compared to other works based on geodesic [Caselles et al., 1997] and other active contours [Xianghua and Mirmehdi, 2003, Zhou et al., 2013] is its ability to overcome local minima when minimizing the energy functional, an issue which has also recently been dealt with in [Appia and Yezzi, 2011]. Looking at the definition of the weighted TV model (4) it can be seen that a global minimizer is given by the trivial solution $C = 0$ (a point). To get meaningful results the previously calculated seed regions have to be incorporated as additional constraints. The final variational image segmentation model is defined as

$$\min_{u\in[0,1]} \left\{ E_{Seg} := \int_\Omega g(x)|\nabla u| dx + \lambda \int_\Omega u f dx \right\}. \tag{5}$$

The first term is the TV formulation weighted by an edge indicator function $g(x)$. In addition, a second term, which is convex in $u$, resembles the incorporation of the seed regions into $f$. $\lambda$ is a trade-off factor between constraints and surface area minimization, while the edge indicator function $g$ reaches a minimum at the edges and is defined as

$$g(I) = e^{-\alpha|\nabla I|^\beta}. \tag{6}$$

The segmentation is initialized with $f = -\infty$ according to the previously found seed regions and 0 otherwise. $f = -\infty$ leads to the propagation of $u = 1$ and for $f = 0$ only the pure Geodesic energy is minimized until the edge indicator $g$ prevents it. A binary segmentation is then obtained by applying a threshold of 0.5 for $u$. Even though this only approximates the original GAC energy, this choice not critical in practice. For numerical optimization of $E_{GAC}$, again the primal-dual algorithm of [Chambolle and Pock, 2011] is used.

# 4 Experiments and Results

To evaluate the presented method, a comparison to the clinical standard [Lohscheller et al., 2007] has been carried out on a set of images from male and female patients of different ages, with and without medical conditions. The ground truth for these image sets was generated through manual glottis annotation by an experienced computer vision researcher together with an expert in voice disorders. Three video sequences that include one or more opening and closing cycles were acquired with an *HRES ENDOCAM 5562* (*Richard Wolf GmbH*, Knittlingen, Germany) and annotated to establish the ground truth. Spatial resolution was $256 \times 256$ pixels and temporal resolution 4000 frames per second. Additionally, from a set of randomly chosen images from a variety of different videos, another set of annotated images was constructed, containing a larger variety of videos.
Four experiments were designed:

- **Experiment 1**: A 60 frames video sequence, every second frame used.
- **Experiment 2**: A 60 frames video sequence.
- **Experiment 3**: A 30 frames video sequence.
- **Experiment 4**: 25 randomly picked frames from a variety of videos.

In all experiments, frames containing manual segmentation annotation were used to compare the proposed method to the implementation of the clinical standard [Lohscheller et al., 2007]. An overlap measure of the ground truth segmentation and the result of the presented and the SRG method was calculated. The Dice coefficient (DC) is a statistical measure for comparing 2D region overlap. It ranges from 0 to 1 and is computed as

$$DC = \frac{2\,|A \cap B|}{|A| + |B|},$$

with $A$ the ground truth and $B$ the automated segmentation.

A number of parameters were set to empirically chosen fixed values for all these experiments. A fixed sized bounding box based on the detected glottis with a width of 80 and a height of 150 pixels was used. For ROF denoising a $\lambda$ value of 25 provided the best trade-off between data fidelity and regularization. In registration, 25 different translation offsets with a step size of 1 pixel and 9 rotation angles between $[-\frac{\pi}{240}, \frac{\pi}{240}]$ were used for exhaustively searching the optimal transformation between frames. For the 3D GAC segmentation, edge computation with $\alpha = 15, \beta = 0.55$ and a weighting factor of $\lambda = 0.01$ empirically proved suitable. All implementations were done in C++ and evaluation was performed on an Intel Core i7 (8 cores/3.4 GHz) with 8GB of memory and running Ubuntu Linux 14.04 as operating system. A number of steps in the proposed method (i.e., denoising, sum of absolute difference computation for registration and 3D GAC segmentation) were accelerated by a parallel implementation using the graphics adapter as a numeric co-processor based on NVidia CUDA. This lead to a significant speedup, thus enabling the method to be used in clinical applications due to the very fast computation time of around 25-30 minutes for a whole sequence of over 8000 frames ($\delta_S = 8$).

The automatically computed quantitative results of the comparison between the proposed method, using $\delta_S = 3$ and $\delta_S = 8$, and the clinical standard are shown in Fig. 8 as box-whisker plots showing the DC. Qualitative results can be found in Fig. 9.

## 5 Discussion

The quantitative results show that the proposed fully automatic segmentation approach yields good qualitative results. It overcomes certain problems of the SRG-based clinical standard [Lohscheller et al., 2007] like leakage or over- and undersegmentation, and outperforms it in all four experiments. The proposed method with saliency step size $\delta_S = 3$ performs better than the reference implementation of the clinical standard with median DC percentage values of 86.1, 61.3, 92.7, 76.7 and 67.4, 41.6, 79.8, 54.5, respectively. The SRG method has problems with leakage as depicted in Figure 9 (a,c,d), while the proposed method does not show these shortcomings. As expected, it performs slightly better than the version with saliency step size $\delta_S = 8$ with median percentage values of 86.1, 61.3, 92.7, 76.7 and 85.9, 60.5, 92.6, 76.4, respectively, but the differences are not significant. Both versions of our method show a significantly better median value than the clinical standard at the 5 % significance level as represented by the notches of the box-whisker plots. As mentioned in Section 3.2, the main benefit of a larger step size is the lower calculation time. Using $\delta_S = 8$ instead of $\delta_S = 3$ results in a speed-up factor around 2, while the quality of the results is nearly the same.The computation of a block with 1500 frames at $\delta_S = 8$ takes approximately 5 minutes

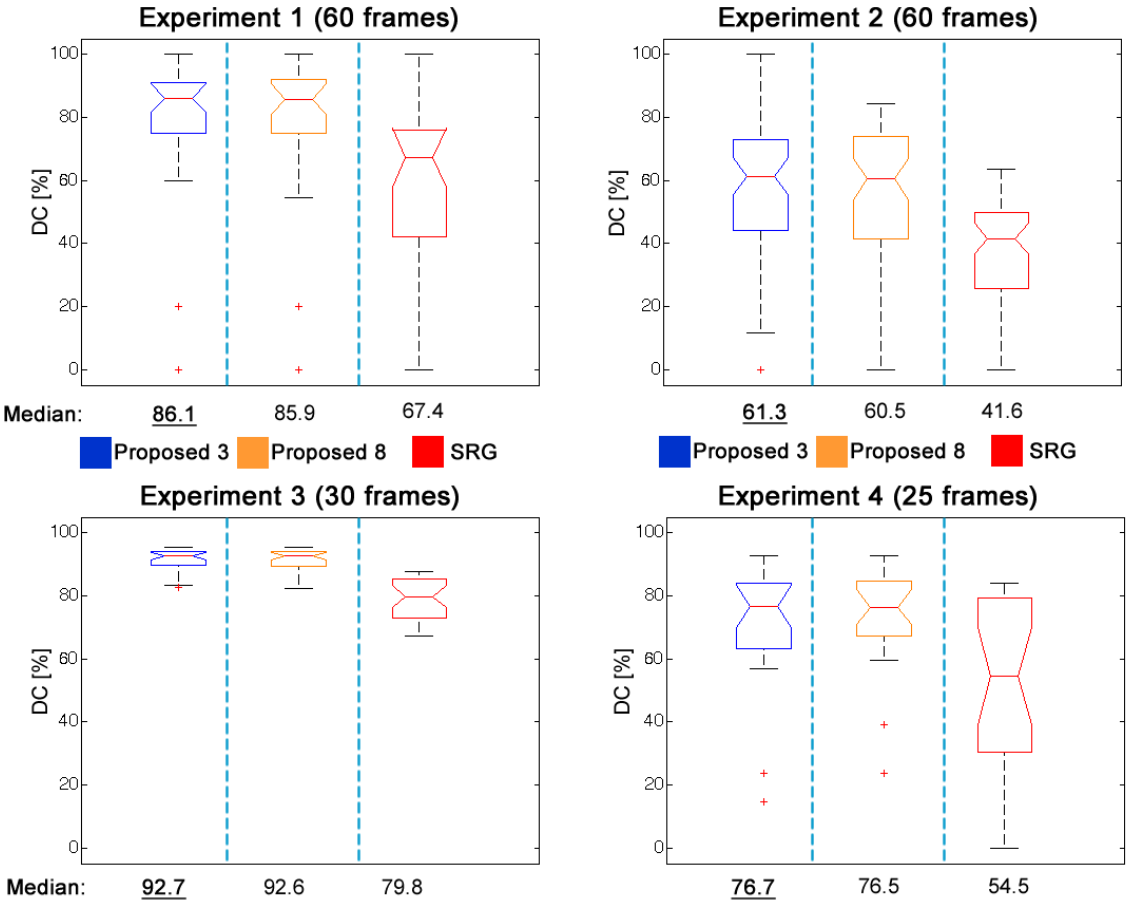


Figure 8: Box-whisker plot with DC results in % of all four experiments. Proposed method in blue for $\delta_S = 3$ and orange for $\delta_S = 8$, SRG [Lohscheller et al., 2007] results in red. The notches of the proposed and SRG method do not overlap, showing a significantly better median value at the 5 % significance level.

and for a whole video of 8000 frames around 30 minutes. Some of the outliers in the box-whisker plots (see Fig. 8) can be explained due to the nature of DC, which over-emphasizes relatively small mistakes for small structures (e.g., missing a ground truth segmentation that only consists of a single pixel would result in $DC = 0$). In Fig. 9 (e) an oversegmentation can be seen in the top right corner when using $\delta_S = 8$. This is not an error due to segmentation, but a problem from ROI detection, because a connection to a nearby dark area exists, which could not be removed by filtering out weak signals in the ROI detection step.

Our dataset already includes some morphological variations due to various recordings from different patients with and without medical conditions. However, to study morphological dissimilarities in more detail an extended evaluation on a larger database is required.

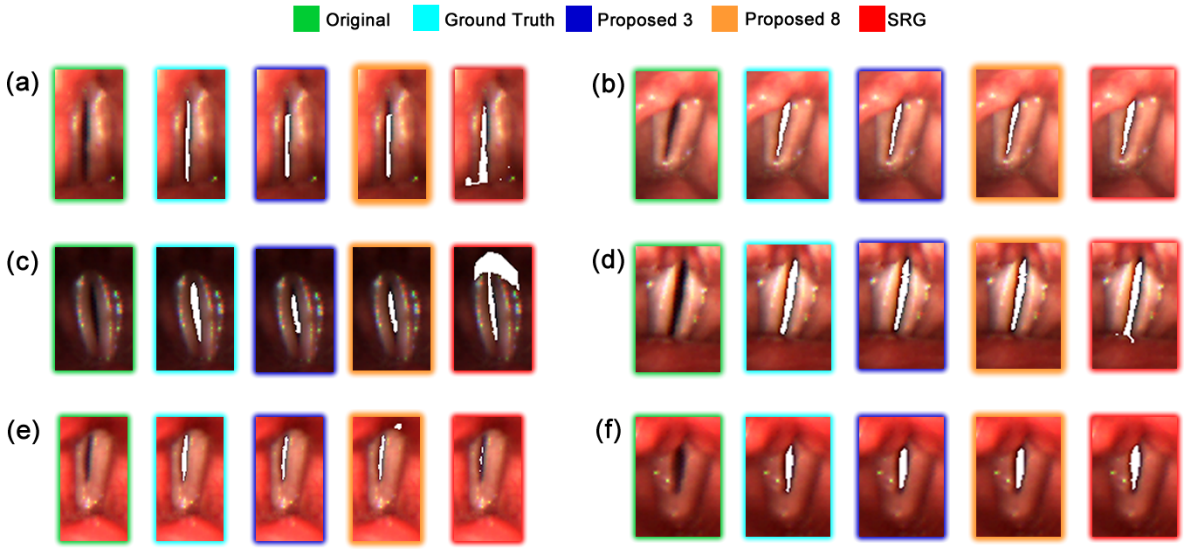


Figure 9: Comparison of the original (green), the ground truth (blue), the proposed method with $\delta_S = 3$ (turquoise) and $\delta_S = 8$ (orange), and the SRG based clinical standard (red).

## 6 Conclusion and Outlook

A fully automatic glottis segmentation method for LHSVs was presented, which shows promising quantitative and qualitative results and is useful for routine clinical application. The method was evaluated by comparing it to the clinical standard on a data set containing annotated ground truth segmentations. Four different experiments were designed, consisting of three video sequences and 25 individual frames, randomly selected from a variety of video recordings. In all experiments, the proposed method outperformed the current clinical standard regarding accuracy while not requiring any user interaction. Despite the promising results, there is still room for improvement. More robust behavior is expected by further refining the ROI and bounding box detection. Additionally, to replace the current clinical standard, in future work a more extensive evaluation on a larger database as well as comparisons to other related methods are needed.

## References

V. Appia and A. Yezzi. Active geodesics: Region-based active contour segmentation with a global edge-based constraint. In *Proc. IEEE International Conference on Computer Vision*, pages 1975–1980, 2011.

A. Borji and L. Itti. State-of-the-art in visual attention modeling. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35(1):185–207, 2013.

X. Bresson, S. Esedoglu, P. Vandergheynst, J.-P. Thiran, and S. Osher. Fast global minimization of the active contour/snake model. *Journal of Mathematical Imaging and Vision*, 28(2): 151–167, 2007.

V. Caselles, R. Kimmel, and G. Sapiro. Geodesic active contours. *Int. Journal of Computer Vision*, 22(1):61–79, 1997.

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

T.F. Chan and L.A. Vese. Active contours without edges. *IEEE Trans. on Image Processing*, 10 (2):266–277, 2001.

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

J. Demeyer, T. Dubuison, B. Gosselin, and M. Remacle. Glottis segmentation with a high-speed glottography: a fully automatic method. In *3rd Advanced Voice Function Assessment Int. Workshop*, 2009.

M. Doellinger, T. Braunschweig, J. Lohscheller, U. Eysholdt, and U. Hoppe. Normal voice production: Computation of driving parameters from endoscopic digital high speed images. *Methods of Information in Medicine*, 3(42):271–276, 2003.

U. Eysholdt, F. Rosanowski, and U. Hoppe. Vocal fold vibration irregularities caused by different types of laryngeal asymmetry. *European Archives of Oto-Rhino-Laryngology*, 260 (8):412–417, 2003.

S. Hertegard, H. Larsson, and T. Wittenberg. High-speed imaging: applications and development. *Logopedics Phoniatrics Vocology*, 28:133–139, 2003.

D.L.G. Hill, P.G. Batchelor, M. Holden, and D.J. Hawkes. Medical image registration. *Physics in Medicine and Biology*, 46(3):R1, 2001.

U. Hoppe, M. Doellinger, S. Schuberth, F. Rosanowski, and U. Eysholdt. Hoarseness caused by laryngeal asymmetries. In *Proc. of 22nd Congress of Union of the European Phoniatricians. Frankfurt/ Main*, 2001.

L. Huang and H. Pashler. A boolean map theory of visual attention. *Psychological Review*, 114:599–631, 2007.

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.

T. Koç and T. Çiloğlu. Automatic segmentation of high speed video images of vocal folds. *Journal of Applied Mathematics*, pages 1–16, 2014.

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.

J. Lohscheller, U. Eysholdt, H. Toy, and M. Doellinger. Phonovibrography: mapping high-speed movies of vocal fold vibrations into 2-d diagrams for visualizing and analyzing the underlying laryngeal dynamics. *IEEE Trans. on Medical Imaging*, 27(3):300–309, 2008.

S.E. Palmer. *Vision science: Photons to phenomenology*. The MIT Press, 1999.

N. Riche, M. Duvinage, M. Mancas, B. Gosselin, and T. Dutoit. Saliency and human fixations: State-of-the-art and study of comparison metrics. In *2013 IEEE Int. Conference on Computer Vision (ICCV)*, pages 1153–1160. IEEE, 2013.

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

L. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992.

F. Schenk, M. Urschler, C. Aigner, I. Roesner, P. Aichinger, and H. Bischof. Automatic glottis segmentation from laryngeal high-speed videos using 3D geodesic active contours. In *Proc. Medical Image Understanding and Analysis (MIUA)*, pages 111–116, 2014.

J.A. Sethian. *Level Set Methods and Fast Marching Methods*. Cambridge University Press, 2nd edition, 1999.

I.R. Titze. On the mechanics of vocal-fold vibration. *The Journal of the Acoustical Society of America*, 60:1366–1380, 1976.

I.R. Titze. *Principles of Voice Production*. Prentice-Hall, 1st edition, 1994.

J.M. Wolfe and T.S. Horowitz. What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience*, 5(6):495–501, 2004.

X. Xianghua and M. Mirmehdi. *Geodesic Colour Active Contour Resistent to Weak Edges and Noise*, pages 399–408. BMVA Press, 2003. Proceedings of the 14th British Machine Vision Conference.

J. Zhang and S. Sclaroff. Saliency detection: a boolean map approach. In *2013 IEEE Int. Conference on Computer Vision (ICCV)*, pages 153–160. IEEE, 2013.

H. Zhou, X. Li, G. Schaefer, M. E. Celebi, and P. Miller. Mean shift based gradient vector flow for image segmentation. *Computer Vision and Image Understanding*, 117:1004–1016, 2013.