Graz University of Technology

Institute for Computer Graphics and Vision

Master's Thesis

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

---

Fabian Schenk

Graz, Austria, September 2014

Thesis supervisor

Prof. Dr. Horst Bischof

Thesis advisor

Dr. Martin Urschler

Ludwig Boltzmann Institute for Clinical Forensic Imaging

To My Mother

Abstract

During the last few decades our economy changed from a manual labor to a service oriented one and speech, as the main form of communication between people, has gained a tremendous economic value. Diagnosis and classification of voice and speech disorders have become important research topics in recent years because generating an appropriate voice signal plays an essential role in verbal communication. In this context laryngeal high-speed videos have emerged as a state of the art method to investigate vocal fold vibration, but the vast amount of data produced prevents it from being used in every day clinical applications.

Segmentation of the glottal opening is an essential, preliminary step for voice and speech disorder research. We present a novel, fully automatic segmentation method involving rigid motion compensation, salient region detection and 3D Geodesic Active Contours segmentation. By using the whole color information and establishing spatio-temporal volumes with time as third axis, our method deals with problems due to low contrast or multiple opening areas. Efficient computation is achieved by a parallelized implementation using modern graphics adapters and NVidia CUDA. A comparison to the seeded region growing based clinical standard on a set of ground truth data shows that we achieve higher segmentation accuracy.

Keywords. laryngeal high-speed videos, geodesic active contours, salient region detection, glottis, segmentation, larynx

Kurzfassung

In den letzten Jahrzehnten fand eine Deindustrialisierung statt und unsere Gesellschaft hat sich zu von einer industriellen zu einer dienstleistungsorientierten gewandelt. Sprache, als wichtigstes Kommunikationsmittel zwischen Menschen, hat daher einen unglaublich hohen wirtschaftlichen Stellenwert. Aus diesem Grund wurden die Diagnose und Klassifizierung von Sprach- und Stimmerkrankungen auch in der Forschung immer wichtiger. In diesem Zusammenhang haben sich Hochgeschwindigkeitsvideos des Kehlkopfs (LHSVs) als wichtigste Methode zur Erforschung von Vibrationen der Stimmlippen etabliert, aber die große Datenmenge verhindert einen vernünftigen Einsatz im klinischen Alltag.

Die Segmentierung der Glottisöffnung ist ein wichtiger, vorbereitender Schritt für die Analyse von Sprach- und Stimmerkrankungen. In dieser Arbeit präsentieren wir eine neue, vollautomatische Segmentierungsmethode, die Kamera- und Patientenbewegungen kompensiert, die Glottis detektieren kann und eine 3D Geodesic Active Contour Segmentierung durchführt. Wir verwenden die gesamte Farbinformation und etablieren Zeit als dritte Dimension, um ein räumlich-zeitliches Volumen zu erstellen, was uns beim Umgang mit geringem Kontrast oder mehreren Öffnungen hilft. Effiziente Berechnungen werden durch eine parallele NVidia CUDA Implementierung auf der Grafikkarte ermöglicht. Ein Vergleich mit dem auf region-growing basierten klinischen Standard auf einer Grundwahrheit zeigt, dass unsere Methode eine höhere Genauigkeit aufweist.

Acknowledgments

Studying here in Graz has been the most amazing and challenging time of my life and on my way to the master's degree a lot of people influenced and supported me. First, I want to thank my supervisor Dr. Martin Urschler, whose guidance and constant help made this thesis possible. His motivation and passion for science really inspired me during the last year. He always helped me when I encountered problems or had any questions. I also want to thank Dipl.-Ing. Philipp Aichinger from the AKH Vienna, who had the initial idea for this thesis, provided us with video material and introduced me the medical part of this work.

Finally, I want to thank my mother, who constantly encourages me to pursue my dreams and is always there for me.

Statutory Declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Place	Date	Signature

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Ort	Datum	Unterschrift

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Contents

1.1 Motivation

During the last few decades our economy changed from a manual labor to a service oriented one and speech, as the main form of communication between people, has gained a tremendous economic value. With e.g. around 60 % of the jobs in the US requiring communication skills and an abundance of 9,5 % of speech and voice disorders, they have become a crucial part in any country's economy [48]. Diagnosis and classification of these disorders have become important research topics in recent years because generating an appropriate voice signal plays an essential role in verbal communication. Voice is usually generated by the glottis in the larynx, which consists of two vocal folds and the opening between them (see Fig. 1.1). The primary voice signal is generated within the larynx by the two opposing vocal folds when set into vibrations by streaming air provided by the lungs [56]. The frequency and intensity of the voice signal can be varied within a wide range by modifying muscle tensions, lengths of vocal folds and lung pressure [55].

Generally, a healthy voice requires symmetric and regular vocal fold oscillations [16, 24]. Perturbations of the acoustic speech signal can be caused by irregular and asymmetric

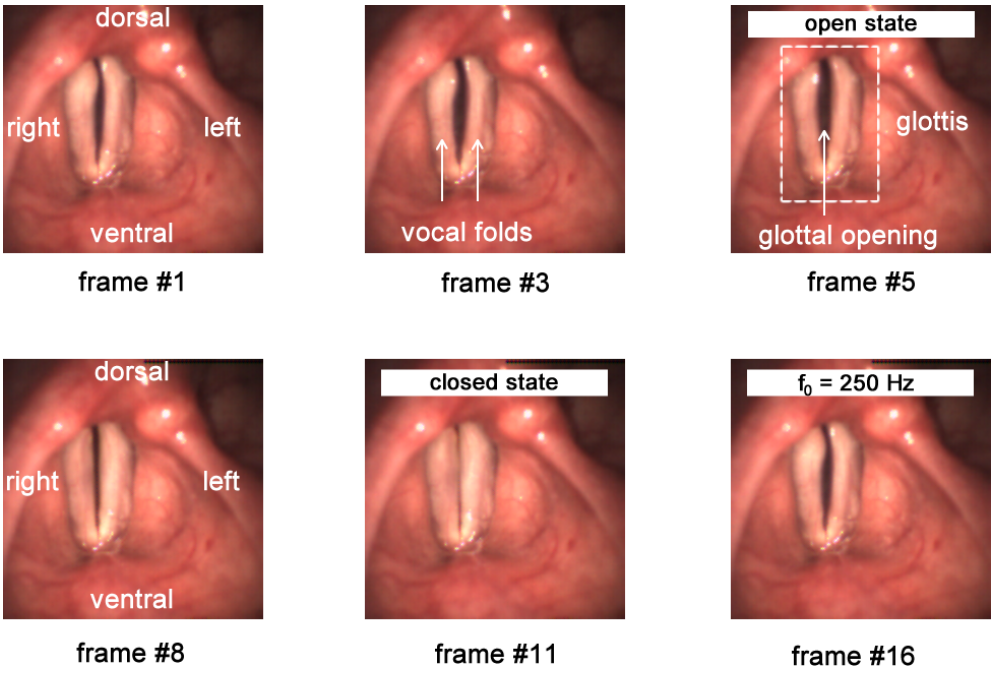


Figure 1.1: Several frames of a laryngeal high-speed video sequence showing a complete cycle of glottis movement including the open and closed states.

vibrations, which results in hoarseness [26]. Eysholdt et al. [17] 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 the World Health Organization 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. Therefore, for an investigation of voice disorders both, the anatomy and vocal fold vibration patterns, have to be analyzed.

Due to the very high frequency of the vocal fold vibrations (e.g. 250 Hz in Fig. 1.1), a high-speed camera with a frame rate of 4000 frames/s is used for the recordings. In this context, laryngeal high-speed videos (LHSVs) have emerged as a very sophisticated tool [13] to accurately record the vocal fold vibrations within the larynx, with the potential of image analysis based post-processing and investigation. Figure 1.1 shows a video recording depicting a typical, complete vocal fold oscillation cycle with a duration of 16 frames, including the open and closed states. The main limitation of LHSVs is the vast amount of video material produced in a single investigation, which makes a manual assess-

*World Health Organization, International Classification of Diseases and Related Health Problems, 2010,http://apps.who.int/classifications/icd10/browse/2010/en#/R49.0, Accessed: 2014-09-12.

ment very time-consuming and renders this method nearly impossible to use in everyday clinical applications. Therefore, methods for automatically processing LHSVs with the aim of detecting voice disorder patterns are highly relevant for speech disorder research.

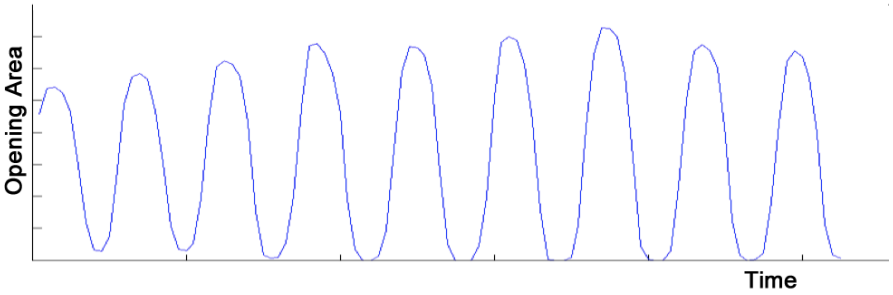


Figure 1.2: This is a diagram of the glottal opening area over time, showing the opening and closing cycles.

The automated segmentation of the glottis area is an important preliminary step for later visual assessment of spatio-temporal plots (i.e., phonovibrograms [40], opening area plots (see Fig. 1.2)). Typical obstacles for this segmentation problem are the drift of the glottis due to patient and camera movements, fluid artifacts, brightness changes and contrast inadequacies during acquisition. State of the art methods are far from ideal when it comes to accuracy and efficiency. Processing speed is also limited 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. Therefore, a method to overcome these problems and provide an automatic and robust way for the glottis segmentation is really important to advance voice and speech disorder research. The presented work was done in cooperation with the *General Hospital Vienna (AKH Wien)* and the *Signal Processing and Speech Communication Laboratory of Graz University of Technology*, where research on LHSV-based voice disorder detection is ongoing.

For this master's thesis the following goals were defined:

- Designing a method to fully automatically detect and segment the glottal opening with the possibility to process multiple recordings consecutively.
- Development of a fast and efficient algorithm, so that it can be used in everyday clinical applications on a standard computer.
- Providing the user with the tools to manually refine the segmentation if necessary.

1.2 Medical Image Data

The medical image data for the experiments and additional material for general testing was provided by the *General Hospital of Vienna*. Most of the recordings were performed by Dipl.-Ing. Philipp Aichinger, who currently establishes a database of LHSVs videos. A schematic setup of the recording process is depicted in Figure 1.3, where it can be seen that an endoscope with a high-speed camera (2) is inserted into the patient's mouth to record the vocal fold vibrations (1) in real time. During the acquisition process the patient

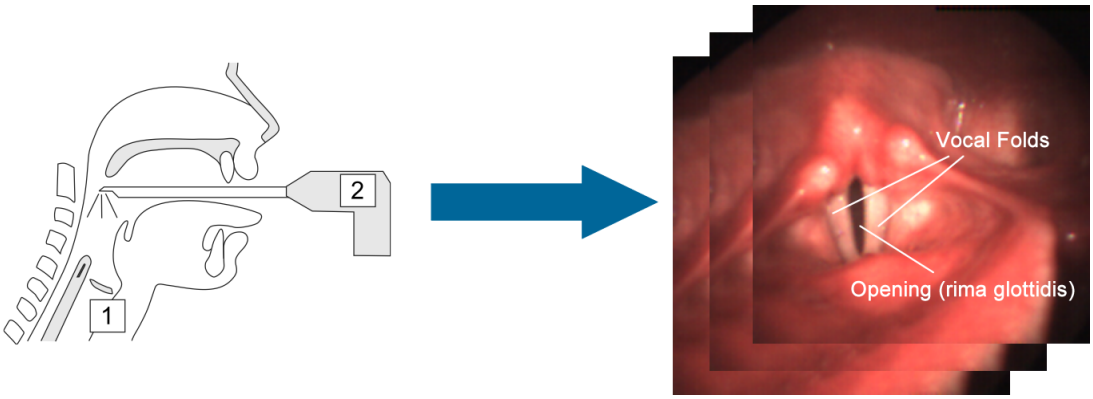


Figure 1.3: Typical recording setup for LHSVs (taken from [41]) and resulting single frames. A high-speed camera (2) is attached to an endoscope in a 90° angle to provide an optimal view of the glottis and the vocal folds (1). The video is then recorded from this position in real time.

is instructed to generate sound to induce vocal fold vibrations. Depending on the research question, the patient varies pitch or intensity of the voice. Usually a few consecutive recordings are performed and afterwards the best ones are analyzed.

In the laboratory in Vienna a *High Speed ENDOCAM 5562* from *Richard Wolf GmbH* is used. It has an active-pixel sensor with a full color resolution of 512 x 384 pixels and includes a microphone to record the voice as well. Different recording modes are possible with the device but only the high speed mode is fast enough to accurately capture the vocal fold vibrations. During high speed recordings it is not possible to use the full resolution because the sensor can only record with 128 x 256 pixels at a very high frame rate. One such color video is usually recorded at 4000 frames per second and has a length of 2,048 seconds, resulting in a total number of 8192 single images. By reading only every second pixel from the charge-coupled device array in the sensor a 256x256 chessboard-like pattern is generated. The resulting missing data points are then interpolated using

bilinear interpolation [21, p. 88].

Our algorithm can only process single images, so the frames have to be extracted from the recording. For the in- and output we chose the Portable Network Graphics (PNG) format, which is a raster graphics file format supporting color and gray-scale images. The main benefit compared to JPEG is the lossless compression, which is extremely important when the resolution is not really high. Color information is stored in three separate channels representing red, green and blue (RGB) as a value between 0 and 255. These images have to be stored in a folder and named using a 4 digit numbering system (e.g.: 0002.png, 6415.png). Memory requirements on the hard disk are usually between 500 and 600 MBytes per video.

1.2.1 Artifacts and Obstacles

The video material is essential for the whole image processing method but a perfect recording is not possible. This is due to the acquisition being exhausting for both, the patient and the doctor, the rather low resolution of the high-speed camera, noise, illumination and anatomical reasons. We decided to exclude recordings where the glottal opening is not visible from the experiments (see Fig. 1.4(a) and (b)).

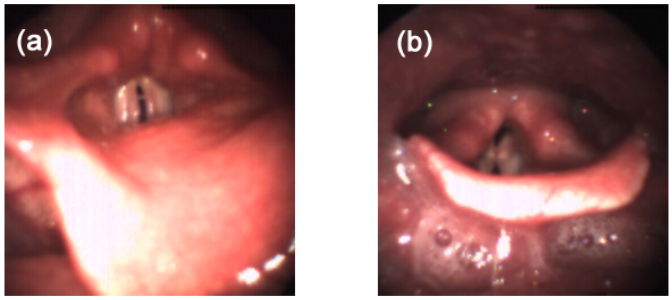


Figure 1.4: In these two recordings the glottal area is not clearly visible and therefore they were excluded from the experiments.

Even though common problems can be overcome by an experienced person controlling the camera, there are still certain obstacles we have to deal with using image processing techniques. Typical problems with the LHSVs will be explained in the following sections.

Resolution

With a resolution of 256x256 the image size is quite small and as depicted in Figure 1.5 the glottal area only occupies a small part of the whole image. The border between the glottal area and the vocal folds is not clearly visible and therefore an exact segmentation with an algorithm or even by experts is very difficult.

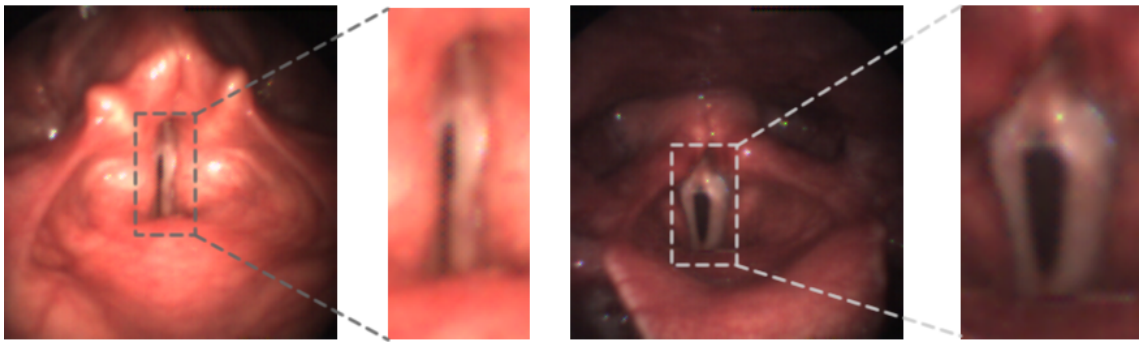


Figure 1.5: The low resolution of the high-speed camera does not show the edges clearly.

Contrast

Poor illumination leads to parts of the recording appearing very dark because not the whole color range (0-255) is used (see Sec. 2.2.1). This leads to rather small contrast differences resulting in difficulties when finding the glottal opening and calculating edges for the segmentation.

Non-homogeneous Background

A homogeneous background is very important to distinguish the glottal opening from its surroundings. As depicted in Figure 1.6, there are typically two factors influencing the homogeneity: anatomical structures and light artifacts. Blood vessels, the epiglottis and other structures are normally visible in every recording. Light artifacts reflected by body fluids like saliva are also very common and can result in small, brightly glowing points in the images. In some videos there is also a reflection of the light source taking up a big part of the image (see Fig. 1.6(b)), which may influence image registration.

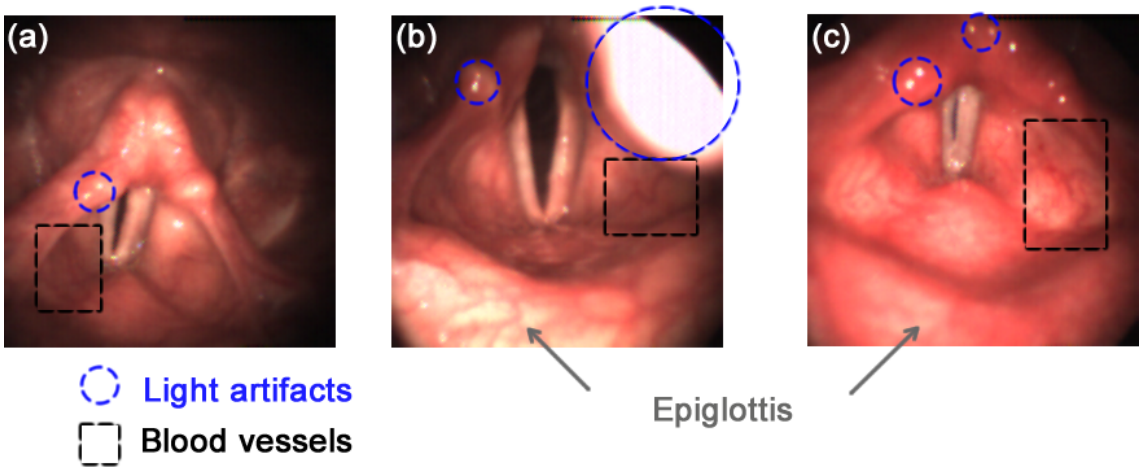


Figure 1.6: These three images depict the typical obstacles causing a non-homogeneous background. Light artifacts caused by reflecting fluids are visible in all of them with (b) showing an extreme case. Blood vessels and other structures are also common in every recording. The epiglottis only appears in (b) and (c).

Patient and Camera Movements

In every recording a global drift occurs caused by breathing or an unsteady camera (see Sec. 2.2.3) Depending on the patient and the experience of the doctor, this global movement is different in all the videos and even unpredictable throughout one sequence. Fortunately, these movements occur rather slowly in comparison to the vocal fold movement due to the high frame rate of the LHSV recording. Therefore, they may be corrected using image registration.

Partly visible Glottis

A completely and well visible glottal area is essential for a correct segmentation. Figure 1.7 shows that in some of the recordings the glottis is only partly visible. The glottal opening in (a) looks very similar to the background because it is covered by saliva. In (b), the reflected light source is covering a large part of the image resulting in difficulties with the glottis detection. Anatomical structures cover the glottal opening in (c) and in Figure 1.4 (a) and (b) rendering a correct segmentation impossible.

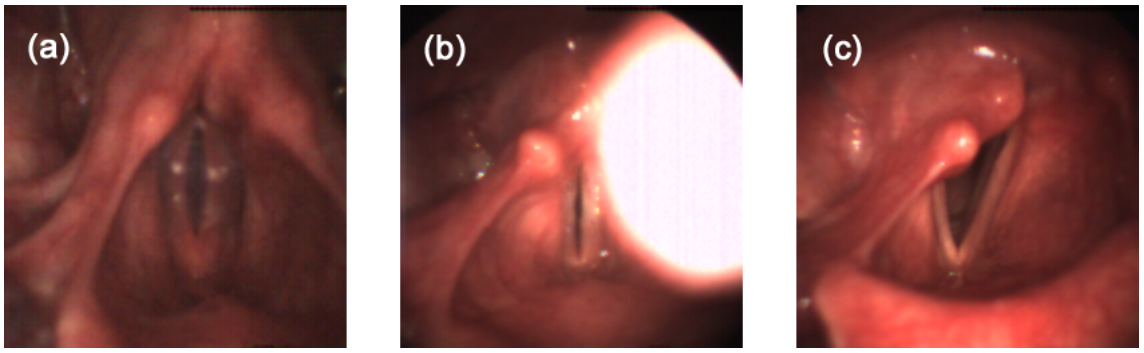


Figure 1.7: In (a) the glottal opening is covered by saliva, whereas in (b) the reflected light source is an obstacle. In (c) an anatomical structure is covering the glottis.

1.3 Related Work

In recent years the number of people studying voice and speech disorders has increased and new recording techniques have greatly extended the research possibilities. With the introduction of LHSVs researchers have received a very powerful tool to analyze vocal fold vibrations. The huge amount of data produced by this method has lead to the development of computer-aided tools to study the video material. Especially the segmentation of the glottis from LHSVs has received a lot of attention in the last decade. In general, the previous approaches can be divided into seeded region growing (SRG) [41, 67], active contour [34], a combination of these two [14] and histogram based methods [37, 43]. For this section we will focus on three well-known and established algorithms and one very recent work.

The first widely known segmentation method was developed by Lohscheller et al. [41], who presented a segmentation approach using a seeded region growing algorithm [21, p. 785]. In this method the necessary initial seed points and a continuous threshold progression have to be manually specified by the user on a large number of frames throughout the video, which is a tedious, 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 a full closure. To account for a glottal drift they divide the video into intervals and use the segmentation of the maximum opening as seed points for the next interval. Even though this method has emerged as the clinical standard, it has a few drawbacks. The biggest problem of

this method is that a lot of user-intervention is necessary and most of the time the final segmentation has to be refined as well because threshold based methods often suffer from leakage or only partially segmented regions. Another drawback of this method is its lack of a proper motion compensation step to account for patient or camera movements, because the propagation of seed points does not always work well.

Demeyer et al. [14] developed a method, which avoids user-interaction by first analyzing the video. The primary goals of this analysis are the detection of interesting frames and getting an estimation of the vocal fold oscillation frequency. They define the most important images per cycle as the ones with the maximum glottal opening, which is usually a dark, big area. This particular frame can be detected by calculating the sum of pixel intensities and taking the one with the lowest value. For the glottis localization they look for dark, elliptical regions bordered by the brighter vocal folds using a *Laplacian of Gaussian*. After a few filter steps they are able to detect the center of the glottis and estimate its size. Then an SRG method is applied with an automatically calculated threshold from the mean value of the already detected area until the segmented region size is greater than previously estimated. This segmentation result is then propagated for- and backward throughout the image cycle using per image level sets, which evolve the seed regions towards the glottis edges. The detection of the glottal opening is not very robust and the authors state that size may be underestimated due to inclination of the glottis. Similar to [41] this algorithm can also have leaking problems due to the SRG approach and motion is supposed to be compensated through result propagation. Another problem is that the edges can not be calculated accurately without a denoising step, which will later be discussed in Section 2.2.2. Further, the empirically selected parameters of the 2D level set are crucial and the propagation heavily depends on the initialization from the SRG step.

Karakozoglou et al. [34] proposed another interesting way for a fully automatic glottis segmentation. First, they look for the frames with the maximum opening similar to [14], which they call landmark frames (LF). Then they search these images for large, nearly vertically oriented areas and apply an edge detection filter. With the following connected component analysis [50] they find the vertically oriented object with the largest area in each cycle. A bounding box is then computed for every frame to reduce calculation time and memory requirements. The glottal drift and endoscope movements are compensated by keeping the bounding box steady throughout every cycle. To avoid segmenting frames without a glottal opening they exclude images with pixel intensities above a certain

threshold and a bounding box far away from the one of the LF. They use an active contour segmentation as proposed by Chan and Vese [10] using information obtained from the LF like shape and area of the object of interest. The segmentation always starts from the LF, where they use either an automatically computed threshold or an elliptic mask for the critical curve initialization. After this process, the segmentation results are propagated using the mask of the (n-1)th or (n+1)th frame as initial mask for the nth frame. Despite showing promising results and being fully automatic, this method has certain limitations. The calculation of the initial curve for the LF can be problematic due to leaking (especially when using thresholding) or when the glottis does not have an elliptic form. Segmentation accuracy of the active contours is also reduced by the noise in the image, which results in imprecise edges (see Sec. 2.2.2). Changes in topology (i.e., two glottal openings) are really hard to correctly segment with 2D active contours, a problem which could be overcome by a 3D active contour approach.

Recently, a novel approach was proposed by Koç, Turgay and Çiloğlu [37], which is based on image histogram modeling and thresholding. They try to detect the glottal opening due to the area changes caused by the opening and closing cycles. The problem is that the relatively small glottal area in most of the recordings does not change the histogram enough to give any useful information but by looking just at the region containing the glottis, the changes are significant. Accurately determining the region of interest (ROI) is essential to make the histogram bimodal and sensitive to changes in the glottal area. To find the glottis in the image they assume that the vocal folds are the most active structures within the sequence and their movement produces pixel changes. By calculating the Total Variation (TV) norm over a time sequence they get the largest TV values where there is the most movement, which coincides with the glottal area. The non-uniform illumination is a big issue when determining a threshold from the intensity histogram but they overcome this problem by using a novel illumination model to calculate a reflectance histogram. As a last step they present an automatic threshold computing method using Gaussian Mixture Models and the reflectance. This work presents some interesting techniques, provides solutions for common segmentation problems in LHSVs and the results look promising but there are still a few limitations. They completely neglect the global drift and therefore do 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. When there is no movement over a longer period of time or the global drift is high the TV calculation leads to wrong results and a wrong ROI detection.

All methods presented in this section neglect the full color information and lack a proper motion compensation step. The clinical standard [41] requires a lot of user-interaction and has problems with leakage due to the SRG method, while other methods [14, 34] have problems with special topologies and leakage due to the 2D nature of the algorithms. With the very recent work [37] it is impossible to detect the glottis when the vocal folds are not moving.

1.4 Overview and Contribution

This master's thesis has its focus on a robust, automatic segmentation of the glottis. For this purpose we developed a new image processing algorithm to overcome the drawbacks and issues seen in previous methods (see Sec. 1.3).

Our method is fully automatic, computationally efficient and is also able to process multiple recording consecutively, which is especially useful when working with different videos from one patient or analyzing a whole database of videos. With a specially designed preprocessing step we address most of the common problems induced by the acquisition process and prepare the video for the segmentation. This includes contrast stretching, a motion compensation step to correct rigid patient or camera drifts and edge-preserving denoising to get clearer edges for the segmentation. We also offer a robust detection of the glottis region in the image to address the difficult problem of finding the ROI. For this purpose we use a novel salient region detection method adapted from eye-tracking, which is also used for finding the initialization for the segmentation. We are also able to extract more information compared to traditional methods based on gray-scale images by taking the full color-information into account. For the image segmentation we establish time as a third axis to generate a spatio-temporal volume. On this volume we perform a 3D Geodesic Active Contour segmentation, which gives an advantage over more traditional 2D active contour approaches [14, 34].

A big drawback in most of the previous works is that the computation time is rather slow and therefore the method can not be used in everyday clinical practice. Most of our algorithm is 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.

1.5 Structure of this Thesis

This thesis is organized as follows. Chapter 2 gives an overview and a detailed description of the different algorithms we use for contrast stretching, denoising, rigid motion compensation, salient region detection and segmentation. Related work to these models and the deductions are also shown in this chapter.

In Chapter 3 we present the results of our method by comparing it to the clinical standard [41] using a previously created ground truth. For that purpose we use an experimental setup of randomly selected single and consecutive frames from different recordings.

The work is summarized and concluded in Chapter 4. Appendix A shows a publication and a list of oral presentations based on this master's thesis. Abbreviations and definitions are described in Appendix B. Appendix C explains the two configuration files used by our software and Appendix D finally gives an overview of the libraries and software dependencies used for the implementation.

Chapter 2

An Automated Glottis Segmentation Approach

Contents

2.1 Overview

An overview of the image processing pipeline we use to tackle the problem of fully automatic glottis segmentation from LHSVs is depicted in Figure 2.1. The high-speed recordings are characterized by a short time span between the frames allowing us to treat the LHSVs as 3D spatio-temporal volumes, where the typical behavior of the glottis is a repeated opening and closing process, which is responsible for producing sound. Our method is split into three major blocks: preprocessing, region of interest (ROI) and seed region detection, and segmentation. In the preprocessing phase we prepare the image material for the segmentation and deal with obstacles (see Sec. 1.2.1) like non-homogeneous background, light-artifacts and global movement by performing denoising and frame-based

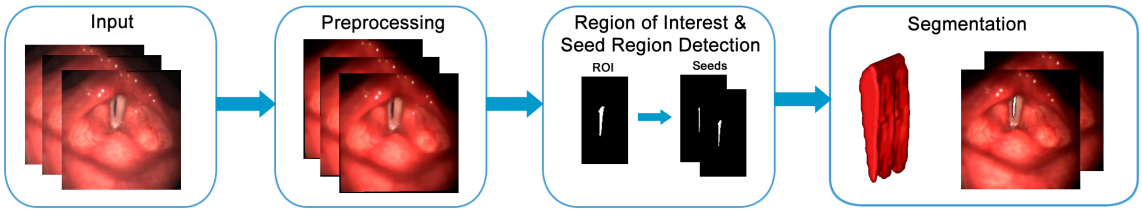


Figure 2.1: Our method can be divided into three major blocks: preprocessing, region of interest and seed region detection and segmentation

image registration. Then the glottis ROI is computed and in a similar fashion the seed regions are calculated as initialization for the final segmentation. As a result we get a 3D spatio-temporal volume of the opening and closing glottis.

After this brief overview, our method will be explained in greater detail as shown in Figure 2.2. First, we utilize a simple contrast stretching operation to scale the color channels to the full range of $[0, 255]$. To get rid of small light and fluid artifacts and generate a homogeneous background, we apply an edge-preserving denoising filter to the 3D spatio-temporal volume. Global motion, which is present in most of the images, is compensated by a rigid intensity-based registration of the single frames. Efficient computation is very important for everyday clinical application, thus requiring a GPU based implementation. Unfortunately, graphics adapter memory is still limited compared to CPU main memory. To save memory, we limit computation to a ROI from the image. This ROI contains the glottis, the interesting part for subsequent analysis, and achieves a reduction of the size of the investigated video data.

The core of our method is the salient region detection using a Boolean Map based approach, which is used for detecting the ROI and the seed regions located in the interior part of the glottis. To find the correct ROI we search for the frame with the maximum opening, compute the interesting areas and eliminate all that are not fulfilling a certain criteria. For the seed region detection we also calculate the interesting areas and simply multiply with the ROI to filter unwanted details. These seed regions are then used as initialization in the following 3D Geodesic Active Contour segmentation. Our final result then resembles the 3D spatio-temporal volume of the opening and closing glottal area. The output can be single images or a volume file, which can be visualized using ITK-Snap [68] or similar tools.

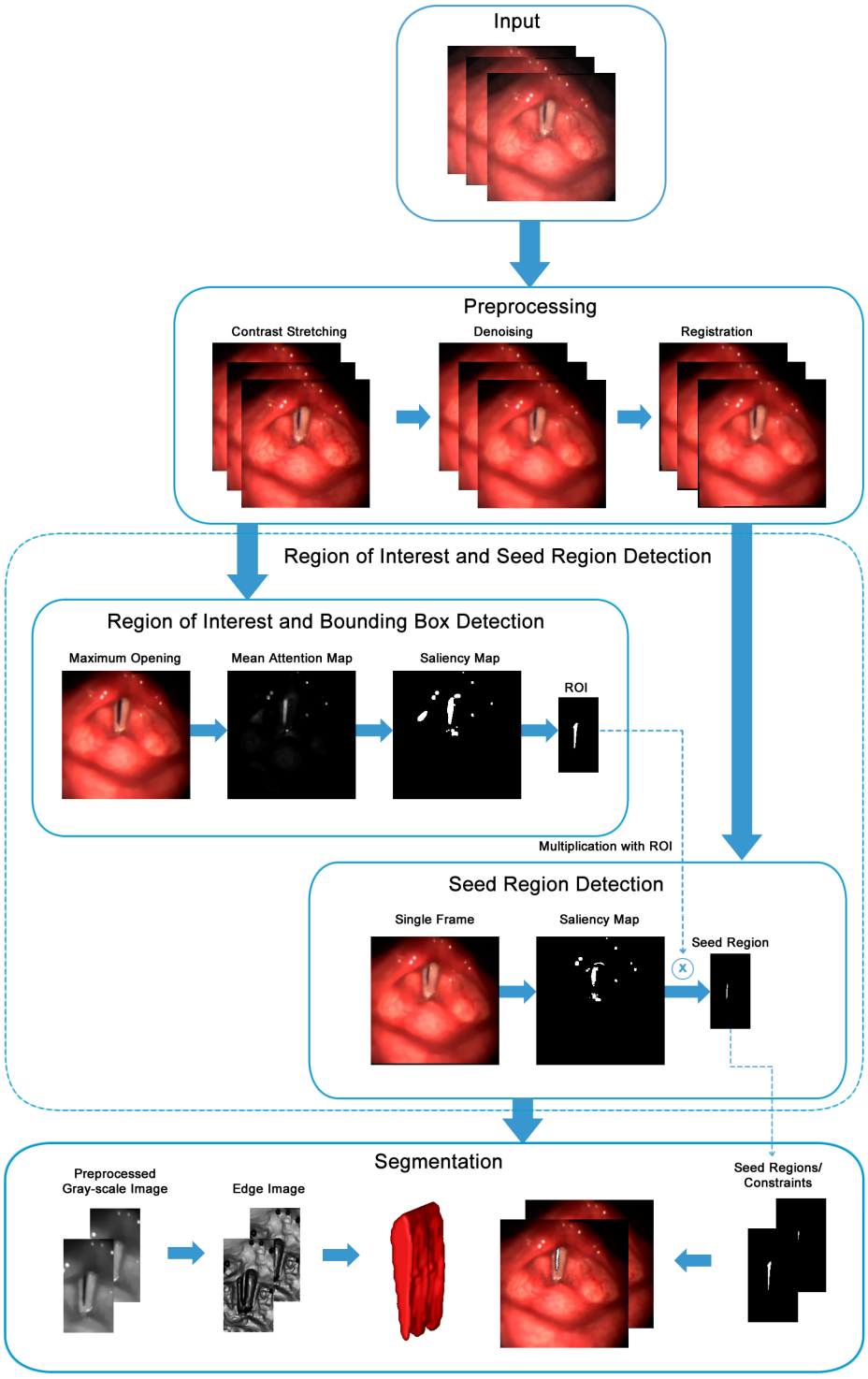


Figure 2.2: Detailed image processing pipeline for the proposed automatic glottis segmentation approach.

2.2 Preprocessing

In the preprocessing phase the images are prepared for the segmentation and we address obstacles like non-homogeneous background, light-artifacts and global movement (see Sec. 1.2.1). Figure 2.3 shows the image processing sequence and the results after each step.

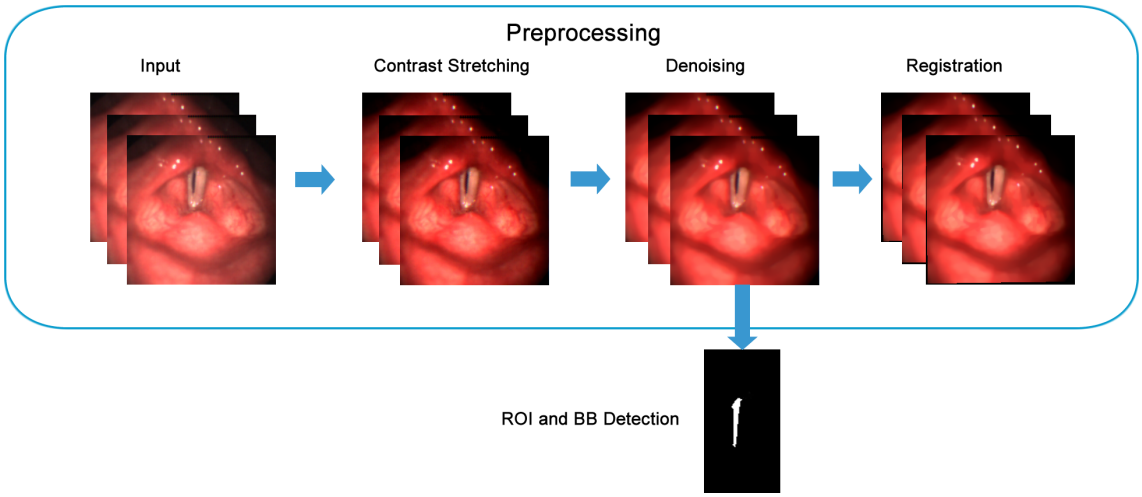


Figure 2.3: The preprocessing pipeline showing the contrast stretching, denoising and registration steps and their results.

First, we reduce contrast inadequacies using a simple contrast stretching operation. The following denoising step generates a homogeneous background by removing unwanted details like blood vessels, small light artifacts and noise while preserving the edges. Global drifts are then compensated by a rigid motion compensation step accounting for translation and rotation. To reduce computation time and memory requirements a ROI is detected and a surrounding bounding box is calculated. After this step only the image information within this bounding box is used for the calculations.

In the following sections every single step will be described in detail.

2.2.1 Contrast Stretching

Part of the source material is very dark, because not the whole contrast range from 0 to 255 is used, resulting in contrast inadequacies. This can lead to rather small differences between glottal opening and background intensities, which makes a glottis detection and edge computation rather difficult. We want to enhance the image quality of these pictures

by performing a contrast stretching operation on all the color channels of the image. In [37] the authors also perform an image enhancement operation before the actual processing steps. For a gray-scale image the contrast stretching operation [21, p. 137] is defined as

$$I_{output} = (I_{input} - oldMin)\frac{newMax - newMin}{oldMax - oldMin} + newMin,$$

where $newMin$ and $newMax$ are the limits the image should be stretched to and $oldMin$ and $oldMax$ are the old limits. For a color image the greatest contrast range among the three color channels is calculated. This can be done by first calculating a histogram for every color channel [21, p. 142], which is basically a data structure representing the number of occurrences of each intensity value in an image. To calculate the new limits we take the color value greater than 5 % of the intensities as $newMin$ and the one greater than 95% as $newMax$ (see Fig. 2.4). The resulting transformation is then applied to all the color channels separately.

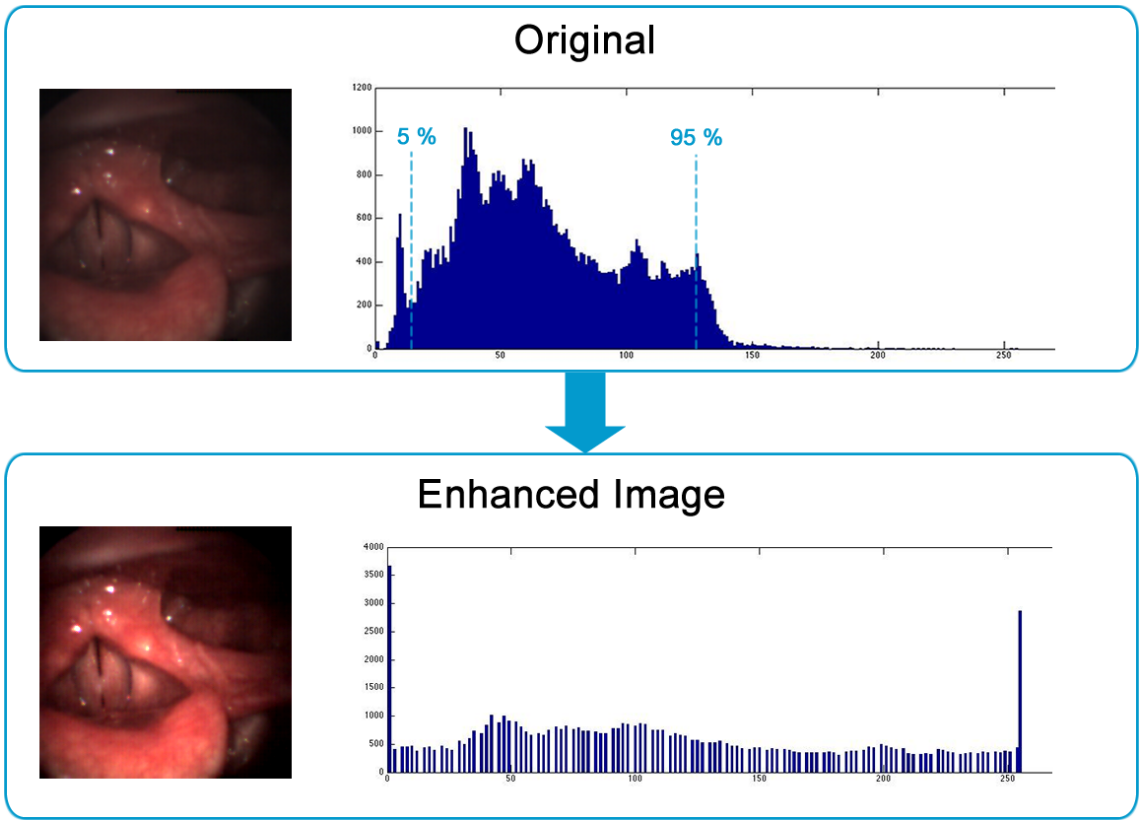


Figure 2.4: Contrast stretching using the 5 and 95 percentiles of the histogram.

2.2.2 Edge-preserving Denoising

Noise is a general problem in digital images and the principal sources arise during image acquisition (digitization). The performance of imaging sensors is also affected by a variety of factors [21, p. 335]. In all of the videos there is noise as well as small structures like disturbing light artifacts and blood vessel greatly influencing following steps like salient region detection (see Sec. 2.3) and edge calculation (see Sec. 2.5.2). The effects of noise and small structures like blood vessels and light artifacts is shown in Figure 2.5. They greatly affect the edge computation for the segmentation and result in more salient regions, thus we have to include a denoising step in our method, which also provides image edges since those are central to later segmentation.

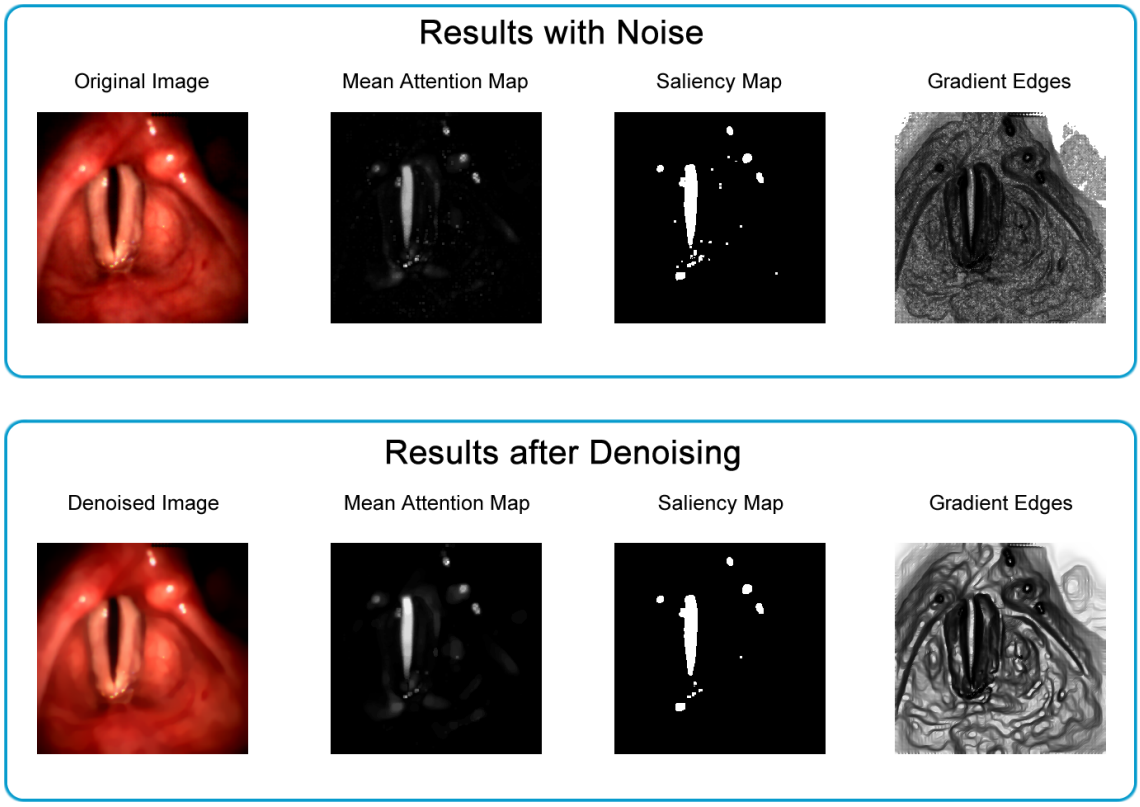


Figure 2.5: The effect of noise and small structures like blood vessels and light artifacts on the salient region detection and the edge detector function compared to a denoised version.

In computer vision one often has to deal with problems like image denoising or restoration, which are typically *ill-posed*. Hadamard [22] defined three conditions that have to

be fulfilled for a problem to be *well-posed*:

- Existence: A solution x exists for the problem.
- Uniqueness: The solution x is unique.
- Stability: The solution x depends continuously on the initial conditions.

Every other problem is *ill-posed*. Solving *ill-posed*, inverse problems is only possible with prior knowledge. A common way is to use the Bayesian approach on inverse problems, which has been extensively studied.

The Bayesian approach leads to an optimization problem, which is given by:

$$\max_u p(u|f),$$

where we have to find a hypothesis u by maximizing the probability based on the observation f. Using Bayes theorem [1] we can define the maximum a posteriori (MAP) as

$$p(u|f) = \frac{p(f|u)p(u)}{p(f)},$$

with the prior $p(u)$, conditional probability $p(f|u)$ and a normalization term $p(f)$, which can be ignored for the optimization. From the Bayesian approach we will now deduce the Tikhonov regularized model [54] for image denoising. The following derivation closely follows the one in Markus Unger's PhD thesis [57]. By assuming that the observed data f is subject to Gaussian noise with a variance of σ^2 and mean μ^2, the likelihood can be represented as

$$p(f|u) = \prod_{x\in\Omega} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(u(x)-f(x))^2}{2\sigma^2}}$$

and the prior probability as

$$p(u) = \prod_{x\in\Omega} \frac{1}{\sqrt{2\pi}\mu} e^{-\frac{|\nabla u(x)|^2}{2\mu^2}}.$$

If we introduce these terms into the MAP formulation we get

$$\begin{aligned}\max_u p(u|f) &= \max_u \left\{ \prod_{x\in\Omega} e^{-\frac{(u(x)-f(x))^2}{2\sigma^2} - \frac{|\nabla u(x)|^2}{2\mu^2}} \right\} \\ &= \max_u \left\{ e^{-\int_\Omega \frac{(u(x)-f(x))^2}{2\sigma^2} + \frac{|\nabla u(x)|^2}{2\mu^2} dx} \right\}. \end{aligned} \tag{2.1}$$

To maximize 2.1 we have to minimize the exponent, which leads to:

$$\min_u \left\{ \int_\Omega \frac{(u(x)-f(x))^2}{2\sigma^2} + \frac{|\nabla u(x)|^2}{2\mu^2} dx \right\} = \max_u p(u|f).$$

Assuming $\frac{1}{\sigma^2} = \lambda$ and $\mu = 1$ we get the convex formulation of Tikhonov denoising [54]:

$$\min_u \left\{ \frac{1}{2} \int_\Omega |\nabla u|^2 + \frac{\lambda}{2} \int_\Omega (u-f)^2 dx \right\},$$

By replacing the quadratic regularization with the L1-norm, we get the ROF model, named after its inventors Rudin, Osher and Fatemi [49]. The convex formulation of the ROF model was introduced by Chambolle and Lions [8] as

$$\min_u \left\{ E_{ROF} = \int_\Omega |\nabla u| dx + \frac{\lambda}{2} \int_\Omega (u-f)^2 dx \right\}. \tag{2.2}$$

with the original, noisy image f, the reconstructed image u and a weighting factor λ. The first term is the regularization or total variation (TV) term, which punishes non-homogeneous regions and the second one is the L2-data fidelity term, which makes differences between f and u expensive. λ determines the influence of the data fidelity term and therefore sets the similarity between the original image f and the reconstructed, denoised image u. The ROF model is giving better results than the Tikhonov-regularization model, because it also preserves image edges. In Figure 2.6 three functions with different step sizes are shown, which illustrate the edge-preserving effect. Smaller steps cost significantly less in the Tikhonov model due to the quadratic regularization term, which favors smooth transitions leading to a blur of the image. For the TV term the cost is always the same and therefore only the data term is important for the final result, which results in a preservation of edges or smooth transitions. TV regularization prefers flat over rippled functions and the higher cost of noise and small structures is responsible for the denoising effect. In the denoised images in Figure 2.7 small plateaus occur, which can be especially well seen

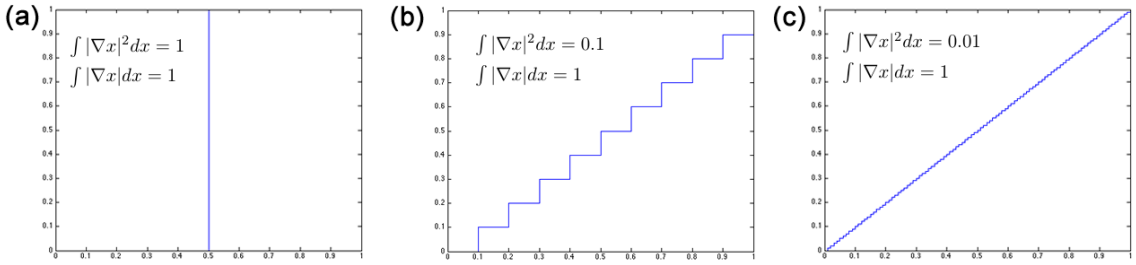


Figure 2.6: Comparison between the quadratic regularization of the Tikhonov model and the total variation (TV) regularization used in the ROF model based on three different functions. Smaller steps cost less in the quadratic regularization because of the square, whereas the TV regularization always costs the sum of all steps.

for $\lambda = 1$. This is a well known side-effect of TV regularization, called stair-casing effect. A higher value reduces the denoising effect because differences become more expensive and a lower value even smoothes over the edges. For our denoising problem we empirically found a value of 25 to yield good results. We achieve efficient computation by using a

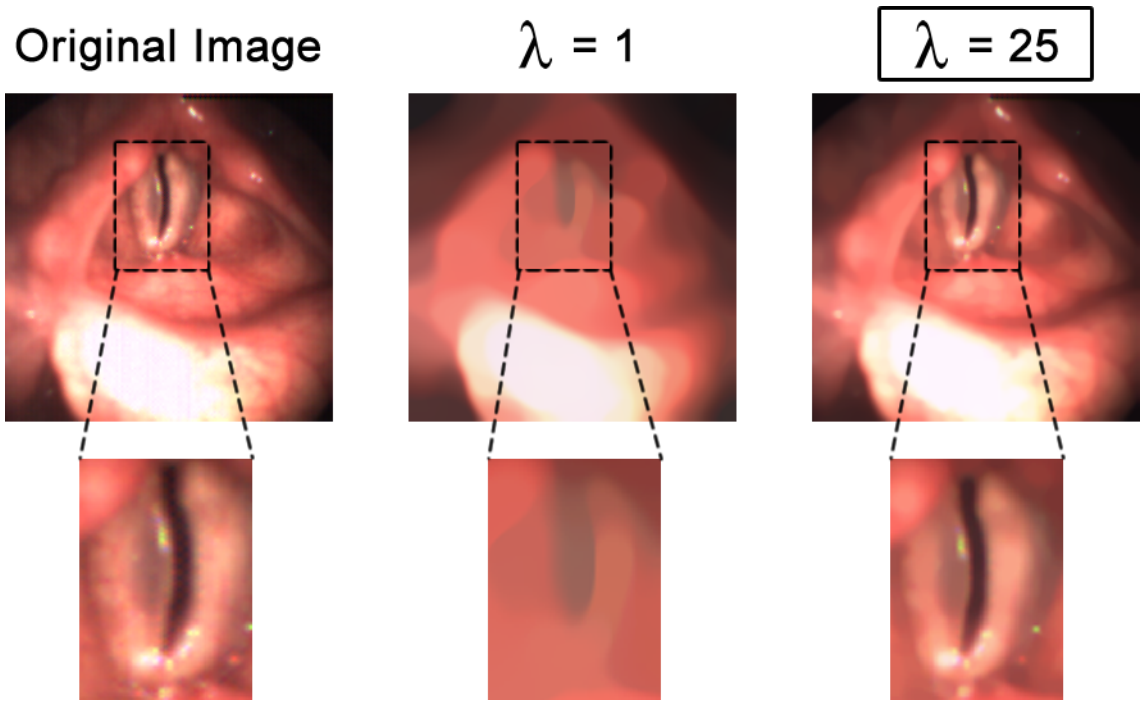


Figure 2.7: Denoising effects of λ, where a higher value reduces the denoising and a lower one smoothes over the edges. We empirically determined a value $\lambda = 25$ for our method.

continuous convex optimization scheme based on the primal-dual algorithm by Chambolle and Pock [9]. This operation is performed on the 3D spatio-temporal volume utilizing the graphics adapter in an NVidia CUDA implementation.

2.2.3 Image Registration

In practice, LHSVs always contain global movement caused by the patient and the camera during the recording process, which can greatly effect the image segmentation quality. This can happen due to breathing and the uncomfortable feeling of having an endoscope in the mouth while creating different pitches and voice patterns.

In Figure 2.8 the difference between the original, fixed image I_0 combined with the contours of a registered and unregistered image I_n are shown. Without a proper registration step, the glottis would not always be in the same position, making it impossible to generate a smooth 3D spatio-temporal volume from the single frames, which is crucial for our method. Most of the previously published methods discussed in Section 1.3 also identify these global drifts as a problem in the segmentation process. The solutions are

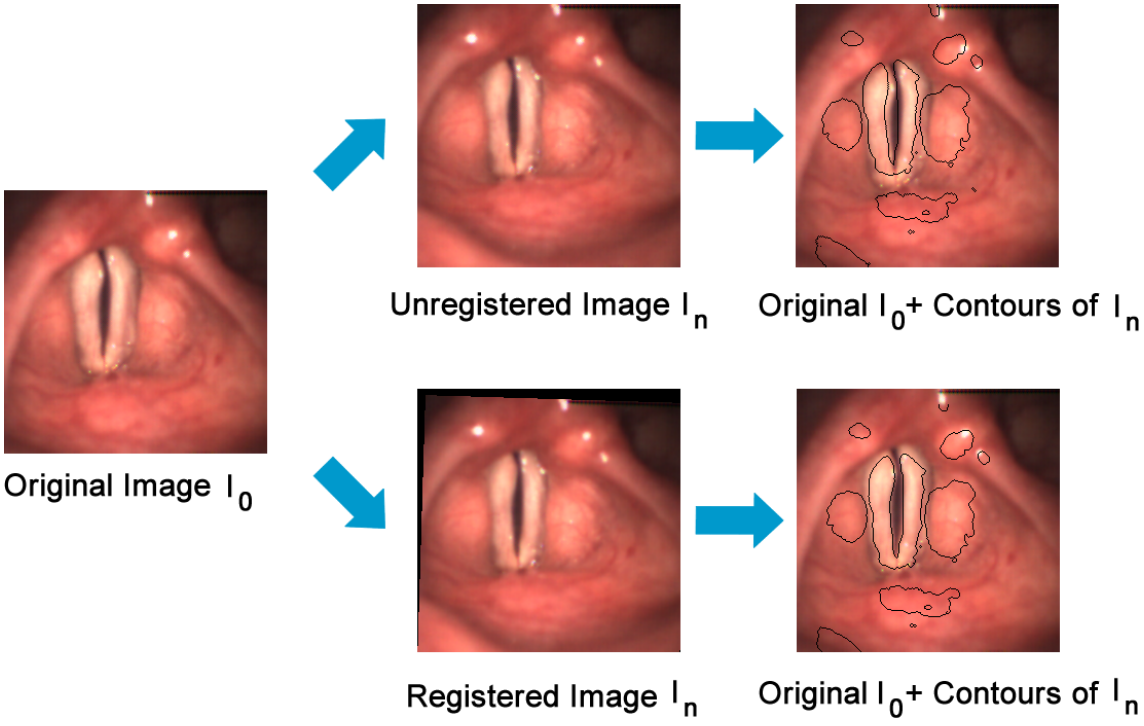


Figure 2.8: This image shows the importance of the registration by showing the contours of the registered and unregistered on top of the original.

quite different in every algorithm, depending on the final segmentation step. In [41] there is already a lot of user-intervention necessary in the overall procedure, so they simply let the user specify seed points in a lot of images and the result is then interpolated in between. Other methods [14, 34] perform calculations for every opening and closing cycle and assume that the movement within one of these cycles is negligible. Motion compensation in LHSVs has been a topic since the development of the first segmentation techniques. Deliyski et al. [12] conducted a study of seven different methods, with two using image cross-correlation, two minimizing the L2-norm and magnitude difference between two images, and three FFT-based methods. The evaluation was based on the computation speed, the mean absolute error and the absolute range of error. While the FFT based methods were by far the fastest, their accuracy was not among the best. They showed that the *magnitude of difference minimization* method is a very good compromise between speed and accuracy.

In our work we use a similar method but we additionally take rotation into account. We use the well studied sum of absolute differences (SAD) [25, 62, 63, 65] as similarity measure derived solely from the color pixel intensities of the three RGB channels. It is an intensity based rigid registration, which means that translation and rotation are compensated.

As depicted in Figure 2.9 a common problem in the videos is that very big structures like the epiglottis ((a) through (d)) or light sources ((e) and (f)) move during the whole recording process. Their movement would look like a global drift for an intensity based algorithm taking the whole image into account because large parts of the image are involved. To overcome this limitation we only use the pixels within the previously calculated bounding box (see Sec. 2.4) for our calculation. We define the first image I_0 in a sequence

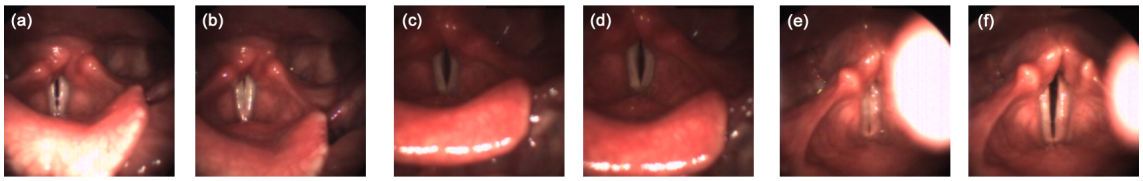


Figure 2.9: In images (a) through (d) the epiglottis takes up huge part of the images, while a light source causes problems in (e) and (f).

as fixed and register all the subsequent ones $I_1, .., I_n$ onto it. Then we calculate a similarity value $C_{i,T}$ between the image I_0 and a following, transformed image $I_{i,T}, i \in [1, n]$ using

the SAD,

$$C_{i,T} = \frac{1}{N} \sum_x \sum_y \left| \begin{array}{l} R_0(x,y) - R_{i,T}(x,y) \\ G_0(x,y) - G_{i,T}(x,y) \\ B_0(x,y) - B_{i,T}(x,y) \end{array} \right|,$$

where R,G,B represent the color channels of I_0 and $I_{i,T}$, N the number of pixels and x and y the pixel coordinates. N is not the same for every frame because due to rotation or translation, pixels are moved outside of the image space leaving blank spaces. These empty areas are filled with a default value to allow a later identification. To provide a meaningful comparison between different results $C_{i,T}$ we normalize each value according to its area.

The optimal registration is found by applying combinations of rotation and translations to I_i until the minimum of $C_{i,T}$ is found. Since individual differences between frames are small, it is feasible to solve this registration problem globally by such an exhaustive search strategy over the two translation and the single rotation parameter in a limited range. In our calculations we cover a translation space of [-2 px, 2 px] with a step size of 1 px in x- and y-direction and a rotation space of $[-\frac{\pi}{240}, \frac{\pi}{240}]$ with a step size of $\frac{\pi}{720}$. This results in 9 rotations and 5 translations in each direction, leading to a total number of 225 different configurations to be tested.

The high frame rate of the recording allows us to speed up the registration process even further because we can assume that there is no movement between two or even a few consecutive frames, since global drifts occur usually more slowly over the videos. Therefore we only register every tenth image to our fixed image I and apply the transformation of the optimal registration to the four frames before and the five after the registered frame (see Fig. 2.10). Efficient computation of this step is achieved by a parallel NVidia

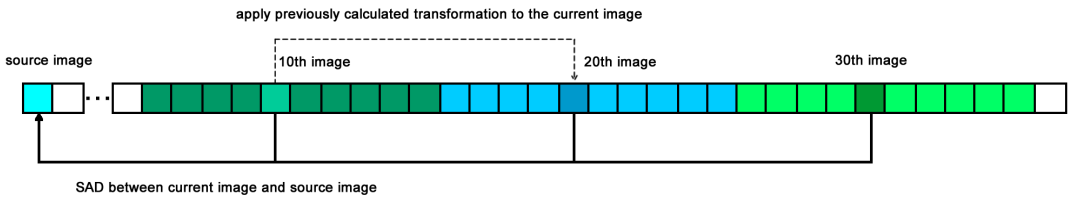


Figure 2.10: Due to the high frame rate, only every 10th frame is registered and the calculated transformation is applied to the images before and after it.

CUDA implementation of the similarity computation on the graphics adapter. For the transformation process we use a bi-cubic interpolation [53] and a prefilter kernel [60] for correct interpolation results. The interpolation is a computationally expensive step and we try to reduce the number of transformation processes to a minimum. In order to reduce the 225 transformations, we only calculate the 9 rotations and perform the translation by simply copying the data of the rotated image into an empty one with an offset according to the translation.

2.3 Salient Region Detection

The salient region detection is an important tool in our method and we use it for two different problems:

- Detection of the region of interest and the bounding box
- Finding the seed regions or constraints for the segmentation

In our case we face a saliency detection problem, where we have to calculate a saliency map from an RGB-image to represent the interesting regions. The basic idea is to find salient areas and show their likelihood to be of importance as intensity value in a mean attention map. Figure 2.11 depicts the mean attention maps of generic image examples, which were taken from [69].

We define an area surrounded by background and not connected to the borders as a salient region. The glottal opening fulfills the requirements for a salient region because it is a dark area surrounded by tissue and usually is in the center of the image. Without the previous denoising step more light artifacts caused by reflecting fluids and blood vessels would be detected as interesting regions as they are also areas surrounded by a homogeneous background (see Sec. 2.2.2).

Recently, saliency detection methods and the different possibilities to use them have received a great amount of attention. Only a few of the important methods will be discussed in this work but for an extensive review of the state-of-the-art see [2, 47]. Most saliency models detect complex or rare salient patches by using center-surround filters or image statistics. In [29] conspicuous regions on multi-scale feature maps are detected applying center-surround difference. Other methods use the negative logarithm of probability [5, 70] or the "Bayesian surprise" [30] to find salient patches by calculating improbabilities. A recent method [19] utilizes a hierarchically whitened feature space, with

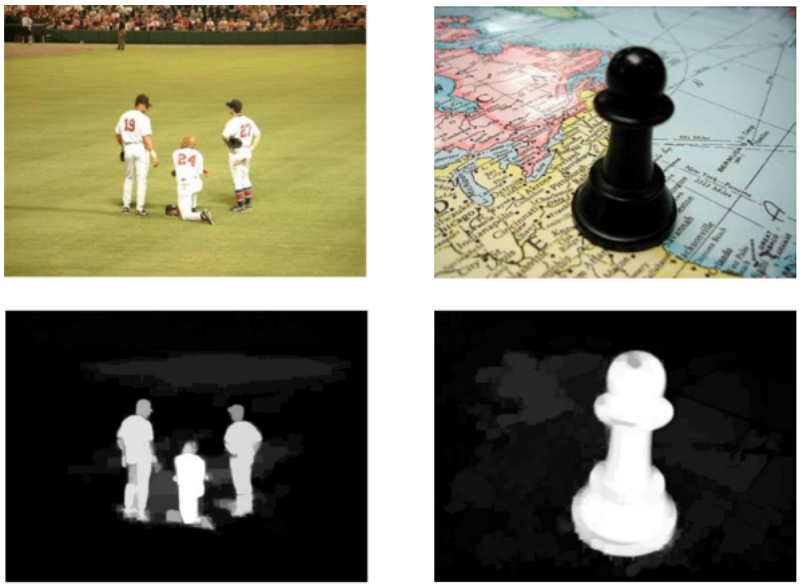


Figure 2.11: With the salient region detection we can find interesting areas and their likelihood to be of importance is shown as intensity value in a mean attention map. More important regions are depicted in white, whereas background is dark-gray or black. Taken from [69].

the square of the vector norms used as saliency metric. There are also models based on spectral domain analysis [27, 39, 51]. In [39] it is shown that a few of these spectral domain methods are equivalent to a local gradient operator and Gaussian blurring and therefore large salient regions can not be detected. They overcome that problem by using spectral scale-space analysis. Another type of model follows a machine learning approach by training a support vector machine (SVM) [33, 36]. A modern approach, called Boolean Map based Saliency (BMS), recently proposed by Zhang and Sclaroff [69] does not use any of the above mentioned techniques but relies on topological structural information, which is known to attract visual attention [11, 64]. This method shows promising results, performs very well compared to other state-of-the-art methods, has great capabilities in salient object detection and is very easy to implement. For the problem of the glottal area detection we slightly modified this approach.

2.3.1 Boolean Map based Saliency (BMS)

A Boolean Map is a spatial representation that partitions a visual scene into two distinct complementary regions [28], where one is the foreground and the other the background. Figure 2.12 shows a typical Boolean Map (right) calculated from a color image (left). As it

can be seen all the color information is lost during the process. Boolean in this case refers to the division into two binary regions. BMS [69] was originally designed for eye-tracking

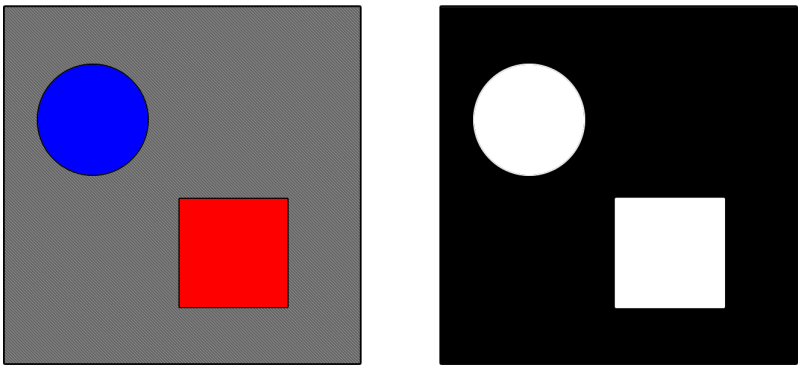


Figure 2.12: The input color image on the left and its Boolean Map on the right, where white represents the foreground and black the background.

and uses the concept of showing an observer's momentary conscious awareness of a scene as a Boolean Map [28]. They assume that Boolean Maps can be generated from randomly selected feature channels, and the influence of a Boolean Map B on visual attention can be represented by an attention map $A(B)$, which highlights regions on B that attract visual attention. The saliency is then represented as mean attention map $\bar{A}$, which can be further processed depending on the task. Figure 2.13 shows that from an image I a set of Boolean Maps $B = B_1, B_2, ..., B_n$ is generated. For each of these maps an attention map

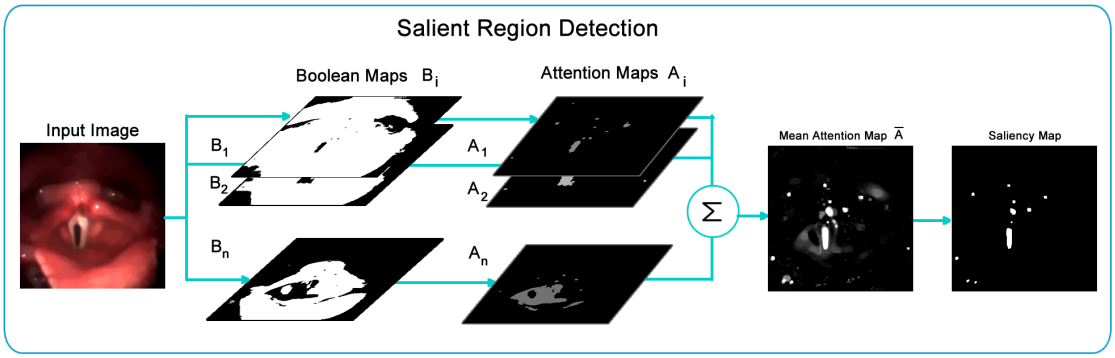


Figure 2.13: Saliency detection based on Boolean Maps, adapted from [69].

A_i is calculated and through linear combination the mean attention map $\bar{A}$ is computed. To generate a Boolean Map we have to randomly threshold the image's feature map $\phi(I)$

at a value θ:

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

The threshold function $\textbf{THRESH}(\phi(I), \theta)$ assigns 0 to every pixel smaller than θ and 1 otherwise. Feature channels are assumed to be in the range between 0 and 255 and can be color, depth, orientation, motion, etc. $\phi \sim p_\phi$ and $\theta \sim p_\theta$ represent the prior distribution of θ and ϕ.

In our case we only use the color information of one frame at a time. When generating Boolean Maps we want salient regions to have a higher chance to be separated from the background. To achieve that we need a uniform distribution of the threshold θ, which is best represented by a color space that reflects the visual differences between colors. We use the *CIE Lab* [18, p. 200] color space due to its perceptual uniformity, where L is lightness, a the position between red and green and b the position between yellow and blue. L is usually in the range between 0 and 100 but a and b do not have a certain range, which is why we limit them to $[-127, +127]$ in our conversion from RGB. For the actual calculation all channels are translated to [0,255] using $L_{new} = 2.55L_{orig}$, $a_{new} = a_{old} + 127$ and $b_{new} = b_{old} + 127$ and assuming that all of them are equally important for visual perception.

In order to generate Boolean Maps we iterate through all the channels and sample at a certain intensity threshold value θ, where 1 is assigned to every value $\geq \theta$ and 0 to all the others. After each iteration θ is increased by a fixed step size δ_S, which can be selected by the user. The final saliency map can vary depending on step size, which is caused by the structures in the image and the color distribution. For our evaluation we used $\delta_S = 3$ and $\delta_S = 8$. A higher δ_S means faster computation because not so many Boolean Maps must be calculated.

From the Boolean Map B an attention map $A(B)$ is computed using a Gestalt principle for figure-ground segregation, which states that surrounded regions are more likely to be perceived as figures [45]. This principle was also used in the salient region definition. A surrounded region in a Boolean Map has a closed contour, which means that all areas connected to the borders are not surrounded. Therefore, the holes of the Boolean Map, which represent the salient parts, are 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.

To additionally emphasize the salient regions we dilate [21, p. 655] them. The resulting attention maps have to be normalized to give small concentrated areas more emphasis. For the normalization we use the Frobenius norm [20, p. 55] instead of the proposed L2-normalization (largest singular value) because of the easier and faster calculation. The Frobenius norm of an image with height h and width w is defined as

$$||A||_F = \sqrt{\sum_{x=0}^{w-1}\sum_{y=0}^{h-1}(a_{xy})^2}.$$

The normalized attention maps A_i are then summed up to a mean attention map $\bar{A}$ and a threshold operation with $\frac{1}{\delta_S}$ filters out weak signals. After that the map is differently processed for the ROI and seed region detection.

2.4 Bounding Box and Region of Interest Detection

The glottis area only takes up around 25 % of the whole image and therefore most of the image is not interesting for the analysis and segmentation. To guarantee a fast calculation most of the data should be stored and processed on the graphics adapters, with memory capacity being a limiting factor. Finding a region of interest (ROI) and a surrounding bounding box therefore is a very important task to reduce both, the computation time and the memory requirements. This step was also incorporated in previous works using different approaches like searching for regions with certain properties [14, 34] or calculating the region with the maximum movement [37].

We use a novel method for this task, which is depicted in Figure 2.14. Throughout the video the glottal area varies resembling an opening and closing cycle (see Sec.1.1). The frame with the largest opening is the best way to detect the ROI and derive the bounding box from it. Due to the glottal opening being a dark area surrounded by tissue it can be assumed that the sum of pixel intensities of the gray-scale image is lowest when the opening is large and highest when there is none. In [34] they use the term landmark frame (LF) for the frames with the maximum opening, which we will use in this thesis in a similar fashion. The only difference is that we just have one LF because due to the image registration we do not have to calculate a new one for each cycle. After finding the LF we convert it to a gray-scale image using the formula for the luminance signal from

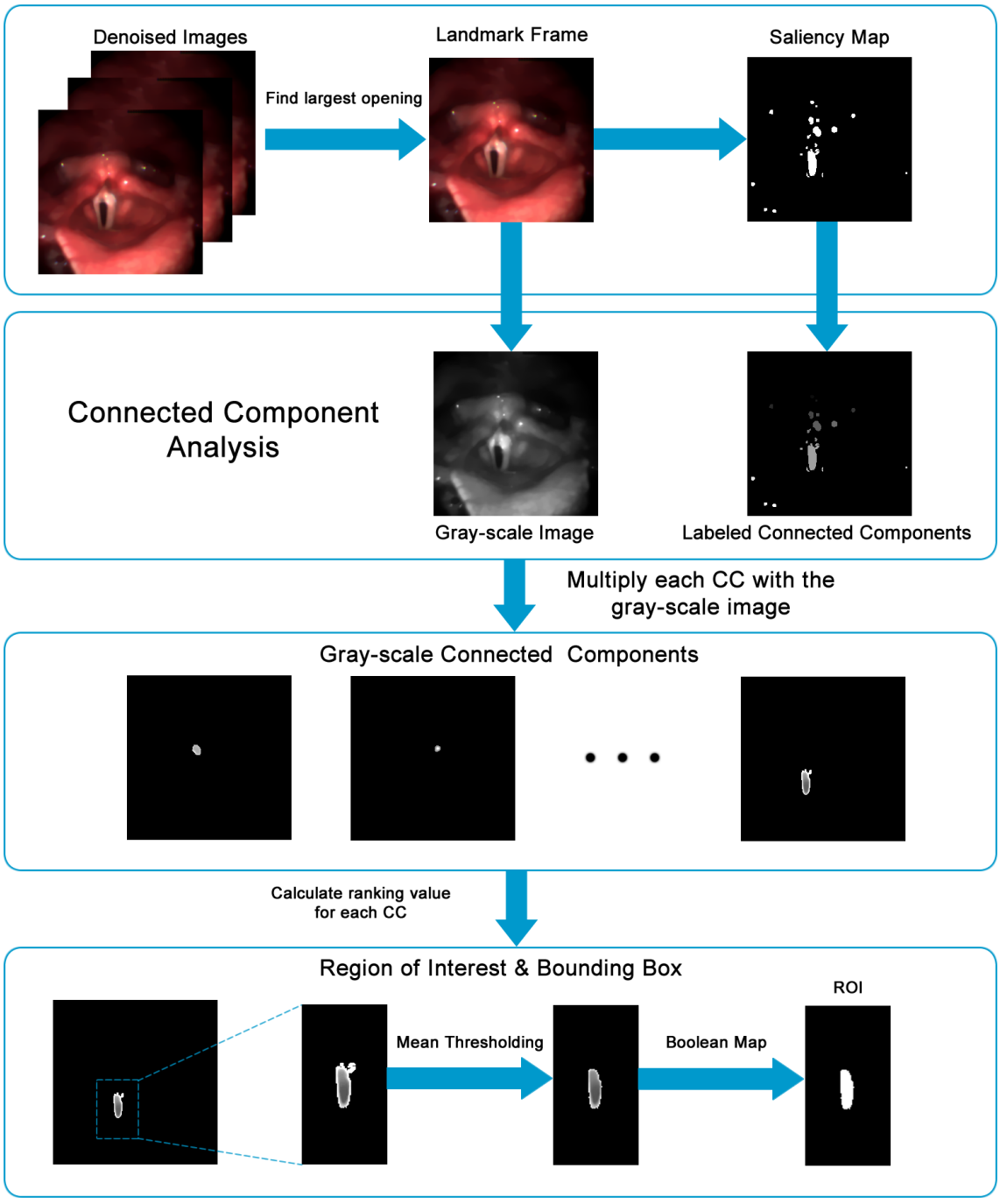


Figure 2.14: Processing pipeline to find the region of interest and bounding box for the largest glottal opening. Images were contrast-enhanced for visualization purposes.

the *ITU-R BT.709* standard for HDTV* leading to

$$I_{gray} = 0.2126 \cdot I_{red} + 0.7152 \cdot I_{green} + 0.0722 \cdot I_{blue},$$

with I_{red}, I_{green} and I_{blue} describing the different color channels.

We also computed a saliency map from the LF and apply a connected component analysis [50]. This method usually gives us many different, potential glottal opening areas and choosing the right one is not an easy task. Requirements for the ROI are that it is a large and dark area and we want the size to be the dominant factor. By multiplying our potential ROIs with the previously calculated gray-scale image we get areas with a certain size and intensity. From these regions we calculate a ranking value for all the potential areas and afterwards choose the highest ranked as ROI. Our ranking value x_i has to get bigger for larger and darker areas and we want size to have a greater influence. Therefore, x_i is defined as

$$x_i = \frac{A_i^2}{|A_i|},$$

where A_i^2 is the squared size to give more emphasis to small concentrated areas and $|A_i|$ is the sum of all the intensity values within the area. $|A_i|$ can never be 0 because a connected component can not have zero size. First, the smallest bounding box BB_1 fully enclosing the ROI is calculated and then a larger one BB_2 is determined by using a fixed offset to the left, right, bottom and top of the center point (see Fig. 2.15). In our application we use an offset of 75 pixels to the top and bottom and 40 pixels to the left and right side leaving us with dimensions of 150 x 80 for the final bounding box BB_2. After the registration process only the information within this bounding box is stored on the graphics adapter to reduce the necessary memory and computation time. The ROI is very important to calculate the seed regions for the final segmentation step.

2.5 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 master's thesis. All the previous steps were necessary to optimally prepare the LHSVs for this most important part of our

*ITU - International Telecommunication Union, Parameter values for the HDTV standards for production and international programme exchange, 2002, http://www.itu.int/dms_pubrec/itu-r/rec/bt/R-REC-BT.709-5-200204-I!!PDF-E.pdf, Accessed: 2014-08-29.

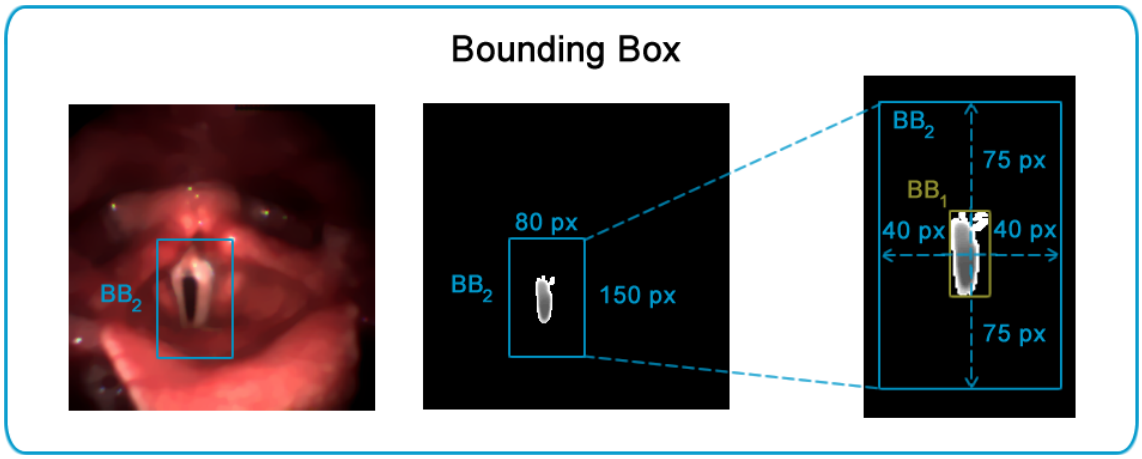


Figure 2.15: First, a small bounding box BB_1 is found for the ROI and then a bigger one BB_2 is computed from center point using a fixed offset.

image processing pipeline. As depicted in Figure 2.16 we calculate the seed regions and edge information from the preprocessed images and use them as initialization. These seed regions then evolve towards the edges to yield the final segmentation.

2.5.1 Seed Region Detection

For the algorithm to know where to start the segmentation process we need initial seed points or rather constraints. Finding this first initialization was also studied in the works presented in Section 1.3. The easiest method was [41], where the user simply specifies seed points. In [34] and [14] they go through the images looking for certain qualities like large, nearly vertical oriented areas or dark, elliptical regions. Another way was presented in [37], where they use the TV-norm to detect the most active regions, which are then used as ROI.

We present a novel way of detecting the glottis by utilizing the previously described salient region detection model (see Sec. 2.3). The seed region detection is performed on each image and is very similar to the extraction of the ROI discussed in Section 2.4 but faster and simpler. After calculating the mean attention map we use a threshold filter with $\frac{1}{\delta_S}$, where δ_S is the saliency step size, to filter out weak signals. Then we only extract the information inside the previously calculated bounding box (2.4) from the image. In the next step we multiply with the ROI to remove all unwanted salient regions in the image, which avoids the time consuming connected component analysis and ranking used in the ROI calculation. The resulting area is still too large to be used for the

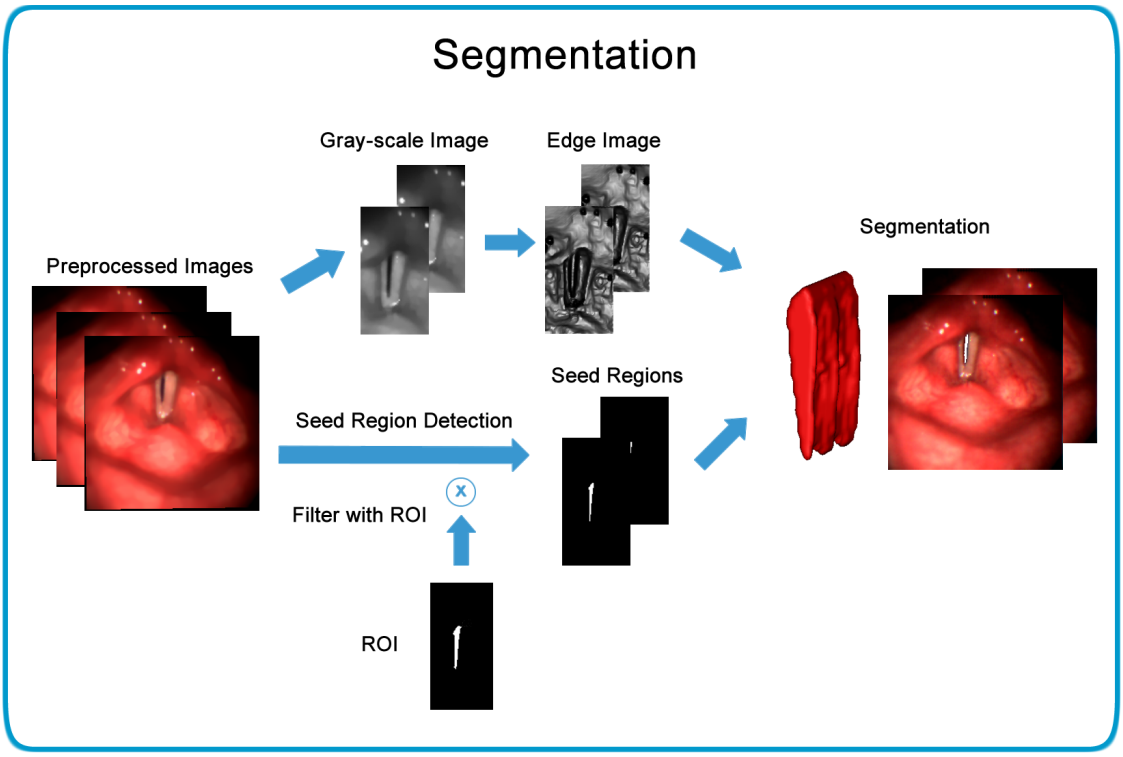


Figure 2.16: First, a bounding box is cut out from the preprocessed images and the edges are computed from the gray-scale image. The seed regions are determined using a salient region detection and the previously calculated ROI.

segmentation, so we use a thresholding operation to set all pixels above a certain intensity to zero. This threshold value is the mean intensity gray-scale value of the previous seed region. It is possible that the initialization is still equal or a little bit larger than the actual segmentation and we would directly start at the outer edges of the glottal area, which can result in a wrong segmentation. We overcome this problem by performing an erosion [21, p. 525] to reduce the initial segmentation. If the initialization is very small, this process sometimes eliminates the complete seed region of this frame but thanks to the 3D segmentation we have the information of the previous and next frame to work with.

2.5.2 3D Geodesic Active Contours

For segmentation we apply the 3D Geodesic Active Contour method. The original Snakes or Active Contour model, which deforms an initial contour C_0 towards the boundaries of an object, was developed by Kass et al. [35]. This deformation is performed by minimizing

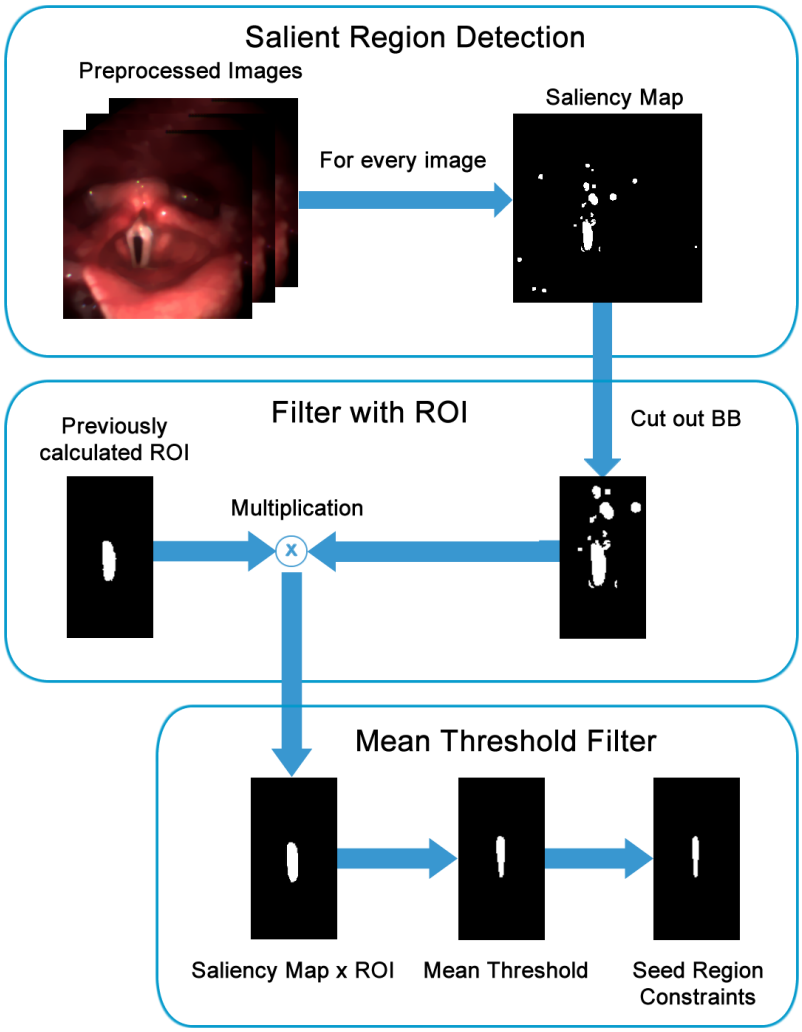


Figure 2.17: The seed region detection is divided into three major blocks: the salient region detection, the filter step with the ROI to get rid of unwanted areas and the final mean threshold filter.

a functional, whose (local) minimum is at the boundaries of the object. For a parametrized planar curve $C(q) : [0, 1] \rightarrow \mathbf{R^2}$ and an image $I : [0, a] \times [0, b] \rightarrow \mathbf{R}^+$, in which we want to detect the object's boundaries, the classical approach is given by

$$E(C) = \alpha \int_0^1 |C^{'}(q)|^2 dq + \beta \int_0^1 |C^{''}(q)|^2 dq - \lambda \int_0^1 |\nabla I(C(q))| dq, \tag{2.3}$$

with α, β and λ being real positive constants. The first two terms represent the internal energy and control the smoothness of the contour, whereas the third term is the external energy, which is responsible for attracting the contour towards the object. To solve the

active contour problem a curve C must be found that minimizes E with respect to the given constants α, β and λ. Although this method has been used for many different applications [42] [66] [32] especially in medical image segmentation, it also has certain limitations. The main limitation is that it is not possible to detect multiple objects in an image because the topology of the initial curve will be the same as the one of the final curve and therefore this method is not able to deal with changes in topology. Another problem is the parameter estimation to determine a trade-off between proximity to the object and smoothness.

To overcome some of the limitations of the classical Snakes model, Caselles et al. [7] introduced the Geodesic Active Contour (GAC) model. The following derivation was established with the help of [23, 57]. In order to derive the GAC model Caselles et al. only considered a certain instance of the original model by setting $\beta = 0$, which leads to

$$E(C) = \alpha \int_0^1 |C^{'}(q)|^2 dq - \lambda \int_0^1 |\nabla I(C(q))| dq. \tag{2.4}$$

This simplification is possible because the regularization effect on the *GACs* comes from curvature based curve flows obtained only from these two terms. To minimize 2.4 the value of $|\nabla I|$ has to be maximized. $-|\nabla I|$ reaches the lowest value at the edges, so it acts like an edge detector and stops at the minimum point of the curve. We can then define an edge detector function, which is high for flat areas and low for strong edges. In mathematical terms this means a strictly decreasing edge detection function $g : [0, +\infty) \rightarrow \mathbb{R}^+$ in a way that $lim_{r\rightarrow\infty} g(r) = 0$. Therefore, $-|\nabla I(C(q))|$ can be replaced by the edge detection function $g(|\nabla I(C(q))|)^2$ leading to

$$E(C) = \alpha \int_0^1 |C^{'}(q)|^2 dq + \lambda \int_0^1 g(|\nabla I(C(q))|)^2 dq. \tag{2.5}$$

After a few calculations using the metric $g_{ij} = 4\alpha\lambda g(|\nabla I(C)|)^2 \delta_{ij}$ and $E_0 = E_{int} - E_{ext} = 0$ we get the length L_R as

$$L_R := \int_0^1 g(|\nabla I(C(q))|)|C'(q)| dq.$$

For details please refer to [7]. The minimization problem (2.4) has been transformed into a geodesic computation in a Riemannian space. From the classical Euclidean length

$L(C) := \oint |C'(q)|dq = \oint ds$ we get $|C'(q)|dq = ds$ and obtain a new definition of the length

$$L_R := \int_0^{L(C)} g(|\nabla I(C(q))|)ds,$$

which weighs the Euclidean length element ds by $g(|\nabla I(C(q))|)$ and therefore takes edge information into account. This leads to the GAC model, which is defined as the following variational problem:

$$\min_C \left\{ E_{GAC} := \int_0^{L(C)} g(|\nabla I(C(q))|)dq \right\}. \tag{2.6}$$

The minimization of 2.6 is equal to finding a geodesic curve in a Riemannian space. Using a steepest-descent method for minimizing, we end up with the curve evolution equation

$$\frac{\partial C(t)}{\partial t} = g(I)\kappa\overrightarrow{N} - (\nabla g(I) \cdot \overrightarrow{N})\overrightarrow{N},$$

where κ is the Euclidean curvature and $\overrightarrow{N}$ the unit inward normal. This is equal to the following level-set representation with $c = 0$ (no constant velocity):

$$\frac{\partial u}{\partial t} = g(c + \kappa)|\nabla u| + \nabla u \cdot \nabla g,$$

One way to globally minimize 2.6 are graph based approaches like the one proposed by Boykov et al. [3], which approximates the Euclidean length element by partitioning a graph based on the image. Bresson et al. [4] introduced a different approach using the weighted Total Variation, which is defined as

$$TV_{g(u)} = \int_\Omega g(x)|\nabla u|d\Omega. \tag{2.7}$$

For u being a characteristic function 1_C, Bresson et al. showed that E_{GAC} (2.6) is equivalent to $TV_{g(u)}$ (2.7). The characteristic function 1_C is a closed set in the image domain Ω and C represents its boundaries. By allowing u to vary continuously between $[0, 1]$, we get a convex functional for (2.7) and therefore we are able to compute a global minimizer. Looking at the definition of the weighted TV model (2.7) we can see that a global minimizer is given by the trivial solution $C = 0$ (a point), thus we need additional constraints to get meaningful results. Although this only approximates the original GAC energy, the choice of a suitable threshold in $[0, 1]$ is not critical in practice. A threshold value of

0.5 was used in our implementation. In [4] they used a TV-L^1 data fidelity term, which resulted in the following minimization problem:

$$\min_C \left\{ E_B := \int_\Omega g(x)|\nabla 1_C| d\Omega + \lambda \int_\Omega |1_C - f| d\Omega \right\},$$

where u is approximated by a binary function 1_C. A similar approach was used by Leung and Osher to unify denoising, segmentation and inpainting, where they used the *weighted TV* (2.7) with a spatially varying L^1 data fidelity term [38] and the continuous formulation of u instead of 1_C giving

$$\min_C \left\{ E_L := \int_\Omega g(x)|\nabla u| d\Omega + \int_\Omega \lambda(x)|u - f| d\Omega \right\}. \tag{2.8}$$

Including various local constraints was also described in [58], showing promising results for different medical gray-scale datasets. We use a Geodesic Active Contour segmentation method similar to Unger et al. [59], but we replace the data term with the one proposed in [46]. The variational image segmentation model is then defined as

$$\min_{u \in [0,1]} \left\{ E_{Seg} := \int_\Omega g(x)|\nabla u| d\Omega + \lambda \int_\Omega u f d\Omega \right\}. \tag{2.9}$$

Predefined constraints can be included into the energy functional because f is provided by user, where $f = -\infty$ indicates foreground and $f = +\infty$ background. To minimize the second term of 2.11 u tends to 0 for the background ($f = +\infty$) and 1 for the foreground ($f = -\infty$). Note that in our implementation we use constants to indicate hard fore- and background, which are then translated to the correct u values as mentioned above. Instead of letting the user choose constraints we calculate seed regions (see Sec. 2.5.1) using the salient region detection described in Section 2.3. These seed regions are then evolved iteratively towards the edges of the image.
λ is a trade-off between the constraints and the contour. For our algorithm we used a λ value of 0.01.

The edge detector function $g(x)$ in 2.11 has not been discussed yet, but is an important part of the segmentation model. In their work [7], Caselles et al. suggested an edge detector function

$$g = \frac{1}{1 + |\nabla \hat{I}|^p},$$

where $\hat{I}$ is a smoothed version of I and $p = 1$ or 2 for the GAC model in 2.6. The authors used Gaussian filtering to compute $\hat{I}$ and stated that other filters and decreasing functions of the gradient are possible as well. Unger et al. [59] proposed an edge detector

$$g(I) = e^{-\alpha|\nabla I|^{\beta}}, \tag{2.10}$$

using $\alpha = 10$ and $\beta = 0.55$. The authors also denoise the image I using the *ROF* model [49] (see Section 2.2.2) before the actual edge computation, to ensure that noise does not influence the edge image and the GAC energy is only computed on significant edges (see Fig. 2.5).
In our work, we use ROF denoising already in our preprocessing phase and then compute the edges (2.10) using $\alpha = 15$ and $\beta = 0.55$ from the denoised gray-scale volume. For the conversion we use

$$V_{gray} = 0.2126 \cdot V_{red} + 0.7152 \cdot V_{green} + 0.0722 \cdot V_{blue},$$

with V_{red}, V_{green} and V_{blue} representing the different color volumes. We also evaluated the edge detector 2.10 on a full color volume, which did not give clear edges. The algorithm is directly used on the 3D spatio-temporal volume and leads to a smooth minimal surface segmentation solution. As the numerical optimization scheme we use the primal-dual algorithm from Chambolle and Pock [9], which is extensively explained in [57]. Efficient calculation is achieved by using the graphics adapter and an NVidia Cuda implementation.

2.6 Manual Segmentation Refinement

We developed a fully automatic method to segment the glottis, but there is always the possibility that the segmentation is not accurate enough. Therefore, one of the goals of this thesis is to provide the user with a tool to correct the segmentation if it should be necessary (see Sec. 1.1). Generally there are two different problems that can arise during the segmentation process:

1. The ROI and bounding box is not detected correctly, which results in a wrong segmentation.

2. The seed regions are not correct, which can be caused by the ROI being too large or miss detections due to artifacts or dark areas.

The first problem can be solved by using a configuration file *blocks.ini* (see Appendix C), where the user can specify the landmark frame and bounding box for a certain part of the recording. Our algorithm then only searches for the ROI inside this predefined bounding box.

For the second problem we use an interactive segmentation tool developed by Urschler et al. [61]. It uses the same GAC segmentation model as we do, which is defined as

$$\min_{u\in[0,1]} \left\{ E_{Seg} := \int_\Omega g(x)|\nabla u| d\Omega + \lambda \int_\Omega u f d\Omega \right\}, \tag{2.11}$$

where f contains the constraints or seed regions. With this software it is possible to load the seed regions calculated by our method and refine them. Completely new ones can also be added if the topology of the glottis is special (e.g. multiple openings). Figure 2.18 shows the typical scenario where the constraints in the left image are not enough for a complete segmentation whereas an additional region leads to a much better result.

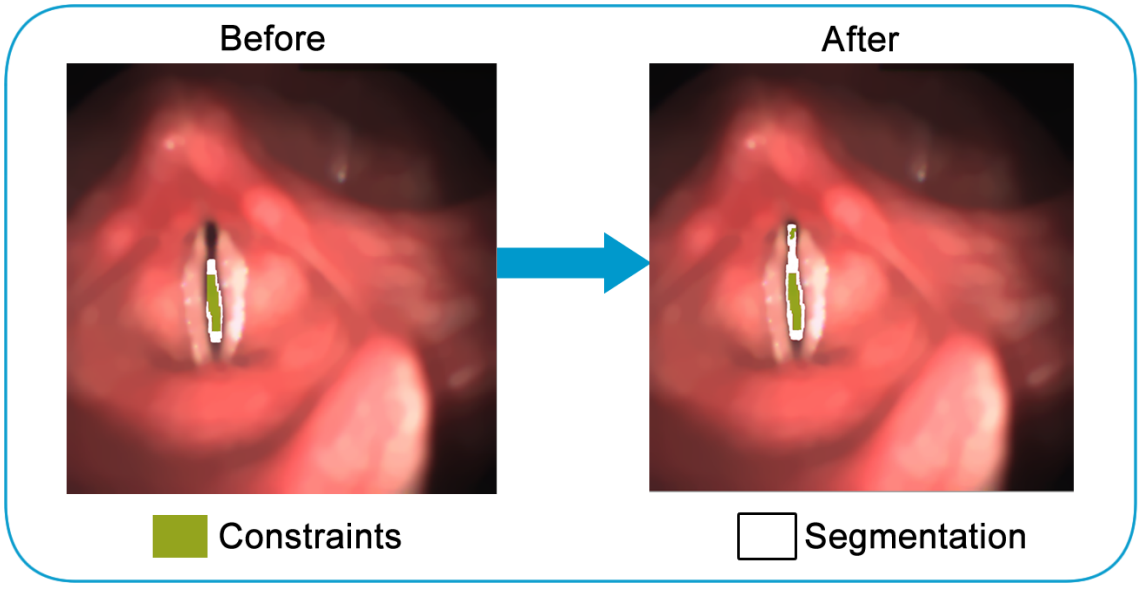


Figure 2.18: On the left side the seed region (green) is not enough for a correct segmentation (red), whereas on the right side an additional region was added with the segmentation tool [61] leading to a much better result.

The tool also has denoising and segmentation capabilities, which allows the user to easily try different parameters for edge calculation, denoising and segmentation.

2.7 Summary

In this section we described the different steps and algorithms used in our work. We start with a preprocessing phase to deal with common problems like non-homogeneous background, light and fluid artifacts, and the global drift in order to prepare the video sequence for segmentation. This phase incorporates contrast stretching, ROF denoising and intensity-based rigid registration. For finding a ROI and seed regions, which are then used as initialization for the segmentation, we use a salient region detection. The final step gives us a segmentation of the 3D spatio-temporal volume of the glottal area using 3D Geodesic Active Contours.

Chapter 3

Experiments and Results

Contents

3.1 Implementation Details

A fast and efficient implementation was very important in order to use the proposed automatic glottis segmentation method in everyday clinical applications. For that purpose we decided to use C++ because it is a very flexible and efficient language. The main speed up in our algorithm comes from the use of parallelized computations on the NVidia graphics adapters. CUDA is a platform developed by NVidia and offers the basic methods to access the GPU. Another abstraction level was introduced by a special framework developed at ICG and LBI, which provides implementations of common image processing algorithms and easily accessible data structures. For reading and writing image files to and from the hard disk we use the QT4-framework, a cross-platform application framework. Due to the memory limitations of the graphics adapter not all the 8192 images of one video can be processed at once and therefore we divide the whole recording into blocks (see Fig. 3.1). Each block is treated independently and has a predefined size BS, which can be configured according to the memory capacity of the graphics adapter. There is a certain overlap OV between the single blocks to prevent errors at the borders during

denoising and segmentation. Even though each block is processed completely, only part of it is actually taken as output. In the first block $BS - \frac{1}{2}OV$ and in all the others $BS - OV$ images are used.

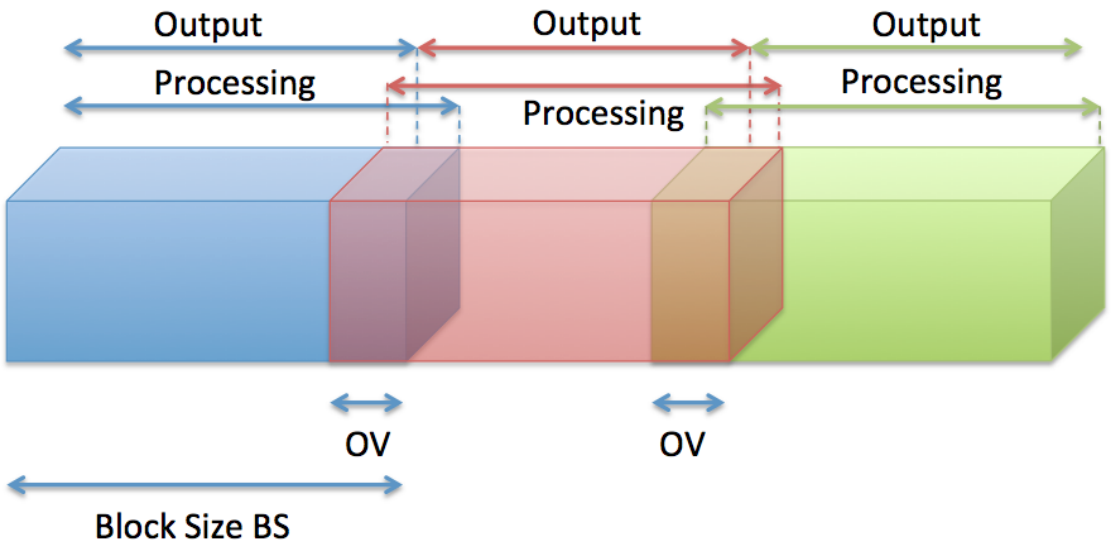


Figure 3.1: Due to memory limitations on the graphics adapater the recording is split into blocks, which are then processed separately.

3.1.1 Configuration

In our method there are a lot of different values to be set and in this section we will state the standard settings we used for the experiments. For every recording a different setup leads to the best results but we chose values best suited for a wide range of videos. These values are fully configurable by the user using two INI files, which will be explained in detail in Appendix C. Basically there is a configuration file for the method itself (*config.ini*) and one for later user-intervention like bounding box or LF specification (*blocks.ini*). Table 3.1.1 shows the standard setup we used for our experiments.

3.2 Evaluation

To evaluate our method we compared our results to an implementation of the clinical standard [41] on a set of ground truth (GT) images. On all the frames an experienced computer vision researcher has performed a manual segmentation of the glottal opening,

Table 3.1: This table shows the standard configuration values, which were used for the evaluation.

Standard configuration values	
Bounding Box	
Width	80 px
Height	150 px
Block settings	
Block size BS	1500 frames
Block overlap OV	10 frames
Denoising settings	
Trade-off factor λ	25
Maximum Iterations	10000
Registration settings	
Angle range	$\frac{\pi}{240}$ (translates to $[-\frac{\pi}{240}, \frac{\pi}{240}]$)
Angle step size	$\frac{\pi}{720}$
Translation range	2 px (translates to [-2 px, 2 px])
Translation step size	1 px
Default fill value	280
Register every ... frame	10th
Salient region detection	
Step size δ_S	3 and 8
Segmentation	
α for edge computation	15
β for edge computation	0.55
Trade-off factor λ	0.01
Maximum Iterations	10000
Landmark frame detection	
Images to check	150 frames

which was investigated and corrected by an expert in kymography and voice disorders. In order not to influence the annotation, it was performed without knowing, which videos were from sick or healthy patients. We decided to evaluate on three video sequences with many consecutive, annotated frames including one or more opening and closing cycles as well as on single frames. For that purpose we designed four different experiments with their own data sets:

- **Experiment 1**: 63 frame video sequence, where only every second frame was annotated.
- **Experiment 2**: 61 frame video sequence.
- **Experiment 3**: 33 frame video sequence.

- **Experiment 4**: 30 randomly picked frames.

Experiment 4 was especially designed to give a broad overview of the segmentation quality, but 5 out of the 30 frames caused major problems, so we excluded them from the evaluation. These problems will be discussed in Section 3.4 and might be addressed in future work. To show the benefits of the 3D segmentation we decided to also segment ± 100 frames around the chosen sequence or single frame.

We compare our proposed method and the reference implementation of the clinical standard [41] using the following measurement methods.

Dice coefficient

To evaluate our method an overlap measure between the ground truth and our results has to be calculated. For that purpose we use the Dice Coefficient (DC) (sometimes also called Sørensen-Dice index) [15, 52], which is a statistical value for comparing the similarity of two samples by calculating an overlap between 0 and 1, which is then converted into a percentage. It is defined as

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

with A the ground truth and B the proposed segmentation. Even though the DC is a well established way to compare two segmentations, it has a limitation, which is especially crucial in our case. Due to the opening and closing cycles of the glottis, the segmentation area varies over time and can get very small or even disappear. If the area is small, even a tiny difference between the two segmentations has a great impact on the overlap measure, even though the absolute error is small. Figure. 3.2 illustrates this, since on the left the absolute segmentation error is 10 pixels and the DC still gives 94 % compared to the right frames with an absolute error of 1 pixel and a DC of 85 %.

Over- and Undersegmentation

A common way to compare a segmentation to a ground truth is the calculation of the Sensitivity (also Recall), the Specificity and the Precision, which are defined as:

$$Sensitivity = \frac{TP}{TP + FN}$$

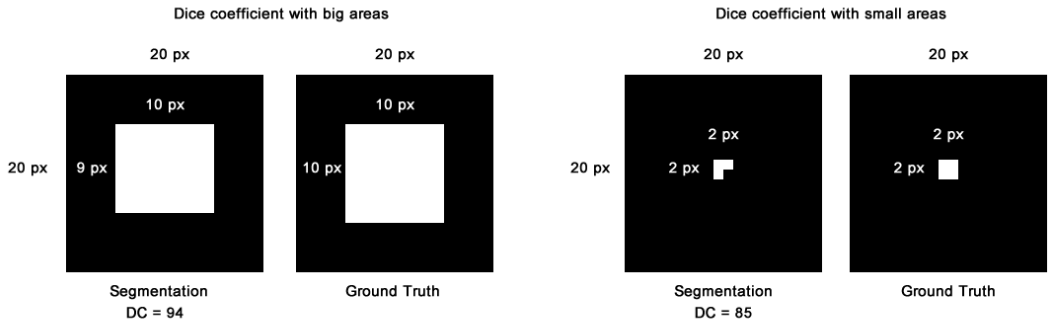


Figure 3.2: This image shows the problems with the DC and small areas. Even though the absolute error in the left segmentation is 10 times higher than the one in the right picture, the DC is better for the bigger area on the left.

$$Specificity = \frac{TN}{TN + FP}$$

$$Precision = \frac{TP}{TP + FP}$$

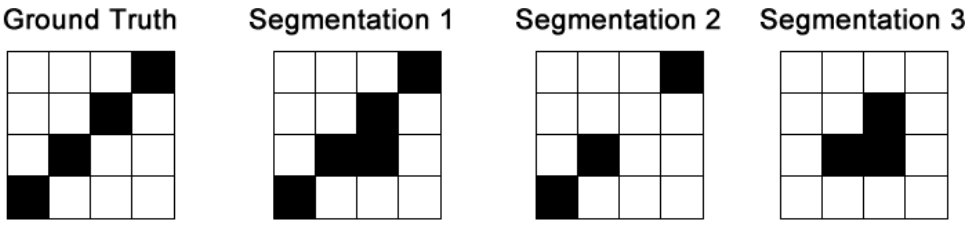
TP = number of true positives, pixels segmented in the GT and by the algorithm.
FP = number of false positives, pixels segmented by the algorithm but not in the GT.
TN = number of true negatives, pixels not segmented by the algorithm and the GT.
FN = number of false negatives, pixels not segmented by the algorithm but in the GT.

The Sensitivity tells us how much of the GT we actually cover with our segmentation whereas the Specificity does the same for the background. A problem here is that an image, where every pixel is part of the segmentation, would give an optimal score for Sensitivity, thus we can not use this value alone. With the Specificity we have the same problem when nothing is segmented. Another obstacle is that the background area is very large compared to the actual foreground and errors of a few pixels are too small to have any effect. By computing the harmonic mean of the Precision and Sensitivity we get the F1 score:

$$F1 = \frac{2TP}{2TP + FP + FN}$$

In [44] it was shown that the DC is equal to F1 score and therefore it is not really useful for our evaluation.

As representation of over- and undersegmentation we will use the absolute values FP and FN. The absolute segmentation error ϵ_{abs} is given as the sum of these two.



	Ground Truth	Segmentation 1	Segmentation 2	Segmentation 3
Area:	4	5	3	3
Dice-Index:	1	$\frac{8}{9}$	$\frac{6}{7}$	$\frac{4}{7}$
FP:	0	1	0	1
FN:	0	0	1	2
Error:	0	1	1	3

Figure 3.3: The ground truth, three different segmentations and the scores we use for the evaluation of the results.

Segmented Area

The most intuitive way of comparing segmentations is a simple comparison of the segmentation area, which gives us a basic idea if the results can be correct. A limitation here is that we do not get any idea if the segmentations are in the same place in the image and that is why we have to take the FN and FP values into account.

3.3 Results

In this section we present the results of our automatic glottis segmentation method and compare them to a previously annotated set of ground truth data. Further, we compare our method to the clinical standard [41] in order to show an improvement to a traditional approach. All the calculations were performed on a standard Linux computer with a GeForce GTX 580, Intel Core i7 (8 cores/3.4 GHz), 8 GB of memory and Ubuntu 14.04 LTS.

For our method we used two different step sizes $\delta_S = 3$ and $\delta_S = 8$ for the salient region detection to show the differences in the outcome. The presented results were created completely automatically without any user corrections or refinement. Figure 3.5 shows

one qualitative results for Experiment 1, 2, 3, while all the single frames of Experiment 4 are depicted in Figures 3.5, 3.6, 3.7 and 3.8.

We use four different values to measure the quality of the segmentation - segmented area, Dice Coefficient (DC), number of false positives (FP) and number of false negatives (FN). A higher DC and lower FP and FN values indicate a better segmentation. The mean $\bar{x}$ and median $\tilde{x}$ values of the four experiments are shown in Table 3.3, with the best scores for the DC highlighted. For Experiment 4 the results for the 25 single frames are shown in Table 3.3 and 3.4. Figure 3.4 shows a separate box-whisker plot of the DC for all four experiments with the proposed method in turquoise ($\delta_S = 3$) and orange ($\delta_S = 8$), and the SRG-based clinical standard in red. Each box has the median as central mark (red line) and its edges are the 25th and 75th percentiles. The whiskers represent the most extreme data points not considered outliers, which are plotted separately (red cross). If the intervals of two data sets described by their notches do not overlap, the medians are significantly different at the 5 % significance level.

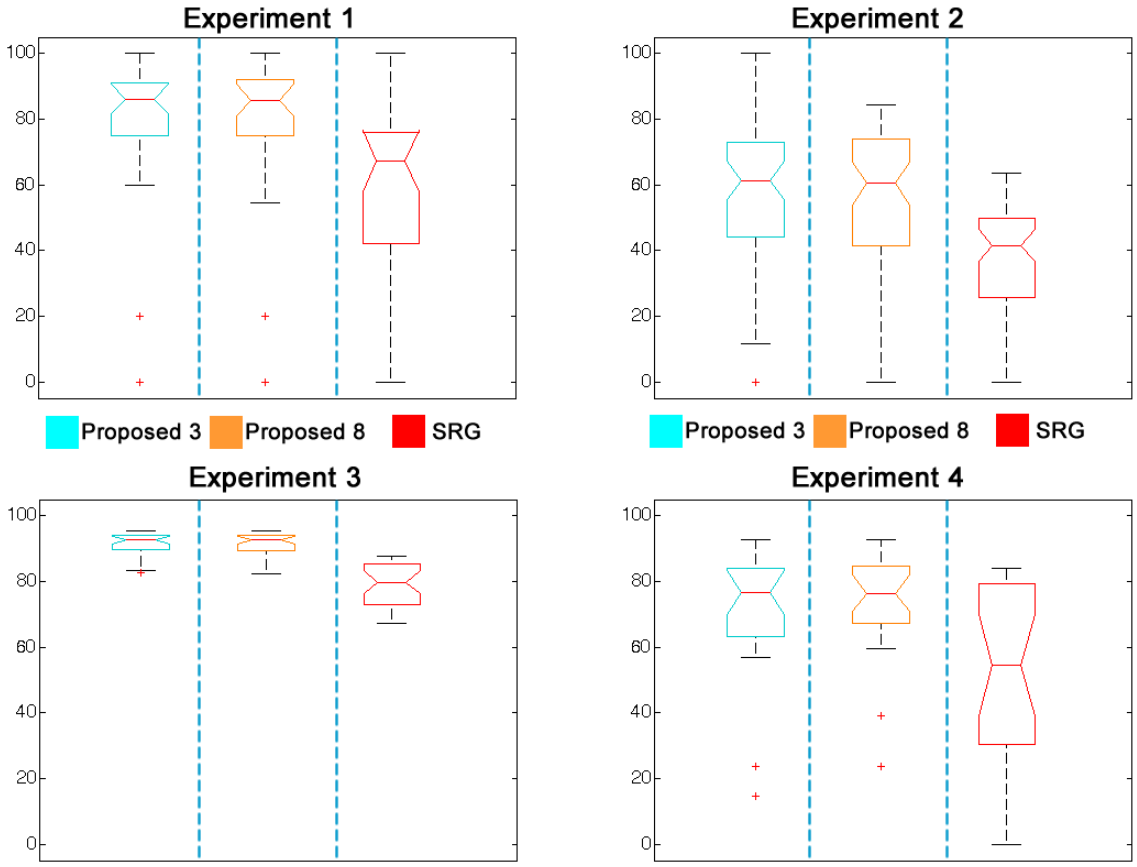


Figure 3.4: This box-whisker plot shows the DC results of the four different experiments with the proposed method in turquoise for $\delta_S = 3$ and orange for $\delta_S = 8$, and the SRG [41] version in red.

Table 3.2: The mean $\bar{x}$ and median $\tilde{x}$ values of the segmentation evaluation of the three video sequences (Experiment 1,2,3) and the randomly picked frames (Experiment 4). We evaluated the Dice Coefficient, false positives (FP), false negatives (FN) and the segmented area.

	DC [%]			FP [px]			FN [px]			Area [px]			
#	δ_3	δ_8	SRG	δ_3	δ_8	SRG	δ_3	δ_8	SRG	GT	δ_3	δ_8	SRG
Experiment 1 - 63 frame video sequence, every second frame annotated													
$\bar{x}$	73.4	**75.5**	56.6	**29.7**	30.3	41.2	5.8	**5.1**	28.7	115.0	138.9	140.2	127.5
$\tilde{x}$	**86.1**	85.9	67.4	**18.5**	21.0	49.0	2.5	**2.0**	28.5	106.0	151.0	154.5	123.5
Experiment 2 - 61 frame video sequence													
$\bar{x}$	**53.7**	52.7	35.5	**10.4**	10.6	72.4	44.8	**43.9**	51.2	92.7	58.3	59.4	113.9
$\tilde{x}$	**61.3**	60.5	41.6	**4.0**	**4.0**	46.0	40.0	**39.0**	45.0	93.0	51.0	55.0	85.0
Experiment 3 - 33 frame video sequence													
$\bar{x}$	**91.5**	91.3	78.8	**5.3**	**5.3**	62.2	**44.0**	44.7	64.7	332.4	293.7	293.0	329.8
$\tilde{x}$	**92.7**	92.6	79.8	**5.0**	**5.0**	62.0	**45.0**	**45.0**	65.0	311.0	288.0	288.0	315.0
Experiment 4 - 30 randomly picked frames													
$\bar{x}$	71.8	**73.9**	53.0	97.8	**45.2**	101.1	45.8	**43.9**	81.5	196.5	248.5	197.7	216.1
$\tilde{x}$	**76.7**	76.4	54.5	**10.0**	14.0	35.0	34.0	**32.0**	62.0	161.0	124.0	138.0	145.0

Table 3.3: The first part of the evaluation results of the randomly picked frames (Experiment 4). We evaluated the Dice Coefficient, false positives (FP), false negatives (FN) and the segmented area.

	DC [%]			FP [px]			FN [px]			Area [px]			
#	δ_3	δ_8	SRG	δ_3	δ_8	SRG	δ_3	δ_8	SRG	GT	δ_3	δ_8	SRG
1	**82.5**	75.7	22.1	3	25	**0**	34	**32**	106	121	90	114	15
2	67.1	**68.7**	28.9	1	**0**	526	96	93	**73**	195	100	102	648
3	81.3	**81.7**	19.0	**13**	14	1039	94	**91**	183	326	245	249	1182
4	73.3	**74.2**	47.8	**21**	33	27	49	**40**	91	145	117	138	81
5	**88.5**	**88.5**	79.0	**7**	**7**	59	**49**	**49**	53	264	222	222	270
6	73.7	73.7	**83.4**	**0**	**0**	50	159	159	**73**	382	223	223	359
7	**86.1**	84.5	80.5	**3**	**3**	33	36	40	**29**	157	124	120	161
8	82.5	**84.5**	66.7	**6**	**6**	11	14	**12**	25	61	53	55	47
9	**92.9**	92.8	84.2	**1**	**1**	108	62	63	**51**	474	413	412	531
10	80.9	**81.5**	55.9	47	47	**4**	24	**22**	105	174	197	199	73
11	71.6	**76.4**	69.1	72	73	**26**	99	**75**	135	315	288	313	206
12	**63.7**	**63.7**	50.4	**0**	**0**	11	**89**	**89**	107	167	78	78	71
13	**83.8**	**83.8**	30.6	**0**	**0**	5	**108**	**108**	317	388	280	280	76

Table 3.4: The second part of the evaluation results of the randomly picked frames (Experiment 4). We evaluated the Dice Coefficient, false positives (FP), false negatives (FN) and the segmented area.

	DC [%]			FP [px]			FN [px]			Area [px]			
#	δ_3	δ_8	SRG	δ_3	δ_8	SRG	δ_3	δ_8	SRG	GT	δ_3	δ_8	SRG
14	**59.5**	**59.5**	52.6	**52**	**52**	119	16	16	**0**	66	102	102	185
15	65.3	**68.0**	43.6	**4**	**4**	50	13	12	**7**	29	20	21	72
16	85.2	**85.3**	58.1	100	102	**81**	23	**21**	190	378	455	459	269
17	**92.8**	**92.8**	83.4	**7**	**7**	50	**40**	**40**	62	343	310	310	331
18	14.6	39.1	**79.7**	1802	477	**23**	**6**	**6**	39	161	1957	632	145
19	**23.7**	**23.7**	0.0	**8**	**8**	14	**50**	**50**	59	59	17	17	14
20	**76.7**	76.5	54.5	173	180	**41**	10	**7**	179	311	474	484	173
21	83.1	**87.5**	54.8	45	**28**	**28**	**12**	**12**	84	152	185	168	96
22	**60.3**	**60.3**	49.1	**19**	**19**	51	31	31	**30**	69	57	57	90
23	56.9	**65.2**	29.5	31	**18**	120	**13**	**13**	14	42	60	47	148
24	61.9	**73.7**	20.0	10	**5**	17	6	**5**	15	19	23	19	21
25	**86.4**	**86.4**	81.7	**20**	**20**	35	**12**	**12**	11	114	122	122	138

3.4 Discussion of Results

The results show that our fully automatic segmentation approach yields good qualitative results, which are promising for voice and speech disorder research. It also overcomes certain problems of the SRG-based clinical standard [41] like leakage, over- and undersegmentation, and outperforms it in all four experimental setups.

The proposed method with salient step size $\delta_S = 3$ performs better than the reference implementation of the clinical standard with mean DC values ($\bar{x}$) of 73.4, 53.7, 91.5, 76.7 compared to 56.6, 35.5, 78.8, 53 and median values ($\tilde{x}$) of 86.1, 61.3, 92.7, 76.7 and 67.4, 41.6, 79.8, 54.5. The SRG method has a problem with leakage as depicted in Figure 3.5 (Exp. 1 (e), Exp. 2 (e) and Exp. 4 (2e, 3e)), while with our method a similar effect only occurs once in Figure 3.7 (18c, d). Both approaches have limitations when the glottal opening is very small as can be seen in Figure 3.8 (19). It also performs slightly better than the version with salient step size $\delta_S = 8$ with mean values ($\bar{x}$) of 73.4, 53.7, 91.5, 76.7 compared to 75.5, 52.7, 91.3, 73.9 and median values ($\tilde{x}$) of 86.1, 61.3, 92.7, 76.7 and 85.9, 60.5, 92.6, 76.4. Both versions of our method show a significantly better median value than the clinical standard at the 5 % significance level because their intervals, represented by the notches of the box plots, do not overlap in all of the experiments. As mentioned in Section 2.3, 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 of around a factor 2, while the results are similar. The computation of a block with 1500 frames at $\delta_S = 8$ takes approximately 5 minutes and for a whole video around 30 minutes. Some of the outliers in the box-whisker plot (see Fig. 3.4) can be explained due to the nature of the Dice Coefficient, 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 $DC = 0$). The segmentation of the two video sequences (Experiment 1 and 3) worked very well, while there were problems in Experiment 2 and 4, which will be discussed in the next section.

3.4.1 Image Segmentation Problems

In Experiment 2 the narrow opening causes problems when detecting the seed regions as initialization for the segmentation. An erode operation is performed on every seed region, which can lead to the vanishing of very small areas. This is usually not a problem because the glottal opening gets bigger and we have the 3D information to work with, but if the initialization vanishes, in most of the cases there are not enough seeds for a correct segmentation.

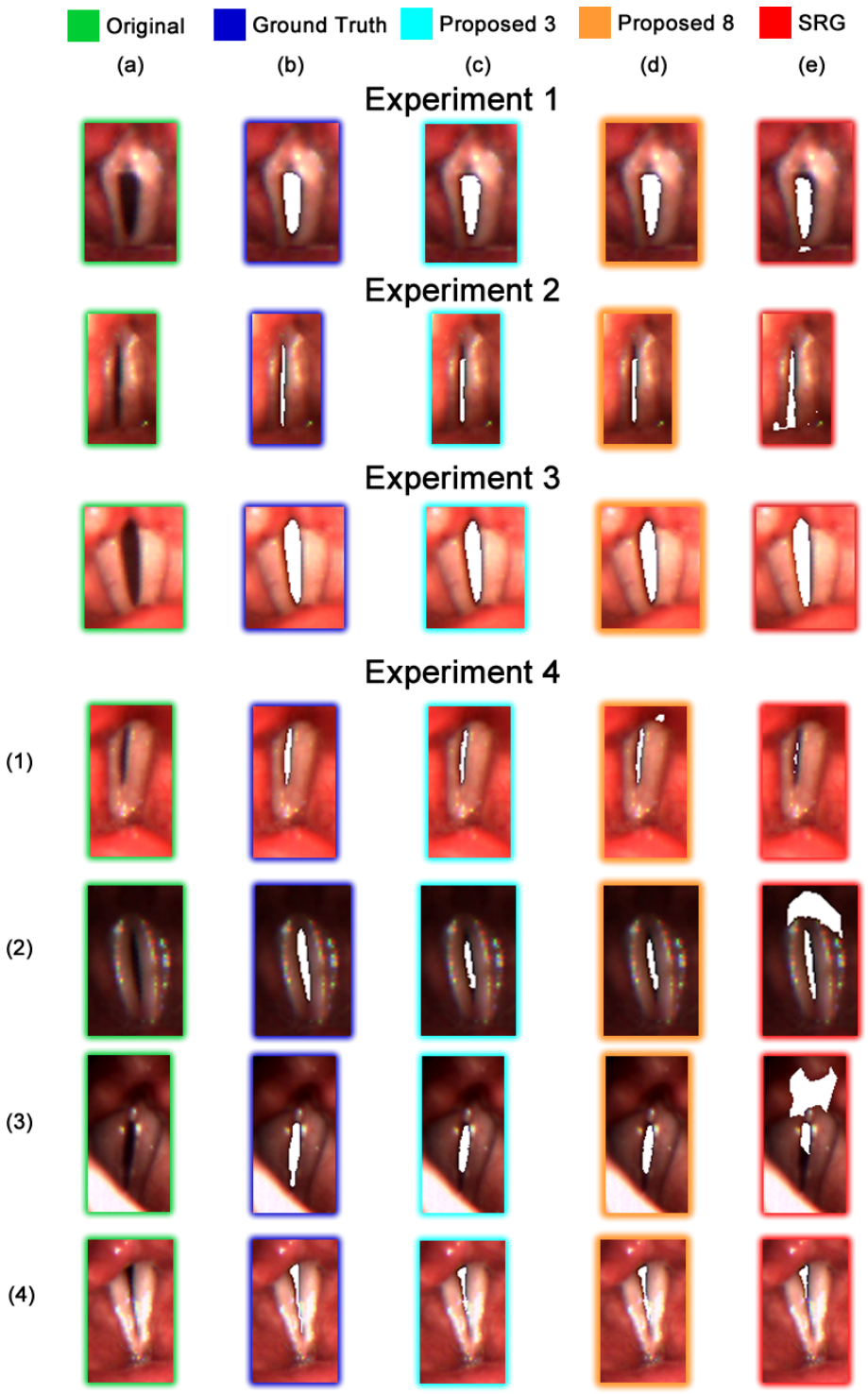


Figure 3.5: This figure shows the first part of the segmentation results of *Experiment 1*. There are five different images for each dataset: original (green), ground truth (blue), segmentation results for the proposed method with $\delta_S = 3$ (turquoise) and $\delta_S = 8$ (orange), and the result of the SRG reference method (red).

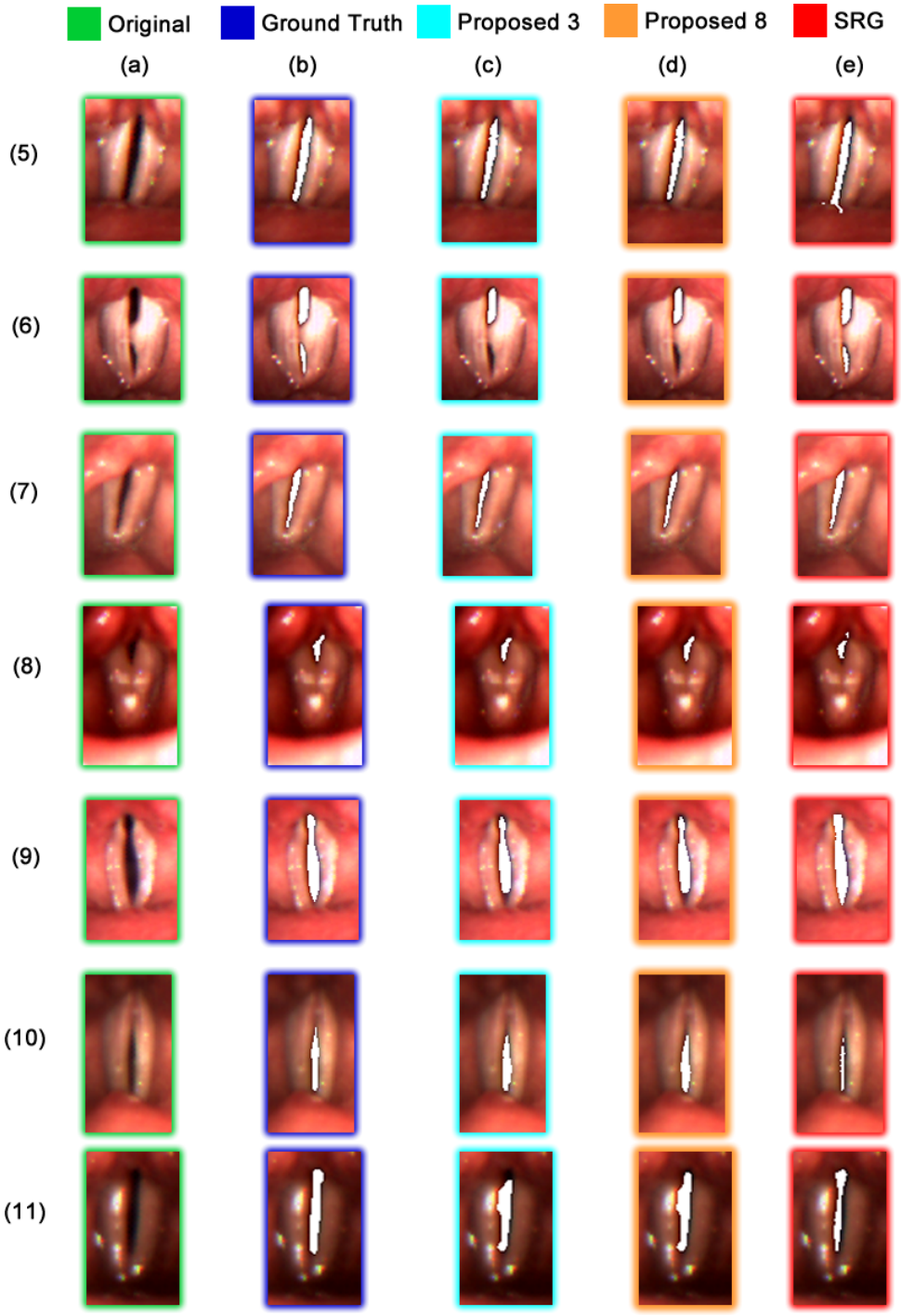


Figure 3.6: This figure shows the second part of the segmentation results of *Experiment 1*. There are five different images for each dataset: original (green), ground truth (blue), segmentation results for the proposed method with $\delta_S = 3$ (turquoise) and $\delta_S = 8$ (orange), and the result of the SRG reference method (red).

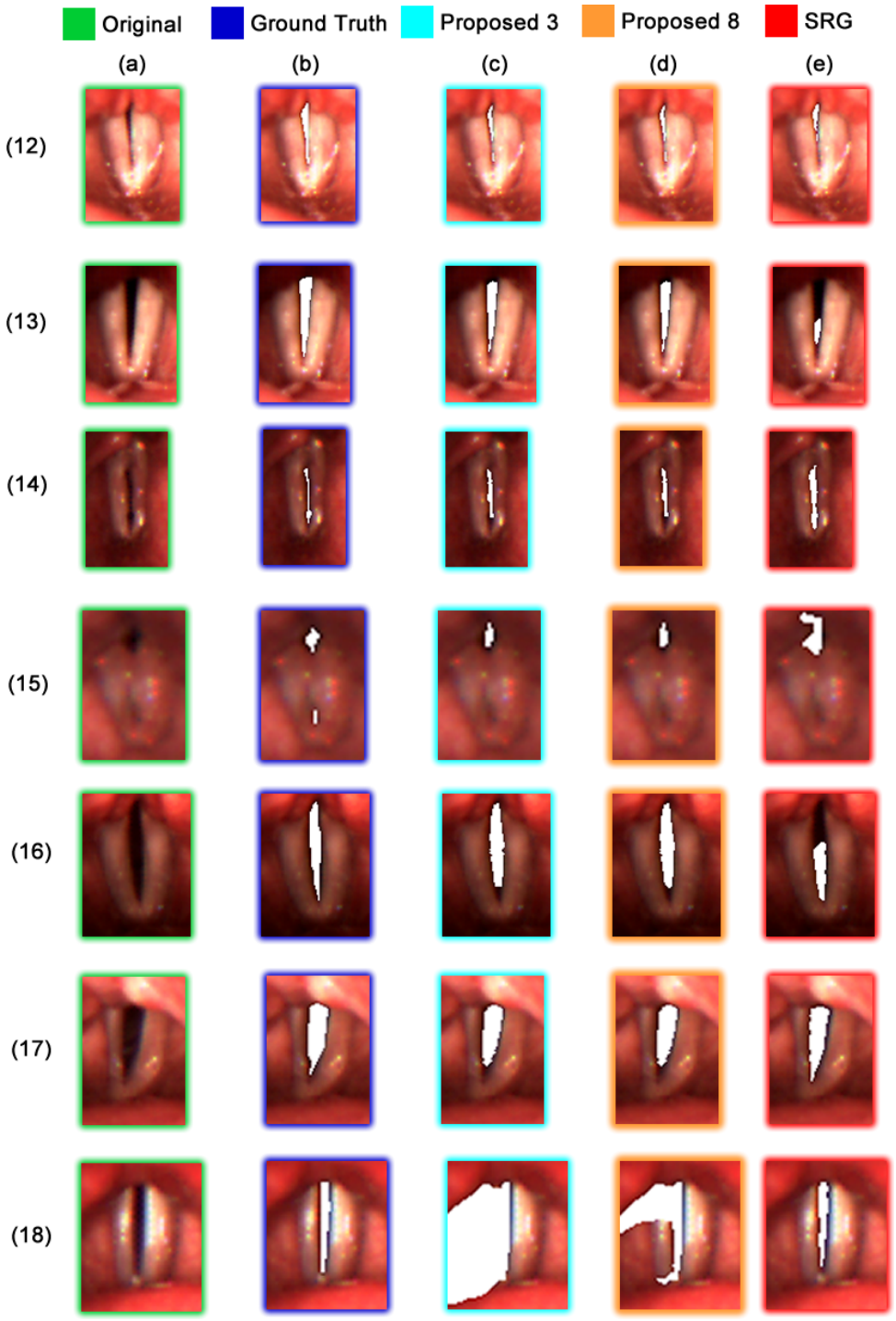


Figure 3.7: This figure shows the third part of the segmentation results of Experiment 4. There are five different images for each dataset: original (green), ground truth (blue), segmentation results for the proposed method with $\delta_S = 3$ (turquoise) and $\delta_S = 8$ (orange), and the result of the SRG reference method (red).

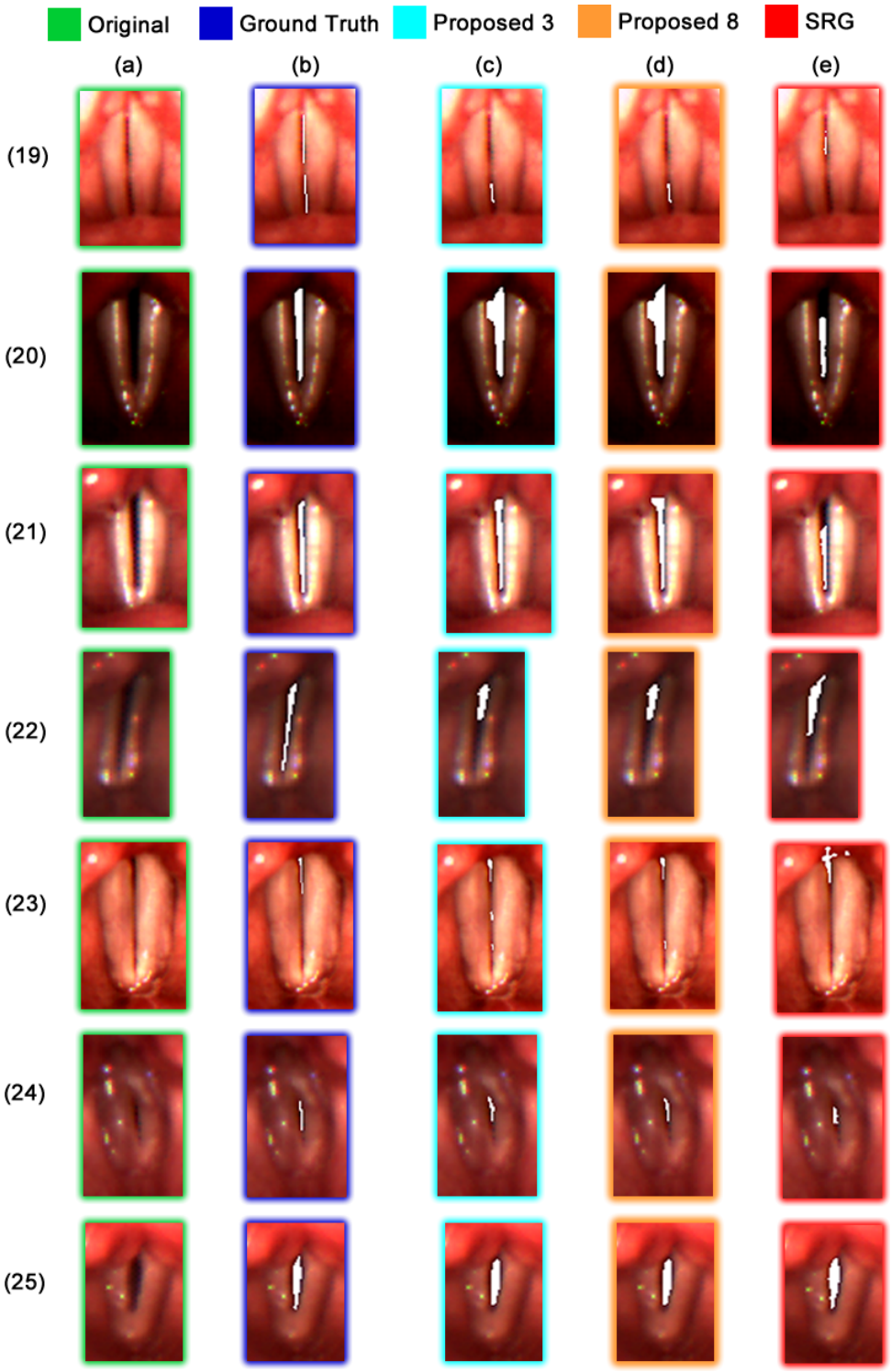


Figure 3.8: This figure shows the third part of the segmentation results of Experiment 4. There are five different images for each dataset: original (green), ground truth (blue), segmentation results for the proposed method with $\delta_S = 3$ (turquoise) and $\delta_S = 8$ (orange), and the result of the SRG reference method (red).

Due to the broad range of video data given by Experiment 4, a few different shortcomings of our method can be seen. In Figure 3.5 (3c, d), 3.6 (11c, d), 3.7 (16c, d) and 3.8 (22c, d) it can be seen that the initialization does not evolve completely towards the upper or lower end of the glottal opening, which is caused by GAC segmentation algorithm. It tries to minimize the surface of the segmentation volume and it stops if it is too costly to grow over certain edges.

Another problem is shown in Figure 3.5 (1d) and 3.7 (18c, d), where the segmentation is bigger than the actual glottis. In this case the ROI is not detected correctly because there is a connection to a nearby dark area, which could not be removed by filtering out weak signals. The wrong saliency maps and ROIs are depicted in Figure 3.9(a,b).

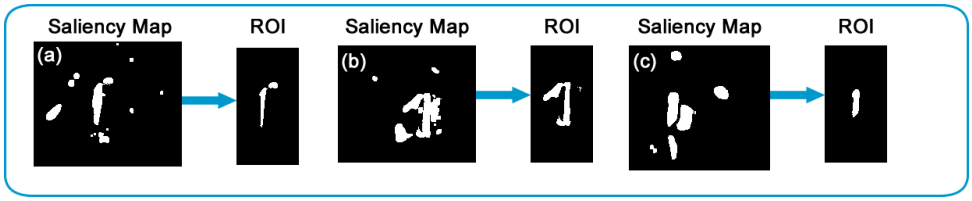


Figure 3.9: In the saliency map a connection between the glottal opening and the surrounding areas can be seen, which causes problem when calculating the seed regions. Experiment 4 (1d) is depicted in (a), (18d) in (b) and (6d) in (c).

Our current ROI detection algorithm only chooses one region from the saliency map and ignores the other ones. Therefore, two glottal openings can never be a ROI and usually the larger one is selected. This can be seen in Figure 3.6(6c,d) and 3.9(c), where only the larger opening is segmented. If the glottis has only one big opening in the beginning and later changes to two or more, the ROI normally covers all the glottal area and the segmentation works.

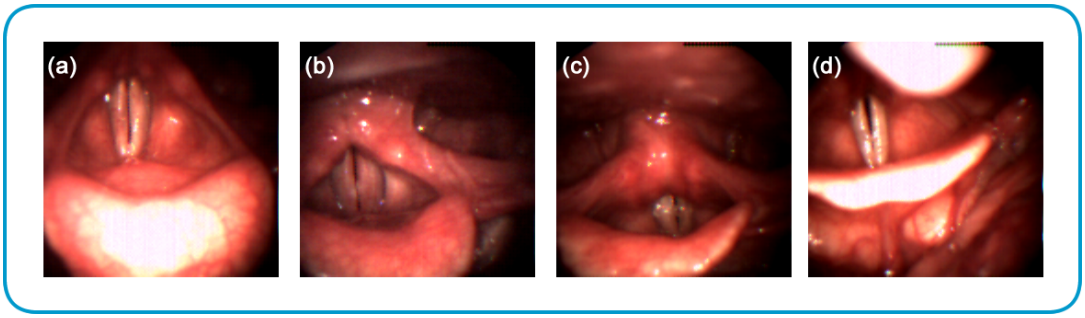


Figure 3.10: These four recordings were excluded from the evaluation because (a,b,c) had too narrow openings to yield meaningful results, while in (d) the ROI detection failed.

As mentioned in Section 3.2, five out of the 30 single frames were excluded from the evaluation because we encountered segmentation problems with these recordings. Three were excluded because of too narrow openings to yield meaningful segmentation results (see Fig. 3.10(a,b,c)). In (d) there are a lot of disturbing anatomical structures giving stronger signals than the actual glottal opening, thus it was automatically filtered out, rendering a segmentation impossible. The fifth video was excluded due to a problem with the seed region detection, which resulted in a segmentation of the vocal folds in the first frames and later no segmentation at all.

Chapter 4

Summary and Conclusion

We presented a fully automatic glottis segmentation method from LHSVs, which shows promising qualitative results and is usable for everyday clinical applications due to very efficient implementation on the graphics adapter. Our algorithm addresses problems in the video recordings by using a combination of contrast stretching, denoising and motion compensation. We tackled the difficult problem of finding the glottis and an initialization for the segmentation by using a salient region detection method. Finally, we use a 3D Geodesic Active Contour segmentation, which overcomes limitations of traditional 2D based methods.

The proposed method was evaluated by comparing it to the clinical standard on a set of ground truth data. We designed four different experiments consisting of three video sequences and 25 single frames, randomly picked from various recordings. In all the experiments our method performed significantly better than the current clinical standard, without requiring any user interaction, thus our expected run-time scales much better to the real-world application of segmenting thousands of video frames or a database of videos. On a standard computer we achieve efficient and fast computation of around 30 minutes for a whole recording ($\delta_S = 8$) by using a parallelized CUDA implementation on an NVidia graphics adapter, making it perfectly applicable in everyday clinical applications.

Even though the achieved results were more accurate than the current clinical standard, there is still room for improvements. Especially the ROI and bounding box detection requires further refinements. An additional filter step could be included to prevent the occasional connection between surrounding dark areas and the glottal opening. Another improvement would be that the ROI detection could also handle special topologies like multiple or narrow openings. A future goal is to make this method usable for as many

videos as possible, including the recordings excluded from evaluation as well as bad ones, where the glottis is occasionally covered by the epiglottis. This could be achieved by first analyzing the video sequence and marking all the frames with a covered glottis and treat them specially.

For those videos, where problems occur, the user can manually specify parameters like the landmark frame or bounding box in configuration files or use an external tool for interactive segmentation editing. However, currently this requires manual screening of the segmentation results. To make the program easily accessible for non-experienced users, a graphical user interface for segmentation refinement and configuration directly integrated into the software would be a great improvement of usability.

Another possible next step is a more extensive evaluation on a larger database and a comparison to different segmentation methods. With the good qualitative results it could also be possible to automatically make a connection between a segmentation and certain voice or speech disorders.

Appendix A

Publications and Presentations

Publications

[1] Schenk, F., Urschler, M., Aigner, C., Roesner, I., Aichinger, P. and Bischof, H. Automatic glottis segmentation from laryngeal high-speed videos using 3D geodesic active contours. In: *Proceedings Medical Image Understanding and Analysis (MIUA); London (2014).*

Oral Presentations

[1] Schenk, F. and Urschler, M. Towards automatic segmentation of the glottal area. *Scientific seminar 2: Detection of Diplophonia; Vienna (2013)*

[2] Schenk, F. and Urschler, M. Automatic glottis segmentation from laryngeal high-speed videos using 3D geodesic active contours. *Scientific seminar 3: Detection of Diplophonia; Vienna (2014).*

[3] Schenk, F., Urschler, M., Aigner, C., Roesner, I., Aichinger, P. and Bischof, H. Automatic glottis segmentation from laryngeal high-speed videos using 3D geodesic active contours. In: *Proceedings Medical Image Understanding and Analysis (MIUA); London (2014).*

Appendix B

Abbreviations and Definitions

BMS	Boolean Map based Saliency proposed by Zhang and Sclaroff [69].
BS	Block Size.
CIE LAB	A color space with perceptual uniformity, where L is lightness, a the position between red and green and b the position between yellow and blue.
CPU	Central Processing Unit.
CUDA	Compute unified device architecture. A parallel programming platform invented by NVidia, which runs on the graphics adapter.
DC	Dice Coefficient.
FN	Number of false negatives.
FP	Number of false positives.
Glottis	The glottis consists of the vocal folds and (glottal) opening between them.
GPU	Graphics Processing Unit.
ICG	Institute for Computer Graphics and Vision.
JPEG	Joint Photographic Experts Group. Is a file format for lossy compression of digital images.
LBI	Ludwig Boltzmann Institute for Clinical Forensic Imaging.
LF	Landmark frame. The frame with the maximal glottal opening in a cycle or a block.
LHSV	Laryngeal high-speed videos.
NVidia	American company that manufactures GPUs.
OV	Overlap between the different processing blocks.

PNG Portable Network Graphics. A raster graphics file format that supports lossless data compression.

RGB Color space, which represents every color by mixing red, green and blue values.

ROF Edge-preserving denoising according to Rudin, Osher and Fatemi [49].

ROI Region of Interest.

SVM Support Vector Machine.

SRG Seeded Region Growing.

SAD Sum of Absolute Differences.

TN Number of true positives.

TP Number of true negatives.

TV Total Variation.

Appendix C

Configuration Files

There are two configuration files, one for general settings (*config.ini*) and one for the separate blocks. We use the very intuitive and simple INI file format. Basically it consists of groups or sections and keys with values:

```
[Group1]
keyA = 123
keyB = true
[Group2]
keyA = 'Hallo'
abc  = 123
name1 =
```

The key value only has to be unique inside one group and can be boolean, integer, string, etc. For *name1* the value part is left empty because only the name of the key is important.

If the file *config.ini* does not exist, the program will automatically create one with the standard values. The settings and values generated by the program are shown in the next paragraphs, with explanations of the more complicated sections.

```
[BlockSettings]
BlockOverlap=100
BlockSize=1500

[BoundingBox]
Height=150
Width=80
```

```
[Denoising]
DenoisingLambda=25
DenoisingMaxSteps=10000
```

In *[GeneralSettings]* the file name of the blocks configuration file has to be specified as well as if it should be used. *UseBlockInfoFile* has to be true to enable user-intervention such as bounding box specification. *DebugModus* gives additional output.

```
[GeneralSettings]
BlockInfoFileName=blocks.ini
DebugModus=false
UseBlockInfoFile=true


[Input]
MainInputFolder=/home/schenk/DA/Kehlkopfvideos/MIUA/
```

The folders in *[InputFolderList]* are read from the *MainInputFolder*. It is important to use the equality sign after the key name.

```
[InputFolderList]
ab123771_A001=
xy123565_A003=
```

If *Output* is false, no images or volumes will be written. *ImagesInVolume* specifies the size of the output volume because the tool for later user-refinement cannot work with extremely large volumes.

```
[Output]
ImagesInVolume=300
MainOutputFolder=/home/schenk/DA/Kehlkopfvideos/MIUA/results/sal8/
Output=true
OutputBBImg=true
OutputConstraintsVolume=true
OutputSegmentationVolume=true
OutputWholeImg=true


[Preprocessing]
ContrastStretching=true
```

ImagesToCheck defines in how many frames the algorithm searches for the maximum glottal opening/landmark frame.

```
[ROIandBB]
ImagesToCheck=150

[Registration]
AngleRange=0.01308996939
AngleSteps=0.00436332313
DefaultFillValue=280
ImagesToSkip=10
TranslationRange=2
TranslationSteps=1

[SaliencyDetection]
SaliencyStepSize=3

[Segmentation]
EdgeAlpha=15
EdgeBeta=0.55
SegmentationLambda=0.01
SegmentationMaxSteps=10000
```

If *UseBlockInfoFile* is true in *[GeneralSettings]* a file will be created with information about each block. It is named according to the value in *BlockInfoFileName*. The file will be created in the first run, saving all automatically calculated information like the landmark frame file name and the bounding box specifications. An example can be seen in the next paragraphs. If the file exists and *UseUserSpecifiedValues* is true, the values will be read instead of calculated and therefore a user can simply change the settings.

```
[ba100771_A001]
UseUserSpecifiedValues=false
```

The information for the first block *B0* contains the bounding box dimensions and the file name of the landmark frame, which can be changed by the user. *StartFrame* and *EndFrame* are fixed values to give the user additional information. *BBHeight* and *BBWidth* have priority over the one specified in config.ini. *LandmarkFrameNumber* is the index in the file list and changes effect the image segmentation process, while *LandmarkFrame* just

shows the file name as additional information.

```
[ba100771_A001B0]
BBHeight=150
BBStartX=77
BBWidth=80
BBstartY=5
EndFrame=5311.png
LandmarkFrame=4458.png
LandmarkFrameNumber=146
StartFrame=4312.png

[ba100771_A001B1]
BBHeight=150
BBStartX=57
BBWidth=80
BBstartY=6
EndFrame=6210.png
LandmarkFrame=5257.png
LandmarkFrameNumber=46
StartFrame=5211.png
```

Appendix D

Software Dependencies and Libraries

We implemented our method in C++ and evaluated under Ubuntu 14.04 LTS. 64 bit versions of all the libraries were used an we developed with Qt Creator. Table D shows the open source libraries and tools we used in our work.

Table D.1: Libraries and tools required for compilation of our program.

Tool/Library	Version
CMake	2.8.12.2
CUDA	6.5.12
GCC	4.8.2
ITK	3.2
Qt Creator	3.0.1
Qt	5.2.1

Bibliography

[1] Bayes, M. and Price, M. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions (1683-1775)*, pages 370–418.

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

[3] Boykov, Y. and Jolly, M.-P. (2000). Interactive organ segmentation using graph cuts. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2000*, pages 276–286. Springer.

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

[5] Bruce, N. and Tsotsos, J. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, 9(3):1–24.

[6] Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:679–698.

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

[8] Chambolle, A. and Lions, P.-L. (1997). Image recovery via total variation minimization and related problems. *Numerische Mathematik*, 76(2):167–188.

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

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

[11] Chen, L. (1982). Topological structure in visual perception. *Science*, 218:699–700.

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

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

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

[15] Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

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

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

[18] Fairchild, M. D. (2013). *Color appearance models*. John Wiley & Sons.

[19] Garcia-Diaz, A., Vidal, X., Pardo, X., and Dosil, R. (2012). Saliency from hierarchical adaptation through decorrelation and variance normalization. *Image and Vision Computing*, 30(1):51–64.

[20] Golub, G. H. and van Van Loan, C. F. (1996). *Matrix computations*. The Johns Hopkins University Press, 3rd edition.

[21] Gonzalez, R. C. and Woods, R. E. (2007). *Digital Image Processing*. Prentice Hall, Boston, MA, USA, 3nd edition.

[22] Hadamard, J. (1902). Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton university bulletin*, 13(49-52):28.

[23] Haghtalab, N. (2014). *Geodesic Active Contours*. Report, University of Waterloo.

[24] Hertegard, S., Larsson, H., and Wittenberg, T. (2003). High-speed imaging: applications and development. *Logoped Phoniatr Vocol*, 28:133–139.

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

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

[27] Hou, X., Harel, J., and Koch, C. (2012). Image signature: Highlighting sparse salient regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1):194–201.

[28] Huang, L. and Pashler, H. (2007). A boolean map theory of visual attention. *Psychological review*, 114:599–631.

[29] Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency- based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259.

[30] Itti, L. and Pierre, B. (2006). Bayesian surprise attracts human attention. In *Neural Information Processing Systems (NIPS 2006).*

[31] Jain, A. K. (1989). *Fundamentals of Digital Image Processing.* Prentice Hall.

[32] Jayadevappa, D., Kumar, S. S., and Murty, D. S. (2011). Medical image segmentation algorithms using deformable models: A review. *Institution of Electronics and Telecommunication Engineers Technical Review*, 28(3):248–255.

[33] Judd, T., Ehinger, K., Durand, F., and Torralba, A. (2009). Learning to predict where humans look. In *International conference on Computer Vision*, pages 2106–2113. IEEE.

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

[35] Kass, M., Witkin, A., and Terzopoulos, D. (1988). Snakes: Active contour models. *International Journal of Computer Vision*, 1:321–331.

[36] Kienzle, W., Wichmann, F., Schölkopf, B., and Franz, M. (2007). A nonparametric approach to bottom-up visual saliency. *Neural Information Processing Systems (NIPS 2007).*

[37] Koç, T. and Çiloğlu, T. (2014). Automatic segmentation of high speed video images of vocal folds. *Journal of Applied Mathematics*, 2014:doi:10.1155/2014/818415.

[38] Leung, S. and Osher, S. (2005). Global minimization of the active contour model with tv-inpainting and two-phase denoising. In *Variational, geometric, and level set methods in computer vision*, pages 149–160. Springer.

[39] Li, J., Levine, M. D., An, X., Xu, X., and He, H. (2013). Visual saliency based on scale-space analysis in the frequency domain. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(4):996–1010.

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

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

[42] McInerney, T. and Terzopoulos, D. (1996). Deformable models in medical image analysis: a survey. *Medical Image Analysis*, 1(2):91–108.

[43] Mehta, D. D., Deliyski, D. D., Quatieri, T. F., and Hillman, R. E. (2011). Automated measurement of vocal fold vibratory asymmetry from high-speed videoendoscopy recordings. *Journal of Speech, Language, and Hearing Research*, 54(1):47–54.

[44] O'Connor, B. (2013). Learning frames from text with an unsupervised latent variable model. *arXiv preprint arXiv:1307.7382*.

[45] Palmer, S. E. (1999). *Vision science: Photons to phenomenology*. The MIT press.

[46] Reinbacher, C., Pock, T., Bauer, C., and Bischof, H. (2010). Variational segmentation of elongated volumetric structures. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3177–3184. IEEE.

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

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

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

[50] Samet, H. and Tamminen, M. (1988). Efficient component labeling of images of arbitrary dimension represented by linear bintrees. *IEEE Transactions onPattern Analysis and Machine Intelligence*, 10(4):579–586.

[51] Schauerte, B. and Stiefelhagen, R. (2012). Quaternion-based spectral saliency detection for eye fixation prediction. In *Computer Vision–ECCV 2012*, pages 116–129. Springer.

[52] Sørensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons. *Biologiske Skrifter*, 5:1–34.

[53] Thévenaz, P., Blu, T., and Unser, M. (2000). Interpolation revisited [medical images application]. *IEEE Transactions on Medical Imaging*, 19(7):739–758.

[54] Tikhonov, A. N. (1963). Regularization of incorrectly posed problems. *Soviet Mathematics Doklady*, 4(6):1624–1627.

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

[56] Titze, I. R. (1994). *Principles of Voice Production*. Prentice-Hall, 1 edition.

[57] Unger, M. (2012). *Convex Optimization for Image Segmentation*. PhD thesis, Graz University of Technology.

[58] Unger, M., Pock, T., and Bischof, H. (2008a). Continuous globally optimal image segmentation with local constraints. In *Computer Vision Winter Workshop*, volume 2008.

[59] Unger, M., Pock, T., Trobin, W., Cremers, D., and Bischof, H. (2008b). TVSeg – interactive total variation based image segmentation. In *Proceedings British Machine Vision Conference*.

[60] Unser, M. (1999). Splines: A perfect fit for signal and image processing. *Signal Processing Magazine, IEEE*, 16(6):22–38.

[61] Urschler, M., Leitinger, G., and Pock, T. (2014). Interactive 2d/3d image denoising and segmentation tool for medical applications. In *Proceedings MICCAI IMIC Workshop on Interactive Medical Image Computing*.

[62] Vanne, J., Aho, E., Hamalainen, T. D., and Kuusilinna, K. (2006). A high-performance sum of absolute difference implementation for motion estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 16(7):876–883.

[63] Vassiliadis, S., Hakkennes, E. A., Wong, J., and Pechanek, G. G. (1998). The sum-absolute-difference motion estimation accelerator. In *Euromicro Conference, 1998. Proceedings. 24th*, volume 2, pages 559–566. IEEE.

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

[65] Wong, S., Vassiliadis, S., and Cotofana, S. (2002). A sum of absolute differences implementation in fpga hardware. In *Euromicro Conference, 2002. Proceedings. 28th*, pages 183–188. IEEE.

[66] Xu, C., Pham, D. L., and Prince, J. L. (2000). *Medical Image Segmentation Using Deformable Models*. SPIE Press.

[67] Yan, Y., Chen, X., and Bless, D. (2006). Automatic tracing of vocal-fold motion from high-speed digital images. *IEEE Transactions on Biomedical Engineering*, 53(7):1394–1400.

[68] Yushkevich, P. A., Piven, J., Hazlett, H. C., Smith, R. G., Ho, S., Gee, J. C., and Gerig, G. (2006). User-guided 3d active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage*, 31(3):1116–1128.

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

[70] Zhang, L., Tong, M., Marks, T., Shan, H., and Cottrell, G. (2008). Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7):1–20.