FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

DOCTORAL THESIS

RNDr. Aleš Zita

Analysing Videokymograms Using Classical and Deep Learning Methods

Institute of Information Theory and Automation,
the Czech Academy of Sciences

Supervisor of the doctoral thesis:	Prof. Ing. Jan Flusser, DrSc.
Study programme:	Computer Science
Study branch:	Visual Computing and Computer Games

Prague 2022

I declare that I carried out this doctoral thesis independently, and only with the cited sources, literature and other professional sources. It has not been used to obtain another or the same degree.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In .... date ....

Author's signature

I would like to express my deepest gratitude to my supervisor, Prof. Ing. Jan Flusser, DrSc., for his guidance and support during my studies, my fellow colleagues for their help and friendly attitude, and the Department of Image Processing at IITA for the resources and exciting research opportunities. Additionally, I thank my dearest wife for all the proofreading and care, and all my family for their patience during my study.

I dedicate this work to my wonderful wife, without whom this would not be possible, my kids, who haven't seen as much of me as they deserve, and my parents.

Thank you.

Title: Analysing Videokymograms Using Classical and Deep Learning Methods

Author: RNDr. Aleš Zita

Institute: Institute of Information Theory and Automation,
the Czech Academy of Sciences

Supervisor: Prof. Ing. Jan Flusser, DrSc., Department of Image Processing

Abstract: Videokymography (VKG) belongs to a family of medical imaging techniques capable of human larynx function visualization. Images produced by this method are ideal for automatic processing. In the last few years, the performance of deep learning systems increased significantly. In some areas, the machine learning approach exceeds the human experts in speed and accuracy. This doctoral thesis focuses on the continuous development of VKG image automatic analysis and touches on the possibility of connecting the classical approach to Videokymographic image processing with the modern computer vision approach.

Keywords: Videokymography, Medical Imaging, Digital Image Processing, Computer Vision, Machine Learning

Contents

1. Human vocal folds

1.1 Computer-aided vocal folds examination

Image processing and computer vision research have long helped automate and process various tasks in everyday life. The advanced knowledge of these domains has found many applications in commercial software, forensics, everyday tools and toys, and more. Medicine is one of the most important fields of human activity affected by these disciplines. In addition to many regular medical applications such as microscopy, ultrasound, magnetic resonance, or computed tomography, they can also be helpful in lesser-known technologies such as image processing of the vocal folds using Videokymography (VKG).

1.2 Anatomy

The human vocal folds are fast-vibrating tissues in the glottis. Their primary function is creating sound through vocalization. The vocal folds produce sound as a result of high-pressure air from the lungs pushing through the closed larynx (refer to Figure 1). If sufficiently high, the difference in air pressure causes the soft tissue to release a quantum of air from the space under the vocal folds to the throat. When this process repeats with adequate frequency, the pressure waves of the released quanta of air form sound. The described process is called phonation. For a detailed illustration, see Figure 2.

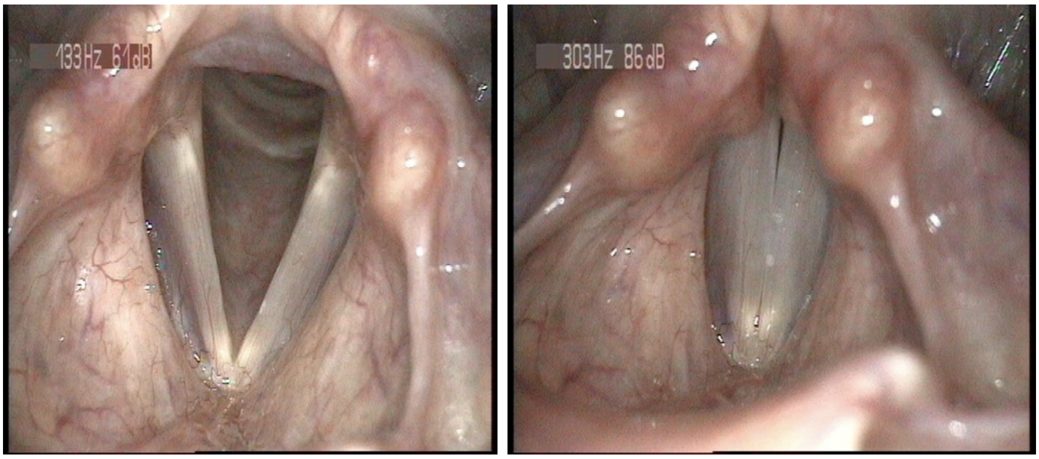


Figure 1: Laryngoscopic views of the human larynx in inspiratory (open) and phonatory (closed) positions. Images taken from [1].

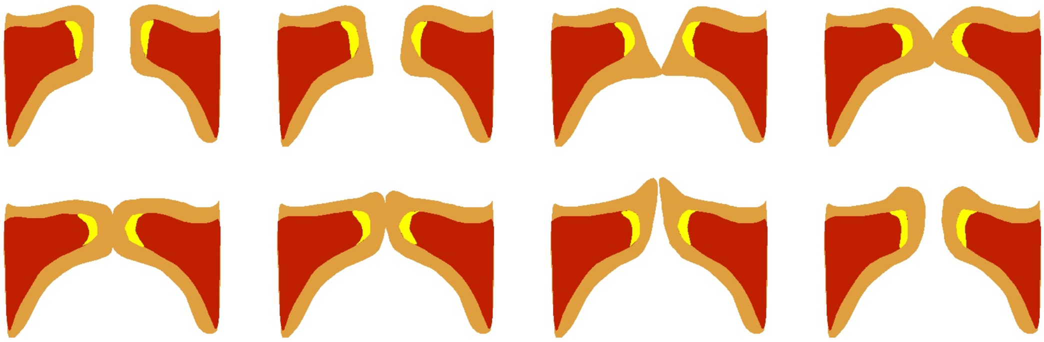


Figure 2: A schema of the glottal cycle during phonation [2]. The air pressure difference causes the soft tissue to release a quantum of air from space under the vocal folds to the throat. Repeating this process with sufficient frequency produce sound waves.

1.3 Examination complexity

From a medical point of view, the vocal folds examination is paramount in the early diagnosis of numerous diseases or dysfunctions, such as:

- Laryngeal cancer that causes hoarseness, a lump in the neck, and difficulty breathing or swallowing;
- Laryngoceles, resulting in hoarseness and airway obstruction;
- Spasmodic dysphonia, causing an inability to speak or may sound whispery, jerky, creaky, or garbled;
- Vocal cord contact ulcers;
- Vocal cord paralysis;
- Vocal cord polyps, nodules, granulomas, papillomas.

Examining vocal folds' function is a complex process. The vocal folds reside deep in the throat in the area called the larynx. For an examiner (physician) to even see the larynx means using a piece of special viewing equipment called a laryngoscope (refer to Figure 3). A laryngoscope is a particular type of endoscope adapted for viewing the larynx. The laryngoscope consists of an eyepiece, a light post connector, and a long tube equipped with a mirror or prism, allowing seeing the larynx. To capture the laryngoscopic image, a digital camera can be mounted, replacing the eyepiece. Figure 1 shows examples of the laryngoscopic image.

Nevertheless, simply viewing the vocal folds can only reveal limited information about the function. For better understanding, the examiner needs to observe the organ in motion. The typical frequency of the human voice ranges from 85 to 155 Hz. This frequency is too high for the movement of the vocal cords to be seen by the human eye; therefore, every vocal fold function visualization is inherently mediated. A system capturing the vocal cords in motion must have at least an order of magnitude higher scanning frequency than the vocal fold vibration frequency.

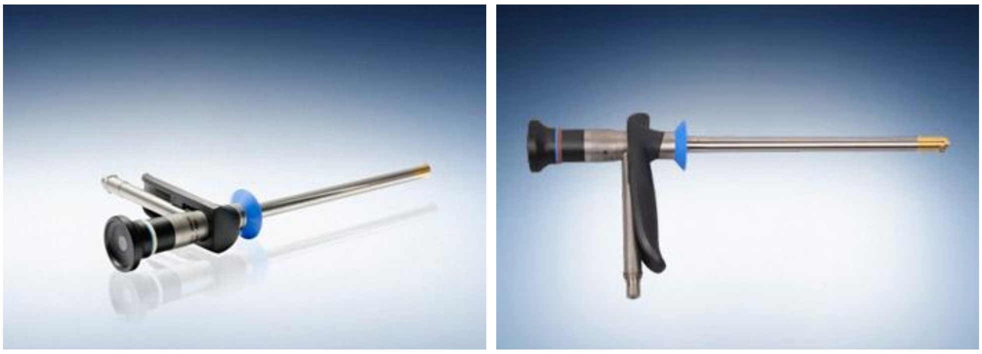


Figure 3: Laryngoscope: A viewing device capable of reaching and visualizing the human larynx, including the vocal fold. The laryngoscope consists of an eyepiece, a light post connector, and a long tube equipped with a mirror or prism, allowing seeing the larynx. The images are taken from [3].

1.4 Current trends in larynx visualization

Nowadays, there are three main techniques to display and analyze the vibration of vocal folds [4].

One way to capture the vocal cord motion with sufficient time resolution is the laryngeal *High-Speed Videoendoscopy* (HSV) [5; 6]. The display frequency of high-speed systems ranges from 2000 frames per second with a resolution of 256×256 pixels [7; 8] to 10000 fps [8]. Although the data captured by an HSV camera contain the entire course of vibrations over the whole length of vocal folds [9], the HSV equipment is expensive; the method is memory- and time-wise demanding and produces large amounts of data. Some of the mentioned disadvantages can be decreased by postprocessing methods such as *Digital Kymography* (DKG) [10; 11; 12; 13] or *Phonovibrography* [14; 15; 16].

Another way to deal with high vocal cord vibration frequency is to achieve an apparent slowdown of the recording. The method based on such a principle is called *Strobolaryngoscopy* [17]. It is a method that uses rapid flashes of light to seemingly undersample the video recording by utilizing the time-aliasing effect. The vocal fold vibrations are visualized in slowed-down motion by synchronizing the image-capturing moments with the frequency of the vocal fold oscillations. Refer to Figure 4 for an illustration of the principle. Although this method is widespread due to its availability, it has several rather substantial disadvantages. One of the drawbacks of this method is the necessity to fine-tune the system's flash rate frequency for every examination and voice pitch to produce the desired effect. It is because the method is directly dependent on the frequency of the vocal cord oscillations. Another considerable disadvantage is that the strobolaryngoscopic system is, due to the under-sampling process, not suitable for capturing aperiodic vibrations and events.

The last method to be mentioned is the *Videokymography* [19] (VKG). This technique is based on visualization of the vibrations of the vocal cords using a line scanner. A line scanner is a specially modified CCD camera that, instead of scanning all the lines, repeatedly scans the same line but with a frequency number-of-CCD-rows times higher. The individual scanned rows are then stacked up and form a vidokymogram (See Figure 5 for examples of VKG images.).

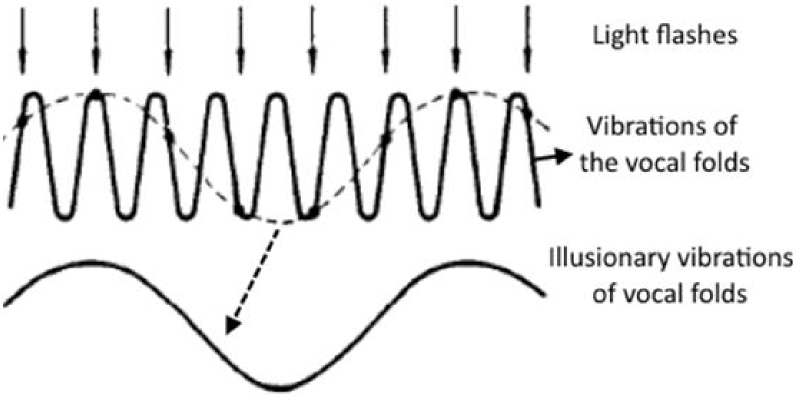

Light flashes
Vibrations of the vocal folds
Illusionary vibrations of vocal folds

Figure 4: Illustration of the principle of Videostroboscopy showing the synchronization of light flashes at frequencies slightly above the fundamental frequency vocal folds of vibration to obtain the new illusionary slow motion of the vocal folds. Image taken from [18].

From all the methods mentioned above, Videokymography has excellent potential to become an achievable and widespread diagnostic tool because it does not have the limitation of the stroboscopic method while being cost-effective.

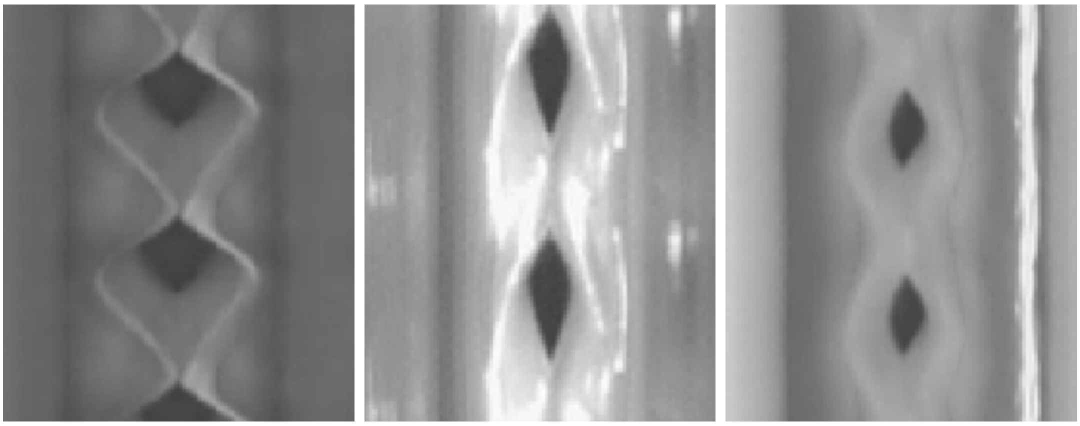


Figure 5: Videokymographic images, where the vertical axis represents the temporal and the horizontal axis denotes the spatial domain.

2. Videokymography

2.1 Hardware

In 1994 in Groningen (NL), the Czech-Dutch team took an ordinary CCD camera outputting a 555×333 image at 25fps and transformed it into a single-line scanner by re-reading the same line 520 times. In this way, they achieved a frequency of 288×25 = 7200 fps. The camera operated in two modes: The standard mode displayed a full-size image at a frequency of 50fps interlaced, and the high-speed line scanner mode achieved the full 7200 fps. For comprehensive visualization, the captured line images were stacked on top of each other, forming a 288×520 monochromatic image. The image's x-axis represents a spatial dimension (the lines), while the y-axis represents the temporal dimension (from top to down, the top row being the oldest). See Figure 5 for illustration.

A videokymogram offers apparent benefits by combining real-time imaging feedback found in Videostroboscopy with the advantages of high-speed imaging, particularly the frame rates sufficient to truthfully document each oscillatory cycle of the vocal folds [19].

The manufacturer, unfortunately, discontinued this camera, but its ancestors are commercially available (Kymocam, CYMO, b.v. Groningen, the Netherlands).

For medical examination of vocal cords, the VKG camera is mounted on an appropriate laryngoscope with an adequate light source. During the examination, a physician can position the laryngoscope using the standard (full-image) mode where the larynx is clearly visible. Helping with the positioning of the line scanner is a depiction of the line on the monitor, denoting the exact position of the scanned line [20; 21; 22] (see Figure 6). With this technique, when the patient phonates, the physician can capture the vocal fold vibrations on the VKG camera in the exact position, usually in the place of the highest opening of the vocal folds

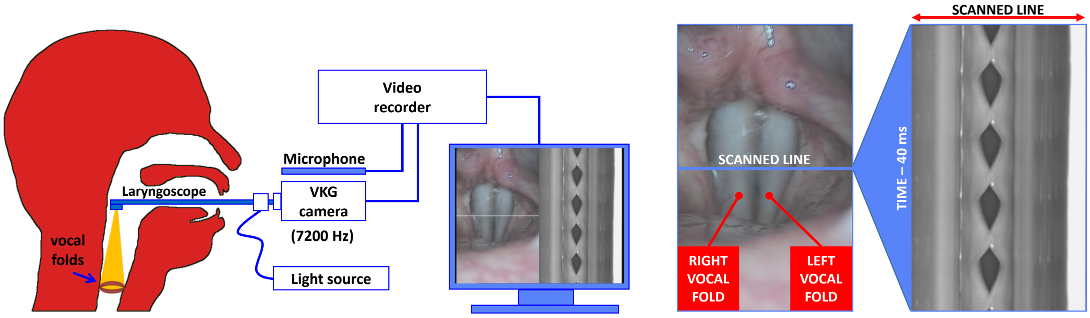


Figure 6: Videokymography: On the left is the examination of vocal fold vibrations by laryngeal endoscopy using a VKG camera. On the right, there are two parallel imaging modes of the VKG camera: standard (left) and videokymographic (right). The videokymographic image is composed of successively acquired scanned lines at the location indicated in the standard mode. The time is mapped onto the vertical axis within the VKG image, going from top to bottom. The standard duration is 40 ms (resulting from the standard rate of 25 frames/s [20; 21]).

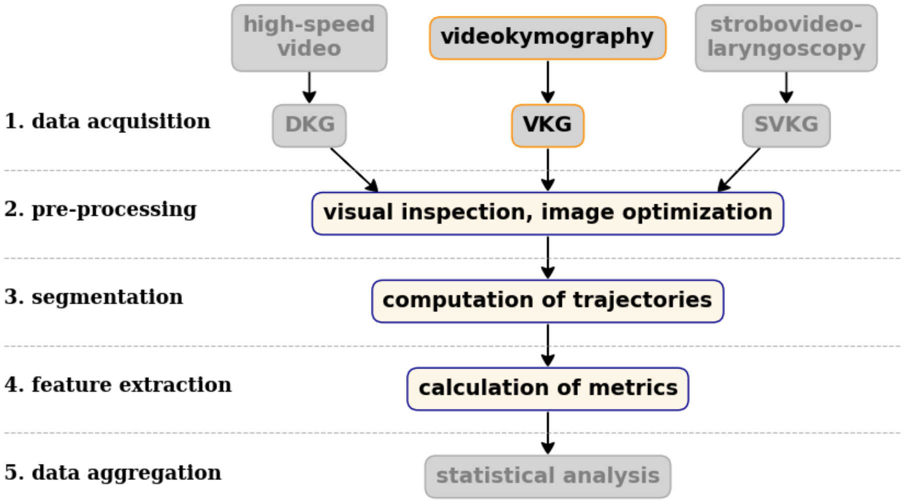


Figure 7: Typical data processing pipeline for kymographic documentation and analysis of vocal fold vibration.

(see Figure 6). A typical examination yields a several-second recording, with a frequency of 25 VKG images per second, producing tens to hundreds of images.

Analyzing all the produced images by hand would take a considerable amount of time; therefore, the VKG data are ideal for automatically extracting the necessary information from the recordings through image processing methods.

2.2 Software

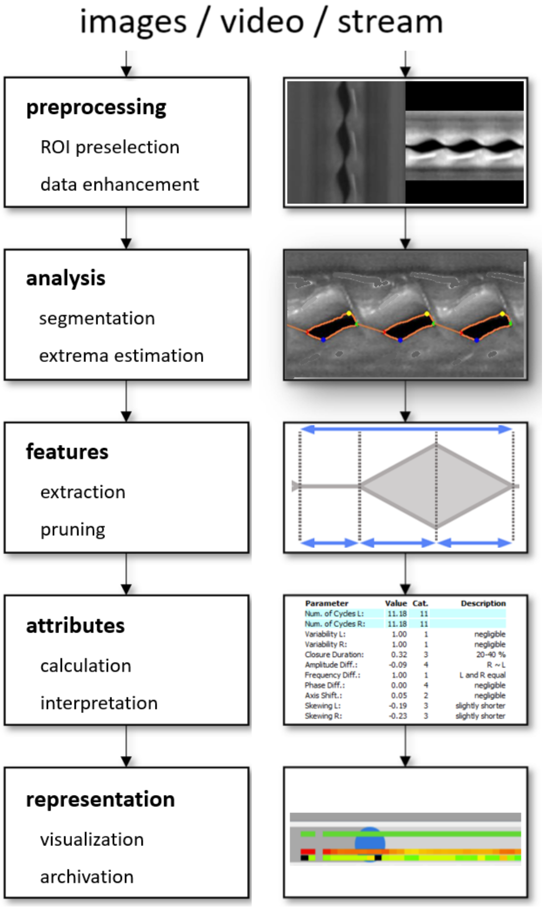
Besides the ability to fast process all the data from an examination, the principal reason behind automatically processing the VKG recordings is the process's objectivity and robustness.

The videokymogram (see Figure 5) is well suited for automatically extracting vibrational features using image processing methods. The main reason for this is the principle from which the VKG image forms. Because one dimension of the VKG image denotes the spatial dimension (the x-axis) and the second dimension is the temporal dimension (the y-axis), the kymogram representation is ideal for spatial-temporal analysis, such as frequency and amplitude estimations. Moreover, although the VKG images are spatial-temporal data, we can still address them using classical image processing algorithms.

Another advantage that plays into the hand of automatic analysis is the high-contrasting feature of the glottal opening (see dark diamond-shaped patches along the middle line of the VKG images in Figure 5). By simplest thresholding methods, one can already achieve crude glottal openings segmentation. From that, extracting the first vibration attributes, such as the number of cycles, the left and the right oscillation frequencies, the amplitudes, the periodicities, the L-R balance, and others, is easy.

A typical design of automated processing of the VKG is as follows (see Figure 7:

1. Acquisition of VKG images using the appropriate hardware.



Parameter	Value	Cat.	Description
Num. of Cycles L:	11.18	11	
Num. of Cycles R:	11.18	11	
Variability L:	1.00	1	negligible
Variability R:	1.00	1	negligible
Closure Duration:	0.32	3	20-40 %
Amplitude Diff.:	-0.09	4	R ~ L
Frequency Diff.:	1.00	1	L and R equal
Phase Diff.:	0.00	4	negligible
Axis Shift.:	0.05	2	negligible
Skewing L:	-0.19	3	slightly shorter
Skewing R:	-0.23	3	slightly shorter

Figure 8: VKG Analyzer Tool [23] processing pipeline schema. The input sequence is processed frame by frame. The first stage focuses on image preprocessing (**layer 2** in Fig.7). Next, the glottal openings are segmented (**layer 3**). The segmentation determines the lateral extrema and opening/closing points. Then the derived vibration features and final attributes are calculated (**layer 4**). Lastly, the software visualizes the results in the graphical user interface.

2. Preprocessing the footage to clean up and enhance the data in preparation for the next steps.

3. Determining the trajectory of each vocal fold.

4. Extraction of the vibrational features.

5. Statistical analysis of gathered data.

The available hardware, its settings, the examining physician, and the patient are the main factors influencing the output quality from the first stage. This stage's sole goal is to acquire the footage in the best quality achievable.

The second stage mainly serves the purpose of preparing data for image processing algorithms so they can perform efficiently. The typical preprocessing methods include de-noising, techniques for improving the contrast of images, processes that modify the content in overexposed places (inpainting algorithms), and similar.

The third stage's goal is to determine the position of the key structures (the vocal folds) in each line of the VKG image. Because the scanner frequency is sufficiently high compared to the frequency of the vocal fold vibrations, we can assume a continuous trajectory for each vocal fold running from the top of the VKG image to the bottom. (For illustration, see Figure 8 - second image on the right.)

The fourth and fifth stages focus on calculating fundamental and derived features from the vocal fold trajectories determined in the third stage.

One of the advantages of the methods working with videokymographic images is, that they can be seamlessly used on images produced by other kymographic techniques that generate the videokymographic images from previously captured recordings [10]: *Digital Kymography* (DKG) [13; 24], operating on HSV recordings and *Strobovideokymography* (SVKG) [25; 26; 27], operating on Videostroboscopic data.

2.3 Research

Because the primary purpose of the VKG technique is medical, and it even overlaps with biophysics, any Videokymography-related research is intrinsically multidisciplinary. That means that published VKG articles are always multi-author publications that typically include image-processing researchers, clinical physicians, biophysicists, and other experts.

The Department of Image Processing at the Institute of Information Theory and Automation has a long history of working with the author who invented the system and patented the hardware. Its researchers have over ten years of experience developing VKG processing methods. Among other activities, the department has cooperated on several studies and developed necessary tools for them, proposed many methods for automatical processing and cleaning the VKG data, produced sophisticated software, which can be used in clinical practice, and created methods for vibrational features extraction and analysis. Until recently, all the proposed techniques were based on conventional image processing methods that used standard, calculated image features. An example of a VKG image processing tool created by our department is the *VKG Analyzer tool* software [23]. Figure 8 depicts the detailed processing pipeline of the product.

One of the goals of this Thesis was to address the same and future VKG processing tasks using a deep-learning approach.

In the next section, I provide a brief literature research and the setting of my work.

3. Related work

This chapter summarizes the published work on Videokymography and automatic systems for processing VKG recordings.

3.1 Beginings

Jan G. Švec's dissertation thesis [28] *Section II: Development and Application of Videokymography* covers the beginnings of the videokymography method with the original publication of the work of J.G. Švec and H.K. Shutte introducing the Videokymographic system [19].

From that time, Videokymography slowly gained the interest of groups of researchers not only from the Czech Republic [19; 28; 29; 30; 31; 32] and the Netherlands [33; 20; 21], where the method originated, but also from Italy [34], China [35; 33; 20; 21], Austria [36], and others.

3.2 Software tools

In 2003, Qiu at al. [33] introduced a quantitative method for eight fundamental parameters (frequency, open and closed quotients, time and amplitude periodicities, phase and amplitude symmetries) of the vocal fold vibration. They based the method on segmenting the glottal gap using an active contour model with a genetic algorithm for further improvement.

In 2012, a team of Italian researchers created an automated software package extracting three selected parameters for diagnosing patients with polyps and cysts and patients submitted to phonosurgical excision of expansile lesions [34]. The parameters were based on left-right vocal fold ratios of amplitude and frequency and open/close quotient.

In the Institute of Information Theory and Automation, Czech Academy of Sciences, the first version of a complex tool for automatic analysis of VKG images was proposed in 2013 [29]. The tool could extract most of the fundamental parameters and vibratory features from the glottal opening segmentation (all the symmetries, frequencies, amplitudes, and quotients) but also included evaluation of the time-varying extent of rima glottidis and the progression of mucosal waves.

In 2017 the cooperation between the Czech Academy of Sciences and the Palacky University in Olomouc in the Czech Republic resulted in the creation of the *Certified Methodology* [31]. The established methodology sets a standard in using Videokymography for diagnostic purposes by defining the discrete categories for each evaluated parameter used in the typical VKG examination. The methodology represents a crucial milestone in the VKG research field, as it helps to alleviate the gap between the analog world of physicians and the digital world of any VKG computerized systems.

3.3 Lateral Peak Sharpness

The sharpness of the lateral peak of the vocal fold trajectory on VKG images is a practical indication of the pliability of the vocal fold tissue. It is, therefore, an important characteristic to evaluate. The computerized approach to lateral peak sharpness estimation from VKG images is as old as the Videokymography itself. But, due to the problem's complexity, only a few teams have addressed this topic.

Jiang et al. in 2000 [37] used an indirect method of peak sharpness estimation by quantifying the vertical phase difference using a sinusoidal model approximation. Although the method is correct in theory, the more complex shapes of the glottal contour are hard to process or interpret.

In 2015, Yamauchi et al. [38] chose a different approach. They defined a *Lateral Peak Index* as an angle formed by two lines between the start of the open phase and the lateral peak; and between the lateral peak and the end of the open phase. They quantified the sharpness of the lateral peak using the defined index. The drawbacks of this approach are that the index is sensitive to unrelated factors and discounts the changes of curvature of the vocal fold waveform that influence peak sharpness.

3.4 Deep Learning in VKG

Deep learning methods influence most current research fields, especially in image processing or computer vision. In many tasks, these modern methods can outperform human experts. It is an implicit next step in the development of VKG processing methods. Some work has been done to use the deep learning approach for the vocal fold diagnosis, but it often relies on directly processing the voice rather than the captured VKG images. For example, in [39], authors proposed a system for detecting various vocal fold diseases through pathological voice recognition using artificial intelligence. Other examples of utilizing deep learning approach for larynx examination and diagnosis work with High-Speed Videoendoscopic recordings as a data source for the processing [40; 41].

To our best knowledge, minimal research has been done using the deep learning approach to process Videokymographic images. Authors in [42] compared a trained convolutional network to a method of fitting synthetic VKG images to the clinical images and achieved competitive results in deriving five of the fundamental vibration parameters.

Using deep learning methods for processing VKG images has great potential. Therefore, we selected it as one of the topics of the Thesis.

4. Goals of the Thesis

Based on the literature search and the demands of otorhinolaryngologists, we formulate the Thesis goals as follows:

- Broaden the current set of methods for automatic evaluation of VKG recordings by proposing new approaches based on user-defined (handcrafted) features. Consolidate all the functionality into a useful software tool to aid laryngologists in vocal fold examination and diagnosis.

- Gain expertise in deep learning methods to be able to apply the acquired knowledge to VKG-specific problems. Develop deep learning solutions for problems with dataset size deficiency to solve VKG problems, which typically lack annotated datasets.

- Use the knowledge gained in deep learning tasks and propose a machine learning system for processing the VKG data to tackle the hard problems in the VKG image processing that are difficult to solve with conventional means.

5. Structure of the Thesis

The Thesis consists of seven papers attached below. Each publication presents the work contributing to declared goals.

The first three publications focus on proposing methods for automatic VKG image evaluation. The following three publications are products of the process of acquiring knowledge about deep-learning techniques with an emphasis on problems with small datasets. Lastly, the final publication represents a successful attempt to apply the gained deep learning expertise on VKG data processing.

The first publication introduces several image processing algorithms for VKG image preprocessing, vocal fold characteristics extraction, and lateral mucosal wave estimations.

The second paper included in the Thesis aims to investigate parameters that can be helpful in objectively quantifying the lateral peak sharpness from the VKG images. The methods introduced in this publication include the estimation of lateral peak sharpness and the algorithm for mucosal wave detection.

The third paper presents a comprehensive software tool for medical praxis to aid physicians with their diagnoses and verifies the robustness of the proposed methods. It follows the previous work in the field and introduces several new approaches. A conducted detailed comparison study supports the robustness of the tool. This paper concludes the first declared goal.

The following three publications are the result of an effort to gain expertise in the field of machine learning with a focus on the cases of small datasets.

The fourth and fifth publications are CLEF conference papers written as a result of our winning the ImageCLEF competitions. Both successes resulted from our understanding of the importance of each class's training dataset statistical distribution and sample representation. The introduced data manipulation techniques, including pseudo-labeling, synthetic datasets, test-time data augmentation, and ensemble methods, proved critical factors in winning the competitions.

The sixth publication focuses on creating an entirely synthetic dataset for a problem where annotated data are completely absent. The main contribution of the conference paper is an engine capable of producing a synthetic video with an incorporated physically accurate motion simulation. The sixth publication concludes the second declared goal.

The ultimate paper covers the last goal by introducing a work connecting both previous endeavors. It utilizes the attained knowledge from the deep learning tasks and applies them to the VKG processing. The resulting work attempts to produce a robust system for mucosal wave lateral peak sharpness estimation achieving a human level of accuracy.

Summaries of the attached papers are in the following sections.

6. Publications

6.1 Paper 1

6.1.1 Citation

A. Novozámský, J. Sedlář, A. Zita, F. Šroubek, J. Flusser, J. G. Švec, J. Vydrová, and B. Zitová, “Image analysis of videokymographic data,” in *2015 IEEE International Conference on Image Processing (ICIP)*, Québec City, Canada, 9 2015, pp. 78–82

6.1.2 Abstract

Videokymography (VKG) is a high-speed medical imaging technique used in laryngology and phoniatrics for the examination of vocal fold vibrations, it offers important characteristics for the diagnosis and treatment of voice disorders. VKG repeatedly scans only a single line from the scene and captures movements of vocal folds in this region of interest. This paper proposes methods for computer-assisted evaluation of diagnostically important vibration features, related to movements of vocal folds and their surroundings. They are derived from existing as well as newly developed methods of digital image processing, mainly based on data segmentation and morphological operations. The performance of the developed methods is compared to expert manual assessments, and it proves to be comparable with clinicians’ conclusions.

6.1.3 Main contribution of the paper

This work introduced a complete set of image processing algorithms for evaluating vocal fold vibrational features from VKG images. In previous attempts, only the most fundamental characteristics (such as amplitude and frequency), usually concerning a particular diagnosis or disease [34; 44], were extracted from the recordings. The set of features proposed in our publication includes: the existence and length of mucosal waves, the left and right variability, left and right phase differences, axis shift, and left and right skewing.

6.1.4 Main contribution of the author

- Mucosal wave detection, tracking, and length estimation methods
- Participation in studies, data evaluation
- Software design and implementation

6.2 Paper 2

6.2.1 Citation

S. P. Kumar, K. V. Phadke, J. Vydrová, A. Novozámský, A. Zita, B. Zitová, and J. G. Švec, "Visual and automatic evaluation of vocal fold mucosal waves through sharpness of lateral peaks in high-speed videokymographic images," *Journal of Voice*, vol. 34, no. 2, pp. 170–178, 2020 (IF=2.06)

6.2.2 Abstract

The sharpness of lateral peaks is a visually helpful clinical feature in high-speed videokymographic (VKG) images indicating vertical phase differences and mucosal waves on the vibrating vocal folds and giving insights into the health and pliability of vocal fold mucosa. This study aims at investigating parameters that can be helpful in objectively quantifying the lateral peak sharpness from the VKG images. Forty-five clinical VKG images with different degrees of sharpness of lateral peaks were independently evaluated visually by three raters. The ratings were compared to parameters obtained by automatic image analysis of the vocal fold contours: Open Time Percentage Quotients (OTQ) and Plateau Quotients (PQ). The OTQ parameters were derived as fractions of the period during which the vocal fold displacement exceeds a predetermined percentage of the vibratory amplitude. The PQ parameters were derived similarly but as a fraction of the open phase instead of a period. The best correspondence between the visual ratings and the automatically derived quotients were found for the OTQ and PQ parameters derived at 95% and 80% of the amplitude, named OTQ95, PQ95, OTQ80 and PQ80. Their Spearman's rank correlation coefficients were in the range of 0.73 to 0.77 ($P < 0.001$) indicating strong relationships with the visual ratings. The strengths of these correlations were similar to those found from inter-rater comparisons of visual evaluations of peak sharpness. The Open time percentage and Plateau quotients at 95% and 80% of the amplitude stood out as the possible candidates for capturing the sharpness of the lateral peaks with their reliability comparable to that of visual ratings.

6.2.3 Main contribution of the paper

In this publication, we proposed a lateral peak sharpness estimation method. Lateral peak sharpness has been recognized as one of the crucial indicators of the vocal folds' healthiness. The work focuses on determining secondary attributes for the indication of sharpness, which can be estimated in a relatively robust and reliable manner. These attributes are variants of the *Plateau Quotients*, defined as the proportion of time during which the vocal fold displacement exceeds R% of vibration amplitude within the open phase (denoted as PQ_R) and the *Open Time Percentage Quotients*, defined as the proportion of time during which the vocal fold displacement exceeds a chosen percentage (R) of the vibration amplitude within a period (denoted as OTQ_R). We have found that the parameters producing the best correlation with the human raters are PQ_{95}, PQ_{80}, OTQ_{95}, and OTQ_{80}. The acquired insight into the relationship between these attributes and the lateral peak sharpness can provide a mechanism for computer algorithms to quantify the sharpness from the VKG images automatically.

6.2.4 Main contribution of the author

- Lateral peak sharpness comparative method based on vocal fold trajectory heuristics and geometrical interpretation
- Study dataset creation and study evaluation

6.3 Paper 3

6.3.1 Citation

A. Zita, A. Novozámský, B. Zitová, M. Šorel, C. T. Herbst, J. Vydrová, and J. G. Švec, "Videokymogram analyzer tool: Human–computer comparison," *Biomedical Signal Processing and Control*, vol. 78, p. 103878, 2022 (IF=3.88)

6.3.2 Abstract

Videokymography (VKG) is a modern video recording technique used in laryngology and phoniatrics to examine vocal fold vibrations. To obtain quantitative information on the vocal fold vibration, VKG image analysis is needed, but no software has yet been validated for this purpose. Here, we introduce a validated software tool that aids clinicians in evaluating diagnostically important vibration characteristics in VKG and other types of kymographic recordings. State-of-the-art methods for automated image evaluation were implemented and tested on a set of videokymograms with a wide range of vibratory characteristics, including healthy and pathologic voices. The automated image segmentation results were compared to the manual segmentation results of six evaluators revealing average differences smaller than one pixel. Furthermore, the automatically categorized vibratory parameters precisely agreed with the average visual assessment in 84 and 91 percent of the cases for pathological and healthy patients, respectively. Based on these results, the newly developed software was found to be a valid, reliable automated tool for quantifying vocal fold vibrations from VKG images, offering several novel features relevant to clinical practice.

6.3.3 Main contribution of the paper

In this paper, we addressed two primary objectives.

Objective I was to develop a software tool for videokymographic data analysis, usable in the clinical praxis. To satisfy this objective, we have created a sophisticated computer program with a user-friendly interface for loading, filtering, visualizing, examining, and processing VKG video streams, VKG video sequences, or VKG still images. Its primary goal is to help a physician diagnose laryngeal diseases. Among other parameters, the analyzer can estimate the vibrational characteristics of each vocal fold, such as frequency variability, opening & closing phase durations, frequency differences, phase differences, phase shift, and skewing. Verifying the robustness and precision of the estimated parameters was a part of Objective II.

Objective II was to present a study validating the performance of the proposed tool, focusing on the glottal segmentation precision and the extracted parameters precision. The first performed study showed good accuracy of the segmentation algorithm, with the mean error of the segmentation border being 0.12±0.79 in the spatial domain (x-axis) and 0.21±1.48 in the temporal domain (y-axis). The second study, which focused on the extracted parameters precision, showed 91% agreement with expert assessments for healthy patients and 84% for patients with various disorders.

6.3.4 Main contribution of the author

- Image preprocessing methods, with a focus on image enhancements
- Attributes estimation methods: Mucosal waves detection and lateral peak sharpness
- Conceptualization and software development (C++), data visualization design and implementation
- Participation in studies, data processing, and evaluation

6.4 Paper 4

6.4.1 Citation

L. Picek, A. Říha, and A. Zita, “Coral reef annotation, localisation and pixel-wise classification using Mask R-CNN and bag of tricks,” in *CEUR Workshop Proceedings : Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*, vol. 2696, no. 83. Thessaloniki, GR: CLEF, 2020 (IS=2.64)

6.4.2 Abstract

This article describes an automatic system for the detection, classification, and segmentation of individual coral substrates in underwater images. It introduces several data preprocessing techniques that eliminate the negative influence of varied dataset acquisition conditions on deep network precision. The proposed system achieved the best performances in both tasks of the second edition of the ImageCLEFcoral competition. Specifically, mean average precision with Intersection over Union (IoU) greater than 0.5 (mAP@0.5) of 0.582 in the case of Coral reef image annotation and localization and mAP@0.5 of 0.678 in Coral reef image pixel-wise parsing. The system is based on Mask R-CNN object detection and instance segmentation framework boosted by advanced training strategies, pseudo-labeling, test-time augmentations, and Accumulated Gradient Normalisation.

6.4.3 Main contribution of the paper

In preparation for incorporating machine learning into medical imaging tasks, we attended a well-known computer vision competition to develop and test our understanding of the effects of insufficient or incomplete data on deep learning performance. This paper covers the winning work of the *ImageCLEF2020* competition in automatic coral reef annotation from underwater images. The dataset for this task had arisen from different places, depths, cameras, and picture quality, which brought quite a challenging condition similar to that of the VKG recordings. We had to use a relatively large number of data augmentation and preprocessing algorithms to achieve the necessary performance. Most algorithms are not novel, but their ensemble was a substantial result of the deeper data understanding. The acquired knowledge about dataset augmentation and normalization as a part of preprocessing was well utilized in the consequent VKG publications.

6.4.4 Main contribution of the author

- Dataset examination, statistical evaluation
- Data augmentation methods
- Testing and evaluation

6.5 Paper 5

6.5.1 Citation

A. Zita, L. Picek, and A. Říha, “Sketch2Code: Automatic hand-drawn UI elements detection with Faster R-CNN,” in *CEUR Workshop Proceedings: Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*, vol. 2696, no. 82. Thessaloniki, GR: CLEF, 2020 (IS=2.64)

6.5.2 Abstract

Using an incomplete dataset for machine learning tasks necessarily yields biased results. We use the DrawnUI ImageCLEF2020 challenge to demonstrate the necessary dataset preparation to support the optimal training results even with incomplete training data. Transcription of User Interface (UI) elements hand drawings to the computer code is a tedious and repetitive task. Therefore, a need arose to create a system capable of automating such a process. This paper describes a deep learning-based method for hand-drawn user interface elements detection and localization. The proposed method scored 1st place in the ImageCLEFdrawnUI competition while achieving an overall precision of 0.9708. The final method is based on the Faster R-CNN object detector framework with ResNet-50 backbone architecture trained with advanced regularization techniques.

6.5.3 Main contribution of the paper

In this task, we tested our insight into the necessary statistical properties of the training dataset input space and distribution of the data. This paper covers the winning work of the ImageCLEF2020 competition in the automatic transcription of hand-drawn user interface design to the programming language. The poor distribution of the dataset classes was the most challenging part of the introduced problem. After a thorough data examination, we had to manually split the data into the deep network's training and evaluation datasets to ensure the proper division of classes with minimal samples. Furthermore, we supplemented the missing data in datasets with synthetically generated data. For that, we created an engine for constructing artificial images of hand-drawn user interfaces to even the dataset class distribution. The experience and knowledge gained about dataset distribution effect on the system performance is directly transferable to future VKG deep learning classification problems and was influential in our effort to use deep learning methods for VKG data processing.

6.5.4 Main contribution of the author

- Dataset examination, statistical evaluation
- Dataset classes distribution augmentation
- Synthetic data generator design and implementation
- Data augmentation and normalization
- Testing and evaluation

6.6 Paper 6

6.6.1 Citation

A. Zita and F. Šroubek, "Tracking fast moving objects by segmentation network," in *2020 25th International Conference on Pattern Recognition (ICPR).* IEEE, 2021, pp. 10 312–10 319 (IS=1.05)

6.6.2 Abstract

Tracking Fast Moving Objects (FMO), which appear as blurred streaks in video sequences, is a difficult task for standard trackers, as the object position does not overlap in consecutive video frames, and the texture information of the objects is blurred. Up-to-date approaches tuned for this task are based on background subtraction with a static background and slow deblurring algorithms. This article presents a tracking-by-segmentation approach implemented using modern deep learning methods that perform near real-time tracking on real-world video sequences. We have developed a physically plausible FMO sequence generator to be a robust foundation for our training pipeline and demonstrate straightforward network adaptation for different FMO scenarios with varying foregrounds.

6.6.3 Main contribution of the paper

We adopt a definition of an FMO (a fast-moving object) stating that an FMO is defined as an object captured on a video sequence so that the object boundaries do not overlap on consecutive images. Although this definition allows it, it does not consider the fast movement of the camera but rather the fast movement of an actual object in the scene having a non-FMO background. Tracking such objects using the existing tracking algorithm fails on the absent texture and missing overlaps.

This paper presents the FMO tracking system based on the tracking-by-segmentation principle. We proposed a deep learning architecture capable of detecting and segmenting FMO in the video sequence. We consequently passed the detected object's positional information to a conventional tracker.

Because no complete annotated dataset for the FMO existed, we proposed a runtime, physically correct FMO data synthesizer. The synthesizer can produce a sequence of images in time for each network training iteration. The synthesizer takes a foreground image representing the object and the background video sequence. To train the network for sports scenes, we used images of different sports balls as a foreground and random video sequences from youtube as a background.

The resulting system can track balls of several spots, such as squash, baseball, tennis, and table tennis.

This research task was essential in understanding the sensitivity of a deep learning system to the quality of generated data. It confirms the importance of sufficient input probability space sampling for the system to learn and generalize well; otherwise, the deep network would quickly overfit the limited training data.

6.6.4 Main contribution of the author

- Neural network architecture
- Procedural synthetic data
- Data gathering and dataset creation
- Testing and evaluation

6.7 Paper 7

6.7.1 Citation

A. Zita, Š. Greško, A. Novozámský, M. Šorel, B. Zitová, J. G. Švec, and J. Vydrová, "Automatic estimation of mucosal waves lateral peak sharpness - modern approach," in *Image Processing: Algorithms and Systems XXI.* Electronic Imaging Symposium, 2023, accepted for oral presentation (IS=1.05 in 2021)

6.7.2 Abstract

Videokymographic (VKG) images of the human larynx are often used for automatic vibratory feature extraction for diagnostic purposes. One of the most

challenging parameters to evaluate is the mucosal wave's presence and its lateral peaks' sharpness. Although these features can be clinically helpful and give an insight into the health and pliability of vocal fold mucosa, the identification and visual estimation of the sharpness can be challenging for human examiners and even more so for an automatic process. This work aims to create and validate a method that can automatically quantify the lateral peak sharpness from the VKG images using a convolutional neural network.

6.7.3 Main contribution of the paper

This publication is the first attempt to converge all the experience and knowledge gained in previous activities, such as deep insight into videokymographic examination techniques and important vibrational attributes, deep learning principles, network architectures, and working with small datasets.

Based on the acquired expertise, we presented a system for an automatic evaluation of mucosal waves from VKG images using deep learning techniques. In particular, we focused on detecting the presence of mucosal waves and estimating the sharpness of the glottal contour lateral peak because previous attempts to extract these attributes by conventional means have proven to be problematic. Additionally, the training dataset consisted of VKG glottal cycles of a limited number of patients, so the data were highly correlated. Again, we could rely on our experience from previous machine learning tasks to sufficiently alleviate the dataset deficiencies by proper preprocessing.

Ultimately, we have achieved a human-expert level of precision and demonstrated the usability of machine-learning systems for automatically processing Videokymograms. The proposed system's performance is good, and the evaluation runs in a superior time compared to previous systems based on a slow cross-correlation technique.

6.7.4 Main contribution of the author

- Architecture proposal
- Data gathering and dataset creation
- Testing and evaluation

7. Main contribution of the Thesis

The main contribution of this thesis is the development of conventional VKG processing methods and the original application of deep learning techniques to the VKG data.

In particular, in the first publication, we broaden the current methods and set a standard for the VKG examination by selecting and processing a complete set of attributes and vibrational characteristics useful for medical examination. Next, in the second paper, we focused on one more, hard to automatically estimate attribute: The Lateral Peak Sharpness. This attribute is a crucial indicator of the general health of the vocal folds because it reflects the vocal fold tissue pliability. The proposed method is an indirect estimation of this parameter using secondary parameters, which are easier to evaluate. Ultimately, we introduced a complex VKG Analyzer software, which physicians can use directly to help them with patient examination and diagnosis. The program uses algorithms published in the first paper and introduces several new methods. This work is covered by the third publication, which also includes two important and comprehensive studies verifying the performance and accuracy of the proposed system.

During the next phase, we proposed deep learning solutions to several problems, typical with their limited datasets, to gain expertise in the matter. This work is covered by the following three publications, two of which are winning solutions to two different ImageCLEF2020 competitions. The third is a solution to the complex problem of tracking fast-moving objects from the video.

In the concluding paper, we introduced a deep learning method of directly estimating the lateral peak sharpness attribute, which is hard to obtain using a conventional approach. We verified the method's performance by a study confirming the human-level accuracy of the proposed system.

Bibliography

[1] J. Vydrová, J. G. Švec, and F. Šram, “Videokymography (vkg) in laryngologic practice,” *J Macrotrends Health Med*, vol. 3, pp. 87–95, 2015.

[2] Reinhard-commonswiki, “”file:vocal fold animated.gif”,” 2007, [Online; accessed 10-November-2022]. [Online]. Available: ”https://upload.wikimedia.org/wikipedia/commons/e/eb/Vocal_fold_animated.gif”

[3] Olympus, “Images:rigid laryngoscpoe,” [Online; accessed 10-December-2022]. [Online]. Available: https://medical.olympusamerica.com/products/rigid-laryngoscope

[4] D. M. Bless, R. Patel, and N. Connor, “Laryngeal imaging: stroboscopy, high-speed digital imaging, and kymography,” *The Larynx*, vol. 1, pp. 181–210, 2009.

[5] P. Woo, “Stroboscopy and high-speed video examination of the larynx,” in *Sataloff's Comprehensive Textbook of Otolaryngology: Head & Neck Surgery: Laryngology*, R. T. Sataloff and M. S. Benninger, Eds., vol. 4. JP Medical Ltd, 2015, p. 193.

[6] D. Deliyski, “Laryngeal high-speed videoendoscopy,” in *Laryngeal Evaluation: Indirect Laryngoscopy to High-Speed Digital Imaging*, K. A. Kendall and R. J. Leonard, Eds. Thieme Medical, New York, 2010, pp. 245–270.

[7] J. Lohscheller, U. Eysholdt, H. Toy, and M. Dollinger, “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*, vol. 27, no. 3, pp. 300–309, 2008.

[8] W. Chen, P. Woo, and T. Murry, “Vocal fold vibratory changes following surgical intervention,” *Journal of Voice*, vol. 30, no. 2, pp. 224–227, 2016.

[9] K. Kendall, K. A. Kendall, and R. Leonard, *Laryngeal Evaluation: Indirect Laryngoscopy to High-speed Digital Imaging*, ser. Thieme Publishers Series. Thieme, 2010. [Online]. Available: ”https://books.google.cz/books?id=ud6TNAEACAAJ”

[10] J. G. Švec and H. K. Schutte, “Kymographic imaging of laryngeal vibrations,” *Current opinion in otolaryngology & head and neck surgery*, vol. 20, no. 6, pp. 458–465, 2012.

[11] D. D. Deliyski, P. P. Petrushev, H. S. Bonilha, T. T. Gerlach, B. Martin-Harris, and R. E. Hillman, “Clinical implementation of laryngeal high-speed videoendoscopy: challenges and evolution,” *Folia Phoniatrica et Logopaedica*, vol. 60, no. 1, pp. 33–44, 2008.

[12] S. Hertegård, H. Larsson, and T. Wittenberg, “High-speed imaging: applications and development,” *Logopedics Phoniatrics Vocology*, vol. 28, no. 3, pp. 133–139, 2003.

[13] T. Wittenberg, M. Tigges, P. Mergell, and U. Eysholdt, "Functional imaging of vocal fold vibration: digital multislice high-speed kymography," *Journal of Voice*, vol. 14, no. 3, pp. 422–442, 2000.

[14] J. Lohscheller, U. Eysholdt, H. Toy, and M. Döllinger, "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*, vol. 27, no. 3, pp. 300–309, 2008.

[15] S.-Z. Karakozoglou, N. H. Bernardoni, C. d'Alessandro, and Y. Stylianou, "Automatic glottal segmentation using local-based active contours," in *9th International Conference on Advances in Quantitative Laryngology, Voice and Speech Research (AQL 2010)*, 2010.

[16] G. Andrade-Miranda, Y. Stylianou, D. D. Deliyski, J. I. Godino-Llorente, and N. Henrich Bernardoni, "Laryngeal image processing of vocal folds motion," *Applied Sciences*, vol. 10, no. 5, p. 1556, 2020.

[17] C. A. Rosen, "Stroboscopy as a research instrument: development of a perceptual evaluation tool," *Laryngoscope*, vol. 115, no. 3, pp. 423–428, Mar 2005.

[18] K. V. Phadke, "Selected topics in laryngeal, perceptual and acoustic assessments of human voice: Videokymographic evaluations of vocal folds and investigations of teachers' voices. Doctoral Thesis, Palacky University, Olomouc, Czech Republic." Doctoral Dissertation, Palacky University Olomouc, Olomouc, Czech Republic, 2018.

[19] J. G. Švec and H. K. Schutte, "Videokymography: high-speed line scanning of vocal fold vibration," *Journal of Voice*, vol. 10, no. 2, pp. 201–205, 1996.

[20] Q. Qiu and H. K. Schutte, "A new generation videokymography for routine clinical vocal-fold examination," *Laryngoscope*, vol. 116, no. 10, pp. 1824–1828, 2006.

[21] Q. Qiu and H. K. Schutte, "Real-time kymographic imaging for visualizing human vocal-fold vibratory function," *Review of Scientific Instruments*, vol. 78, no. 2, p. 024302, 2007.

[22] J. G. Švec and F. Šram, "Videokymographic examination of voice," in *Handbook of Voice Assessments*, 3rd ed., E. P. M. Ma and E. M. L. Yiu, Eds. San Diego, CA: Plural Publishing, 2011, pp. 129–146.

[23] A. Zita, A. Novozámský, B. Zitová, M. Šorel, C. T. Herbst, J. Vydrová, and J. G. Švec, "Videokymogram analyzer tool: Human–computer comparison," *Biomedical Signal Processing and Control*, vol. 78, p. 103878, 2022.

[24] A. Yamauchi, H. Imagawa, H. Yokonishi, K.-I. Sakakibara, and N. Tayama, "Multivariate analysis of vocal fold vibrations in normal speakers using high-speed digital imaging," *Journal of Voice*, 2021.

[25] Y. Isogai, "Analysis of the vocal fold vibration by the laryngo-strobography-improvements of the analytic function," *Larynx Jpn*, vol. 8, pp. 27–32, 1996.

[26] M. W. Sung, K. H. Kim, T. Y. Koh, T. Y. Kwon, J. H. Mo, S. H. Choi, J. S. Lee, K. S. Park, E. J. Kim, and M. Y. Sung, “Videostrobokymography: a new method for the quantitative analysis of vocal fold vibration,” *Laryngoscope*, vol. 109, no. 11, pp. 1859–1863, 1999.

[27] P. Krasnodebska, A. Szkiełkowska, B. Miaśkiewicz, and H. Skarżyński, “Characteristics of euphony in direct and indirect mucosal wave imaging techniques,” *Journal of Voice*, vol. 31, no. 3, pp. 383–e13, 2017.

[28] J. G. Švec, “On vibration properties of human vocal folds: voice registers, bifurcations, resonance characteristics, development and application of videokymography,” Doctoral Dissertation, University of Groningen, Groningen, the Netherlands, 2000.

[29] A. Novozámský, J. Sedlář, A. Zita, C. T. Herbst, J. G. Švec, B. Zitová, and J. Flusser, “VKFD: Computerized analysis of videokymographic data,” in *PEVOC - Pan Europian Voice Conference*, T. Domagalský, Ed., vol. 10, 2013, pp. 293–294.

[30] B. Zitová, A. Novozámský, A. Zita, M. Šorel, J. G. Švec, and J. Vydrová, “VKGanalyzer: objective analysis of vocal fold vibrations,” in *Artificial Intelligence in Medicine 2017*, vol. 2017, 2017.

[31] J. Vydrová, J. G. Švec, B. Zitová, A. Novozámský, A. Zita, and M. Šorel, “(19-12-2017) methodology of evaluation of voice disorders from videokymographic recordings. (certified methodology no. 1170896),” Certificate of Czech Electrotechnical Testing Institute, Prague, 12 2017.

[32] J. G. Švec, S. P. Kumar, K. V. Phadke, J. Vydrová, A. Novozámský, A. Zita, and B. Zitová, “Evaluation of mucosal waves through sharpness of lateral peaks in videokymographic images,” in *Pan-European Voice Conference 2019*, I. Jenny and T. S. Løvind, Eds., 2019, pp. 90–90.

[33] Q. Qiu, H. K. Schutte, L. Gu, and Q. Yu, “An automatic method to quantify the vibration properties of human vocal folds via videokymography,” *Folia Phoniatr Logop*, vol. 55, no. 3, pp. 128–136, 2003.

[34] C. Piazza, S. Mangili, F. Del Bon, F. Gritti, C. Manfredi, P. Nicolai, and G. Peretti, “Quantitative analysis of videokymography in normal and pathological vocal folds: a preliminary study,” *European Archives of Oto-Rhino-Laryngology*, vol. 269, no. 1, pp. 207–212, 2012.

[35] J. J. Jiang, Y. Zhang, M. P. Kelly, E. T. Bieging, and M. R. Hoffman, “An automatic method to quantify mucosal waves via videokymography,” *Laryngoscope*, vol. 118, no. 8, pp. 1504–1510, 2008.

[36] C. T. Herbst, “DKG plugin for FIJI,” Available: www.christian-herbst.org (last accessed: August 18, 2014).

[37] J. J. Jiang, C. I. Chang, J. R. Raviv, S. Gupta, F. M. Banzali, and D. G. Hanson, “Quantitative study of mucosal wave via videokymography in canine larynges,” *Laryngoscope*, vol. 110, pp. 1567–1573, 2000.

[38] A. Yamauchi, H. Yokonishi, H. Imagawa, K.-I. Sakakibara, T. Nito, N. Tayama, and T. Yamasoba, “Quantitative analysis of digital videokymography: A preliminary study on age- and gender-related difference of vocal fold vibration in normal speakers,” 2015. [Online]. Available: 'http://dx.doi.org/10.1016/j.jvoice.2014.05.006'

[39] H.-C. Hu, S.-Y. Chang, C.-H. Wang, K.-J. Li, H.-Y. Cho, Y.-T. Chen, C.-J. Lu, T.-P. Tsai, O. K.-S. Lee *et al.*, “Deep learning application for vocal fold disease prediction through voice recognition: preliminary development study,” *Journal of Medical Internet Research*, vol. 23, no. 6, p. e25247, 2021.

[40] A. M. Yousef, D. D. Deliyski, S. R. Zacharias, A. de Alarcon, R. F. Orlikoff, and M. Naghibolhosseini, “A deep learning approach for quantifying vocal fold dynamics during connected speech using laryngeal high-speed videoendoscopy,” *Journal of Speech, Language, and Hearing Research*, pp. 1–16, 2022.

[41] A. M. Yousef, D. D. Deliyski, S. R. Zacharias, and M. Naghibolhosseini, “Deep-learning-based representation of vocal fold dynamics in adductor spasmodic dysphonia during connected speech in high-speed videoendoscopy,” *Journal of Voice*, 2022.

[42] S. Bulusu, S. Kumar, J. Švec, and P. Aichinger, “Neural-network-based estimation of vocal fold kinematic parameters from digital videokymograms,” *AQL 2021 BOOK OF ABSTRACTS*, p. 25.

[43] A. Novozámský, J. Sedlář, A. Zita, F. Šroubek, J. Flusser, J. G. Švec, J. Vydrová, and B. Zitová, “Image analysis of videokymographic data,” in *2015 IEEE International Conference on Image Processing (ICIP)*, Québec City, Canada, 9 2015, pp. 78–82.

[44] C. Manfredi, L. Bocchi, G. Cantarella, and G. Peretti, “Videokymographic image processing: objective parameters and user-friendly interface,” *Biomedical Signal Processing and Control*, vol. 7, no. 2, pp. 192–201, 2012.

[45] S. P. Kumar, K. V. Phadke, J. Vydrová, A. Novozámský, A. Zita, B. Zitová, and J. G. Švec, “Visual and automatic evaluation of vocal fold mucosal waves through sharpness of lateral peaks in high-speed videokymographic images,” *Journal of Voice*, vol. 34, no. 2, pp. 170–178, 2020.

[46] L. Picek, A. Říha, and A. Zita, “Coral reef annotation, localisation and pixel-wise classification using Mask R-CNN and bag of tricks,” in *CEUR Workshop Proceedings : Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*, vol. 2696, no. 83. Thessaloniki, GR: CLEF, 2020.

[47] A. Zita, L. Picek, and A. Říha, “Sketch2Code: Automatic hand-drawn UI elements detection with Faster R-CNN,” in *CEUR Workshop Proceedings: Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*, vol. 2696, no. 82. Thessaloniki, GR: CLEF, 2020.

[48] A. Zita and F. Šroubek, “Tracking fast moving objects by segmentation network,” in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 10 312–10 319.

[49] A. Zita, Š. Greško, A. Novozámský, M. Šorel, B. Zitová, J. G. Švec, and J. Vydrová, “Automatic estimation of mucosal waves lateral peak sharpness - modern approach,” in *Image Processing: Algorithms and Systems XXI*. Electronic Imaging Symposium, 2023, accepted for oral presentation.

Fulltext papers

IMAGE ANALYSIS OF VIDEOKYMOGRAPHIC DATA

Adam Novozámský[a], *Jiři Sedlář*[a], *Aleš Zita*[a], *Filip Šroubek*[a], *Jan Flusser*[a],
Jan G. Švec[b], *Jitka Vydrová*[c], *Barbara Zitová*[a] *

[a]Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic, Prague, Czech Republic
{novozamsky,sedlar,zita,zitova,sroubekf, flusser}@utia.cas.cz
[b]Voice Research Lab, Department of Biophysics
Faculty of Sciences, Palacký University, Olomouc, Czech Republic
svecjang@gmail.com
Voice Centre Prague, Medical Healthcom, Ltd., Czech Republic
vydrova@medico.cz

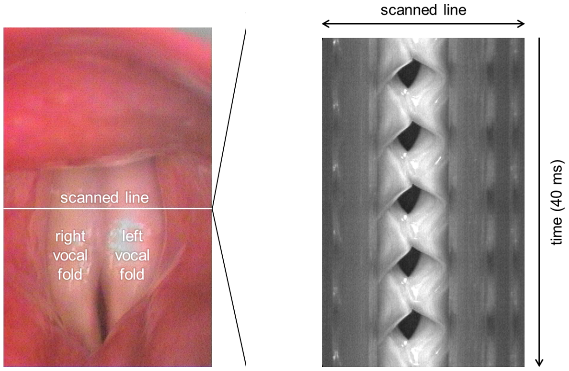
ABSTRACT

Videokymography (VKG) is a high-speed medical imaging technique used in laryngology and phoniatrics for examination of vocal fold vibrations, it offers important characteristics for diagnosis and treatment of voice disorders. VKG repeatedly scans only a single line from the scene and captures movements of vocal folds in this region of interest. This paper proposes methods for computer assisted evaluation of diagnostically important vibration features, related to movements of vocal folds and their surroundings. They are derived from existing as well as newly developed methods of digital image processing, mainly based on data segmentation and morphological operations. Performance of the developed methods is compared to expert manual assessments and it proves to be comparable with clinicians conclusions.

Index Terms— videokymography, medical imaging, data segmentation

1. INTRODUCTION

Digital image processing methods form an integral part of medical data analysis and evaluation. Our paper addresses an analysis of videokymographic data (videokymograms), which are collected using videokymography (VKG) - a high-speed imaging technique convenient for observation of vocal fold vibrations. VKG is used in laryngology and phoniatrics for diagnosis of vibration parameters of vocal folds. Our aim is to complement visual evaluation of videokymograms, which can be tedious and clinician-dependent. We proposed automatic software tools for VKG preprocessing and detection of important features.



(a) standard mode (b) VKG mode

Fig. 1. Two modes of videokymographic camera data acquisition: (a) standard and (b) videokymographic. The videokymogram (b) is composed of successively acquired scanned lines at the location indicated in (a).

There are several techniques of capturing human vocal fold vibrations for assessment of their functionality. The most commonly used are the videostroboscopy, high-speed videoendoscopy and the latest, videokymography (VKG). The VKG is an original Czech-Dutch method, developed in 1994 in Groningen (NL) as an alternative to high-speed video recording [1]. The system consists of specially adapted CCD camera, which operates in two modes - standard (50 fps - interlaced, Figure 1(a)) and in high speed (currently 7200 lines per second), when the system records images of a single horizontal line of the selected camera row and stacks them below each other (Figure 1 (b)). The method allows efficient recording of vibration patterns of vocal folds. An application of digital image processing methods for analysis of vocal fold

*The work was supported by the Technology Agency of the Czech Republic under the project no TA04010877 and by GACR agency project GA13-29225S.

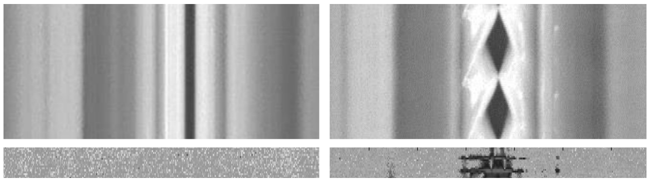


Fig. 2. The VKG data without (top left) and with (top right) an activity. Respective representations in the column-wise Fourier spectral domain close-ups are shown in the bottom line.

data have been attracting an attention for some time. Most of them have been oriented on high speed videoendoscopy [2], while VKG attracted lesser interest [3], eventhough this modality offers many benefits due to its efficient data and vocal fold characteristics representation [4]. Our approach broaden proposed methodology for VKG data and make them more robust to VKG data variability. To facilitate the evaluation of VKG data we focus ourselves on three phases of VKG analysis: (I) data preprocessing, (II) vocal folds characteristics extraction and (III) auxilliary features extraction.

2. VKG PREPROCESSING

Typical videokymogram is a gray-scale image capturing several openings and closings of vocal folds (see Figure 1(b)). The data can be noisy, with low contrast and with reflections caused by present mucus. All these factors can negatively influence the performance of further software analysis tools. Moreover, due to the manual examination procedure of the data acquisition, the laryngoscope can be randomly shifted from the optimal position. An important factor which influences the data examination is the patient's discontinued phonation, when the vibrations are missing in VKG data at all.

To ensure the best possible outcome of the automatic analysis of VKG data we apply data preprocessing steps such as median denoising, locally-adaptive contrast enhancement and, if needed, mucus reflection removal using adaptive thresholding followed by diffusion inpainting. The effect of unexpected patient movements and his discontinued phonation is handled by selection of meaningful data subsequences only. This method was developed to select only these parts of VKG recording where the vocal folds are approximately in the center of the VKG image and are active. The Fourier transform of fixed width columns is analyzed and these VKGs with the spectral response under the given threshold are omitted from further processing (see Figure 2).

The attention is paid also to VKG data taken in the standard mode. They can be blurred due to the wrong camera focus and movements of the patient. We proposed to apply multichannel blind deconvolution method [6] to improve the sharpness of the data, even that they are only for visual inspection and are not used in the further automatic analysis.

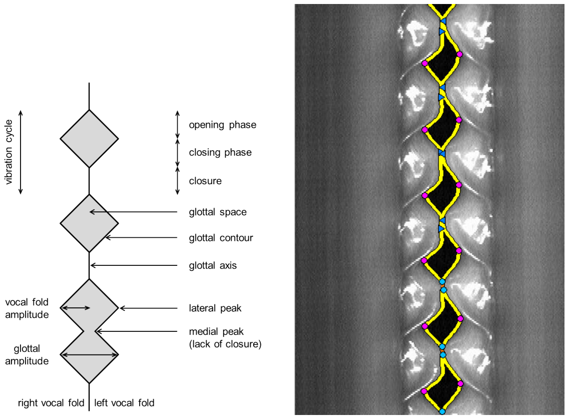


Fig. 3. (left) - Vibration features in videokymograms. (right) - The shape of glottal space and detected base glottal features – opening and closing points (light blue), medial peaks (dark blue), and lateral peaks (magenta).

feature	notation
opening points	O_i
closing points	C_i
lateral peaks	A_i^R, A_i^L
medial peaks	M_i^R, M_i^L

Table 2. Base glottal features in videokymograms (see Figure 3); upper indices R and L denote the right and left vocal folds, respectively, and lower index i denotes the number of the vibration cycle in the videokymogram.

3. VOCAL FOLD CHARACTERISTICS

Proposed methods for analysis of VKG data are based on vocal folds / glottal contours and detected base features (see Figure 3) which are key elements for computation of established vibration parameters [5]. The primary step for all further VKG evaluation is the detection of glottal contour (Figure 3 - (right), yellow curves), which is realized by means of an thresholding segmentation with an optimized threshold estimated by normalized graph cuts [7, 8]. This approach maximizes dissimilarity between two parts of the scene according to both spatial and gray-level relations of their pixels. The detected glottal space is then used for estimation of elementary glottal features - opening and closing points, lateral and medial peaks, vibration cycles and their opening, closing, open and closed phases, and glottal and vocal fold amplitudes (see Figure 3). The opening and closing points, and the lateral and medial peaks (see Table 2) are the base features and are used for derivation of the other aforementioned features called derived glottal features. Their derivation from the base features is listed in Table 1. They all are used for computation of es-

feature	notation and definition
generalized opening points	$\tilde{O}_i^j = \{O_i, M_i^j\}$
generalized closing points	$\tilde{C}_i^j = \{C_i, M_i^j\}$
opening phase duration	$t_i^{oj} = A_i^j(y) - \tilde{O}_i^j(y)$
closing phase duration	$t_i^{cj} = \tilde{C}_i^j(y) - A_i^j(y)$
open phase duration	$T_i^{oj} = t_i^{oj} + t_i^{cj} = \tilde{C}_i^j(y) - \tilde{O}_i^j(y)$
closed phase duration	$T_i^{cj} = \tilde{O}_{i+1}^j(y) - \tilde{C}_i^j(y)$
vibration cycle duration	$T_i^j = T_i^{oj} + T_i^{cj} = t_i^{oj} + t_i^{cj} + T_i^{cj} = \tilde{O}_{i+1}^j(y) - \tilde{O}_i^j(y)$
vocal fold amplitudes	$a_i^j = \max(\lvert A_i^j(x) - \tilde{O}_i^j(x)\rvert, \lvert A_i^j(x) - \tilde{C}_i^j(x)\rvert)$
glottal amplitudes	$a_i = A_i^L(x) - A_i^R(x)$

Table 1. Derived glottal features in videokymograms [5]; upper index $j \in \{R, L\}$ denotes the right and left vocal folds, respectively, and lower index i denotes the number of the vibration cycle in the videokymogram.

tablished vocal fold vibration parameters [5]. Their detailed definition and discussion can be found in [9].

The proposed base and derived glottal features and set of vocal fold vibration parameters [9, 5] were evaluated on the testing dataset of 50 videokymograms and compared to manual evaluations [10], done by evaluators with different level of experience and resulting in 18 assessment sets in total (18 × 50 videokymograms). The comparison was realized in an automatic–visual and visual–visual manner. The method introduces the following notation. Let P denotes the set of both automatically and visually evaluated parameters, n the number of evaluated videokymograms, and m the number of visual evaluations. Let $E_A(p; i)(p \in P; i = 1; ...; n)$ denotes the result of automatic evaluation of parameter p in videokymogram i, and $E_V(p; i; j)(p \in P; i = 1; ...; n; j = 1; ...; m)$ the result of j^{th} visual evaluation of parameter p in videokymogram i. Let $V^+(p; i)$ denote set of indices of visual evaluations of parameter p in videokymogram i with defined result (non-NA)

$$V^+(p; i) = \{j \mid j \in \{1...m\} \wedge E_V(p; i; j) > 0\}$$

Then the consensus result is defined as NA if and only if the result of at least half of corresponding visual evaluations was NA; otherwise, the definition estimates it by the most frequent non-NA result. For each parameter $p \in P$ and videokymogram $i \in \{1, ..., n\}$ the proposed method compares the consensus result of visual evaluations $E_V(p; i)$ with the result of automatic evaluation $E_A(p; i)$ (automatic–visual match) and with the results of visual evaluations $E_V(p; i; j), (j = 1, ..., n)$ (visual–visual match).

The automatic–visual match compared the automatically estimated parameter categories with the category most frequently selected by the visual evaluators whereas the visual–visual match estimated reliability of the visual evaluations (how often the assessments of the visual evaluators were in agreement). In all cases two evaluations are set to be matching if their respective results fall into the same category or into directly neighboring non-NA categories. The results can be seen in Table 3. Figure 5 illustrates the variability of the VKG data that the proposed algorithms must to be able to cope.

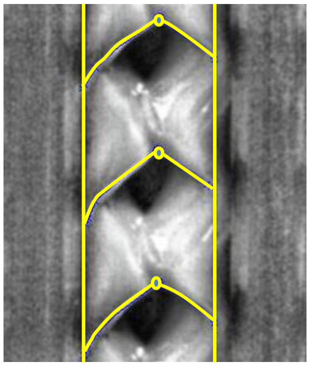


Fig. 4. Detected lateral mucosal waves with the starting points (circles) and their detected extent.

The experiments showed consistency between automatic and visual evaluations. The similarity in comparative statistics demonstrates that the performance of the automatic evaluation is comparable with visual evaluations and thus the proposed approach in the computer-aided evaluation is found applicable in clinical practice.

4. LATERAL MUCOSAL WAVES EXTRACTION

Besides of the established set of vibratory features [5] the research was focused on the auxiliary characteristics of vocal chords and their vibrations – laterally travelling mucosal waves. Mucosal waves are tissue waves propagating across located on the upper surface of vocal folds. They propagate laterally across the surface until they disappear or reaches the lateral border of the vocal fold. Their presence and extent can indicate how pliable a vocal fold is and can may indicate problems with stiffness of vibrating tissue.

In videokymograms, mucosal waves are demonstrated as diagonal, sometimes slightly bended lines on vocal folds running in the direction of the opening movement. Their detec-

vibration parameter	automatic–visual match	visual–visual match
NumberOfCyclesR	98%	95%
NumberOfCyclesL	98%	96%
VariabilityR	92%	93%
VariabilityL	88%	93%
ClosureDuration	98%	93%
AmplitudeDifferences	100%	88%
FrequencyDifferences	100%	95%
PhaseDifferences	88%	85%
AxisShift	88%	79%
SkewingR	86%	87%
SkewingL	90%	85%

Table 3. Comparison of results of automatic and visual evaluations on a set of 50 videokymograms by the automatic–visual and visual–visual match with tolerance between closely neighboring classes. The similarity in comparative statistics for each parameter indicates that the performance of the automatic evaluation is comparable and often better than visual evaluations.

tion can be complicated by reflections, low contrast and their small extent. To solve the problem we introduced the *iterated masked cross-correlation* method. It is based on the detection of self-similarity of the data around the expected position of the wave, which starts at lateral peaks and runs in the direction of the connector of the opening points and the lateral peaks.

The respective cross-correlation kernel is established, positioned at the beginning of the wave and shaped as a tilted rectangle with its size proportional to the glottal space. The kernel is then iteratively updated as the cross-correlation is processed in the given direction, till any progress is made or the steps are shorten below certain level. In this way an approximate path of a glottal wave is constructed and its main direction is then estimated using Fourier spectral analysis of the detected wave path. In Figure 4 there is an illustration of the VKG data with the detected directions and extent of the lateral mucosal waves (circles are representing starting points of the method).

5. CONCLUSION

New image processing methods for analysis of vocal fold vibrations in videokymograms were developed. The motivation was to create the tools for the computer-aided evaluation of vibratory patterns to be used by laryngologists and phoniatricians in clinical practice for more detailed diagnosis of voice disorders. The introduced algorithms provide data preprocessing, detection of the glottal space by normalized graph cuts thresholding and extraction of glottal vibration features and parameters proposed in [5]. Comparison of the performance of the developed methods with subjective visual evaluations indicated good match making them promising for future use in clinical practice. In addition to the glottal features an original methodology was also developed for detecting laterally travelling mucosal waves. These features which are revealing on the pliability of the vocal folds and their surrounding are detected using *iterated masked cross-correlation* method. In the future the research will be focused on new sets of auxiliary features describing the close surrounding of vocal folds and their interpretability with respect to the diagnosis and treatment of voice disorders.

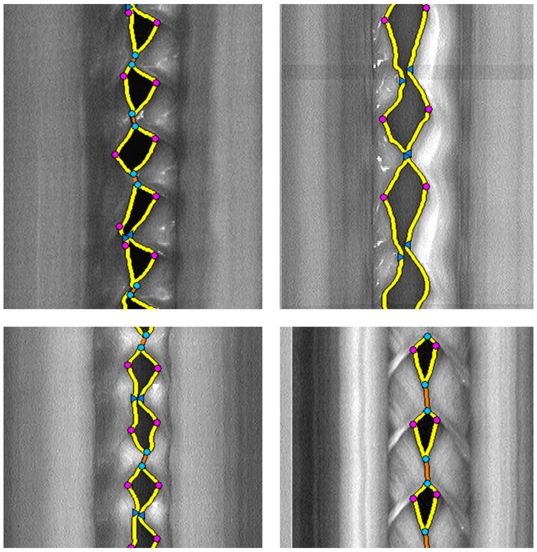


Fig. 5. The variability of the VKG data with detected features.

6. REFERENCES

[1] J. G. Svec and H. K. Schutte, “Videokymography: high-speed line scanning of vocal fold vibration,” *J Voice*, vol. 10, no. 2, pp. 201–205, 1996.

[2] J. Lohscheller, U. Eysholdt, H. Toy, and M. Dollinger, “Phonovibrography: Mapping high-speed movies of vocal fold vibrations into 2-d diagrams for visualizing and analyzing the underlying laryngeal dynamics,” *Medi-*

cal Imaging, IEEE Transactions on, vol. 27, no. 3, pp. 300–309, March 2008.

[3] Q. Qiu, H. K. Schutte, Gu L., and Q. Yu, "An automatic method to quantify the vibration properties of human vocal folds via videokymography," *Folia Phoniatr Logop*, vol. 55, no. 3, pp. 128–136, 2003.

[4] J. G. Svec and H. K. Schutte, "Kymographic imaging of laryngeal vibrations. current opinion," *Otolaryngology & Head and Neck Surgery*, vol. 20, no. 6, pp. 458–465, 2012.

[5] J. G. Svec, F. Sram, and H. K Schutte, "Videokymography in voice disorders: what to look for?," *The Annals of otology, rhinology, and laryngology*, vol. 116, no. 3, pp. 172180, March 2007.

[6] F. Sroubek and J. Flusser, "Multichannel blind deconvolution of spatially misaligned images," *Image Processing, IEEE Transactions on*, vol. 14, no. 7, pp. 874–883, July 2005.

[7] W. Tao, H. Jin, Y. Zhang, L. Liu, and D. Wang, "Image thresholding using graph cuts," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. 38, no. 5, pp. 1181–1195, Sept 2008.

[8] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[9] J. Sedlar, "Image analysis in microscopy and videokymography," *Ph.D. thesis, Charles University, Prague, Czech Republic*, 2012.

[10] V. Hampala, "Visual evaluation of videokymographic features in voice disorders (in czech)," *Master's thesis, Palacky University, Olomouc, Czech Republic*, 2011.

Visual and Automatic Evaluation of Vocal Fold Mucosal Waves Through Sharpness of Lateral Peaks in High-Speed Videokymographic Images

[*]S. Pravin Kumar, [*]Ketaki Vasant Phadke, [†]Jitka Vydrová, [‡]Adam Novozámský, [‡]Aleš Zita, [‡]Barbara Zitová, and [*]Jan G. Švec, [*]*Olomouc, and* [†,‡]*Prague, Czech Republic*

Abstract: Introduction. The sharpness of lateral peaks is a visually helpful clinical feature in high-speed videokymographic (VKG) images indicating vertical phase differences and mucosal waves on the vibrating vocal folds and giving insights into the health and pliability of vocal fold mucosa. This study aims at investigating parameters that can be helpful in objectively quantifying the lateral peak sharpness from the VKG images.

Method. Forty-five clinical VKG images with different degrees of sharpness of lateral peaks were independently evaluated visually by three raters. The ratings were compared to parameters obtained by automatic image analysis of the vocal fold contours: *Open Time Percentage Quotients* (OTQ) and *Plateau Quotients* (PQ). The OTQ parameters were derived as fractions of the period during which the vocal fold displacement exceeds a predetermined percentage of the vibratory amplitude. The PQ parameters were derived similarly but as a fraction of the open phase instead of a period.

Results. The best correspondence between the visual ratings and the automatically derived quotients were found for the OTQ and PQ parameters derived at 95% and 80% of the amplitude, named OTQ_{95}, PQ_{95}, OTQ_{80} and PQ_{80}. Their Spearman's rank correlation coefficients were in the range of 0.73 to 0.77 ($P < 0.001$) indicating strong relationships with the visual ratings. The strengths of these correlations were similar to those found from inter-rater comparisons of visual evaluations of peak sharpness.

Conclusion. The Open time percentage and Plateau quotients at 95% and 80% of the amplitude stood out as the possible candidates for capturing the sharpness of the lateral peaks with their reliability comparable to that of visual ratings.

Keywords: Mucosal waves−Lateral peak sharpness−Kymography−Vocal fold vibration−Image analysis−Quantification.

INTRODUCTION

The occurrence of mucosal waves on the vibrating vocal folds has been generally recognized as a crucial indicator for healthy voice. Mucosal waves originate at the inferior surface of the vocal fold mucosa, propagate vertically along the medial surface, and then horizontally along the superior surface, creating a wave-like motion on the vocal folds.[1-8] A soft and pliable superficial layer of the lamina propria is necessary for their occurrence.[1,9] In other words, health and pliability of vocal fold mucosa may be indicated by the presence of mucosal waves.[10] Reduced mucosal wave amplitude is clinically observed in cases of increased mucosal stiffness due to, eg, lesions or scarring.[9,11]

Observations on excised hemilarynges[5,7,12-15] and lately also ultrasonic laryngeal observations *in vivo*[16] have shown that mucosal waves are associated with the phase-delayed movements of the upper vocal fold margin (lip or edge) trailing the lower margin. This delay is termed "vertical phase difference", and it facilitates the delivery of airflow energy to vocal fold tissue.[2,3,10,17-19] Titze et al (1993)[5] stroboscopically tracked the fleshpoints in excised larynges to quantify the phase delay and demonstrated its relationship with mucosal wave propagation velocity.

In vivo laryngoscopic imaging techniques such as videostroboscopy and high-speed videoendoscopy (HSV) have enabled easier visualization and quantitative evaluation of the presence, absence, or reduction of mucosal waves in clinical practice.[1,20-28] An alternative view for clinical evaluation of the mucosal waves has been offered by kymographic (ie, single-line) imaging techniques such as videokymography (VKG), digital kymography (DKG) or strobovideokymography (SVKG).[29]

Kymography assesses mucosal waves based on (1) vertical phase differences and (2) laterally traveling mucosal waves.[10,30,31] Vertical phase differences show up as sharp lateral peaks in kymograms, and laterally running mucosal

Accepted for publication August 30, 2018.

A part of the work was presented at the Voice Foundation's 47th Annual Symposium − Care of the Professional Voice, May 30-June 3, 2018, Philadelphia, USA. At this symposium S.P.K. received the Hamdan International Presenter Award 2018 and Sataloff New Investigator Award 2018 for this work. The research of the authors was supported by Technological Agency of the Czech Republic project no. TA04010877 (2014-2017) and the Czech Science Foundation (GA CR) project no. GA16-01246S (2016-2018). S. P. K. thanks Sri Sivasubramaniya Nadar (SSN) College of Engineering, Kalavakkam, India, for granting the leave of absence during his sabbatical stay (2016-2018) at Palacký University in Olomouc, Czech Republic.

From the *Voice Research Lab, Department of Biophysics, Faculty of Science, Palacký University, Olomouc, Czech Republic; †Voice and Hearing Centre, Medical Healthcom Ltd., Prague, Czech Republic; and the ‡Department of Image Processing, Institute of Information Theory and Automation of the Czech Academy of Sciences, Prague, Czech Republic.

Address correspondence and reprint requests to S. Pravin Kumar or Jan G. Švec, Voice Research Lab, Department of Biophysics, Faculty of Science, Palacký University, 17. listopadu 12, Olomouc 77416, Czech Republic. E-mail: pravin.subbaraj@upol.cz; jan.svec@upol.cz

Journal of Voice, Vol. ■■, No. ■■, pp. 1−9
0892-1997

https://doi.org/10.1016/j.jvoice.2018.08.022

waves appear on the kymogram as lines running obliquely sidewards along the upper margin during the medial excursion of the vocal fold[10,30-32] (Figure 1).

The sharpness and roundedness as observed from the shape of the lateral peaks are resulting from the vertical phase differences between the lower and upper margins of the vibrating vocal folds[30,33] (Figure 2). Looking from above the vocal folds, the boundary between the glottis and the vocal fold is created by the most medial part of the vocal fold. Due to the vertical phase differences, during the opening phase, this boundary is formed by the position of the upper margin of the vocal fold, whereas during the closing phase the boundary is normally formed by the position of the lower margin of the vocal fold. At the point of transition from opening to the closing phase, the glottal edge shifts from the upper to the lower margin (Figure 2A). When the vertical phase differences are large, the shift from upper to lower vocal fold margin is abrupt. In the kymogram, this sudden transition results in a sharp lateral peak within the oscillating vocal fold contour. In smaller vertical phase differences this transition happens rather gradually, causing the lateral peak to be rounded (Figure 2B).

The shape of the lateral peak has been found to be a clinically useful parameter revealing the vocal fold vibration characteristics that are not easily observable in non-kymographic imaging methods.[34] It has gained attention due to its diagnostic importance in assessing various voice disorders such as mucosal inflammations, scarring or tumors related to increased mucosal stiffness.[10,30,31,34-37] Increased vertical thickness of the vocal folds and increased pliability of the mucosa likely lead to larger vertical phase differences producing sharper lateral peaks in kymography. In contrast, increased stiffness of the mucosa is expected to reduce vertical phase differences, thus producing more rounded lateral peak shapes.[10,30,32] The magnitude of the vertical phase differences and the related sharpness of lateral peaks can also reveal on vocal fold vibratory behavior in different voice tokens, such as vocal registers.[33,38]

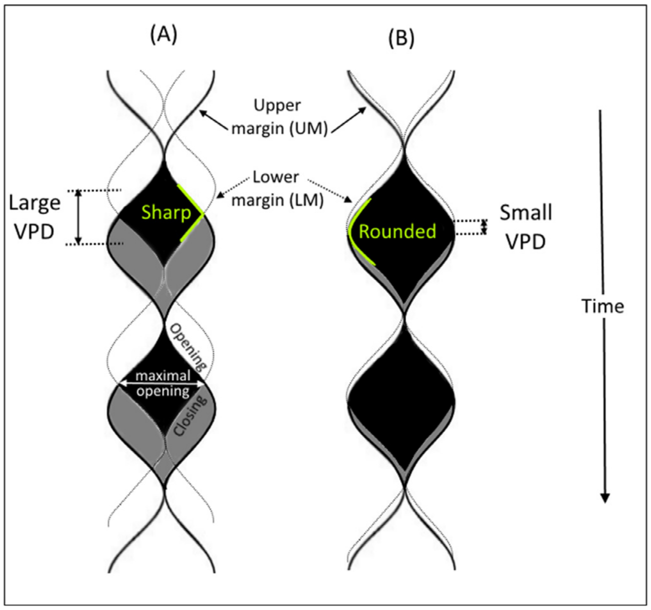


FIGURE 2. Formation of sharp (A) and rounded (B) lateral peaks in the kymogram. Movements of the lower and upper margins of the vocal folds are indicated by thin-dotted and thick-solid curves, respectively. The vibratory displacement of the lower margin precedes that of the upper margin, thus creating a vertical phase difference between their respecive motions. During the opening phase, the motion of the lower margin is invisible—this is indicated by the thin-dotted line; it becomes only visible during the closing phase. A sharp lateral peak (A) is seen when the vertical phase difference is large and a rounded lateral peak (B) is seen when the vertical phase difference is small (indicated in green). LM, lower margin; UM, upper margin; VPD, vertical phase difference. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

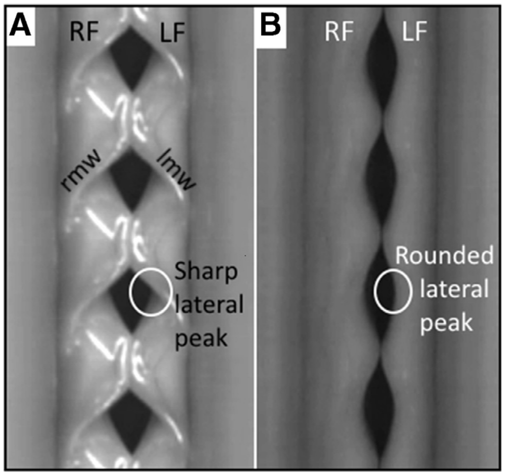


FIGURE 1. Videokymographic images (four vibratory cycles each) showing (A) sharp lateral peaks (encircled) and laterally running mucosal waves (rmw, lmw) on the right and left vocal fold, respectively; (B) rounded lateral peaks (encircled) with no mucosal waves. RF, LF right and left vocal fold. Total time displayed in the kymograms: 17.6 ms (time direction from top to bottom).

Efforts have been made to assess vertical phase differences and laterally traveling mucosal waves using image analysis methods. Shaw and Deliyski (2008)[39] used mucosal wave playback and qualitatively assessed the variations in mucosal wave magnitude and symmetry. Voigt et al (2010)[23] managed to detect the laterally traveling mucosal waves in high-speed endoscopic videos using automated image analysis techniques. Lately, Andrade-Miranda et al (2017)[40] used the optical flow method to detect mucosal wave propagation from high-speed endoscopic videos.

Chen, Woo, and Murry[41-43] applied spectral analysis to vocal fold waveforms obtained from digital kymograms and reported its usefulness in quantifying the waveforms and their changes due to different vocal tokens, pathologies, and surgical interventions. In principle, the spectral features can be expected to reflect the sharpness of the lateral peaks through

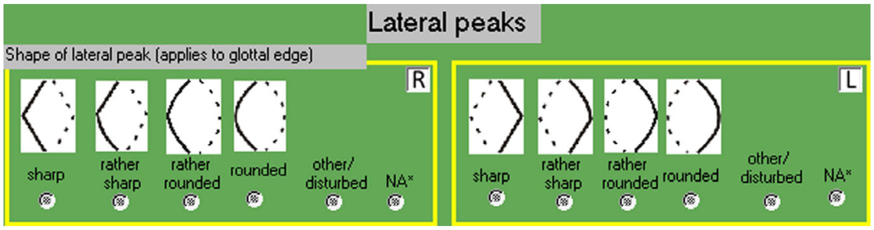


FIGURE 3. The form for visual evaluation of the VKG images with the descriptive pictograms representing varied degrees of lateral peak sharpness in the right (R) and left (L) vocal folds.

increased energy in upper harmonics, but such spectral changes can occur also due to, eg, the occurrence of closed phase; thus the spectral analysis of kymographic waveform makes it difficult to clearly distinguish the sharpness of the lateral peaks from other factors.

According to our knowledge, two methods have tried to quantify the shape of the lateral peak from kymograms so far.[44,45] Jiang et al estimated the shape of the peak indirectly by quantifying the vertical phase difference from kymographic images using a sinusoidal model approximation.[44,46-51] While this method is mathematically elegant, it becomes troublesome and difficult to interpret when the vocal fold motion becomes rather complex. The second method by Yamauchi et al (2015) quantified the peak sharpness from digital kymograms by the “lateral peak index”, defined as an angle formed by two lines between the start of open phase and lateral peak, and between the lateral peak and the end of open phase.[45] However, this index disregards the changes of curvature of the vocal fold waveform that influence the peak sharpness. Its value is additionally influenced also by the closed quotient and the vibratory amplitude, thus making it also sensitive to other factors than vertical phase differences.

Therefore, there is a need to search for other parameters that could help to improve the reliability of visual evaluation of clinical kymographic images of the vocal folds. Due to limited inter- and intrarater reliability, the approach of subjective rating limits the comparability of quantitative parameters on a large set of data. It further prevents the acquisition of reliable standard reference values for clinicians whose treatment decisions are dependent on assessment of such parameters. In contrast, objectification helps to find the accuracy of visual ratings.[52] The purpose of this study was therefore to investigate parameters which could quantify the lateral peak sharpness seen in the kymographic images and could easily be measured automatically from the detected contours of the vibrating vocal folds.

The work was done in the following steps: (1) A set of clinically obtained videokymographic images was evaluated visually to obtain ratings of the lateral peak sharpness. (2) The same images were subjected to automatic image analysis, in order to detect and compute the contours of the vibrating vocal folds as waveforms. (3) The resulting waveforms were quantified in order to obtain numerous parameters expected to reflect the lateral peak sharpness. (4) The obtained values of the parameters were compared to the visual ratings from step (1), in order to determine the parameters that show the best correlation with the visual ratings.

METHODS

Dataset

The dataset used in this work consisted of 45 videokymographic (VKG) images retrospectively selected from clinical records of patients examined for voice complaints at the Voice and Hearing Centre, Medical Healthcom, Ltd, Prague. The VKG recordings were obtained with the second generation VKG camera (Kymocam, CYMO, b.v. Groningen, the Netherlands, image rate 7200 lines/s), which was connected to a laryngoscope (Xion Medical, Germany, 10 mm diameter, 90° angle) using a C-mount objective adapter (R. Wolf, Germany, type 85261.272, 27 mm focal length). The larynx was illuminated by a 300 W endoscopic xenon light source (type FX 300 A, Fentex Medical, Germany). The VKG recordings were stored digitally by means of an EndoSTROB video capturing unit (Xion Medical, Germany). The images were extracted from the video records using the recently developed VKG Analyzer software.[53] The images were selected so that they demonstrated varied degrees of sharpness of lateral peaks.

Visual rating

Three raters independently evaluated the sharpness of the lateral peaks from the VKG images using a visual form (Figure 3).[54] The raters used the pictogram descriptions of the sharpness features as a reference for evaluation. The rating was done on a four points rating scale (1-sharp; 2-rather sharp; 3-rather rounded; 4-rounded) for left and right vocal folds separately, thus making a total of 90 ratings per rater from 45 images.

In order to assess the intra-rater reliability, each rater performed the evaluation twice, with a pause of 7-10 days in between. During the second evaluation, the order of the images was changed to minimize the memory effect. The ratings from the two evaluations, for the three raters, were consolidated, and an average (visual average VA) was obtained. A common consensus (visual consensus VC)

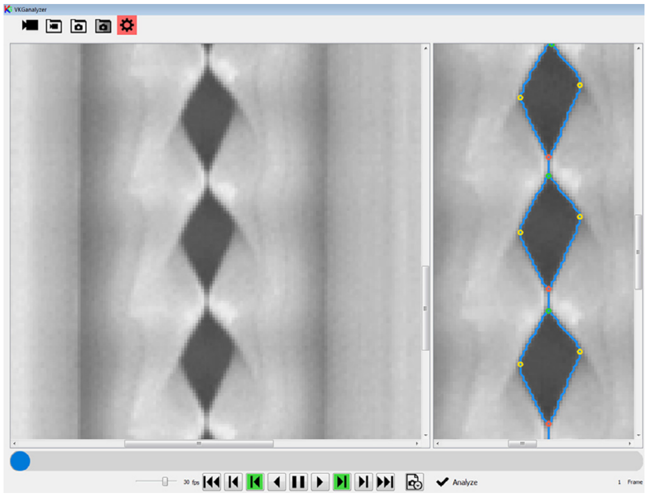


FIGURE 4. The screenshot of the VKG analyzer software showing the VKG image on the left and the detected glottal edge contours on the right.

was also arrived through the discussion among the three raters afterwards.

Image analysis

The recently developed VKG analyzer software[53] was used to detect and extract the contours defining the glottal edge boundary of both the left and right vocal folds (Figure 4). The image brightness and contrast were manually adjusted to improve the accuracy of the edge detection whenever required. The contours extracted from each of the VKG images were saved in a text file as a set of data defining the glottal edges of the left and right vocal folds, along with their respective time instances. A custom MATLAB script was then used to process the vocal fold contours and to obtain parameters capturing lateral peak sharpness, which could be included in the VKG analyzer software in future versions.

Quantification of lateral peak sharpness

Two kinds of parameters were defined for their simplicity in quantifying the vocal fold waveforms and their expected capability of reflecting the sharpness of the lateral peaks: the *Open Time Percentage Quotients* (OTQ) and *Plateau Quotients* (PQ).

The *Open Time Percentage Quotients* (OTQ_R) were inspired by the OT50 parameter published by Woo (1996),[55] who investigated the time for which the glottal area waveform exceeded 50% of the amplitude. Here, we defined the OTQ_R parameter as the proportion of time during which the vocal fold displacement exceeds a chosen percentage (R) of the vibration amplitude within a period (Figure 5):

$$OTQ_R = \frac{D_R}{T}$$

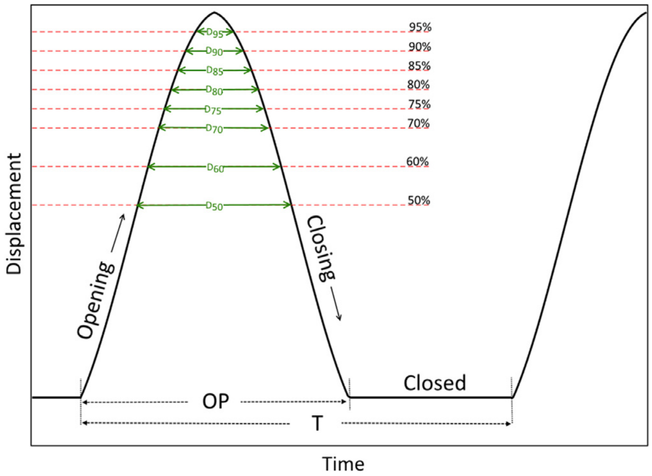


FIGURE 5. Parameterization of the vocal fold waveform for obtaining the Open Time Percentage Quotients (OTQ_R) and the Plateau Quotients (PQ_R) as indicators for peak sharpness. OP is the open phase, T is the period, and D_R is the duration of the phase during which the waveform exceeds a specified R percentage of the amplitude. The R percentages are indicated by the dashed red lines.

where D_R is the duration of the phase where the lateral displacement is greater than R% of the vibration amplitude and T is the period of the vocal fold vibratory cycle. The vibration amplitude was determined as the difference between the most lateral and most medial position of the vocal fold during the open phase.

The *Plateau Quotients* (PQ_R) used here were inspired by the work of Mehta et al,[56] who investigated the proportion of open phase for which the glottal area was larger than 95% of its maximum. Here, we defined PQ_R as the proportion of time during which the vocal fold displacement

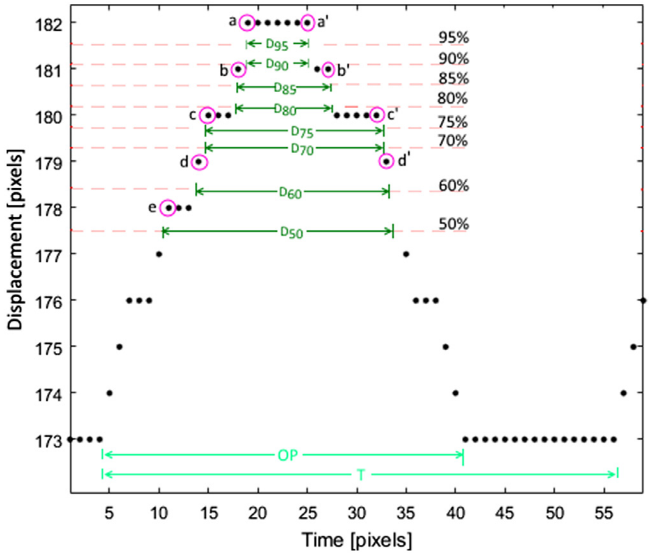


FIGURE 6. Implementation of the parameterization of the waveform illustrating the procedure followed to calculate the D_R durations from the discrete samples.

exceeds R% of vibration amplitude within the open phase (Figure 5):

$$PQ_R = \frac{D_R}{OP}$$

where *OP* is the duration of the open phase.

When implementing the automatic analysis procedure, it was necessary to deal with the fact that the waveforms were not continuous, but consisted of samples of limited temporal and spatial resolution. While the contour samples are defined by integer pixel coordinates, the R% levels usually correspond to noninteger subpixel coordinates. An example of the procedure adapted to calculate the OTQ and PQ parameters from the discrete samples is shown in Figure 6. When digitized, the discrete contour data points were located at specific pixels with coordinates defined by integer numbers. Therefore, sometimes the same pixel coordinates pertained to multiple consecutive time points (see Figure 6). In order to measure the time intervals at which the vocal fold displacement exceeds the criterion level R%, the first and last samples with the values above the R% criterion level in the opening and closing phases, respectively, were selected (marked by circles and indicated as a, b, c, d, e in the opening phase, and a', b', c', d' in the closing phase in Figure 6). Thus, for the R% levels at 95%, 90%, 85%, 80%, 75%, 70%, 60% and 50%, the intervals between $a-a'$, $a-a'$, $b-b'$, $b-b'$, $c-c'$, $c-c'$, $d-d'$ and $e-d'$, respectively, were considered to calculate the D_R durations in the example shown in Figure 6.

Statistical analysis

Statistical analysis was performed using the SPSS (version 24) software. Spearman's rank correlation coefficient was computed to determine the inter- and intrarater reliability of the visual ratings. The intrarater reliability was also tested with Cronbach's Alpha value. To estimate the correlation between the objective measures and the visual ratings, Spearman's rank correlation coefficient was again used.

RESULTS

Visual rating

Results from the repeated visual evaluations of the lateral peak sharpness in VKG images by the three raters were compared to find the intrarater and inter-rater reliability. The intrarater comparisons between the two repeated evaluations resulted in the Cronbach's Alpha values around 0.92 for all three raters, indicating excellent reliability of the raters. The intrarater Spearman's rank correlation coefficients for the individual raters varied between 0.84 and 0.85 ($P < 0.001$, N = 90) indicating very strong and significant correlations between the repeated evaluations.

The inter-rater comparisons showed Spearman's rank correlation coefficients in the range of 0.67 to 0.82, with a mean value of 0.73. These coefficients indicated strong and significant correlations ($P < 0.001$, N = 90) between the evaluations of the different raters, but also hinted at some discrepancies among the raters. Therefore, a consensus among the raters was established by mutual discussions.

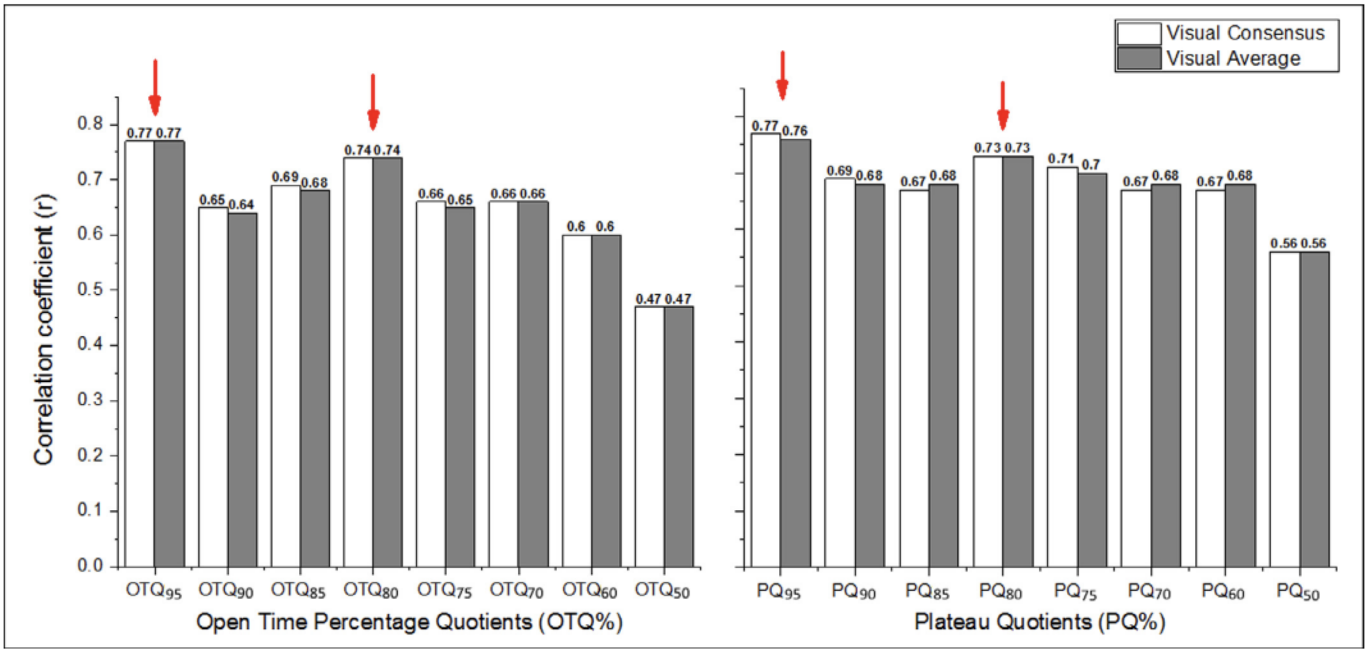


FIGURE 7. Spearman's rank correlation coefficients indicating the agreement between the visual ratings and the measured parameters OTQ and PQ. The highest correlation coefficients are indicated by red arrows. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

The visual consensus versus visual average comparison revealed very strong Spearman's rank correlation ($r = 0.99$, $P < 0.001$). Furthermore, both the visual consensus and visual average values very strongly correlated with the values of all three raters ($r = 0.81\text{-}0.91$, $P < 0.001$) in both evaluations. Therefore, the visual consensus and visual average values were deemed appropriate for further analysis of the correlations between the visual and automatic image analysis.

Correlation between visual ratings and the analyzed parameters

The correlations between the different OTQ and PQ parameters with the visual consensus and visual average ratings are shown in Figure 7. All correlations had a significance level of $P < 0.001$, indicating that all parameters were well related to the visual ratings. Highest correlations were found for the parameters measured at 95% amplitude (OTQ_{95}, PQ_{95}) and at 80% amplitude (OTQ_{80}, PQ_{80}). In Figure 7, these are indicated by arrows. There were minimal differences between the OTQ and PQ parameters measured at the same percentage. Also, there were minimal differences between the visual average and visual consensus. Lowest correlations were found for the parameters measured at 50% amplitude (OTQ_{50}, PQ_{50}).

The relationships between the values of the four best correlating parameters OTQ_{95}, OTQ_{80}, PQ_{95}, and PQ_{80}, and the visual ratings are revealed in Figure 8. As expected, all these quotients clearly increase their values when the peak shape changes from sharp to rounded. There is, however, some spread of the measured data around the best fit line, which indicates that some discrepancies exist between the visual and automatic evaluations. The Spearman's rank correlation values between

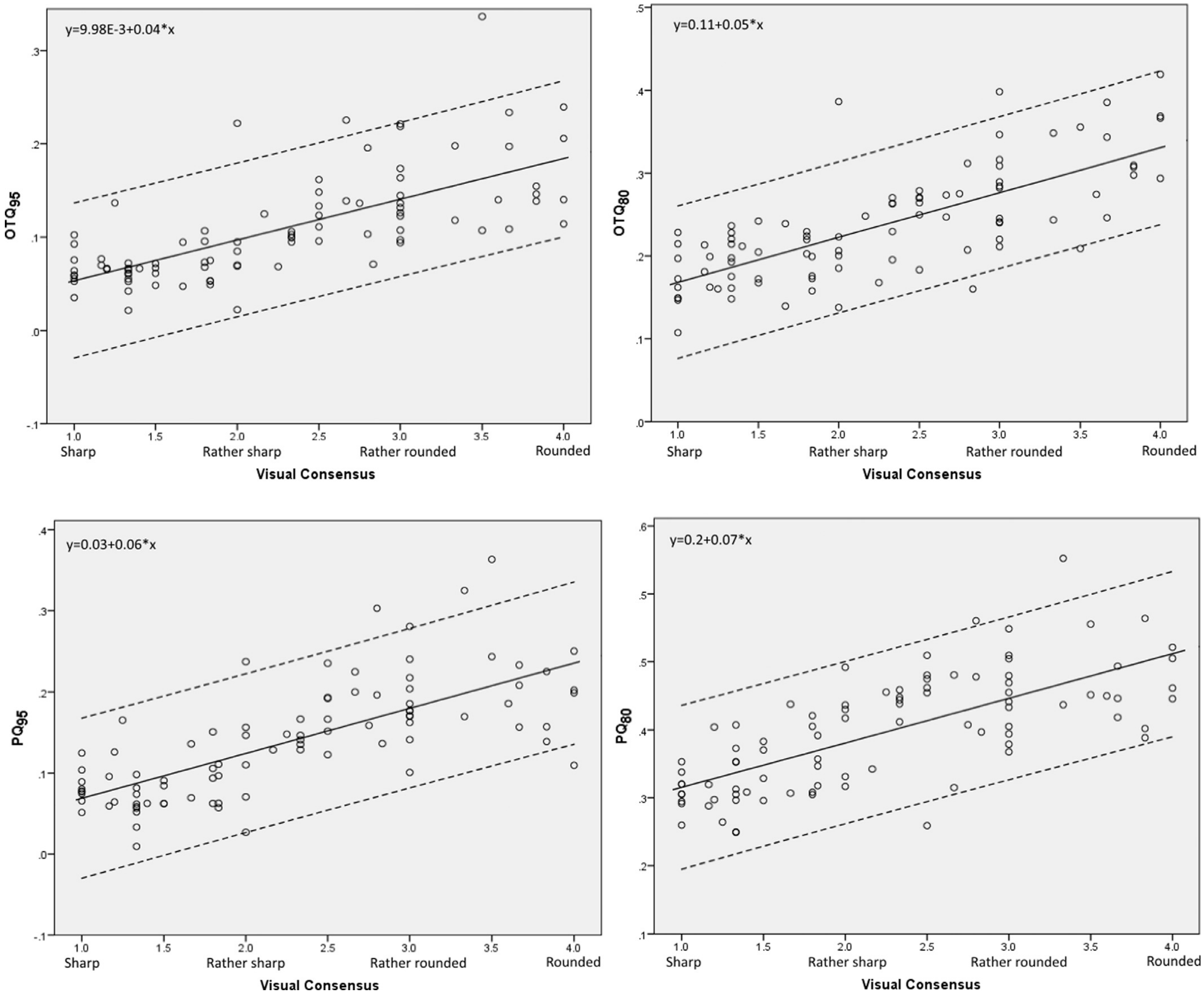


FIGURE 8. The relationship between the measured values and the visual ratings for the four parameters with the highest correlations OTQ_{95}, OTQ_{80}, PQ_{95}, and PQ_{80}. The lines indicate the best fit linear relationship (solid) and 95% confidence intervals (dashed).

these analyzed quotients and the visual ratings (0.73-0.77, as shown in Figure 7) were comparable to those found between different raters (0.67-0.82) indicating that the discrepancies in the automatic-to-visual comparisons are similar to those found in inter-rater comparisons.

DISCUSSION

Sharpness of lateral peaks has been recognized previously as a useful visual feature that can indicate pliability and health of the vocal fold mucosa.[30,45] In a recent study, the lateral peak sharpness has been identified as one of the most helpful visual features for clinical evaluation of voice disorders using videokymography.[34] The peak sharpness is directly related to vertical phase differences between the motions of the upper and lower margin of the vocal folds and results from projection of the vocal fold motion into the laryngoscopic view from above of the vocal folds.[10,30,57] Biomechanically, stiffening of the mucosa leads to increased mucosal wave speed[6] and decreased vertical phase differences, causing the peak to become more rounded.[30,32,35] Apart from physiological factors related, eg, to pitch increase and voice registration, stiffening of the mucosa is considered to be a direct result of pathological processes on the vocal folds. Therefore, evaluation of peak sharpness can help clinicians to better diagnose the health of the vocal fold mucosa, particularly in phonations produced at comfortable pitch in modal/chest register where the mucosa is expected to be pliable.

Visual evaluation, however, is subjective and differences among evaluations of different raters can be expected. This can be spotted also in our results: while the intraindividual Spearman's rank correlations were very strong ($r = 0.84$-0.85), the inter-rater Spearman's rank correlations were lower ($r = -0.67$-0.82) indicating more disagreements between the visual evaluations of different raters than between repeated evaluations of the same rater.

This study searched for objective parameters that are related to the visual ratings of peak sharpness in kymograms and can be used as “peak sharpness indicators”. For this purpose, the OTQ and PQ were defined by relating the durations of different phases of the vibratory cycle to each other, applying the same concept as used for the well-established traditional parameters such as the Closed Quotient (CQ), Open Quotient (OQ) or Speed Quotient (SQ).[58-60] As such, these parameters are relatively simple to measure. As far as their interpretation is concerned, smaller OTQ and PQ values correspond to sharper lateral peaks of the vocal fold waveform detected in the kymogram (recall Figure 8).

The OTQ and PQ parameters measured from the time intervals at different percentages of vibratory amplitude were compared to the visual ratings of the peak sharpness in order to evaluate the congruence of these two approaches. The OTQ and PQ parameters showed very similar correlations to the visual ratings which indicate that they quantified the visual impressions similarly. The best correlations with the visual evaluations were found for the OTQ and PQ parameters measured at 95% and 80% of the amplitude, the worst correlations appeared at 50% of the amplitude. Since the peak corresponds to 100% of the amplitude, it appears logical that the best correlations for peak sharpness should be obtained for the measurements made as closely to the peak as possible this explains the finding of the worst correlations at 50% and best correlations at 95% of the amplitude (recall Figure 7). However, the correlations at 90% and 85% of the amplitude were worse than those at 80%. This seemingly contradictory finding could be attributed to the contour artifacts due to the limited pixel and temporal resolution (compare the ideal waveform in Figure 5 with the real detected waveform in Figure 6). The clinical videokymographic images analyzed here showed the average vocal fold vibratory amplitudes around 8 pixels (range 5-15 pixels). A change of 1 pixel, in this case, corresponds to the spatial resolution of 12.5% of the amplitude (range 7-20%). This means that it is hardly possible to reliably distinguish levels that are close together, such as those at 85%, 90%, and 95% of the amplitude.

Preliminary investigations using synthetic kymograms generated by a kinematic model of the vocal folds[61] with known vertical phase differences (not included here for brevity reasons) showed that the limited spatial and temporal resolution of the kymographic images can influence the accuracy of the results, particularly of those quotients measured at the proximity of the peak, and these artifacts need to be taken into account. Thus the measurements at 80% amplitude could potentially be used as a compromise to reduce the influence of the possible waveform artifacts, but still reflect the peak sharpness and vertical phase differences reasonably well. The waveform artifacts present a general limitation which is inherent in the laryngeal kymographic techniques. Increased spatial resolution of the kymographic images is desirable for improving the quantification accuracy of the vocal fold vibratory patterns in future.

In principle, the OTQ and PQ parameters can be implemented also for analyzing the glottal area waveforms (GAWs) obtained from full high-speed endoscopic videos, as done by Mehta et al (2011).[56] GAWs offer better pixel resolution than kymography due to the fact that the glottal area is distributed over multiple image lines and thus over considerably more pixels. In this respect, GAWs may possibly offer better accuracy than kymographic waveforms in measuring the OTQ and PQ parameters as defined here. However, a more detailed study is needed to elucidate these factors and to better understand the influence of limited spatial and temporal resolution on the accuracy of these parameters.

The detailed comparisons between the visual ratings and the OTQ and PQ parameters shown in Figure 8 reveal that the relationship is not perfect and some discrepancies exist here. Besides of the influence of the limited spatial and temporal resolution of the images (7200 kymographic lines per second with 720 pixels per line used here), these discrepancies could possibly be also due to contour detection artifacts resulting from the image analysis procedure. Furthermore, it is known that the visual perception process is rather complex and visual judgments of the peak shape may also be influenced by, eg, the grayscale shadings which are not captured in the contours. All these factors may contribute to the differences between the automatic analysis and the visual ratings. Nevertheless, the Spearman's rank correlations between the visual ratings and the OTQ and PQ parameters measured at 80% and 95% amplitude (r = 0.73-0.77, recall Figure 7) are similar to those found between different raters. Therefore the reliability of the parameters, although not perfect, is considered acceptable here.

While the shape of the lateral peak appears as a useful clinical feature, ultimately it should be related to the vertical phase differences. These differences cannot be exactly measured laryngoscopically *in vivo*. Therefore, we were not able to establish their direct relationship with the defined parameters, which poses another potential limitation of this study. However, this relationship may be derived and investigated using synthetic kymograms obtained from a mathematical model of the vocal folds with known vertical phase differences[61], which is planned to be addressed in a future study.

CONCLUSION

The PQ_{95}, PQ_{80}, OTQ_{95} and OTQ_{80} parameters stood out as the possible candidates for capturing the sharpness of the lateral peaks. The reliability of these parameters appears comparable to the inter-individual reliability of visual ratings. The results provide basic insights into developing the computer algorithms to automatically quantify the sharpness of lateral peaks from the VKG images.

REFERENCES

1. Hirano M. *Clinical Examination of Voice*. Wien, Austria: Springer-Verlag; 1981.
2. Titze IR. The physics of small−amplitude oscillation of the vocal folds. *J Acoust Soc Am*. 1988;83:1536–1552.
3. McGowan R. An analogy between the mucosal waves of the vocal folds and wind waves on water. *Haskins Lab Status Rep Speech Res*. 1990;101:243–249.
4. Yumoto E, Kurokawa H, Okamura H. Vocal fold vibration of the canine larynx: observation from an infraglottic view. *J Voice*. 1991;5:299–303.
5. Titze IR, Jiang JJ, Hsiao T-Y. Measurement of mucosal wave propagation and vertical phase difference in vocal fold vibration. *Ann Otol Rhinol Laryngol*. 1993;102:58–63.
6. Berke GS, Gerratt BR. Laryngeal biomechanics: an overview of mucosal wave mechanics. *J Voice*. 1993;7:123–128.
7. Boessenecker A, Berry DA, Lohscheller J, et al. Mucosal wave properties of a human vocal fold. *Acta Acust united Ac*. 2007;93:815–823.
8. Krausert CR, Olszewski AE, Taylor LN, et al. Mucosal wave measurement and visualization techniques. *J Voice*. 2011;25:395–405.
9. Hirano M, Bless DM. *Videostroboscopic Examination of the Larynx*. San Diego, California: Singular Publishing Group; 1993.
10. Švec JG, Šram F, Schutte HK. Videokymography. In: Fried M, Ferlito A, eds. 3 ed*The Larynx*. Vol 1, San Diego, CA: Plural Publishing; 2009:253–271.
11. Bless DM, Hirano M, Feder RJ. Videostroboscopic evaluation of the larynx. *Ear Nose Throat J*. 1987;66:289–296.
12. Hiroto I. Vibration of vocal cords: an ultra high-speed cinematographic study (film). Kurume, Japan: Department of otolaryngology, Kurume University; 1968.
13. Berry DA, Montequin DW, Tayama N. High-speed digital imaging of the medial surface of the vocal folds. *J Acoust Soc Am*. 2001;110:2539–2547.
14. Döllinger M, Berry DA, Kniesburges S. Dynamic vocal fold parameters with changing adduction in ex-vivo hemilarynx experiments. *J Acoust Soc Am*. 2016;139:2372–2385.
15. Herbst CT, Hampala V, Garcia M, et al. Hemi-laryngeal setup for studying vocal fold vibration in three dimensions. *J Vis Exp*. 2017; 129:e55303. http://dx.doi.org/10.3791/55303.
16. Jing B, Ge Z, Wu L, et al. Visualizing the mechanical wave of vocal fold tissue during phonation using electroglottogram-triggered ultrasonography. *J Acoust Soc Am*. 2018;143:EL425–EL429.
17. Ishizaka K, Flanagan J. Synthesis of voiced sounds from a two-mass model of the vocal cords. *Bell Syst Tech J*. 1972;51:1233–1268.
18. Titze IR. Comments on the myoelastic-aerodynamic theory of phonation. *J Speech Hear Reas*. 1980;23:495–510.
19. Titze IR. *Principles of Voice Production (Second Printing)*. Iowa City, IA: National Center for Voice and Speech; 2000.
20. Dejonckere PH, Bradley P, Clemente P, et al. A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques. *Eur Arch Otorhinolaryngol*. 2001;258:77–82.
21. Poburka BJ. A new stroboscopy rating form. *J Voice*. 1999;13: 403–413.
22. Deliyski DD, Petrushev PP, Bonilha HS, et al. Clinical implementation of laryngeal high-speed videoendoscopy: challenges and evolution. *Folia Phoniatr Logop*. 2008;60:33–44.
23. Voigt D, Döllinger M, Eysholdt U, et al. Objective detection and quantification of mucosal wave propagation. *J Acoust Soc Am*. 2010;128: EL347–EL353.
24. Kaneko M, Shiromoto O, Fujiu-Kurachi M, et al. Optimal duration for voice rest after vocal fold surgery: randomized controlled clinical study. *J Voice*. 2017;31:97–103.
25. Poburka BJ, Patel RR, Bless DM. Voice-vibratory assessment with laryngeal imaging (VALI) form: reliability of rating stroboscopy and high-speed videoendoscopy. *J Voice*. 2017;31:513e1–513.e14.
26. El-Demerdash A, Fawaz SA, Sabri SM, et al. Sensitivity and specificity of stroboscopy in preoperative differentiation of dysplasia from early invasive glottic carcinoma. *Eur Arch Otorhinolaryngol*. 2015;272:1189–1193.
27. Zacharias SRC, Deliyski DD, Gerlach TT. Utility of laryngeal high-speed videoendoscopy in clinical voice assessment. *J Voice*. 2018;32:216–220.
28. Patel RR, Awan SN, Barkmeier-Kraemer J, et al. Recommended minimum protocols for instrumental assessment of voice: American Speech-Language Hearing Association Committee on Instrumental Voice assessment protocols. *Am J Speech Lang Pathol*. 2018;27:887–905.
29. Švec JG, Schutte HK. Kymographic imaging of laryngeal vibrations. *Curr Opin Otolaryngol Head Neck Surg*. 2012;20:458–465.
30. Švec JG, Šram F, Schutte HK. Videokymography in voice disorders: what to look for. *Ann Otol Rhinol Laryngol*. 2007;116:172–180.
31. Švec JG, Frič M, Šram F, Schutte HK. Mucosal waves on the vocal folds: conceptualization based on videokymography. *Fifth International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*. Firenze, Italy: Firenze University Press; 2007:171–172.
32. Švec JG, Šram F. Videokymographic examination of voice. In: Ma EPM, Yiu EML, eds. *Handbook of Voice Assessments*. San Diego, CA: Plural Publishing; 2011:129–146.
33. Sundberg J, Högset C. Voice source differences between falsetto and modal registers in counter tenors, tenors and baritones. *Logoped Phoniatr Vocol*. 2001;26:26–36.

34. Phadke KV, Vydrová J, Domagalská R, et al. Evaluation of clinical value of videokymography for diagnosis and treatment of voice disorders. *Eur Arch Otorhinolaryngol*. 2017;274:3941–3949.
35. Vydrová J, Švec JG, Šram F. Videokymography (VKG) in laryngologic practice. *J Macrotrends Health Med*. 2015;3:87–95.
36. Yamauchi A, Yokonishi H, Imagawa H, et al. Quantification of vocal fold vibration in various laryngeal disorders using high-speed digital imaging. *J Voice*. 2016;30:205–214.
37. Yamauchi A, Yokonishi H, Imagawa H, et al. Visualization and estimation of vibratory disturbance in vocal fold scar using high-speed digital imaging. *J Voice*. 2016;30:493–500.
38. Švec JG, Sundberg J, Hertegard S. Three registers in an untrained female singer analyzed by videokymography, strobolaryngoscopy and sound spectrography. *J Acoust Soc Am*. 2008;123:347–353.
39. Shaw HS, Deliyski DD. Mucosal wave: a normophonic study across visualization techniques. *J Voice*. 2008;22:23–33.
40. Andrade-Miranda G, Bernardoni NH, Godino-Llorente JI. Synthesizing the motion of the vocal folds using optical flow based techniques. *Biomed Signal Process Control*. 2017;34:25–35.
41. Chen W, Woo P, Murry T. Spectral analysis of digital kymography in normal adult vocal fold vibration. *J Voice*. 2014;28:356–361.
42. Chen W, Woo P, Murry T. Vocal fold vibratory changes following surgical intervention. *J Voice*. 2016;30:224–227.
43. Chen W, Woo P, Murry T. Vocal fold vibration following surgical intervention in three vocal pathologies: a preliminary study. *J Voice*. 2017;31:610–614.
44. Jiang JJ, Chang CIB, Raviv JR, et al. Quantitative study of mucosal wave via videokymography in canine larynges. *Laryngoscope*. 2000;110:1567–1573.
45. Yamauchi A, Yokonishi H, Imagawa H, et al. Quantitative analysis of digital videokymography: a preliminary study on age- and gender-related difference of vocal fold vibration in normal speakers. *J Voice*. 2015;29:109–119.
46. Jiang JJ, Zhang Y, Kelly MP, et al. An automatic method to quantify mucosal waves via videokymography. *Laryngoscope*. 2008;118: 1504–1510.
47. Chodara AM, Krausert CR, Jiang JJ. Kymographic characterization of vibration in human vocal folds with nodules and polyps. *Laryngoscope*. 2012;122:58–65.
48. Krausert CR, Ying D, Zhang Y, et al. Quantitative study of vibrational symmetry of injured vocal folds via digital kymography in excised canine larynges. *J Speech Lang Hear Res*. 2011;54:1022–1038.
49. Li L, Zhang Y, Maytag AL, et al. Quantitative study for the surface dehydration of vocal folds based on high-speed imaging. *J Voice*. 2015;29:403–409.
50. Regner MF, Robitaille MJ, Jiang JJ. Interspecies comparison of mucosal wave properties using high-speed digital imaging. *Laryngoscope*. 2010;120:1188–1194.
51. Zhang Y, Huang N, Calawerts W, et al. Quantifying the subharmonic mucosal wave in excised larynges via digital kymography. *J Voice*. 2017;31:123.e7–123.e13.
52. Bonilha HS, Deliyski DD, Gerlach TT. Phase asymmetries in normophonic speakers: visual judgments and objective findings. *Am J Speech Lang Pathol*. 2008;17:367–376.
53. Novozamsky A, Sedlar J, Zita A, et al. Image analysis of videokymographic data. *2015 IEEE International Conference on Image Processing (ICIP)*, 2015,78−82.
54. Švec JG, Švecová H, Herbst C, et al. Evaluation protocol for videokymographic images. (Ms Access software application). Groningen, the Netherlands: Groningen Voice Research Lab, University of Groningen. 2007.
55. Woo P. Quantification of videostrobolaryngoscopic findings−measurements of the normal glottal cycle. *Laryngoscope*. 1996;106:1–27.
56. Mehta DD, Zañartu M, Quatieri TF, et al. Investigating acoustic correlates of human vocal fold vibratory phase asymmetry through modeling and laryngeal high-speed videoendoscopy. *J Speech Lang Hear Res*. 2011;130:3999–4009.
57. Hiroto I. The mechanism of phonation; its pathophysiological aspects. *Nippon Jibiinkoka Gakkai Kaiho*. 1966;69:2097–2106.
58. Timcke R, von Leden H, Moore P. Laryngeal vibrations - measurements of the glottic wave .I. The normal vibratory cycle. *AMA Arch Otolaryngol*. 1958;68:1–19.
59. Qiu Q, Schutte H, Gu L, et al. An automatic method to quantify the vibration properties of human vocal folds via videokymography. *Folia Phoniatr Logop*. 2003;55:128–136.
60. Lohscheller J, Švec JG, Döllinger M. Vocal fold vibration amplitude, open quotient, speed quotient and their variability along glottal length: kymographic data from normal subjects. *Logoped Phoniatr Vocol*. 2013;38:182–192.
61. Subbaraj PK, Švec JG. Kinematic model for simulating mucosal wave phenomena on vocal folds. In: Manfredi C, ed. *MAVEBA 2017: Models and Analysis of Vocal Emissions for Biomedical Applications. 10th International Workshop*. Firenze: Firenze University Press; 2017:115–118.

Biomedical Signal Processing and Control 78 (2022) 103878

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Videokymogram Analyzer Tool: Human–computer comparison

Aleš Zita [a,*], Adam Novozámský [a], Barbara Zitová [a], Michal Šorel [a], Christian T. Herbst [b,c], Jitka Vydrová [d], Jan G. Švec [b,d]

[a] The Czech Academy of Sciences, Institute of Information Theory and Automation, Pod vodárenskou věží 4, Prague 8, 182 08, Czech Republic
[b] Palacký University, Faculty of Sciences, Department of Experimental Physics, Voice Research Lab, 17. listopadu 12, Olomouc, 771 46, Czech Republic
[c] Mozarteum University, Mirabellplatz 1, Salzburg, 5020, Austria
[d] Voice Centre Prague, Medical Healthcom, Ltd., Národní 11, Prague, 110 00, Czech Republic

ARTICLE INFO

Keywords:
Image analysis
Videokymography
Vocal fold vibrations

ABSTRACT

Videokymography (VKG) is a modern video recording technique used in laryngology and phoniatrics to examine vocal fold vibrations. To obtain quantitative information on the vocal fold vibration, VKG image analysis is needed but no software has yet been validated for this purpose. Here, we introduce a validated software tool that aids clinicians to evaluate diagnostically important vibration characteristics in VKG and other types of kymographic recordings. State-of-the-art methods for automated image evaluation were implemented and tested on a set of videokymograms with a wide range of vibratory characteristics, including healthy and pathologic voices. The automated image segmentation results were compared to manual segmentation results of six evaluators revealing average differences smaller than one pixel. Furthermore, the automatically categorized vibratory parameters precisely agreed with the average visual assessment in 84 and 91 percent of the cases for pathological and healthy patients, respectively. Based on these results, the newly developed software was found to be a valid, reliable automated tool for the quantification of vocal fold vibrations from VKG images, offering a number of novel features relevant for clinical practice.

1. Introduction

The vibration characteristics of the laryngeal tissues – particularly those of the vocal folds – are critical for the evaluation of voice disorders by laryngologists and phoniatricians [1,2]. The vocal folds are a pair of elastic tissues in larynx (see Fig. 1) and their vibrations produce phonation. The vibrating folds gradually close and open the space between them (*rima glottidis*, or glottis) with the frequency range of about 60–1000 Hz [3]. To visualize the vocal folds, laryngeal endoscopy is routinely used in clinical practice (Fig. 1, left). There are three standard techniques to display and evaluate the vibration of the vocal folds using laryngeal endoscopy: *videostroboscopy*, *high-speed laryngeal videoendoscopy*, and *videokymography* [4,5].

Videostroboscopy displays an illusory slowed-down motion of the vocal folds in real time by temporarily synchronizing the images, captured by a standard endoscopic video camera, with vocal fold vibratory cycles [6,7]. While this method is most frequently used in clinical practice, it is not suitable for documenting and quantifying irregular vibrations typical for disordered voices.

High-speed videoendoscopy (HSV) captures laryngeal images with a high-speed camera at frame rates well above the fundamental frequency of phonation, typically exceeding 1000 frames per second [4,8,9]. This method accurately documents each oscillatory cycle of the vocal fold oscillations and produces large amounts of data which are beneficial for research purposes and can be used for various types of analyses, such as glottal segmentation and extraction of glottal area waveforms, digital kymography, phonovibrography, laryngotopography, etc. [10–14]. However, the HSV method has not yet been widely implemented in clinical practice, mainly because it does not provide real time visual feedback and is time-wise demanding, due to the large volume of acquired data [15]. Addressing this, a viable strategy to diminish the vast amount of data generated by HSV is to reduce the two spatial image dimensions to a single one. This is achieved via the videokymographic imaging, which is the main subject of the present study.

In *videokymography* (VKG) special cameras are utilized to capture images of the vibrating vocal folds at a single line perpendicular to glottal axis with the rate of 7200 line images per second and allow simultaneous observation in standard and videokymographic modes (see Fig. 1). This allows the clinician to flexibly orient the camera in order to record the desired line of interest [16–18]. VKG method offers apparent benefits by combining real time imaging feedback

[*] Corresponding author.
E-mail address: zita@utia.cas.cz (A. Zita).

https://doi.org/10.1016/j.bspc.2022.103878
Received 30 December 2021; Received in revised form 8 May 2022; Accepted 4 June 2022
Available online 8 July 2022

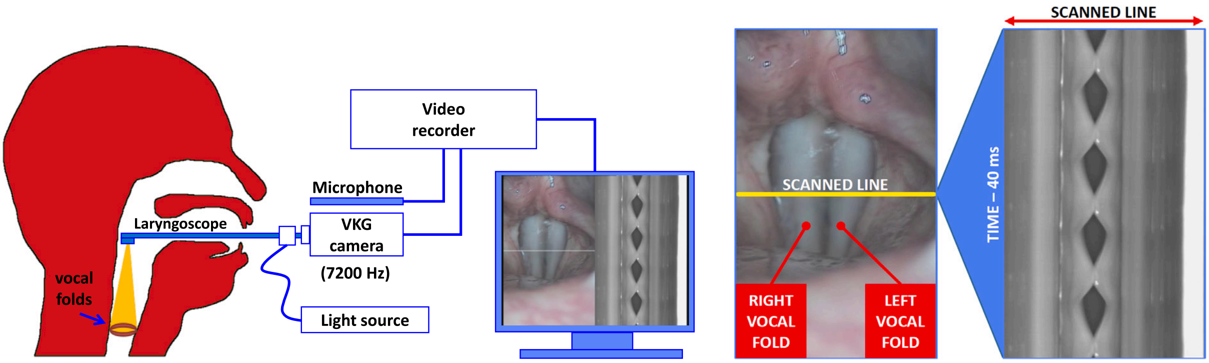


Fig. 1. Videokymography: On the left there is the examination of vocal fold vibrations by laryngeal endoscopy using a videokymographic (VKG) camera. On the right there are two parallel imaging modes of the videokymographic (VKG) camera: standard (left) and videokymographic (right). The videokymographic image is composed of successively acquired scanned lines at the location indicated in the standard mode. The time is mapped onto the vertical axis within the VKG image, going from top to bottom. The standard duration is 40 ms (resulting from the standard rate of 25 frames/s [16,17]). This also applies to the other figures with VKG images in this article.

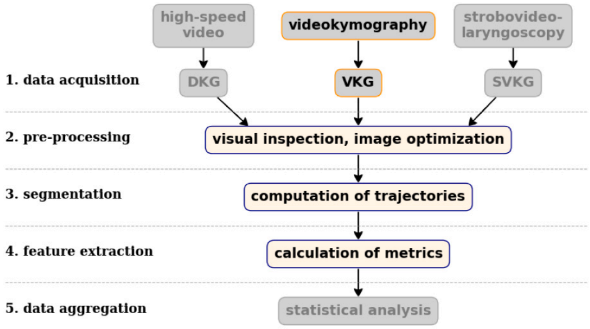


Fig. 2. Proposed data processing pipeline for kymographic documentation and analysis of vocal fold vibration. This study particularly focuses on layers 2 through 4 (i.e., pre-processing, segmentation, and feature extraction). The algorithms and software presented in this manuscript have been validated with videokymographic footage.

found in videostroboscopy with the advantages of high-speed imaging, i.e., sufficient frame rates to truthfully document each oscillatory cycle of the vocal folds [19].

In a VKG system, data acquisition and kymographic image generation is facilitated simultaneously on the hardware layer. In contrast, two further strategies offer the possibility to generate surrogate kymographic images from previously recorded endoscopic laryngeal footage [5]: (a) digital kymography (DKG) [14,20], operating on HSV data; and (b) strobovideokymography (SVKG) [21–23], operating on videostroboscopic data.

Here we propose a system for data analysis in which either of these three types of kymographic image (VKG, DKG, or SVGK) is processed in a pipeline that is schematically illustrated in Fig. 2. In particular, the available kymographic images are visually inspected, optimized and selected for further treatment (see **layer 2: pre-processing** in Fig. 2). In analogy to HSV data, the vibrating glottal edge is segmented, thus computing the time-varying medio-lateral deflections of the vocal folds (**layer 3: segmentation**). However, while in HSV the segmentation operates on two spatial dimensions, VGK images have a reduced dimensionality, thus requiring a fundamentally different segmentation approach. The resulting data is then subjected to extraction of dedicated metrics that allow for quantitative assessment (**layer 4: feature extraction**) and can finally be used for statistical analysis or intra-subjective comparison if so required (layer 5: data aggregation).

While exclusively manual treatment within this analysis pipeline has been partially pursued for scientific exploration [24–26], this is rather time-consuming and thus not feasible in clinical practice. Ideally, layers 2 through 4 should be automated and completed by computer-aided (semi)automatic software algorithms and a supporting graphical user interface (GUI).

Addressing this, some systems have been developed previously. For instance, the commercially marketed Kay Elemetrics Image Processing System (KIPS) offers features to produce DKGs from HSV, segmentation of these DKGs, as well as limited analysis of the segmented DKG contours in the form of metrics addressing glottal opening (glottal width), mean fundamental frequency (f_o), the vibratory amplitudes of left and right vocal folds, as well as the percentage of time when the glottis is closed with respect to the duration of the analysis. Another system, proposed by Manfredi et al. [27] operates directly on VKG images and offers the image segmentation and extraction of basic quantitative parameters, such as the left-to-right amplitude and period ratios, open-to-closed phase ratios, and phase symmetry index [28,29].

Despite the existence of these previously established systems, which constitute commendable groundbreaking work in their own right, a comprehensive coverage of layers 2–4 in the proposed analysis pipeline (Fig. 2) is still missing. Neither of the developed software tools has been available for analyzing the sets of existing clinical VKG recordings: the tool of Manfredi et al. [27] has not been released for external use and its exploration has been limited to preliminary or case studies [29,30], whereas the KIPS software allows analyzing only DKG and not VKG recordings. Furthermore, performance of neither of these software tools has been validated against visual assessment.

The goal of this project was to fundamentally address these issues, targeting two particular objectives:

Objective I was to develop and test a user-friendly software tool for automated analysis of clinical videokymographic recordings that can be used in a clinical setting. This software predominantly targets layers 2–4 in the proposed analysis pipeline (see Fig. 2). In this context, we present (a) a GUI for acquiring kymographic videos or images and their pre-processing; (b) a segmentation algorithm that supports user-defined image adaptations; and (c) implementation of a number of clinically relevant metrics.

Objective II was constituted by a rigorous validation of the implemented segmentation and feature extraction algorithms. This is achieved by comparing the proposed toolset with manual visual assessments, in order to verify the accuracy and applicability of the proposed solution. For this, we took advantage of a previously developed protocol for visual analysis of videokymograms using pictograms [24,31], transferring the visual pictogram features into quantitative vibratory parameters. This was achieved for clinically relevant features, such as: the relative duration of glottal closure, left–right differences in vibratory amplitudes, frequencies and phases, left–right axis shifts, opening versus closing durations, and cycle-to-cycle variability [1,32].

2. Materials and data

2.1. Datasets

All the data used in our study were acquired in cooperation with the Voice and Hearing Centre Prague, a medical institution specialized

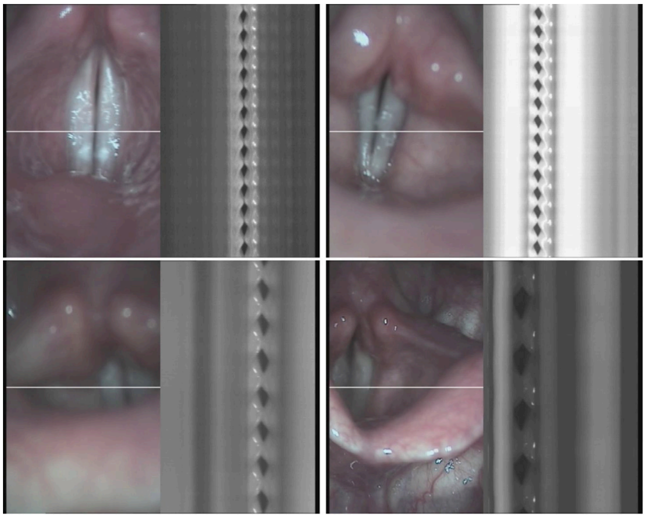


Fig. 3. Examples of video frames extracted from routine clinical VKG recordings demonstrate the variability of laryngeal settings, vibratory patterns, and image quality. The vibrations may show lack of glottal closure or can be asymmetrical. The glottal opening can contain specular reflections, be over-saturated, blurred, or even off-center. Such variability had to be taken into account for developing the VKG Analyzer software.

in voice diagnostics. Routinely acquired videokymographic recordings of patients with and without voice disorders were used; no special recordings were made for this study. No exclusion criteria were applied here on the subjects; our primary goal was to obtain and analyze recordings showing the largest possible variability of findings and image quality, regardless of the clinical diagnosis, gender or age. Healthy as well as disordered patients were included to ensure the robustness of the proposed methods. The most common diagnoses included were: laryngitis chronica, hyperfunctional dysphonia, oedema laryngis, hemorrhage, vocal fold atrophy, paresis mm. interni and voice fatigue. However, since clinical diagnoses have traditionally been based mostly on structural rather than vibrational features, our goal was not to obtain clinical diagnoses but rather to accurately capture the vibratory features of the vocal folds which provide crucial additional information on the functionality of the vocal folds [32]. For the evaluations, we have therefore selected images containing large variation of vibratory patterns and having various levels of image quality in order to test the robustness of image segmentation and vibration analysis. Examples of the clinical VKG images subjected to our analysis are shown in Fig. 3.

Three datasets were used to create and validate the software. The first dataset consisted of 500 randomly chosen images from clinical VKG examinations of healthy as well as voice-disordered subjects. It was used for a heuristic adaptation of the algorithms and fine-tuning of the parameters. We named this dataset the *“Training Dataset”*. This dataset was used in the **layer 2: pre-processing and layer 3: segmentation** of the processing pipeline as depicted in Fig. 2.

A second dataset, the *“Segmentation Validation Dataset”*, was created to test the performance of the segmentation algorithm (**layer 3: segmentation** in the processing pipeline depicted in Fig. 2). The dataset consisted of manual annotations of 834 key points, i.e., the opening, closing, lateral and medial extrema for left and right vocal fold movement contour, performed by 6 raters, yielding the total of 5004 annotations. Details on this dataset are provided in Section 3.6 devoted to the validation studies.

A third dataset, the *“Attributes Validation Dataset”*, was used to evaluate of the accuracy of the extracted vibration attributes (**layer 4: feature extraction** in diagram displayed in Fig. 2), testing the overall analyzer performance. This dataset contained the total of 13500 visually-based manual evaluations of 9 vibratory features obtained from ten evaluators. These evaluations were performed on 50 VKG images from 50 patients with various voice disorders showing the largest possible range of pathological vibratory patterns and 200 VKG images from 40 healthy patients. Further details on this dataset are also provided in Section 3.6 devoted to the validation studies.

2.2. *Videokymographic data acquisition/voice recording*

In order to acquire the videokymographic images, we used a commercially available 2nd generation videokymography camera (Cymo, Netherlands) connected to a 90° rigid laryngoscope (type 130310529, Xion, Germany) with a bright light source (300-W xenon, type FX 300 A, Fentex, Germany) (Fig. 1). Examples of the videokymographic images from different patients can be seen in Fig. 3. Audio signal has also been captured together with the videokymographic data using an electret microphone (Xion) for perceptual monitoring of the recorded voices.

2.3. *Software tool implementation*

Initial development has been realized in *Image Processing Toolbox for Matlab* [33]. The final application is programmed in C++, complemented by the *openCV* library [34] for image and video handling and by the *Qt* library [35] for the graphical user interface. The *SQLite* database system [36] was used for data storage.

3. Method

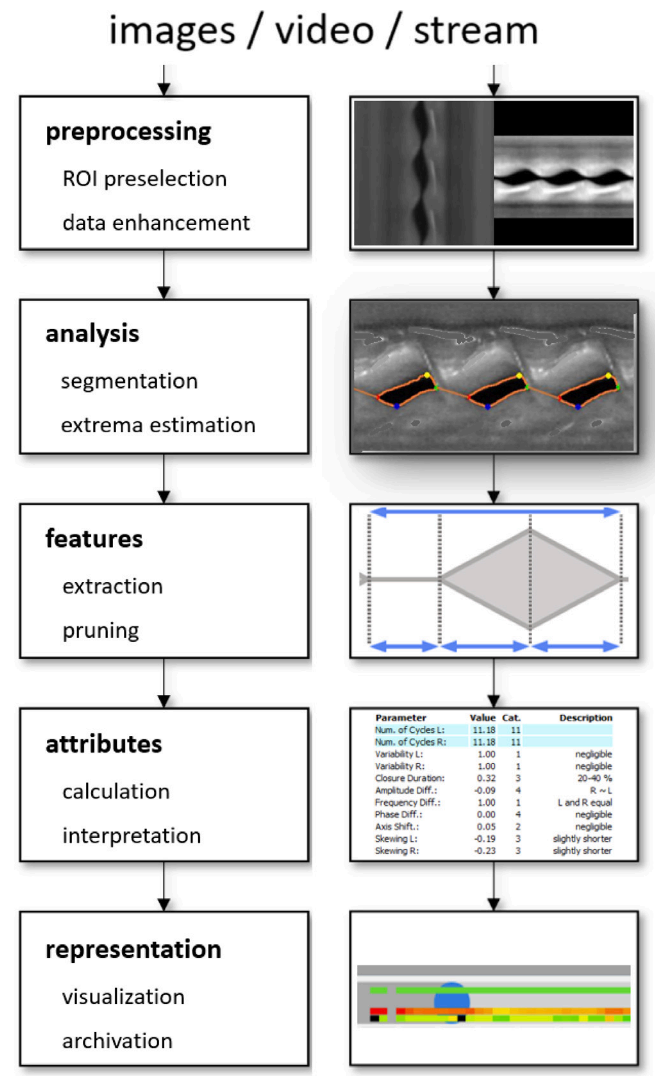
The proposed software solution, addressing the **objective I**, consists of five main building blocks, as shown in Fig. 4. The input data can be in the form of single kymographic images (from DKG, VKG or SVKG modality), a VKG video file, or a live video stream of the VKG examination session. Following the initial information rich frames detection and preprocessing, the software localizes the fundamental vibration structures for every vibration cycle — the contours of glottal openings, the lateral movement extrema, and the opening/closing points. These basic features are used for the derivation of advanced features, and ultimately for computation of the final vibration attributes. Finally, the software visualizes the results in the graphical user interface. In the following paragraphs, the individual pipeline blocs from Fig. 4 are described.

3.1. *Information-rich-frames detection*

A typical VKG video data acquisition process produces many frames containing irrelevant data (the cases where the patient moved, did not phonate, or the resulting images are off-center or low quality). As a part of the preprocessing, the image content richness of every frame is estimated. The content richness detection is based on searching for the vocal fold oscillations' amplitude using the column frequency analysis. The absolute values of the first 32 coefficients of each column's Fourier transformation determine the vibration amplitudes for the relevant frequencies (see Fig. 5). The maximum of calculated amplitudes therefore signifies the level of vibrations in the VKG image. Frames achieving a higher maximum value than the empirically defined threshold are marked. The process of preselecting the content-rich frames helps the physician to focus on the relevant parts of the VKG video, where the vocal folds are visible and vibrating. An interactive visualization tool (the representation part in Fig. 4) helps the user to select the information-rich-frames of interest.

3.2. *Preprocessing of VKG images*

The acquired data contain various degradations (examples are depicted in Fig. 3). Primary goal of Image preprocessing (**layer 2** Fig. 2)



Parameter	Value	Cat.	Description
Num. of Cycles L:	11.18	11	
Num. of Cycles R:	11.18	11	
Variability L:	1.00	1	negligible
Variability R:	1.00	1	negligible
Closure Duration:	0.32	3	20-40 %
Amplitude Diff.:	-0.09	4	R ~ L
Frequency Diff.:	1.00	1	L and R equal
Phase Diff.:	0.00	4	negligible
Axis Shift.:	0.05	2	negligible
Skewing L:	-0.19	3	slightly shorter
Skewing R:	-0.23	3	slightly shorter

Fig. 4. Streamline processing pipeline schema. The input sequence is processed frame by frame. The first stage focuses on image preprocessing (layer 2 in Fig. 2). Next, the glottal openings are segmented (layer 3). The segmentation determines the lateral extrema and opening/closing points. Then the derived vibration features and final attributes are calculated (layer 4). Lastly, the software visualizes the results in the graphical user interface.

is to reduce unwanted artifacts caused by the data acquisition process and normalize the input for further processing.

After loading an image containing the VKG data, the system performs adaptive histogram equalization [37] to normalize the image. In this phase, the operator can further adjust the image contrast and brightness using the controls in the program interface.

When the user initiates the automatic extraction of attributes, the algorithm first removes any specular light reflections caused by the mucosal secretions. Here, the pixels having values higher than the preset threshold (set to 200) are replaced by the mean value of all pixels in the same column. This rudimentary inpainting approach is sufficient for the segmentation process. Next, the algorithm cuts out the image borders, which have no informational value. By default, the cutout is set to 1/4th of the image width from both sides. For the rest of the pipeline, only the middle part of the image containing the relevant vibration structures is kept.

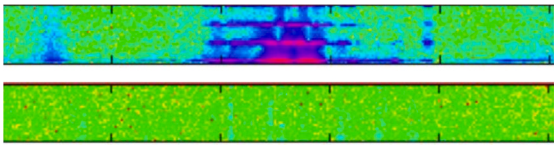


Fig. 5. Spectral analysis for the selection of content-rich images. The example shows the spectral analysis of each column of the VKG image with (top) and without (bottom) pronounced vibrations. The x-axis of the graph shows the VKG image spatial domain; the y-axis shows the first 32 Fourier coefficients' absolute values.

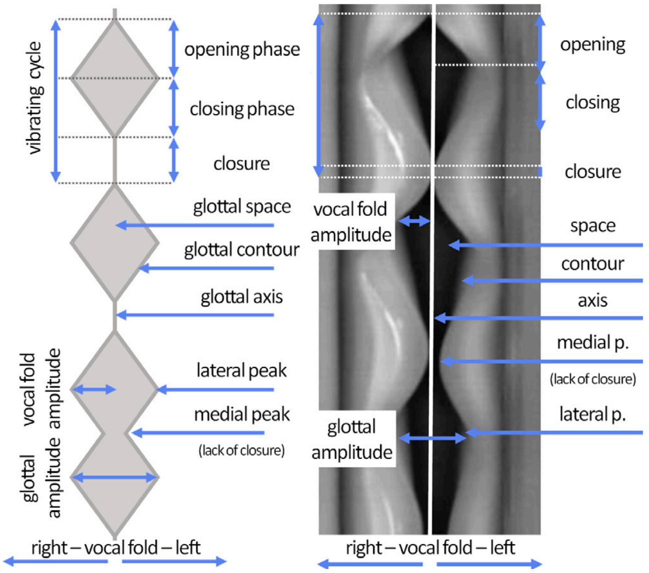


Fig. 6. Vibration features in videokymograms. (a) Schema and (b) real case.

3.3. Segmentation and extrema estimation

Vibration characteristics of the vocal folds in the acquired VKG images are computed from the glottal openings (see Fig. 6). The segmentation of the openings outlines the border between the laryngeal tissue and the open glottis. This part of the processing pipeline covers the **layer 3: segmentation** in the schematic overview (recall Fig. 2).

First, the algorithm determines the active part of the VKG image using the intensity variations in every column. It finds the left-most block of 5 columns of image pixels, all having the standard deviation higher than 0.5. Then it finds the right-most columns of pixels with the same property. The found columns define the part of the VKG image with pronounced vibrations containing the glottal openings for segmentation. In the next phase of the processing, the algorithm executes a segmentation of the glottal area using pixel intensity thresholding. In order to find the global threshold for the image segmentation, the algorithm first estimates the middle line of the glottal opening. Next, the global threshold is estimated using the sorted middle line in the next phase. (see the Algorithm 1). The procedure performs the final segmentation by selecting pixels having an intensity lower than the calculated threshold. All mentioned parameters in both steps of the algorithm were established and fine-tuned empirically on the randomly selected data of 500 VKG frames from healthy and unhealthy patients (the *Training Dataset* defined in Section 2.1).

The global segmentation method can produce unwanted artefacts such as false 'holes' in dark areas of vocal folds. To remove these incorrectly segmented areas, the algorithm performs a morphological opening [38] using a rectangular morphological element of size 3 × 3.

The achieved segmentation determines the contour pixels of the glottal openings (refer to Fig. 7(b)). The extremal points of the contours in the temporal domain (up or down on image) denote the *glottal*

Algorithm 1: Find the Global Threshold

```
sorted_image := SortColumns(image)
new_height := Height(image) * 0.55
sorted_image := sorted_image[1:new_height, :]
column_sums := Sum(sorted_image, 1)
middle_idx := ArgMin(column_sums)
sorted_column = Sort(image[:, midddle_idx])
sorted_column := Filter1D(sorted_column, GAUSS)
for i in 1:Length(sorted_column) do
    if sorted_column[i] $\leq$ 0.1 then
        min_idx := i
    end
    if sorted_column[i] $\leq$ 0.22 then
        max_idx := i
    end
end
sorted_column := sorted_column[min_idx:max_idx]
gradient_vector := sorted_column[1:end-1] - sorted_column[2:end]
for i in 1:Length(sorted_column) do
    if gradient_vector[i] $>$ 0.03 then
        global_threshold := sorted_column[i]
        return global_threshold
    end
end
```

opening (top) and *closing* (bottom) *points*. Between the glottal cycles (defined by the segmented glottal opening(s)), the glottis can be closed (Fig. 7(b)). In such a case, the line connecting the closing of one cycle to the opening of the following cycle is used to approximate the position of the boundary between the left and the right vocal fold during glottal closure. The final tracing contours of the movements of the vocal folds are formed as curves running from the first glottal opening, each on one side, including the connecting lines when the vocal folds are closed, repeatedly for every glottal cycle, until the end of the last glottal opening on the analyzed image frame. These tracing lines are then used for finding the lateral and medial peaks (see Figs. 6 and 7(b)).

Each of the established tracing lines (left and right) can be viewed as a continuous curve. Therefore, we can use the first derivative test to find the lateral peaks (the violet points in Fig. 7(b)) as well as the medial peaks when glottal closure is missing (Fig. 6). Places where the first-order derivative is zero signify the places of either extreme or saddle point. A second-order derivative is used to distinguish between an extreme and the saddle point. The lateral peaks are used for finding the vibration amplitudes. The medial peaks are important to localize for the cases, where the vocal folds do not close completely.

3.4. Features

The main calculation pipeline (**layer 4**, Fig. 2) of the proposed software starts with extracting the basic and advanced (derived) features. The basic vibration features targeted here are: the frequency and regularity of vocal fold vibration, the relative duration of glottal closure, opening versus closing duration, and the left–right vibratory asymmetry. These features are calculated directly from the detected extrema points — namely the frequency, lateral amplitude (vocal fold vibration amplitude), and the maximum opening point (lateral peak). From the combination of the left and right extrema points, we can also determine the phases where the glottis is closed and open. The lateral peaks and the closed phase (or the medial peak when there is no closed phase) are used to detect the opening and closing points. (See Table 1 for reference; the upper index R or L denotes correspondence to the right or left vocal fold; the lower index $i \in \{1, \ldots, n\}$ denotes the number of the corresponding vibration cycle, where n is the number of cycles in the videokymogram.).

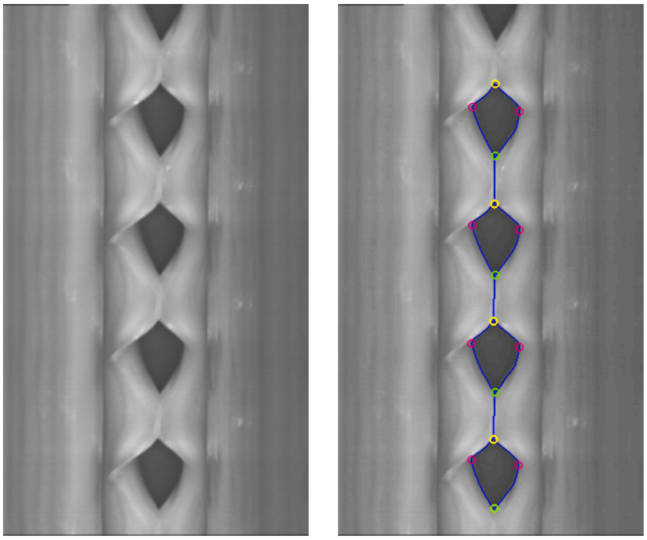


Fig. 7. Glottal space contouring and subsequent main features detection. (a) Original VKG image; (b) Detected glottal opening border with the main feature points: Opening Points, Closing Points, and Lateral Peaks.

Table 1
Basic glottal features in videokymograms (see Fig. 6); upper indices R and L denote the right and left vocal folds and lower index i denotes the number of the vibration cycle in the videokymogram.

Basic feature	Notation
Opening points	O_i
Closing points	C_i
Lateral peaks	A_i^R, A_i^L
Medial peaks	M_i^R, M_i^L

Table 2
Derived glottal features in videokymograms (see Fig. 6); upper index $j \in \{R, L\}$ denotes the right and left vocal folds and lower index i denotes the number of the vibration cycle in the videokymogram.

Advanced features	Notation and definition
Generalized opening points	$\tilde{O}_i^j = \{O_i, M_i^j\}$
Generalized closing points	$\tilde{C}_i^j = \{C_i, M_i^j\}$
Opening phase duration	$t_i^{oj} = A_i^j(y) - \tilde{O}_i^j(y)$
Closing phase duration	$t_i^{cj} = \tilde{C}_i^j(y) - A_i^j(y)$
Open phase duration	$T_i^{oj} = t_i^{oj} + t_i^{cj} = \tilde{C}_i^j(y) - \tilde{O}_i^j(y)$
Closed phase duration	$T_i^{cj} = \tilde{O}_{i+1}^j(y) - \tilde{C}_i^j(y)$
Vibration cycle duration	$T_i^j = T_i^{oj} + T_i^{cj} = t_i^{oj} + t_i^{cj} + T_i^{cj} = \tilde{O}_{i+1}^j(y) - \tilde{O}_i^j(y)$
Vocal fold amplitudes	$a_i^j = \mathrm{mean}(\lvert A_i^j(x) - \tilde{O}_i^j(x)\rvert, \lvert A_i^j(x) - \tilde{C}_i^j(x)\rvert)$
Glottal amplitudes	$a_i = A_i^L(x) - A_i^R(x)$

The derived glottal features are computed from the basic features using the definitions in Table 2. The generalized opening points are defined as the union of opening points and medial peaks, and similarly, generalized closing points are defined as the union of closing points and medial peaks. The generalized opening and closing points enclose open phases, while the generalized opening points separate vibration cycles.[1]

3.5. Attributes

The set of vocal fold vibration attributes used by clinicians was previously implemented by the authors into a visually-perceptual VKG

[1] Depending on the definition, vibration cycles can be separated by the generalized opening points, by the lateral peaks, or by the generalized closing points.

Table 3
Cycle-to-cycle amplitude variability: correspondence between the numerical values of the Amplitude Periodicity Index (API) and categories of the parameter in the VKG visual evaluation form [24,31].

Category	Description	VariabilityR, VariabilityL
1	Negligible	(0.85, 1]
2	Small	(0.61, 0.85]
3	Medium	(0.5, 0.61]
4	Large	[0, 0.5]

evaluation sheet [24,31]. In our software, we aimed at evaluating these visual attributes automatically using a set of parameters derived from the detected glottal features. The intervals for the parameters' discretization (see Tables 3–9) were obtained by manually measuring the pictograms depicting the typical idealized VKG waveforms — those served as visual anchors for the previous visual VKG evaluation studies [24,31]. The discretization of the calculated values is mandatory for a backward reference to manual annotations performed using the VKG visual evaluation tool. Additionally, the human-to-computer comparative study utilizes the discretized values for more straightforward performance evaluation.

The meaning of the parameters and their relation to the VKG features was defined as follows:

Number of cycles

(1) $\text{NumberOfCyclesR} = \frac{y_{max}}{\overline{T}^{R}}$

(2) $\text{NumberOfCyclesL} = \frac{y_{max}}{\overline{T}^{L}}$

The "*Number of cycles*" parameter is defined by the duration of the recorded videokymogram y_{max} and the average length of the vibration cycle $\overline{T}^{j} = \frac{1}{n_j} \sum_{i=1}^{n^j} T_i^j$, where $j = R, L$ and n^R and n^L denote the number of full cycles of the right and left vocal fold in the videokymogram determined from the total number of detected opening points O^R and O^L, respectively.

This parameter is directly related to the fundamental frequency of oscillations of the vocal folds and consequently to the produced fundamental frequency of voice.

Cycle-to-cycle variabilities

(3) $\text{VariabilityR} = \underset{i=1,\dots,n-1}{\text{median}}\ API(i, R)$

(4) $\text{VariabilityL} = \underset{i=1,\dots,n-1}{\text{median}}\ API(i, L)$

The cycle-to-cycle amplitude variability indicates how much the vocal fold vibration amplitudes deviate from ideal periodic vibrations. This feature is related to the degree of voice roughness [39]. The "*Cycle-to-cycle amplitude variability*" parameter is defined by the Amplitude Periodicity Index (API) [40] $API(i, j) = \frac{\min\{a_i^j, a_{i+1}^j\}}{\max\{a_i^j, a_{i+1}^j\}}$, where $i = 1, \dots, n-1, \quad j = R, L$. Analogously, the "*Cycle-to-cycle period variability*" can be defined through the Time Periodicity Index (TPI) [40] $TPI(i, j) = \frac{\min\{T_i^j, T_{i+1}^j\}}{\max\{T_i^j, T_{i+1}^j\}}$, where $i = 1, \dots, n-1, \quad j = R, L$.

Duration of closure

(5) $\text{ClosureDuration} = \underset{i=1,\dots,n}{\text{median}}\, CQ(i)$

The relative duration of glottal closure is a classic feature that indicates how well the vocal folds close during phonation [32,41]. The relative duration of the closure is defined by the Closed Quotient (CQ) [40] as $CQ(i) = \frac{T_i^c}{T_i}$, $i = 1, \dots, n$.

Amplitude differences

(6) $\text{AmplitudeDifferences} = \underset{i=1,\dots,n}{\text{median}}\, ASI(i)$

Table 4
Duration of closure: correspondence between the numerical values of the Closed Quotient (CQ) and categories of the parameter in the VKG visual evaluation form [24,31].

Category	Description	Closure duration
1	No closure	[0, 0.01]
2	1–20	(0.01, 0.2]
3	20–40	(0.2, 0.4]
4	40–60	(0.4, 0.6]
5	>60	(0.6, 1]

Table 5
Amplitude differences: correspondence between the numerical values of the Amplitude Symmetry Index (ASI) and categories of the parameter in the VKG evaluation form [24,31].

Category	Description	Amplitude difference
1	R much larger	[−1,−0.6)
2	R larger	[−0.6,−0.31)
3	R slightly larger	[−0.31,−0.1)
4	R ~ L	[−0.1, 0.1]
5	L slightly larger	(0.1, 0.31]
6	L larger	(0.31, 0.6]
7	L much larger	(0.6, 1]

Table 6
Frequency differences: correspondence between numerical values and categories of the parameter in the VKG evaluation sheet [24,31].

Category	Description	Frequency difference
1	R faster than L	(0, 0.91)
2	L and R equal	[0.91, 1.1)
3	L faster than R	[1.1,1)

The difference in vibration amplitude of the left and right vocal folds shows the vocal fold asymmetry and can help clinicians discover unilateral pathologies hindering the vibratory ability of the vocal folds [32,41]. The "*Amplitude difference*" parameter is defined by the Amplitude Symmetry Index (ASI) [40] $ASI(i) = \frac{a_i^L - a_i^R}{a_i^L + a_i^R}$, $i = 1, \dots, n$.

Frequency differences

(7) $\text{FrequencyDifferences} = \frac{\text{NumberOfCyclesL}}{\text{NumberOfCyclesR}}$

This parameter allows discovering differences in the fundamental frequencies of the left and right vocal folds. In normal phonation, the left and right vocal folds are expected to vibrate at the same fundamental frequencies. In the case of left–right frequency differences, the voice may become biphonic or diplophonic [32,41,42]. The "*Frequency difference*" parameter is defined as a ratio between the number of left and right cycles (see parameters (1)–(2)).

Phase differences

(8) $\text{PhaseDifferences} = \underset{i=1,\dots,n}{\text{median}}\, PSI(i)$

The "*Phase difference*" parameter is defined by the Phase Symmetry Index (PSI) as [40] $PSI(i) = \frac{A_i^L(y) - A_i^R(y)}{T_i}$, $i = 1, \dots, n$. This parameter provides information on the possible asymmetry between the tension of the left and right vocal folds.

Axis shifts

(9) $\text{AxisShift} = \underset{i=1,\dots,n}{\text{median}}\, AS(i)$

The "*Axis shift*" parameter is the third parameter revealing the left–right asymmetry of the vocal fold vibration [32]. In contrast to the phase differences, which are mainly visible during the open phase of the glottal vibratory cycle, the axis shift allows discovering the left–right asymmetries during the closed phase of the glottal vibratory cycle [32,41]. The "*Axis shift*" parameter (AS) is defined as [43] $AS(i) = \frac{O_{i+1}(x) - C_i(x)}{a_i}$, $i = 1, \dots, n$.

Table 7
Phase differences: correspondence between the numerical values of the Phase Symmetry Index (PSI) and categories of the parameter in the VKG visual evaluation form [24,31].

Category	Description	Phase differences
1	R ahead of L: large	(0.3, 1]
2	R ahead of L: medium	(0.15, 0.3]
3	R ahead of L: small	(0.05, 0.15]
4	Negligible	[−0.05, 0.05]
5	L ahead of R: small	[−0.15,−0.05)
6	L ahead of R: medium	[−0.3,−0.15)
7	L ahead of R: large	[−1,−0.3)
14	lambada: large	Yet to be quantified

Table 8
Axis shift: correspondence between numerical values and categories of the parameter in the VKG evaluation form; the evaluation sheet denotes the “R → L” category by 2, the “negligible” category by 1, and the “complex” category by 4.

Category	Description	Axis shift
1	R → L	(0.1,1)
2	Negligible	[−0.1, 0.1]
3	L → R	(−1,−0.1)
6	Complex	Yet to be quantified

Table 9
Opening versus closing duration: correspondence between the numerical values of the Speed Index (SI) and categories of the parameter in the VKG evaluation form.

Category	Description	SkewingR, SkewingL
1	Much shorter	[−1,−0.75)
2	Shorter	[−0.75,−0.35)
3	Slightly shorter	[−0.35,−0.05)
4	Equal	[−0.05, 0.05]
5	Slightly longer	(0.05, 0.35]
6	Longer	(0.35, 0.75]
7	Much longer	(0.75, 1]

Opening versus closing durations, cycle skewing

$$(10)\ \ \text{SkewingR} = \underset{i=1,\dots,n}{\text{median}}\ SI(i,R)$$

$$(11)\ \ \text{SkewingL} = \underset{i=1,\dots,n}{\text{median}}\ SI(i,L)$$

The opening and closing phases of the vibration cycle of the vocal folds can have different duration. These differences appear as a skewing of the vocal fold vibratory pattern and provides clinically interesting information [32,41,44]. The skewing can differ for the left and right vocal fold and reveals the vocal fold vibration’s detailed dynamics.

The “*Opening versus closing duration*” / “*Skewing*” parameter can be quantified by the Speed Index (SI) [45] $SI(i,j) = \frac{t_i^{oj}-t_i^{cj}}{T_i^o} = \frac{t_i^{oj}-t_i^{cj}}{t_i^o+t_i^c} = \frac{SQ(i,j)-1}{SQ(i,j)+1}$, $i = 1,\dots,n, j = R, L$, which is derived from the Speed Quotient (SQ) [46,47] $SQ(i,j) = \frac{t_i^{oj}}{t_i^{cj}}$, $i = 1,\dots,n, j = R, L$.

3.6. Verification studies

To address the **objective II**, two studies were done to compare the performance of the proposed image analysis tool with the clinician visual assessments and verify the usability of the proposed algorithms. The first study evaluated the accuracy of the segmentation process, which is the critical tool for further feature extraction. The second study focused on the estimated vibration attributes and their comparison to the clinician visual assessments.

The first verification study addressed the segmentation accuracy of our algorithm (**layer 3** Fig. 2) using the *Segmentation Validation Dataset* described in Section 2.1. It consisted of annotated key extrema points of the vibration waveforms, i.e., the opening and closing points, and the left and right lateral and medial peaks (recall Figs. 6 and 7(b)). The validation procedure was analogous to the one used by Lohscheller et al. [48]: using an auxiliary manual annotation tool, six expert examiners denoted 834 key points on the set of clinical VKG images yielding the total of 5004 annotations. The images were selected to represent different degradation levels (e.g., noise, blur, or presence of specular reflections) and various types of healthy and pathologic vocal fold vibrations that could influence the segmentation accuracy, regardless of particular clinical diagnoses. The annotated key point positions were compared to the mean of the other annotators’ key points to verify the robustness of the annotations. The ground truth for the 834 key points was then established as the mean of the manually detected coordinates. This procedure ensured the quality of the annotators’ performance. Examples of the annotated points, together with the locations of the points detected automatically by the VKG Analyzer are shown in Fig. 8. The segmentation accuracy of the tool was assessed as the distance errors between the automatically detected key points and those obtained by the manual procedure. We analyzed the errors in both the spatial and temporal domains.

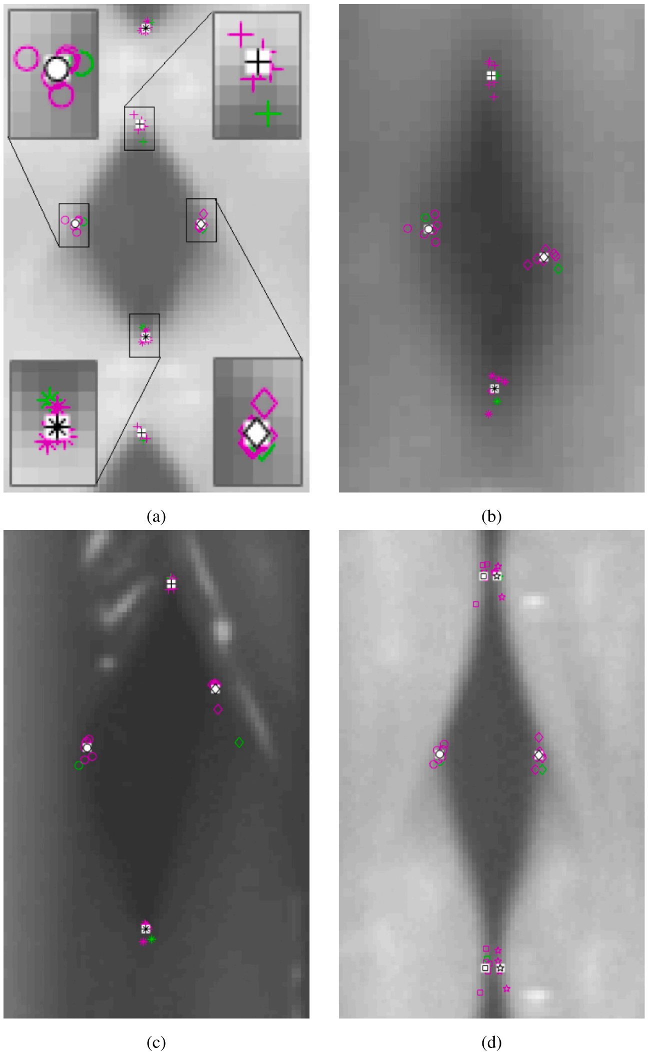
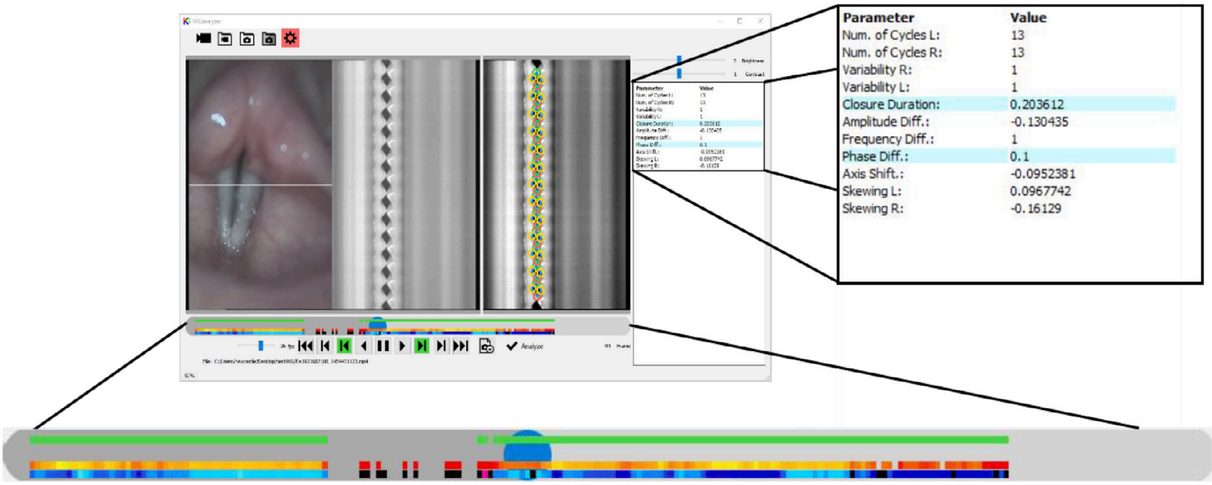


Fig. 8. Examples of the segmentation comparison study. The plus sign, star, diamonds, and circles denote the key points of opening, closing, left, and right lateral extremes, respectively. The magenta color codes positions selected by examiners, the white color codes the average of all examiners’ positions, and the green color denotes the points automatically estimated by the algorithm. In image a), the selected areas are also magnified so that the pixelization of the images is clearly visible.



Parameter	Value
Num. of Cycles L:	13
Num. of Cycles R:	13
Variability R:	1
Variability L:	1
Closure Duration:	0.203612
Amplitude Diff.:	-0.130435
Frequency Diff.:	1
Phase Diff.:	0.1
Axis Shift.:	-0.0952381
Skewing L:	0.0967742
Skewing R:	-0.16129

Fig. 9. Program layout: the left pane shows the dual image as produced by the camera system; next to it, there is the VKG image after contrast and brightness normalization and image de-noising, together with the detected glottal contours and key points; the right pane shows the calculated features of the VKG recording; the bottom pane displays the slide-bar for video scrubbing, as well as the color-coded feature visualization bars.

The second verification study aimed to compare the automatically calculated vibratory features (**layer 3-4** Fig. 2) with those that were manually evaluated by human examiners using the pictogram-based VKG visual evaluation tool [31]. The visual evaluation data previously gathered by Hampala [24] were used for this purpose, partially forming the third dataset, i.e., the *Attributes Validation Dataset*. The original trial involved 50 VKG images obtained from 50 patients with various voice disorders, again showing the largest possible range of healthy and pathological vibratory patterns to which we added another 200 VKG images from 40 healthy subjects. Ten evaluators manually labeled the 50 images of pathological patients using the VKG visual evaluation tool [31], paying attention to 33 vibratory features per image. Eight of these evaluators performed the visual analysis twice for test–retest comparison purposes. This resulted in the total of 29 700 (50 × 18 × 33) manual evaluations. Nine out of the 33 features, thus 8 100 manual evaluations were selected for the purpose of this study. The 200 images of healthy patients were evaluated by three experts annotating the same nine features forming another set of 5400 (200 × 3 × 9) manual evaluations.

The resulting compilation of 250 visually-evaluated clinical images was then subjected to the objective image analysis by the VKG Analyzer tool. The feature values quantified by the software tool were discretized into the visually-based categories using the conversion tables defined in Section 3.5. The individual test–retest comparison results were divided into three categories – *Correct, Partially correct or indecisive, and Incorrect* [24]. Since the evaluation is fundamentally subjective, the label *Partially correct/indecisive* was introduced. In our study, it means that the algorithm misclassified the result to the neighboring category. E.g., “slightly larger amplitude difference” instead of “larger amplitude difference”, etc. This decision was a result of previous observations, that this level of error is common within human evaluation even for repeated evaluation by the same expert [24].

4. Results

4.1. VKG Analyzer tool — representation

The developed user interface (UI) of the tool (addressing **objective I**) is shown in Fig. 9. The emphasis was on visualization clarity and ease of use so clinicians could easily use the VKG Analyzer tool during routine patient examinations. The user interface is divided into 4 areas. The main part of the UI is used to visualize the kymographic image. The left top part shows the original recorded image. In Fig. 9 it is the dual image produced by the VKG camera providing the standard and VKG views, but it could also be a DKG or SVKG image. Next to it, there is the processed kymographic image together with the detected borders and estimated features. The top right pane shows a list of computed vibration attributes.

The bottom part of UI is designed to visualize the video timeline with color-coded values of relevant vibration attributes. A user can select a set of parameters for visualization in the right pane (see Fig. 9). This is helpful particularly for evaluations of video recordings performed with the VKG camera. The interactive visualization timeline slide-bar helps clinicians to find instances of interest directly, eliminating the time-consuming process of frame-by-frame visualization and analysis of the whole video recording. Additionally, the green line at the top of the bar indicates the information-rich video frames where oscillations were detected and which were marked during the pre-selection phase. All the processed data can be stored for later analysis, and the stored records can be analyzed repeatedly. Furthermore, the analyzed data, e.g., the analyzed frames with the segmented contours and the extracted parameters can be exported and used for further external analyses.

4.2. Segmentation precision

The results of the first validation study addressing **objective II**, which focuses on the accuracy of the automatic segmentation tool, are revealed in Tables 10 and 11 showing the mean and standard deviations of the key point positions with respect to the human ground truth. These data are also visualized in Fig. 10.

In spatial domain (left–right accuracy), the mean difference between the software-detected individual key points and their average manual locations was always less than one pixel (refer to Table 10, last row, and to the horizontal differences between the average manual and software results shown in the individual graphs of Fig. 10). The smallest manual vs. automatic average difference was found for the right lateral peak (0.05 pixels) and largest one for the left medial peak (−0.73 pixels). The manual vs. automatic differences show considerable standard deviations, however (up to ±1.08 pixels for the right lateral peak), revealing that the software vs. average manual positions differed across different vibratory cycles. This variability is, nevertheless, comparable to the uncertainty of the manual location of the key points (the largest standard deviation was ± 0.92 pixels for the left lateral peak, see Table 10, second last row), thus suggesting that the software inaccuracy is similar to that of the manual evaluations. Considering all the key points together, the average difference between their automatic and manual locations was 0.12 ± 0.79 pixels (Table 10, last row, last

Table 10
Manual and automatic segmentation accuracy in spatial domain (left–right accuracy), expressed in pixels. Average manual locations of the key points (refer to Fig. 6)) were used as the reference (zero) points. Mean differences from the reference points and their variability (i.e., standard deviation) are shown for the individual raters (rows 1-6) and for the automatic (SW row) segmentation results for each key point. The mean row shows the uncertainty (i.e., standard deviation) of the manual location of the reference points. Last column provides the results for all the key points pooled together. For the up-down accuracy, see Table 11 and for graphical representation of these results, see Fig. 8 and Fig. 10.

	L lateral	L medial	Opening	Closing	R lateral	R medial	All Points
1	0.13 ± 0.66	−0.08 ± 0.43	−0.03 ± 0.45	0.01 ± 0.43	−0.44 ± 0.72	−0.01 ± 0.60	−0.07 ± 0.62
2	−0.39 ± 0.71	−0.29 ± 0.45	−0.01 ± 0.47	−0.04 ± 0.44	0.13 ± 0.73	0.32 ± 0.37	−0.05 ± 0.63
3	−0.68 ± 0.97	−0.19 ± 0.52	−0.22 ± 0.44	−0.22 ± 0.46	0.33 ± 0.77	0.42 ± 0.51	−0.09 ± 0.79
4	1.17 ± 1.24	−0.13 ± 0.66	0.17 ± 0.45	0.27 ± 0.56	−0.57 ± 1.03	−0.09 ± 0.41	0.14 ± 1.08
5	0.51 ± 0.98	1.15 ± 1.07	0.01 ± 0.51	−0.05 ± 0.49	−0.09 ± 0.81	−0.79 ± 0.64	0.12 ± 0.81
6	−0.75 ± 0.80	−0.47 ± 0.55	0.09 ± 0.42	0.04 ± 0.43	0.63 ± 0.86	0.14 ± 0.36	−0.05 ± 0.83
Mean	**±0.92**	**±0.65**	**±0.46**	**±0.47**	**±0.83**	**±0.49**	**±0.81**
SW	**−0.2 ± 0.91**	**−0.73 ± 0.34**	**0.13 ± 0.51**	**−0.17 ± 0.49**	**0.05 ± 1.01**	**0.29 ± 1.08**	**−0.12 ± 0.79**

Table 11
Manual and automatic segmentation accuracy in temporal domain (up–down accuracy), expressed in pixels. The organization of the Table is identical to that in Table 10. For graphical representation of these results, see Fig. 8 and Fig. 10.

	L lateral	L medial	Opening	Closing	R lateral	R medial	All Points
1	0.02 ± 1.10	−0.01 ± 1.57	−0.5 ± 0.94	−0.63 ± 0.92	−0.13 ± 0.98	−0.06 ± 1.71	−0.22 ± 1.06
2	0.86 ± 1.07	−0.19 ± 1.74	0.05 ± 1.22	−0.04 ± 1.12	0.4 ± 1.01	−0.45 ± 2.72	0.11 ± 1.24
3	−0.62 ± 1.24	1.36 ± 1.92	−0.28 ± 1.35	0.04 ± 1.04	−0.24 ± 0.89	0.91 ± 2.18	0.2 ± 1.25
4	0.45 ± 1.33	0.84 ± 1.96	1.72 ± 1.56	−0.55 ± 1.26	0.23 ± 1.23	1.48 ± 2.35	0.7 ± 1.60
5	−0.41 ± 1.39	−2.25 ± 2.75	0.41 ± 1.10	−0.3 ± 0.96	−0.16 ± 1.07	−1.85 ± 2.39	−0.76 ± 1.34
6	−0.3 ± 0.89	0.25 ± 1.44	−1.41 ± 1.09	1.47 ± 1.22	−0.11 ± 0.88	−0.04 ± 1.95	−0.02 ± 1.45
Mean	**±1.18**	**±1.94**	**±1.23**	**±1.09**	**±1.02**	**±2.24**	**±1.33**
SW	**0.18 ± 1.19**	**0.03 ± 1.35**	**0.87 ± 1.19**	**−1.15 ± 1.04**	**0.9 ± 1.44**	**0.43 ± 1.63**	**0.21 ± 1.48**

column), revealing that the performance of the automatic segmentation is very similar to the manual one, even though there is variability across individual vibratory cycles and different key points.

In the time domain (up-down accuracy), the differences between the software and manual locations of the key points were mostly larger than those in the spatial (left–right) domain (Table 11). Also, the standard deviations were larger here, revealing larger variability of the differences between the manual vs. automatic locations of the key points as well as larger uncertainty of the manual location of the key points. This is visually reflected also in the plots of Fig. 10, mostly showing the error bars to be longer in vertical than in horizontal direction. The largest uncertainty was found for the manual location of the medial peaks (±1.9 and ±2.2 pixels for the left and right medial peak, respectively, see Table 11, second last row). The largest differences of the automatic positions from the manual averages were found for the Opening and Closing Point (0.9±1.2 and −1.2±1.0 pixels, respectively), and for the Right Lateral Peak (0.9±1.4 pixels, Table 11, last row). Considering all the key points together, however, the average difference between the manual and automatic locations was only 0.2 pixels with the standard deviation of ±1.48 pixels (Table 11, last row, last column) again revealing that the performance of the automatic segmentation is, on average, similar to the manual one.

4.3. Precision of attributes

The results of the second validation study addressing **objective II** and comparing the estimated vibration attributes to the clinician visual assessments are depicted in Fig. 11. For the healthy subjects' data, 91% of cases were in agreement with the human assessment. For the disordered patients, the software tool's performance agreed with the manual annotation assessments in more than 84% of cases.

5. Discussion

The goals of this work were to develop and test a user-friendly software tool for automated analysis of clinical videokymographic recordings (**objective I**) and to perform a rigorous validation of the implemented segmentation and feature extraction algorithms (**objective II**).

Both objectives were fulfilled. The developed VKG Analyzer tool facilitates selecting and exporting individual frames from the video recordings (see layer 2 in Fig. 2) and provides the means for automatically segmenting the vibrating glottal contours and for detecting key points of vocal fold vibration (layer 3 in Fig. 2). These data can then be subjected to automated feature extraction (layer 4 in Fig. 2).

Because videokymographic data have a different structure than standard laryngoscopic images, novel algorithms had to be developed and tested in order to facilitate proper kymographic image processing. During algorithm design, software implementation and validation, a number of noteworthy issues arose, which are being discussed in the following paragraphs.

5.1. Segmentation method

The approach utilized here differs from the previously explored image segmentation algorithms. While numerous segmentation methods have been developed to process high-speed videolaryngoscopic images [49], these cannot be utilized in VKG recording processing because the input images have different formats and meanings. To achieve the best segmentation results, we have experimented with the Active Contours (Snakes) approach (used for example in [28,50]) but ultimately opted not to use it due to the higher time demands and dependency on good initialization. Other methods we experimented with were the Region Growing methods [48,51,52], Graph-cuts [53], classical thresholding approaches like Otsu thresholding [54], watershed [55], and others. The Region Growing methods were found to be slow and dependent on good initialization. Graph-cut algorithms were promising initially, but in the end, they were hard to initialize correctly. Finally, the standard thresholding methods were fast but did not produce satisfactory results. The problem of correct initialization of certain methods is a circular one. Usually, the initialization consists of identifying pixels inside the glottal opening, but when known, the segmentation is not needed in the first place.

Our final solution is based on a handcrafted segmentation algorithm for finding the best threshold for segmentation. This approach has proven to be both robust and fast. In contrast to the Snakes algorithm (≈0.5 s per frame [50]), our implementation runs in real-time (< 0.04 s per frame). The image pre-processing and segmentation methods use

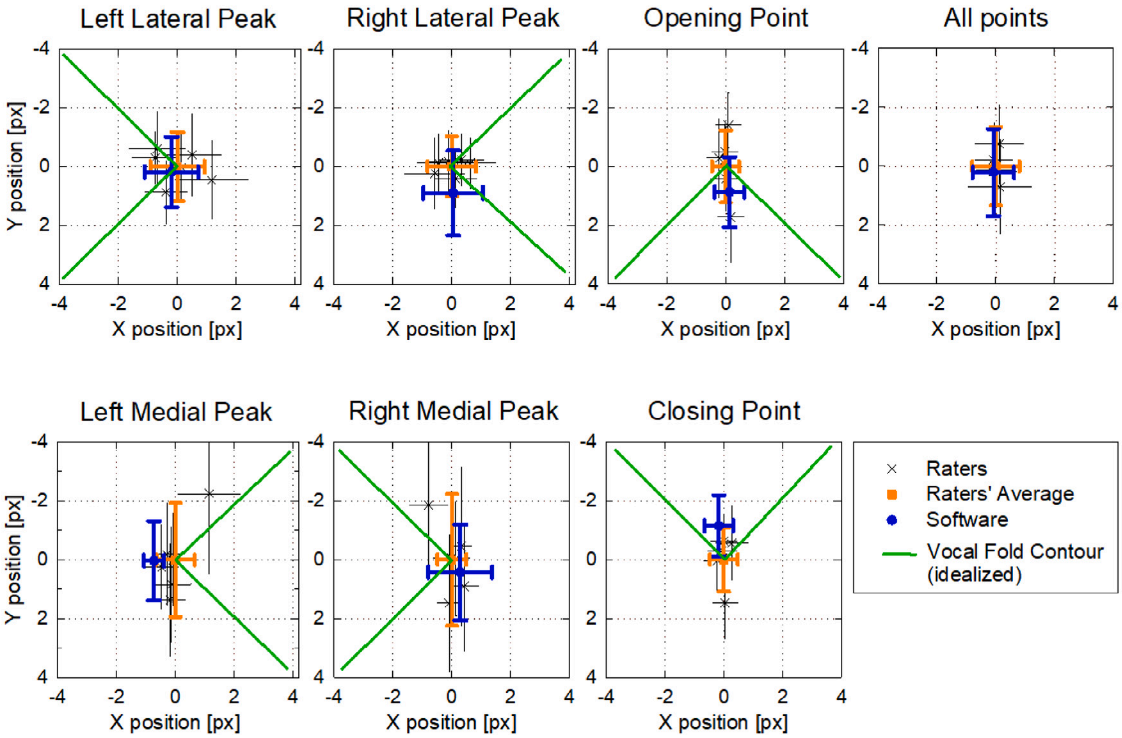


Fig. 10. Segmentation study — graphical representation of the results in Tables 10 and 11. The depicted crosses show the relative mean positions and their sizes in the x-axis and y-axis represent the standard deviations in the spatial and temporal domains, respectively. The overall human average value is colored orange and represents the ground truth, or golden standard, to which the software results (in blue) were compared. The black crosses represent the six human annotators.

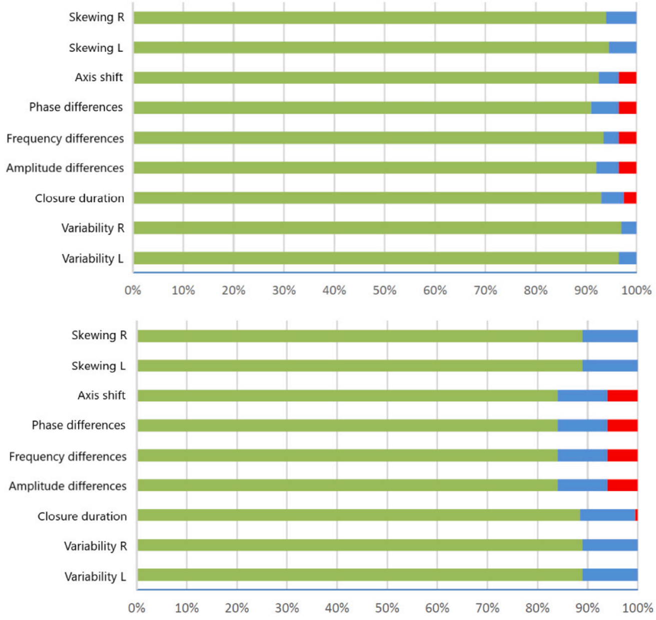


Fig. 11. Results of automatic features extraction by our program compared to visual evaluation by human experts for healthy (top) and disordered patients (bottom). The bars show percentage representation of correct results (green), partial correct (blue) and incorrect results (red) as evaluated by machine vs. human experts.

parameters and thresholds that needed to be determined empirically, however. To fine-tune these parameters, we used the *Training Dataset* (defined in Section 2.1) and performed a parameter search optimizing the resulting algorithm performance.

5.2. Segmentation accuracy

To test the correctness and robustness of our segmentation method, the automatic segmentation results were compared to the key point coordinates segmented manually. Considering the results across all the key points together, there were negligible differences between their automatic locations and the raters' manual average (recall Fig. 10, plot for All points). Furthermore, comparison of the error bars in the same plot reveals that the variability of the manual-to-automatic differences was very similar to the uncertainty of the manual location of the key points, suggesting that the performance of the automatic segmentation algorithm is comparable to the manual segmentation.

Nevertheless, there is a tendency of the software to locate the Opening Points about 1 pixel later, and the Closing Points about 1 pixel earlier, than the raters (Fig. 10, plots "Opening point" and "Closing point"). Taking into account the time running towards the bottom of the VKG image, this makes the duration of the open phase to be slightly shorter than when evaluated manually. This case is also reflected in annotations of the Opening and Closing points in Fig. 8(a) suggesting that manual annotators considered slightly different threshold between the vocal fold and the glottis — the software tends to locate the glottal boundary at slightly darker pixels inside the glottis than the raters. This tendency is detectable also in the plots for the Lateral and Medial peaks in Fig. 10 showing analogous, but much smaller shifts of the automatic key point locations to the right or to the left side, always towards the glottis (see the shifts of the blue versus the orange crosses in Fig. 10 horizontally). Nevertheless, considering the theoretical inaccuracy limit of 1 pixel, the observed differences between the manual and automatic evaluations smaller than c. 1 pixel are deemed acceptable. The tool can therefore be considered as a valid alternative to the manual procedures.

In this respect, it should be noted that manual annotators are not always consistent in their evaluations. Differences among the individual raters are visible in the spread of their annotations for the different key points (black crosses in the plots of Fig. 10). More specifically, Fig. 8(d) demonstrates the low precision of human experts particularly in the temporal (y-coordinate) domain for determining the Medial Peaks in cases of missing closure points. The low precision can be attributed to the roundedness of the contours making it difficult to locate the exact position of the peaks. The raters' uncertainty is reflected also in the large standard deviations (close to ±2 pixels) for the Left and Right Medial Peaks in Table 11 (second last row) and in the correspondingly large orange vertical error bars of the respective plots in Fig. 10.

The segmentation precision study underlines the strengths and weaknesses of our approach. The influence of the input data quality on the segmentation precision is shown in Fig. 8(c). The shift in left (right on image) lateral opening is caused by an imprecise segmentation threshold estimation due to the low image contrast. For purposes of the study, the contrast of the input image was unchanged, although the software tool allows a manual correction of contrast and brightness. The algorithm performed as expected for images with sufficient input image contrast (example can be viewed in Fig. 8(a)).

5.3. Feature extraction

Our VKG Analyzer tool implemented a larger amount of parameters than preceding tools for VKG analysis [28,29]. To enable easier interpretation of the numerical results for the clinicians, we derived empirical ranges for relating the numerical results of the quotients to descriptive categories defined in [31]. This made it possible to perform a comparative study, that aimed to address the precision and objectivity of extracted characteristics.

The result of the study (Fig. 11) show good software agreement with human examiners, namely in more than 91% of the cases for healthy patients (top graph) and more than 84% of the cases for disordered patients (bottom graph), depicted by the green segments of the graphs. We find this result satisfactory.

Additionally, the study revealed that in many cases, for the same image, the same examiner evaluated some attributes differently when the tests were performed several days apart. This experiment underlines the subjectivity of the task, and consequently, the difficulty of obtaining objective ground truth. To incorporate inconsistencies of the human evaluations into the study, we marked the mis-classifications to the neighboring categories as "Partially Correct" (see Fig. 11 blue segments of the graphs). A disadvantage of this approach is that it considers misclassifications on different attributes as equally significant, although different attributes have different interpretations and importance. Nevertheless, this approach allows good insight into the accuracy of the visual as well as visual versus automatic image assessment.

5.4. Additional software features

A noteworthy feature of the presented software is that it is designed to process not only VKGs, but also DKG and SVGK images. Furthermore, it allows to export the extracted glottal contours to a file in order to be analyzed by another means. These exported contour data, created by our tool, have already been successfully used in other detailed studies providing good applicability of the developed software framework [26, 56,57].

5.5. Overall assessment

Results of both the validation studies indicate that the developed software is a valid, fast and robust automatic tool for vocal fold vibration analysis with minimal hardware requirements.

The comparison of the objectively measured attributes, which are automatically estimated by the developed software to visual assessments of ten evaluators makes this study unique. To the best of our knowledge, this is the first study that relates visual perception of such videokymographic features to objectively measured parameters. This rigorous and thorough validation ensures reliable application of the developed tool.

6. Summary

In the context of this study, we have developed and introduced a novel software tool for automated segmentation and feature extraction of all sorts of kymographic data (VKG, DKG, and even SVKG). The software is capable of automatically calculating the vocal folds' fundamental and derived vibration attributes. Additionally, it helps clinicians to focus on the information-rich sections of the VKG video recording by automatically pre-selecting such images from the recorded VKG examination session.

The software and its algorithms have been subjected to a rigorous validation at unprecedented scope, ensuring robust and reliable application in both a clinical and a research setting. Based on comparative results, the vibration attribute estimation demonstrated agreement with manual annotation in more than 91% (healthy patients) and 84% (disordered patients) cases. Owing to these outstanding validation results, the software is expected to become a robust and reliable state-of-the art tool for clinical and scientific examination of vocal fold vibrations and laryngeal function.

CRediT authorship contribution statement

Aleš Zita: Conceptualization, Methods research, Software, Writing. **Adam Novozámský:** Methods research, Software, Investigation, Writing. **Barbara Zitová:** Conceptualization, Supervision, Investigation, Writing. **Michal Šorel:** Methods research. **Christian T. Herbst:** Software, Writing. **Jitka Vydrová:** Investigation, Data acquisition, Medical expert and consultant, Validation. **Jan G. Švec:** Voice researcher and consultant, Data validation, Data acquisition, Writing.

Declaration of competing interest

One or more of the authors of this paper have disclosed potential or pertinent conflicts of interest, which may include receipt of payment, either direct or indirect, institutional support, or association with an entity in the biomedical field which may be perceived to have potential conflict of interest with this work. For full disclosure statements refer to https://doi.org/10.1016/j.bspc.2022.103878.

Acknowledgments

The research was supported by the Technology Agency of the Czech Republic projects no. TA04010877 (2014–2017) and no. TH04010422 "VKG 3.0" (2019–2021) and Czech Science Foundation, grant no. GA21-03921S (2021–2023). .

References

[1] W. Angerstein, G. Baracca, P. Dejonckere, M. Echternach, U. Eysholdt, F. Fussi, A. Geneid, T. Hacki, K. Karmelita-Katulska, R. Haubrich, et al., Diagnosis and differential diagnosis of voice disorders, in: Phoniatrics I, Springer, 2020, pp. 349–430.

[2] R.R. Patel, M.S. Harris, S.L. Halum, Objective voice assessment, in: Sataloff's Comprehensive Textbook of Otolaryngology: Head & Neck Surgery: Laryngology, Vol. 4, JP Medical Ltd, 2015, p. 155.

[3] R.R. Patel, S.N. Awan, J. Barkmeier-Kraemer, M. Courey, D. Deliyski, T. Eadie, D. Paul, J.G. Švec, R. Hillman, Recommended protocols for instrumental assessment of voice: American speech-language-hearing association expert panel to develop a protocol for instrumental assessment of vocal function, Am. J. Speech-Lang. Pathol. 27 (3) (2018) 887–905.

[4] D.M. Bless, R. Patel, N. Connor, Laryngeal imaging: stroboscopy, high-speed digital imaging, and kymography, in: The Larynx, Vol. 1, Plural Publishing San Diego, CA, Oxford, and Brisbane, 2009, pp. 181–210.

[5] J.G. Švec, H.K. Schutte, Kymographic imaging of laryngeal vibrations, Curr. Opin. Otolaryngol. Head Neck Surgery 20 (6) (2012) 458–465.

[6] C.A. Rosen, Stroboscopy as a research instrument: development of a perceptual evaluation tool, Laryngoscope 115 (3) (2005) 423–428.

[7] D.D. Mehta, R.E. Hillman, Current role of stroboscopy in laryngeal imaging, Curr. Opin. Otolaryngol. Head Neck Surgery 20 (6) (2012) 429.

[8] P. Woo, Stroboscopy and high-speed video examination of the larynx, in: R.T. Sataloff, M.S. Benninger (Eds.), Sataloff's Comprehensive Textbook of Otolaryngology: Head & Neck Surgery: Laryngology, Vol. 4, JP Medical Ltd, 2015, p. 193.

[9] D. Deliyski, Laryngeal high-speed videoendoscopy, in: K.A. Kendall, R.J. Leonard (Eds.), Laryngeal Evaluation: Indirect Laryngoscopy To High-Speed Digital Imaging, Thieme Medical, New York, 2010, pp. 245–270.

[10] G. Andrade-Miranda, Y. Stylianou, D.D. Deliyski, J.I. Godino-Llorente, N. Henrich Bernardoni, Laryngeal image processing of vocal folds motion, Appl. Sci. 10 (5) (2020) 1556.

[11] A.M. Kist, P. Gómez, D. Dubrovskiy, P. Schlegel, M. Kunduk, M. Echternach, R. Patel, M. Semmler, C. Bohr, S. Dürr, et al., A deep learning enhanced novel software tool for laryngeal dynamics analysis, J. Speech Lang. Hearing Res. 64 (6) (2021) 1889–1903.

[12] P. Gómez, A. Kist, P. Schlegel, D.A. Berry, D.K. Chhetri, M. Döllinger, Bagls, a multihospital benchmark for automatic glottis segmentation, Sci. Data 7 (1) (2020) http://dx.doi.org/10.1038/s41597-020-0526-3.

[13] M.K. Fehling, F. Grosch, M.E. Schuster, B. Schick, J. Lohscheller, Fully automatic segmentation of glottis and vocal folds in endoscopic laryngeal high-speed videos using a deep convolutional LSTM network, Plos One 15 (2) (2020) e0227791.

[14] A. Yamauchi, H. Imagawa, H. Yokonishi, K.-I. Sakakibara, N. Tayama, Multivariate analysis of vocal fold vibrations in normal speakers using high-speed digital imaging, J. Voice (2021) http://dx.doi.org/10.1016/j.jvoice.2021.08.002.

[15] K.A. Kendall, High-speed digital imaging of the larynx: recent advances, Curr. Opin. Otolaryngol. Head Neck Surgery 20 (6) (2012) 466–471.

[16] Q. Qiu, H.K. Schutte, A new generation videokymography for routine clinical vocal-fold examination, Laryngoscope 116 (10) (2006) 1824–1828.

[17] Q. Qiu, H.K. Schutte, Real-time kymographic imaging for visualizing human vocal-fold vibratory function, Rev. Sci. Instrum. 78 (2) (2007) 024302.

[18] J.G. Švec, F. Šram, Videokymographic examination of voice, in: E.P.M. Ma, E.M.L. Yiu (Eds.), Handbook of Voice Assessments, third ed., Plural Publishing, San Diego, CA, 2011, pp. 129–146.

[19] J.G. Švec, H.K. Schutte, Videokymography: high-speed line scanning of vocal fold vibration, J. Voice 10 (2) (1996) 201–205.

[20] T. Wittenberg, M. Tigges, P. Mergell, U. Eysholdt, Functional imaging of vocal fold vibration: digital multislice high-speed kymography, J. Voice 14 (3) (2000) 422–442.

[21] Y. Isogai, Analysis of the vocal fold vibration by the laryngo-strobography-improvements of the analytic function, Larynx Jpn. 8 (1996) 27–32.

[22] M.W. Sung, K.H. Kim, T.Y. Koh, T.Y. Kwon, J.H. Mo, S.H. Choi, J.S. Lee, K.S. Park, E.J. Kim, M.Y. Sung, Videostrobokymography: a new method for the quantitative analysis of vocal fold vibration, Laryngoscope 109 (11) (1999) 1859–1863.

[23] P. Krasnodębska, A. Szkiełkowska, B. Miaśkiewicz, H. Skarżyński, Characteristics of euphony in direct and indirect mucosal wave imaging techniques, J. Voice 31 (3) (2017) 383–e13.

[24] V. Hampala, Vizuální hodnocení videokymografických snímků u hlasových poruch [Visual evaluation of videokymographic features in voice disorders], (Master's thesis), Palacký University, Olomouc, Czech Republic, 2011.

[25] H.S. Bonilha, D.D. Deliyski, J.P. Whiteside, T.T. Gerlach, Vocal fold phase asymmetries in patients with voice disorders: a study across visualization techniques, 21, (1) 2012, pp. 3–15.

[26] S.P. Kumar, K.V. Phadke, J. Vydrová, A. Novozámský, A. Zita, B. Zitová, J.G. Švec, Visual and automatic evaluation of vocal fold mucosal waves through sharpness of lateral peaks in high-speed videokymographic images, J. Voice 34 (2) (2020) 170–178.

[27] C. Manfredi, L. Bocchi, S. Bianchi, N. Migali, G. Cantarella, Objective vocal fold vibration assessment from videokymographic images, Biomed. Signal Process. Control 1 (2) (2006) 129–136, Voice Models and Analysis for Biomedical Applications.

[28] C. Manfredi, L. Bocchi, G. Cantarella, G. Peretti, Videokymographic image processing: objective parameters and user-friendly interface, Biomed. Signal Process. Control 7 (2) (2012) 192–201.

[29] C. Piazza, S. Mangili, F. Del Bon, F. Gritti, C. Manfredi, P. Nicolai, G. Peretti, Quantitative analysis of videokymography in normal and pathological vocal folds: a preliminary study, Eur. Arch. Oto-Rhino-Laryngol. 269 (1) (2012) 207–212.

[30] P.H. Dejonckere, J. Lebacq, L. Bocchi, S. Orlandi, C. Manfredi, Automated tracking of quantitative parameters from single line scanning of vocal folds: A case study of the 'messa di voce' exercise, Logopedics Phoniatr. Vocology 40 (1) (2015) 44–54.

[31] J. Švec, M. Frič, F. Šram, H. Švecová, H. Schutte, Visually-based evaluation protocol for laryngeal videokymographic images, in: Proceedings AQL, 2006.

[32] J.G. Švec, F. Šram, H.K. Schutte, Videokymography in voice disorders: what to look for? Ann. Otol. Rhinol. Laryngol. 116 (3) (2007) 172–180.

[33] MATLAB Image Processing Toolbox, URL http://www.mathworks.com/products/image/.

[34] Open Source Computer Vision Library, URL http://opencv.org/.

[35] Qt, URL http://www.qt.io/.

[36] R.D. Hipp, Sqlite, 2020, URL https://www.sqlite.org/index.html.

[37] S.M. Pizer, E.P. Amburn, J.D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J.B. Zimmerman, K. Zuiderveld, Adaptive histogram equalization and its variations, Comput. Vis. Graph. Image Process. 39 (3) (1987) 355–368.

[38] J. Serra, Image Analysis and Mathematical Morphology, Academic Press, London, 1988.

[39] N. Isshiki, Recent advances in phonosurgery, Folia. Phoniatr. (Basel) 32 (2) (1980) 119–154.

[40] Q. Qiu, H.K. Schutte, L. Gu, Q. Yu, An automatic method to quantify the vibration properties of human vocal folds via videokymography, Folia. Phoniatr. Logop. 55 (3) (2003) 128–136.

[41] J.G. Švec, F. Šram, H.K. Schutte, Videokymography, in: M.P. Fried, A. Ferlito (Eds.), The Larynx, third ed., Plural Publishing, San Diego, CA, 2009, pp. 253–274.

[42] P. Aichinger, F. Pernkopf, Synthesis and analysis-by-synthesis of modulated Diplophonic Glottal Area waveforms, IEEE/ACM Trans. Audio Speech Lang. Process. 29 (2021) 914–926.

[43] D.D. Mehta, D.D. Deliyski, T.F. Quatieri, R.E. Hillman, Automated measurement of vocal fold vibratory asymmetry from high-speed videoendoscopy recordings, J. Speech Lang. Hear. Res. 54 (1) (2011) 47–54.

[44] E. Dunker, B. Schlosshauer, Irregularities of the laryngeal vibratory pattern in healthy and hoarse persons, in: D.W. Brewer (Ed.), Research Potentials in Voice Physiology, State University of New York, Syracuse, NY, 1964, pp. 151–184.

[45] M. Hirano, Clinical Examination of Voice, Springer-Verlag, Wien, Austria, 1981.

[46] P. Moore, H. von Leden, Dynamic variations of the vibratory pattern in the normal larynx, Folia. Phoniatr. (Basel) 10 (4) (1958) 205–238.

[47] R. Timcke, H. von Leden, P. Moore, Laryngeal vibrations: measurements of the glottic wave. I. The normal vibratory cycle, AMA Arch. Otolaryngol. 68 (1) (1958) 1–19.

[48] J. Lohscheller, H. Toy, F. Rosanowski, U. Eysholdt, M. Döllinger, Clinically evaluated procedure for the reconstruction of vocal fold vibrations from endoscopic digital high-speed videos, Med. Image Anal. 11 (2007) 400–413, http://dx.doi.org/10.1016/j.media.2007.04.005.

[49] Y. Maryn, M. Verguts, H. Demarsin, J. van Dinther, P. Gomez, P. Schlegel, M. Döllinger, Intersegmenter variability in high-speed laryngoscopy-based glottal area waveform measures, Laryngoscope 130 (11) (2020) E654–E661.

[50] T. Shi, H.J. Kim, T. Murry, P. Woo, Y. Yan, Tracing vocal fold vibrations using level set segmentation method, Int. J. Numer. Methods Biomed. Eng. 31 (6) (2015) e02715.

[51] T. Wittenberg, P. Mergell, M. Tigges, U. Eysholdt, Quantitative characterization of functional voice disorders using motion analysis of high-speed video and modeling, in: 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 3, 1997, pp. 1663–1666, http://dx.doi.org/10.1109/ICASSP.1997.598831.

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

[53] Y.Y. Boykov, M.-P. Jolly, Interactive graph cuts for optimal boundary & region segmentation of objects in ND images, in: Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001, Vol. 1, IEEE, 2001, pp. 105–112.

[54] N. Otsu, A threshold selection method from gray-level histograms, IEEE Trans. Syst. Man Cybern. 9 (1) (1979) 62–66.

[55] V. Osma-Ruiz, J. Godino Llorente, N. Saenz-Lechon, R. Fraile, Segmentation of the glottal space from laryngeal images using the watershed transform, Comput. Med. Imag. Graph.: Official J. Comput. Med. Imag. Soc. 32 (2008) 193–201, http://dx.doi.org/10.1016/j.compmedimag.2007.12.003.

[56] Z. Štanclová, Relationships between the vocal fold vibration parameters and voice intensity: A laryngeal high speed videoendoscopic study of a healthy woman. (In Czech), (Bachelor's thesis), Palacky University, Faculty of Science, Olomouc, the Czech Republic, 2021.

[57] H. Lehoux, L. Popeil, J.G. Švec, Laryngeal and acoustic analysis of chest and head registers extended across a three-octave range: a case study, J. Voice (2022) http://dx.doi.org/10.1016/j.jvoice.2022.02.014.

Coral Reef annotation, localisation and pixel-wise classification using Mask R-CNN and Bag of Tricks

Lukáš Picek[1,5], Antonín Říha[2], and Aleš Zita[3,4]

[1] Dept. of Cybernetics, Faculty of Applied Sciences, University of West Bohemia
[2] Faculty of Information Technology, Czech Technical University
[3] The Czech Academy of Sciences, Institute of Information Theory and Automation
[4] Faculty of Mathematics and Physics, Charles University
[5] PiVa AI

Abstract. This article describes an automatic system for detection, classification and segmentation of individual coral substrates in underwater images. The proposed system achieved the best performances in both tasks of the second edition of the ImageCLEFcoral competition. Specifically, mean average precision with Intersection over Union (IoU) greater then 0.5 (mAP@0.5) of 0.582 in case of Coral reef image annotation and localisation, and mAP@0.5 of 0.678 in Coral reef image pixel-wise parsing. The system is based on Mask R-CNN object detection and instance segmentation framework boosted by advanced training strategies, pseudo-labeling, test-time augmentations, and Accumulated Gradient Normalisation. To support future research, code has been made available at: https://github.com/picekl/ImageCLEF2020-DrawnUI.

Keywords: Deep Learning, Computer Vision, Instance Segmentation, Convolutional Neural Networks, Machine Learning, Object Detection, Corals, Biodiversity, Conservation

1 Introduction

The ImageCLEFcoral [4] challenge was organized in conjunction with the ImageCLEF 2020 evaluation campaign [12] at the Conference and Labs of the Evaluation Forum (CLEF[1]). The main goal for this competition was to create such an algorithm or system that can automatically detect and annotate a variety of benthic substrate types over image collections taken from multiple coral reefs as part of a coral reef monitoring project with the Marine Technology Research Unit at the University of Essex.

[1] http://www.clef-initiative.eu/

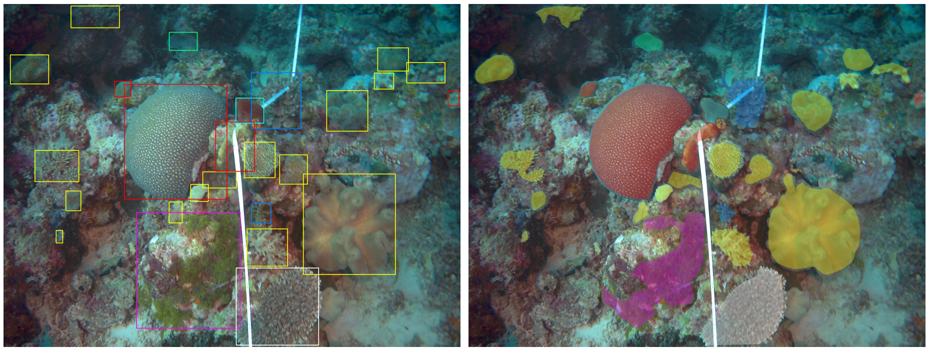


Fig. 1. Example training images showing different types of annotations - Bounding Boxes and Segmentation Masks. Every colour represents one substrate type, e.g. yellow represents *Soft Coral* and red belongs to *Hard Coral Boulder*.

1.1 Motivation

Live corals are an important biological class that has a massive contribution to the ocean ecosystem biodiversity. Corals are key habitat for thousands of marine species [5] and provide an essential source of nutrition and yield for people in the developing countries [3,2]. Therefore, automatic monitoring of coral reefs condition plays a crucial part in understanding future threats and prioritizing conservation efforts.

1.2 Datasets

This section will briefly describe the provided data and their subsets: an annotated dataset that contains 440 images, and a testing dataset with 400 images without annotations. Additionally, we introduce an precisely engineered training/validation split of the annotated dataset for the training purposes.

Annotated dataset - The annotated dataset is a combination of 440 images containing 12,082 individual coral objects. Each coral was annotated with expert level knowledge, including segmentation mask, bounding box, and class that represents 1 out of 13 substrate types. The dataset is heavily unbalanced (refer to Table 1), having almost 50% of objects from a single class (Soft Coral) and approximately 8% for the eight least frequent classes. Moreover, images have different colour variations, are heavily blurred, and came from different locations and geographical regions. Furthermore, coral substrates belonging to the same class can be observed in different morphology, colour variations, or patterns. Finally, some images contain a measurement tape that partially covers objects of interest.

For the network training process evaluation, the annotated dataset needed to be divided into two parts. One used for network optimization and the second for network performance validation. To create these subsets, every tenth image was designated for validation set, the rest was used for training. As the validation set class distribution did not match the training one, particular images from the validation set needed to be replaced by carefully cherry-picked images from the training set. This resulted in an almost perfect split with similar distributions for both, the training and the validation set. This similarity ensured a representative validation process.

Testing dataset - The testing dataset contains 400 images from four different locations. Namely, the same location as is in the training set, similar location to the training set, geographically similar location to the training set, and geographically distinct location from the training set.

Table 1. Dataset class distribution including training and validation split description. 396 images were used for training; 44 for validation.

Dataset distribution			**Train. / Val. split**	
Substrate type	# Bboxes	Fraction [%]	Train. Boxes	Val. Boxes
Soft Coral	5,663	46.87	5,035	628
Sponge	1,691	13.99	1,472	219
Hard Coral – Boulder	1,642	13.59	1,513	129
Hard Coral – Branching	1,181	9.774	1,084	97
Hard Coral – Encrusting	946	7.829	831	115
Hard Coral – Mushroom	223	1.845	199	24
Hard Coral – Submassive	198	1.845	162	36
Hard Coral – Foliose	177	1.464	144	33
Sponge – Barrel	139	1.150	124	15
Algae - Macro or Leaves.	92	0.761	81	11
Soft Coral – Gorgonian	90	0.745	70	20
Hard Coral – Table	21	0.175	17	4
Fire Coral – Millepora	19	0.157	15	4

1.3 The System

The proposed object detection and instance segmentation system extends recent state-of-the-art Convolutional Neural Network (CNN) object detection framework (Mask R-CNN [8]) with additional **Bag of Tricks** that considerably increased the performance. The TensorFlow Object Detection API[2] [11] was used as a deep learning framework for fine-tuning the publicly available checkpoints. All bells and whistles are further described in Section 2. Additionally, approaches that did not contribute positively but could have some potential for future editions of the ImageCLEFcoral competition are discussed.

[2] https://github.com/tensorflow/models/blob/master/research/object_detection

2 Methodology

This section describes all approaches and techniques used in the benthic substrate detection, annotation and segmentation tasks. The modern object detection and instance segmentation methods are summarized, followed by the description of the chosen system and its configuration. Furthermore, all the used bells and whistles (Bag of Tricks) are introduced and described.

2.1 Object Detection

Although conventional digital image processing methods are capable of detecting particular local features, modern object detectors based on Deep Convolutional Neural Networks (DCNN) achieve superior performance in object detection and instance segmentation tasks. Several network architectures were pre-selected based on study published by Huang et al. [11], namely the Faster R-CNN [18], SSD [15] and Mask R-CNN [8]. The initial performance experiment was to train these detection frameworks with default or recommended configurations. This experiment revealed the most suitable framework for both the tasks within the ImageCLEFcoral competition - the Mask R-CNN.

2.2 Network parameters

Experiments on the validation set, reveled the best optimizer settings for the framework. These settings were shared between all of our experiments, unless stated otherwise. For detailed description refer to Table 2.

Table 2. Training and network parameters shared among all experiments.

Parameter	Value	Parameter	Value
Optimizer	RMSprop	Gradient Clipping	12.5
Momentum	0.9	Input size	1000 × 1000
Initial and min LR	0.032 - 0.00004	Feature extractor stride	8
LR decay type	Exponential	Pretrained Checkpoints	COCO
LR decay factor	0.975	Num epochs	50
Batch size	1	Gradient accumulation	16

2.3 Bag of Tricks

Augmentations - The provided dataset contains 440 images. Considering that 44 were used for validation, 396 images is too few for robust network optimization. To alleviate this issue, multiple data augmentation techniques were utilized. The following methods were included in the final training pipeline:

Colour Distortions - Brightness variations with max delta of 0.2, contrast and saturation variations scale each by random value in range of 0.8 - 1.25, hue variations offsets by random value of up to 0.02, and random RGB to grayscale conversion with 10% probability.

- **Image Flips** - Random horizontal and vertical flip, and 90 degree rotations. Each with 50% chance.
- **Random Jitter** - Every bounding box corner can be randomly shifted by amount corresponding up to 2% of the bounding box width and height in x and y coordinates, respectively.
- **Cut Out [6]** - Random black square patches are added into the image. More precisely, add up to 10 patches with 50% occurrence probability and each with side length corresponding to 10% of the image height or width, whichever is smaller.

By utilizing techniques mentioned above, we have increased the model mAP@0.5 performance by **0.0392** as measured on the validation set.

Input Resolution - In the task of object detection, primarily where a small object occurs, input resolution plays a crucial role. Theoretically, the higher the resolution is, the more objects will be detected. Unfortunately, the detection of high resolution images is GPU memory-limited. Hence, it always is a trade-off between performance and hardware requirements.

Backbone - To find the best backbone architecture for Mask R-CNN framework. We performed an experiment over 3 different backbone models including ResNet-50 [9], ResNet-101 [9], and Inception-ResNet-V2 [20]. Detailed performance comparison is included in Table 3.

Table 3. Effect of input resolution and backbone architecture on model performance.

Backbone	Input Resolution	mAP@0.5	mAP@0.75
ResNet-50	600 × 600	0.1826	0.0956
ResNet-50	800 × 800	0.2077	0.1017
ResNet-50	1000 × 1000	0.2227	0.1260
ResNet-50	1200 × 1200	**0.2380**	**0.1579**
ResNet-101	800 × 800	**0.2381**	**0.1453**
Inception-ResNet-V2	800 × 800	0.2362	0.1361

Pseudo Labels - Performance of DCNN's heavily depends on the size of the training set. To facilitate this issue, we have developed a naive pseudo-labelling approach inspired by [1]. In short, already trained network is used to label the unlabelled testing data with so-called weak labels. Only the overconfident detections were used; the rest of the image was blurred out. Even though there is a high chance of overfitting to incorrect pseudo-labels due to the confirmation bias, pseudo-labels can significantly improve the performance of the CNN if pseudo-labelled images are added sensitively.

Transfer Learning - Big-transfer [13] or transfer learning is a fine-tuning technique commonly used in deep learning. Rather then initialize the weights of neural network randomly, pretrained weights are used. Furthermore, final model could benefit from similar domain weights. To evaluate a potential of such approach for the purposes of this competition, we experimented with fine-tuning of the publicly available checkpoints, including ImageNet[3], iNaturalist[3], COCO [14], PlantCLEF2018 [19] and PlanCLEF2019 [17]. The idea was that fine-tuning checkpoints trained on nature-oriented datasets would outperform the non-nature oriented ones. One could assume, that this is caused by significant difference when compared to other domains. Based on that it has been decided to use the COCO pretrained checkpoint which includes both the backbone and region proposed weights.

Table 4. Transfer Learning experiment - Effect of pretrained weights on model performance. For this experiment, the Mask R-CNN with ResNet-50 backbone and input size of 800 × 800 was used.

Pretrained weights	mAP@0.5	mAP@0.75
ImageNet (only backbone)	0.1826	0.0956
COCO (All Mask R-CNN weights)	0.2077	0.1017
iNaturalist (only backbone)	0.2091	0.0854
PlantCLEF2018 (only backbone)	0.1991	0.0914
PlantCLEF2019 (only backbone)	0.1895	0.0932

Test Time Augmentations - Test time augmentation is a method of applying transformations on a given image to generate its several slightly different variations that are used to create predictions that, when combined, can improve final prediction. Our submissions utilized augmentations consisting of simple horizontal and vertical flips of the image. Their combinations produced four sets of detections for each image. These sets were then joined using voting strategy described in [16] by Moshkov et al..

Ensembles - Ensemble methods combine predictions from multiple models to obtain final output [21]. These methods can be used to improve accuracy in machine learning tasks. In our work, we utilize a simple method for combining outputs from multiple detection networks based on voting [16]. Detections describing one object are grouped together by size of the overlap region belonging to the same class. Instances, where majority of the detectors agree on class label and position are replaced by single detection with the highest score.

Accumulated Gradient Normalization - In order to achieve the best performance possible, we aimed to maximize the resolution of input data. Therefore,

[3] https://github.com/tensorflow/models/blob/master/research/object_detection/g3doc/tf1_detection_zoo.md

we have decided to train the network on mini-batches of size 1. To overcome disadvantages that comes with using minimal mini-batch size [7], the Accumulated Gradient Normalization [10] technique was utilized. This approach resulted in a considerable performance gain.

3 Submissions

For evaluation of the participants submissions, the AICrowd platform[4] was used. Each participating team was allowed to submit up to 10 submission files following specific requirements for both tasks. We have used allowed maximum for both tasks. Because we have utilized single architecture for both the detection and segmentation tasks, multiple submissions were produced using the same network. Therefore in the following part, we denoted annotation and localisation task submissions by **D** and pixel-wise parsing task submissions by **S**. Finally, thresholding was used to discard predictions with low confidence.

Baseline configuration - As a baseline for all our experiments we used Mask R-CNN with ResNet-50 as a backbone. For training we used parameters and augmentations described in Table 2.2 and Section 2.3, respectively. Input resolution was 1000 × 1000 pixels.

Submission 1D/1S - Baseline experiment using a confidence threshold that corresponded to the best F1 score on our validation dataset (0.58).

Submission 2D - Submission 1D with a fixed programming bug that resulted in few detections being incorrectly generated.

Submission 3D - Submission 2D with confidence threshold set to 0.95.

Submission 4D/2S - Baseline configuration that used Pseudo-labels as described in Section 2.3. The confidence threshold was set to 0.95.

Submission 5D/3S - Baseline configuration that utilized test time augmentations as described in Section 2.3 with confidence threshold of 0.9.

Submission 6D/4S - Submission 5D/3S with confidence threshold of 0.999.

Submission 7D/5S - Ensemble of two checkpoints of baseline configuration model. Taken after 40 epochs and 50 epochs. Confidence threshold of 0.9.

Submission 8D/6S - Submission 7D/5S with confidence threshold of 0.999.

Submission 9D/8S - Submission 7D/5S with test time augmentations and with confidence threshold of 0.999.

Submission 10D/10S - Submission 7D/5S with confidence threshold of 0.95.

Submission 7S - Submission 9D/8S with confidence threshold of 0.9.

Submission 9S - Submission 9D/8S with modified voting ensemble. Only one detection is sufficient as opposed to majority voting.

[4] https://www.aicrowd.com

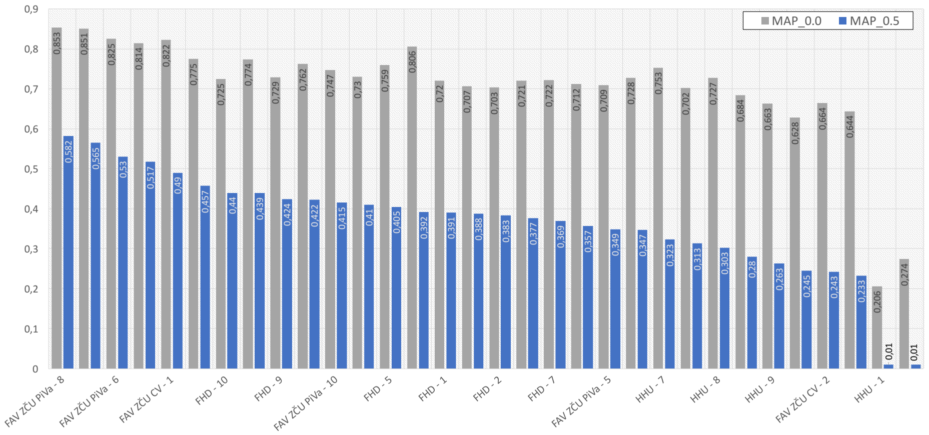


Fig. 2. Results for all runs submitted in annotation and localisation task by the competition participants, including mAP@0.0 and mAP@0.5 metrics.

4 Competition Results

The official competition results are shown in Figure 2 for annotation and localisation task, and in Figure 3 for pixel-wise parsing. Our System achieved the best performances in both tasks of the second edition of the ImageCLEFcoral competition. Specifically, mAP@0.5 of **0.582** in case of Coral reef image annotation and localisation (Run ID 68143), and mAP@0.5 of **0.678** in Coral reef image pixel-wise parsing (Run ID 67864). Results of all our submissions are listed in Table 5. Table 6 illustrates the performance over different subsets of the test dataset. The system performed comparably over the Same Location (SL), Similar Location (SiL) and Geographically Similar Location (GS) subsets. The performance significantly drops in Geographically Distinct Location (GD). This is probably caused by a lack of diverse training data.

The best scoring submission for pixel-wise parsing task was a single Mask R-CNN with ResNet-50 backbone architecture and input resolution of 1000 × 1000. The system was trained for 50 epochs while using heavy augmentations as described in Section 2.3. Additionally, the pseudo-labeling (refer to Section 2.3) was used to increase the training dataset size with overconfident detections from the test set. Finally, the predictions were filtered with confidence threshold of 0.95 to maximize the official mAP metric while still having decent recall score.

The best scoring submission for annotation and localisation task was an ensemble of two checkpoints of the same Mask R-CNN model with ResNet-50 backbone architecture and input resolution of 1000×1000, one taken after 40 and other one after 50 epochs. The system was trained using heavy augmentations. Furthermore, the predictions were filtered with confidence threshold of 0.999 to maximize the official metric of mAP.

Table 5. Submission scores achieved over test set. Official competition metrics.

Annotation and localisation task submissions										
Submission	1D	2D	3D	4D	5D	6D	7D	**8D**	9D	10D
mAP@0.5	0.347	0.357	0.439	0.565	0.349	0.530	0.377	**0.582**	0.517	0.415
mAP@0.0	0.728	0.712	0.774	0.851	0.709	0.825	0.721	**0.853**	0.814	0.747
Run ID	67857	67858	67862	67863	68093	68094	68138	68143	68145	68146

Pixel-wise parsing task submissions										
Submission	1S	**2S**	3S	4S	5S	6S	7S	8S	9S	10S
mAP@0.5	0.441	**0.678**	0.434	0.629	0.470	0.664	0.407	0.624	0.617	0.507
mAP@0.0	0.694	**0.845**	0.689	0.817	0.701	0.842	0.675	0.813	0.807	0.727
Run ID	67856	67864	68092	68095	68137	68139	68140	68142	68144	68147

Table 6. Submission results achieved over 4 subsets of the testing set: Same Location (SL), Similar Location (SiL), Geographically Similar Location (GS), Geographically Distinct Location (GD).

Annotation and localisation task submissions										
Submission	1D	2D	3D	4D	5D	6D	7D	8D	9D	10D
SL mAP@0.5	0.401	0.417	0.489	0.614	0.410	0.566	0.434	**0.648**	0.547	0.475
SiL mAP@0.5	0.234	0.247	0.322	**0.440**	0.230	0.431	0.254	0.343	0.438	0.258
GS mAP@0.5	0.470	0.446	0.508	0.562	0.453	0.516	0.516	**0.627**	0.533	0.527
GD mAP@0.5	0.225	0.230	0.280	0.292	0.231	**0.346**	0.210	0.329	0.344	0.242
Run ID	67857	67858	67862	67863	68093	68094	68138	68143	68145	68146

Pixel-wise parsing task submissions										
Submission	1S	2S	3S	4S	5S	6S	7S	8S	9S	10S
SL mAP@0.5	0.527	**0.744**	0.513	0.670	0.545	0.742	0.480	0.663	0.656	0.583
SiL mAP@0.5	0.312	0.516	0.309	**0.553**	0.335	0.448	0.284	0.529	0.546	0.34
GS mAP@0.5	0.476	**0.588**	0.493	0.537	0.553	0.627	0.493	0.586	0.546	0.573
GD mAP@0.5	0.276	0.403	0.283	0.439	0.266	0.386	0.267	**0.446**	0.418	0.291
Run ID	67856	67864	68092	68095	68137	68139	68140	68142	68144	68147

5 Conclusion and Discussion

The proposed system designed for automatic pixel-wise detection of 13 coral substrates achieved impressive mAP@0.5 of **0.582** in localization task and **0.678**, for instance segmentation task of the ImageCLEFcoral competitions. The system is wrapped up around the Mask R-CNN, the state-of-the-art instance segmentation framework, and additional known as well as some unique techniques, e.g., detection ensemble, test time data augmentations, accumulated gradient normalization, and pseudo-labelling. Surprisingly, results for pixel-wise parsing are considerably better. This is unexpected mainly because the test set is the same for both tasks, and our submissions used the same set of detections. Therefore, more similar scores were expected. This led us to believe that annotations for both tasks are not the same.

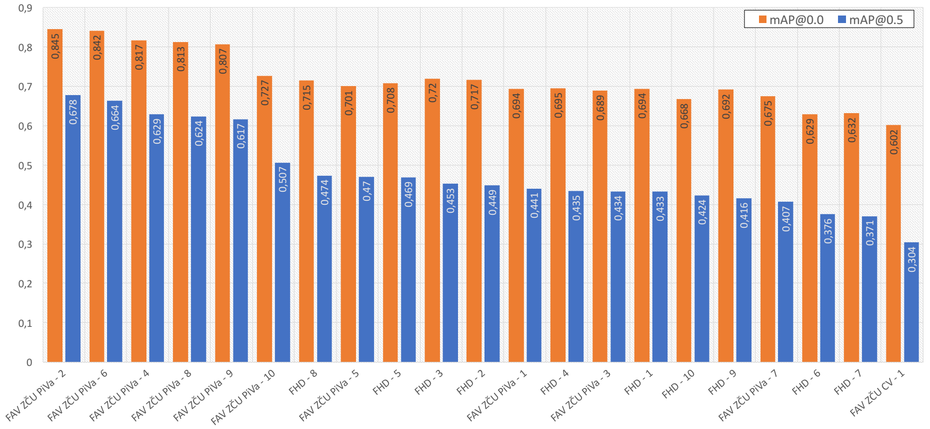


Fig. 3. Results for all runs submitted in pixel-wise parsing task by the competition participants, including mAP@0.0 and mAP@0.5 metrics.

More in-depth performance examination of our submissions revealed a small regularisation capability related to geographical regions and specific locations. This is indication that the network could be over-fitted on the training dataset location, which have specific distribution of coral species. The system could achieve better performance with class priors corresponding to desired location. If the location transfer is essential, location generalisation should be main goal for the future challenges.

While comparing the model performance with the top results from the previous edition of this challenge (mAP@0.5 of 0.2427 and 0.0419), our model achieved superior performance. Even though the test datasets are not identical, such difference shows the increasing trend of machine learning model performance. This increase is probably related to a higher number of training images.

Lastly, due to our GPU memory constraints we were limited to an input image resolution of 1000 × 1000 combined with ResNet-50 backbone. Conducted experiments showed that input resolution of 1200 × 1200 and ResNet-101 would yield better results, therefore usage of GPUs with more memory would lead to a considerable increase of the system's performance.

Acknowledgements

Lukáš Picek was supported by the Ministry of Education, Youth and Sports of the Czech Republic project No. LO1506, and by the grant of the UWB project No. SGS-2019-027.

References

1. Arazo, E., Ortego, D., Albert, P., O'Connor, N.E., McGuinness, K.: Pseudo-labeling and confirmation bias in deep semi-supervised learning. arXiv preprint arXiv:1908.02983 (2019)
2. Birkeland, C.: Global status of coral reefs: In combination, disturbances and stressors become ratchets pp. 35–56 (2019)
3. Brander, L.M., Rehdanz, K., Tol, R.S., Van Beukering, P.J.: The economic impact of ocean acidification on coral reefs. Climate Change Economics **3**(01), 1250002 (2012)
4. Chamberlain, J., Campello, A., Wright, J.P., Clift, L.G., Clark, A., García Seco de Herrera, A.: Overview of the ImageCLEFcoral 2020 task: Automated coral reef image annotation. In: CLEF2020 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org> (2020)
5. Coker, D.J., Wilson, S.K., Pratchett, M.S.: Importance of live coral habitat for reef fishes. Reviews in Fish Biology and Fisheries **24**(1), 89–126 (2014)
6. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)
7. Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., He, K.: Accurate, large minibatch sgd: Training imagenet in 1 hour. arXiv preprint arXiv:1706.02677 (2017)
8. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask r-cnn. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
10. Hermans, J., Spanakis, G., Möckel, R.: Accumulated gradient normalization. arXiv preprint arXiv:1710.02368 (2017)
11. Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., et al.: Speed/accuracy trade-offs for modern convolutional object detectors. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7310–7311 (2017)
12. Ionescu, B., Müller, H., Péteri, R., Abacha, A.B., Datla, V., Hasan, S.A., Demner-Fushman, D., Kozlovski, S., Liauchuk, V., Cid, Y.D., Kovalev, V., Pelka, O., Friedrich, C.M., de Herrera, A.G.S., Ninh, V.T., Le, T.K., Zhou, L., Piras, L., Riegler, M., l Halvorsen, P., Tran, M.T., Lux, M., Gurrin, C., Dang-Nguyen, D.T., Chamberlain, J., Clark, A., Campello, A., Fichou, D., Berari, R., Brie, P., Dogariu, M., Ştefan, L.D., Constantin, M.G.: Overview of the ImageCLEF 2020: Multimedia retrieval in medical, lifelogging, nature, and internet applications. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 11th International Conference of the CLEF Association (CLEF 2020), vol. 12260. LNCS Lecture Notes in Computer Science, Springer, Thessaloniki, Greece (September 22-25 2020)
13. Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., Houlsby, N.: Big transfer (bit): General visual representation learning (2019)
14. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
15. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European conference on computer vision. pp. 21–37. Springer (2016)

16. Moshkov, N., Mathe, B., Kertesz-Farkas, A., Hollandi, R., Horvath, P.: Test-time augmentation for deep learning-based cell segmentation on microscopy images. Scientific reports **10**(1), 1–7 (2020)
17. Picek, L., Sulc, M., Matas, J.: Recognition of the amazonian flora by inception networks with test-time class prior estimation. In: Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum (2019)
18. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems 28, pp. 91–99. Curran Associates, Inc. (2015)
19. Sulc, M., Picek, L., Matas, J.: Plant recognition by inception networks with test-time class prior estimation. In: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum (2018)
20. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-first AAAI conference on artificial intelligence (2017)
21. Zhang, C., Ma, Y.: Ensemble machine learning: methods and applications. Springer (2012)

Sketch2Code: Automatic hand-drawn UI elements detection with Faster R-CNN

Aleš Zita[1,2], Lukáš Picek[3,5], and Antonín Říha[4]

[1] Czech Academy of Sciences, Institute of Information Theory and Automation
[2] Faculty of Mathematics and Physics, Charles University
[3] Dept. of Cybernetics, Faculty of Applied Sciences, University of West Bohemia
[4] Faculty of Information Technology, Czech Technical University
[5] PiVa AI

Abstract. Transcription of User Interface (UI) elements hand drawings to the computer code is a tedious and repetitive task. Therefore, a need arose to create a system capable of automating such process. This paper describes a deep learning-based method for hand-drawn user interface elements detection and localization. The proposed method scored 1st place in the ImageCLEFdrawnUI competition while achieving an overall precision of 0.9708. The final method is based on Faster R-CNN object detector framework with ResNet-50 backbone architecture trained with advanced regularization techniques. The code has been made available at: https://github.com/picekl/ImageCLEF2020-DrawnUI.

Keywords: Web Design, Object Detection, Convolutional Neural Networks, Machine Learning, Computer Vision, User Interface, Deep Learning

1 Introduction

The ImageCLEFdrawnUI [3] challenge was organized in connection with the ImageCLEF 2020 evaluation campaign [7] at the Conference and Labs of the Evaluation Forum (CLEF). The Main goal of this competition was to create an algorithm or system which can automatically recognise and localize UI elements on high resolution pictures of their drawings. The desired outcome of the detection process are localized bounding boxes with corresponding classes assignments of the UI elements.

1.1 Motivation

The main motivation for this task is to simplify the process of websites creation by enabling people to create websites by drawing UI elements on a whiteboard or on a piece of paper to make the web page building process more accessible. In this context, the detection and recognition of hand drawn UI elements task addresses the problem of automatically transcribing the UI to computer code.

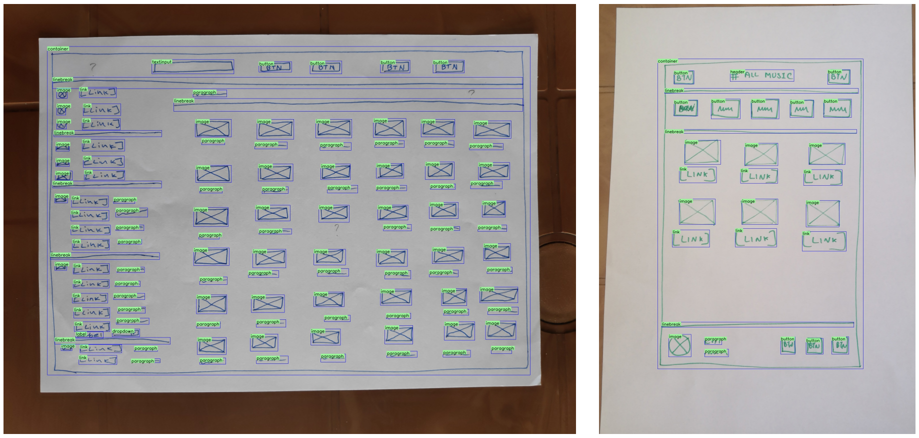


Fig. 1. Example images with annotations from ImageCLEFdrawnUI competition training dataset.

1.2 Dataset

The complete dataset consists of 1,000 hand drawn templates captured multiple times with different cameras, resulting in 2,950 high-resolution images. These data were further randomly split into 2,363 training and 587 test images. The training part includes 65,993 UI elements belonging to 21 classes. All images were annotated with bounding boxes and class labels by human experts. More detailed class distribution description is listed in Table 1. Example images are depicted in Figure 1.

1.3 Solution

The proposed solution is based on utilization of a standard object detection network architecture and coherent data preparation and augmentation. In particular, the Faster R-CNN [10] framework with the ResNet-50 [5] feature extractor was used. The system was implemented and fine-tuned using TensorFlow Object Detection API[1] [6] from publicly available checkpoints. All networks in our experiments shared the optimizer settings - RMSProp [13] with momentum of 0.9. The initial architecture was based on our work [8] submitted to ImageCLEFcoral competition [2]. This included for instance the data augmentation methods or Accumulated Gradient Normalization technique [4]. During our followup research, we considered and tested several approaches including new data synthesis, different network architectures as well as network ensemble variants.

[1] https://github.com/tensorflow/models/blob/master/research/object_detection

Table 1. Class distribution description including number of UI elements and their number in training and validation set.

Dataset distribution			Train. / Val. split		
Class Name	# Boxes	Fraction[%]	Train. Boxes	Val. Boxes	Fraction[%]
button	18,704	28.34	16,841	1,863	9.96%
paragraph	10,367	15.71	9,342	1,025	9.89%
image	7,683	11.64	7,020	663	8.63%
link	6,809	10.32	6,140	669	9.83%
linebreak	5,798	8.786	5,267	531	9.16%
container	4,678	7.089	4,233	445	9.51%
header	4,356	6.601	3,947	409	9.39%
textinput	1,732	2.624	1,577	155	8.95%
label	1,691	2.562	1,539	152	8.99%
dropdown	1,472	2.231	1,350	122	8.29%
list	798	1.209	702	96	12.03%
checkbox	758	1.148	694	64	8.44%
video	360	0.545	323	37	10.28%
radiobutton	279	0.422	246	33	11.83%
toggle	178	0.249	159	19	10.67%
datepicker	91	0.138	83	8	8.79%
rating	75	0.114	62	13	17.33%
slider	75	0.114	65	10	13.33%
textarea	47	0.071	42	5	10.64%
table	29	0.043	25	4	13.79%
stepperinput	13	0.019	10	3	23.08%

2 Methodology

2.1 Data analysis and preparation

Dataset splitting for validation - To create a set for continuous network performance evaluation, the provided dataset needed to be split into training and validation sets. After careful examination of the content, it became apparent that a random split of the dataset could cause discrepancies between the validation and training sets performances. The reason being, that less frequent classes could end up not having comparable representations in both the training and validation sets. Therefore the split had to be carefully engineered and resulted in the final approximate ratio of **11:1** for training and validation sets, respectively.

Data distribution - To better understand the problem at hand, we have performed a frequency analysis on UI element type distribution and concluded, that some of the element types are represented by very few occurrences in the training dataset, namely the *'stepper input'*, *'text area'* or *'table'* (See Table 1). Reviewing the training dataset further revealed that it contains multiple images of the same drawings. This is caused by the fact that the whole dataset (training and testing) consists of 2,950 images of only 1,000 templates, i.e., the templates were each captured by several different cameras. Following the random splitting of the dataset to the training and testing part caused some rarer elements to go to the training set multiple times and others not at all. This worsens the uneven distribution of the UI element classes in such a way that, for example, the rarest element is contained only on two templates (6 images) in the training dataset. For the deep network to learn to recognize such an element, a much higher number of examples is needed.

Synthetic dataset - To compensate for the uneven distribution of the UI element types, we decided to expand the training dataset with synthetic data containing such elements. The data were generated using augmentations of segmented UI elements, which were consequently pasted on random size paper of very light random color. The augmentation consisted mainly of constrained random affine transformations. We have added 500 synthetically generated images with the least frequent classes. Examples of the synthetic data are depicted in Figure 2. UI element classes which were artificially added are: *datepicker*, *rating*, *slider*, *textarea*, *table* and *stepperinput*.

The experiment performed with ResNet-50 backbone and grayscale data with 1000 × 1000 input size was evaluated over the validation set and showed interesting improvement in all measured scores on RGB images. Specifically, mean average precision with Intersection over Union (IoU) greater than 0.5 (mAP0.5) by **0.0081**, mAP by **0.0222**, and by Recall@100 (Recall calculated using best 100 detections) **0.0315**. Although we were able to flatten the UI elements distribution curve, the overall performance of the original network was marginally better on grayscale images.

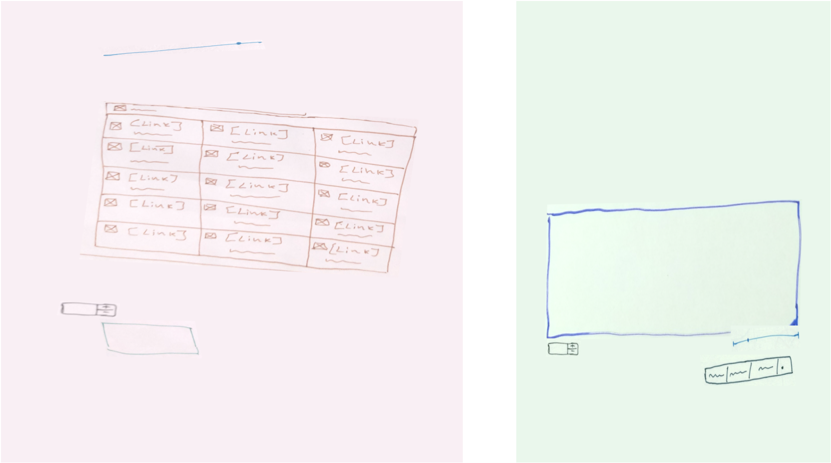


Fig. 2. Synthetic data example images. For classes with very few instances we injected the training set with artificial data containing augmented instances of these classes.

Table 2. Results of input format experiment. Trained for 50 epochs on 1000 × 1000 grayscale images with Faster R-CNN

Input format	mAP0.5	mAP	Recall@100
RGB	**0.9151**	0.5814	0.6594
Grayscale	**0.9151**	**0.5855**	**0.6689**
Gray ++	0.9087	0.5788	0.6705

Data preprocessing - While testing, we experimented with three different approaches to input preprocessing. First, images without any augmentations. Second, images were converted to grayscale. Third, where grayscale images without borders around captured drawing and with dilated lines were used. This effect was achieved by finding contours in the image using OpenCV [1]. These contours were then sorted by area, and the largest of them was used to crop the image. In some cases, this approach did not work correctly, especially where parts of the image were covered with a shadow, so this crop was only applied when the area of the contour was at least 70% of the image. After the crop, we utilized CLAHE [9] for histogram equalization and applied topological erosion to pronounce lines on paper. This approach is described as Gray++ in our results. Refer to Table 2, where it can be seen that Gray++ was worse in our testing, so our submissions mainly used grayscale images.

2.2 Object Detection Networks

There are several network architectures that were taken into the consideration for this task, in particular the Faster R-CNN [10] and EfficientDet [12]. The initial performance test for each particular network architecture was to train these networks with recommended configuration. These tests revealed the overall architecture suitability for the task at hand. The best performance was achieved with the Faster R-CNN architecture that used comparable backbones. Refer to Table 4.

Next, we needed to decide on the object detection network backbone. We have tested several widely used backbone architectures, namely the ResNet-50 [5], ResNet-101 [5], Inception V2 and Inception-ResNet-V2 [11] (Table 3).

Table 3. Results of backbone architecture experiment. Trained for 50 epochs on 1000 × 1000 grayscale images with Faster R-CNN.

Backbone	mAP0.5	mAP	Recall@100
Inception-V2	0.8974	0.5850	0.6689
ResNet-50	0.9151	0.5855	0.6689
ResNet-101	0.9035	0.6035	0.6835
Inception-ResNet-V2	**0.9176**	**0.6095**	**0.6904**

Table 4. Results of detection framework experiment. Trained for 50 epochs on 1000 × 1000 RGB images.

Approach	mAP0.5	mAP	Recall@100
EfficientDet-B3	0.583	0.416	0.458
Faster R-CNN ResNet-50	**0.9151**	**0.5814**	**0.6594**

Table 5. Training and network parameters shared among all experiments.

Parameter	Value	Parameter	Value
Optimizer	RMSprop	Gradient Clipping	10.0
Momentum	0.9	Input size	1000 × 1000
Initial and min LR	0.032 - 0.000032	Feature extractor stride	16
LR decay type	Exponential	Pretrained Checkpoints	COCO
LR decay factor	0.975	Num epochs	50
Batch size	2	Gradient accumulation	12

3 Submissions

In this competition, the AICrowd platform[2] was used to evaluate participants submissions. Each participating team was allowed to submit up to 10 text files with detection bounding-boxes in a specific format for each image. We have created 7 submission using configurations listed below.

Baseline configuration - As a baseline for all our experiments we used Faster R-CNN with ResNet-50 as a backbone. For training we used parameters and augmentations described in Table 5 and [8], respectively. Finally, thresholding was used to select only detection with high confidence.

Submission 1 - Baseline experiment trained on RGB images. Tested on original-size RGB images. Detection confidence threshold was set to 0.8.

Submission 2 - Submission 1 trained and tested on the grayscale images.

Submission 3 - Submission 2 trained on whole training set with no data for validation with confidence threshold of 0.95.

Submission 4 - Submission 3 trained for 80 epochs.

Submission 5 - Submission 1 with Inception-ResNet-V2 as backbone. Trained and tested on grayscale images with confidence threshold of 0.8.

Submission 6 - Voting ensemble created by combining models used in Submissions 2, 3 and 5 with confidence threshold of 0.8.

Submission 7 - Submission 6 with confidence threshold of 0.45.

4 Competition Results and Discusion

The official ImageCLEFdrawnUI competition results are displayed in Figure 3. The proposed system achieved the best Overall Precision score of **0.9709** and outperformed 2 other participating teams as well as the baseline solution proposed by organizers. The best scoring submission was produced by Mask R-CNN model with ResNet-50 backbone architecture and input resolution of 1000×1000 trained for 80 epochs with parameters and augmentations described in Table 5

[2] https://www.aicrowd.com

Table 6. Submission results achieved over test set.

Submission	1	2	3	4	5	6	7
Overall Precision	0.939	0.956	**0.971**	0.944	0.956	0.956	0.942
mAP@0.5	0.688	0.676	0.583	0.647	0.695	0.694	**0.755**
Recall@0.5	0.536	0.517	0.445	0.472	0.520	0.519	**0.555**
Run ID	67733	67814	67816	67991	68003	68014	68015

and [8], respectively. The resulting predictions were filtered with confidence threshold of 0.95 to maximize the official metric of mAP.

In our opinion, the winning submission is not the best of our submissions. According to the widely accepted performance metrics (mAP@0.5 and Recall@0.5), our Submission 5 (run ID 68003), which scored 3rd place overall, is superior to the winning submission. It diminishes ImageCLEF Overall Precision only by **0.0144**, while it increases mAP@0.5 by **0.111** and Recall@0.5 by **0.074**.

5 Conclusion

In this paper, we have presented a system for automatic hand-drawn UI element detection and localization. To achieve this goal, we had to gain a deep understanding of the provided dataset and perform many experiments to craft the best data preprocessing and augmentation methods, as well as objectively adjust the network parameters.

The final methods were based on the Faster R-CNN detection network with ResNet-50 used as a backbone architecture. The presented method scored first place in ImageCLEFdrawnUI competition, with an overall precision of **0.9708**.

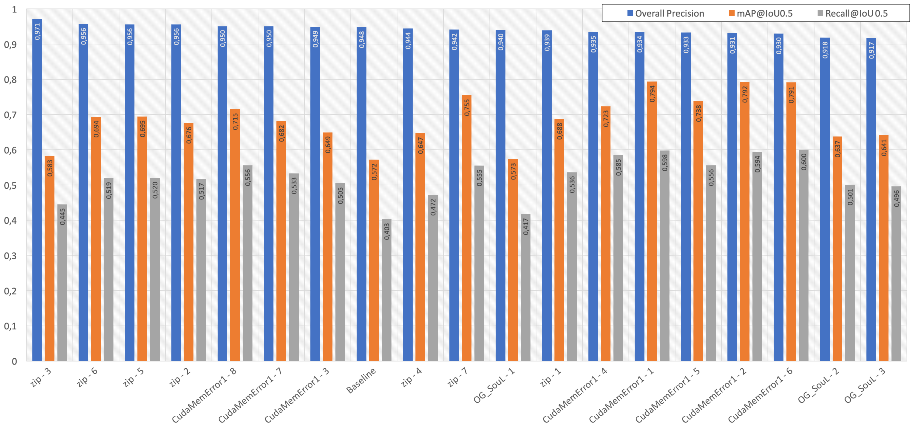


Fig. 3. Results for all runs submitted by the competition participants. Including additional metrics e.g. mAP@IoU0.5 and Recall@IoU0.5.

6 Acknowledgements

Lukáš Picek was supported by the Ministry of Education, Youth and Sports of the Czech Republic project No. LO1506, and by the grant of the UWB project No. SGS-2019-027.

References

1. Bradski, G.: The OpenCV Library. Dr. Dobb's Journal of Software Tools (2000)
2. Chamberlain, J., Campello, A., Wright, J.P., Clift, L.G., Clark, A., García Seco de Herrera, A.: Overview of the ImageCLEFcoral 2020 task: Automated coral reef image annotation. In: CLEF2020 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org> (2020)
3. Fichou, D., Berari, R., Brie, P., Dogariu, M., Ştefan, L.D., Constantin, M.G., Ionescu, B.: Overview of ImageCLEFdrawnUI 2020: The Detection and Recognition of Hand Drawn Website UIs Task. In: CLEF2020 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Thessaloniki, Greece (September 22-25 2020)
4. Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., He, K.: Accurate, large minibatch sgd: Training imagenet in 1 hour. arXiv preprint arXiv:1706.02677 (2017)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
6. Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., et al.: Speed/accuracy trade-offs for modern convolutional object detectors. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7310–7311 (2017)
7. Ionescu, B., Müller, H., Péteri, R., Abacha, A.B., Datla, V., Hasan, S.A., Demner-Fushman, D., Kozlovski, S., Liauchuk, V., Cid, Y.D., Kovalev, V., Pelka, O., Friedrich, C.M., de Herrera, A.G.S., Ninh, V.T., Le, T.K., Zhou, L., Piras, L., Riegler, M., l Halvorsen, P., Tran, M.T., Lux, M., Gurrin, C., Dang-Nguyen, D.T., Chamberlain, J., Clark, A., Campello, A., Fichou, D., Berari, R., Brie, P., Dogariu, M., Ştefan, L.D., Constantin, M.G.: Overview of the ImageCLEF 2020: Multimedia retrieval in lifelogging, medical, nature, and internet applications. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 11th International Conference of the CLEF Association (CLEF 2020), vol. 12260. LNCS Lecture Notes in Computer Science, Springer, Thessaloniki, Greece (September 22-25 2020)
8. Picek, L., Říha, A., Zita, A.: Coral reef annotation, localisation and pixel-wise classification using mask-rcnn and bag of tricks. In: CLEF (Working Notes). CEUR-WS.org <http://ceur-ws.org>, Thessaloniki, Greece (September 22-25 2020)
9. Pizer, S.M., Amburn, E.P., Austin, J.D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J.B., Zuiderveld, K.: Adaptive histogram equalization and its variations. Computer vision, graphics, and image processing **39**(3), 355–368 (1987)
10. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems 28, pp. 91–99. Curran Associates, Inc. (2015)

11. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-first AAAI conference on artificial intelligence (2017)
12. Tan, M., Pang, R., Le, Q.V.: Efficientdet: Scalable and efficient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10781–10790 (2020)
13. Tieleman, T., Hinton, G.: Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural networks for machine learning **4**(2), 26–31 (2012)

2020 25th International Conference on Pattern Recognition (ICPR)
Milan, Italy, Jan 10-15, 2021

Tracking Fast Moving Objects by Segmentation Network

Aleš Zita
The Czech Academy of Sciences
Institute of Information Theory and Automation
Pod Vodárenskou věží 4
Email: http://www.utia.cas.cz/people/zita

Filip Šroubek
The Czech Academy of Sciences
Institute of Information Theory and Automation
Pod Vodárenskou věží 4
Email: http://www.utia.cas.cz/people/sroubek

Abstract—**Tracking Fast Moving Objects (FMO), which appear as blurred streaks in video sequences, is a difficult task for standard trackers, as the object position does not overlap in consecutive video frames and texture information of the objects is blurred. Up-to-date approaches tuned for this task are based on background subtraction with a static background and slow deblurring algorithms. In this article, we present a tracking-by-segmentation approach implemented using modern deep learning methods that perform near real-time tracking on real-world video sequences. We have developed a physically plausible FMO sequence generator to be a robust foundation for our training pipeline and demonstrate straightforward network adaptation for different FMO scenarios with varying foreground.**

I. Introduction

Object tracking is a well-explored field of computer vision. The majority of object tracking algorithms starting from basic correlation trackers up to state-of-the-art deep network trackers utilize texture-based correlation or feature-based methods. Modern video capturing devices with built-in processing algorithms are capable of producing sharp images of moving objects. Moreover, the person capturing the object in motion typically tracks the moving object, hence it predominantly stays in the center of the image and in-focus. For such tasks, the correlation-based trackers are sufficient.

The situation changes dramatically when an object moves so fast, that it appears blurry on individual video frames. Such object in motion is called *'FMO'*, short for *Fast Moving Object* [1], and is loosely defined as an object traveling a distance larger than its diameter within one frame of the video sequence (Figure 1). The inter-frame object overlap is negligible, which causes problems to many conventional trackers.

A typical manifestation of the FMO in video frames is a prolonged streak without any particular texture, colored with the object prevailing color, or a combination of object colors; see Figure 1. The lack of any sharp texture of the object renders most of the texture-based correlation trackers inapplicable. Situation is even worse for very small objects moving fast relative to their sizes, such as ping-pong or squash balls.

The first tracking algorithm specifically designed for FMOs uses a method based on background subtraction [1]. This technique requires a static background, static camera, and large prominent foregrounds. It is also prone to object miss-tracking, which then requires a time-consuming object re-detection.

More recent approaches deal with the problem of FMO tracking by running a de-blurring algorithm [2], [3], [4]. These methods perform considerably better, but are extremely slow, as they require a full-blown de-blurring optimization pipeline. Therefore, they are not suitable for real-time video stream processing.

Our primary goal is to provide a method operating in real-world scenarios such as tracking of ping pong, squash balls, badminton shuttlecocks, and similar objects. This problem is very specific in sense, that the objects that need to be detected are very small and move very fast. This means that existing state-of-the-art tracking, object detection, or segmentation methods can not be used directly. Figure 2 shows the results from state-of-the-art semantic segmentation tool DeepLab3+ [5].

In this work we demonstrate that FMO tracking can be successfully solved with a machine learning approach. The proposed method uses a segmentation convolutional neural network (CNN) with real-time performance in videos with a resolution around 320x240. Network architecture is based on state-of-the-art tracking by segmentation methods rather than on object detection networks. We experimentally prove that tracking by segmentation outperforms tracking by correlation. In addition, we propose an on-demand synthetic FMO data generator to tackle the problem of producing annotated data automatically. Even though the network is trained solely on synthetic data, it can successfully be used in real-life applications, like processing YouTube sport video sequences. The proposed solution focuses predominantly on small bright foregrounds, yet we demonstrate the possibility of fast model fine-tuning for different foreground types.

Resulting segmentation can be further used for trajectory prediction and down the pipeline even for the trajectory estimation in de-blurring algorithms.

The method is evaluated on the FMO dataset [1] on which it shows competitive results. and investigate cases where the proposed algorithm outperforms or under-performs current methods both in precision and execution time.

978-1-7281-8808-9/20/$31.00 ©2020 IEEE

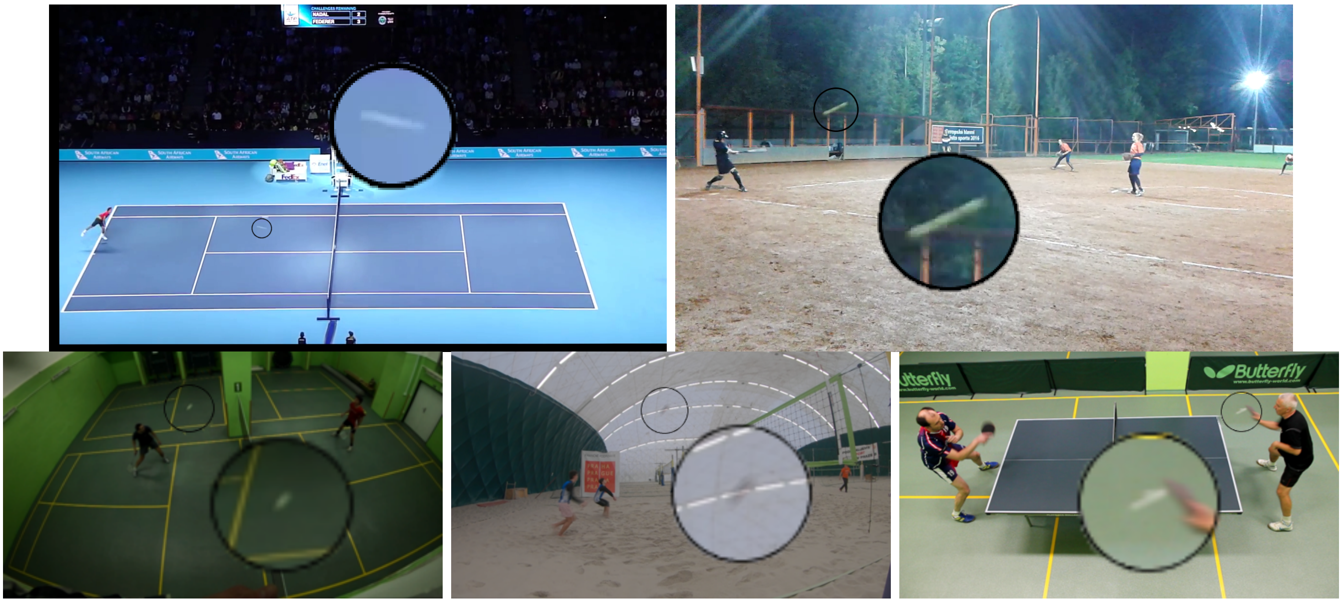


Fig. 1: Examples of Fast Moving Objects in real-world videos.

II. Related Work

Video Object Tracking

Object tracking is a well-established field of research in computer vision. Many methods have been proposed for tracking single or multiple objects in video sequences. Namely tracking by detection [6], [7], tracking by features [8], [9] tracking by correlation [10] and others. All of the approaches mentioned are based on either object detection using texture information of the tracked object or features extracted from it. This assumes that the object image contains some minimum level of details. Also, many of the conventional trackers perform best when the tracked object bounding boxes largely overlap in the consecutive frames. Both of the mentioned assumptions do not hold in sequences containing a FMO.

FMO Tracking

FMO tracking has been attracting the attention of researchers lately. Initial work in this field was done in [1], where the authors introduced the theme and proposed the first tracker based on background subtraction. In the heart of the method lies a tracker capable of tracking the background changes. When the tracker fails a time-consuming re-detection is executed to resume tracking.

Recently, interesting work was done in [2] where the tracking problem was defined as a de-blurring optimization problem. In another similar approach [4], authors show intra-frame tracking capability of the de-blurring approach. Albeit the results are promising in both mentioned publications, these methods focus on videos with static camera and background, and additionally, their algorithm cannot be used in real-time due to the high processor time demands of the optimization algorithm.

Deep Semantic Segmentation

There are many deep segmentation methods currently available, mainly based on encoder-decoder architecture. In [11] researchers introduced an interesting approach by using multiple stacked deconvolution blocks. Impressive results were achieved in DeepLabv3+ [12], where authors use the depth-wise separable convolution to both Atrous Spatial Pyramid Pooling and decoder modules to achieve high scores in both the PASCAL VOC 2012 and the Cityscapes datasets. Because the solutions often incorporate very deep networks, many of them have longer inference times [11] or are GPU memory demanding [13].

III. Method

First we give a brief introduction to the overall framework of the proposed method, then we investigate various strategies and finally, we describe the proposed method in depth.

A. Overview

The work of [1] inspired us to tackle the problematic cases on which the method performs poorly, namely tracking of very small objects. Larger moving objects in sports videos are often sharp because modern acquisition devices have short exposure and the cameraman actively tracks the object of interest. However for small objects that are moving very fast, typically balls in sports such as tennis, softball, or badminton, this is not true.

After some failed attempts to solve the tracking of small FMOs by conventional means, we have turned our attention towards deep learning methods. Learning-based approaches achieve top results in many segmentation tasks in terms of both computation time and precision. We researched several state-of-the-art segmentation networks and we achieved the best results with u-net-type architecture with inception bottleneck modules called ENet [14]; refer to Figure 3 for more details.

We choose the publicly available FMO dataset as a benchmark, to be able to compare the performance of the proposed

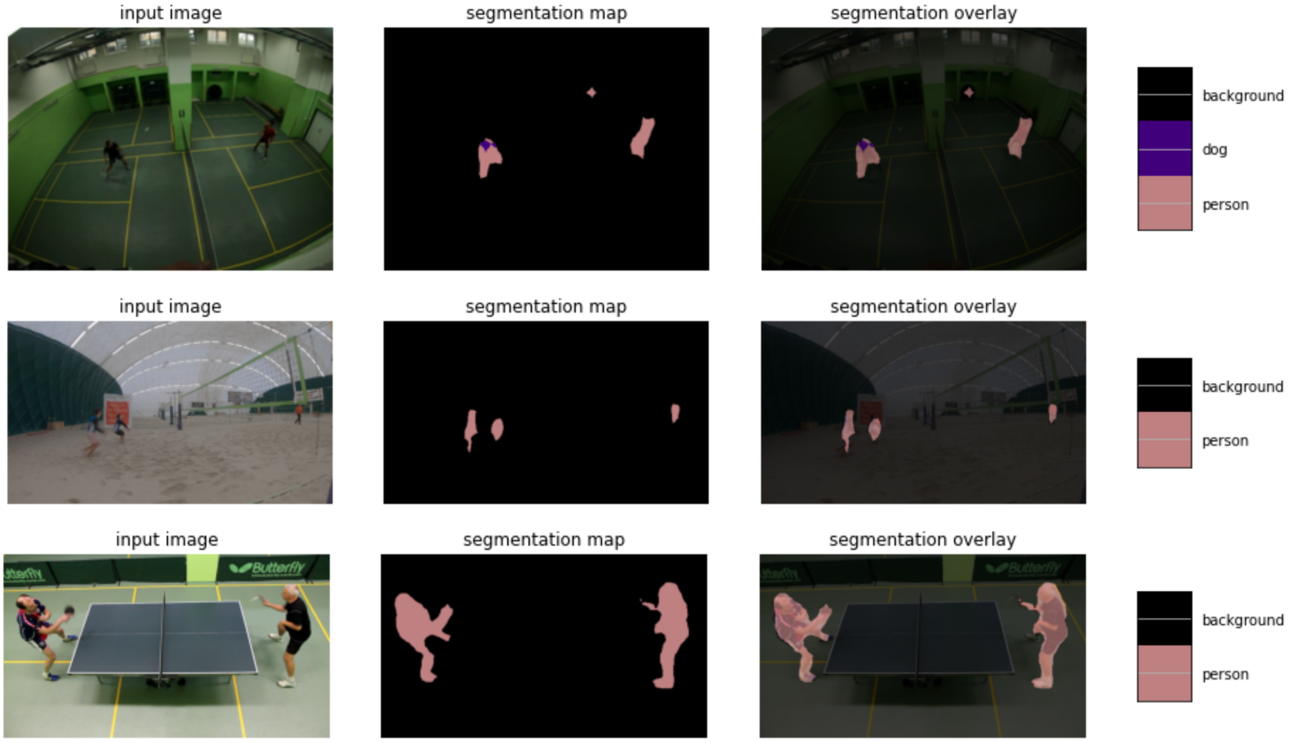


Fig. 2: Examples of FMO images evaluated on publicly available DeepLab3+ [5] semantic segmentation framework. State-of-the-art segmentation methods are not trained for detecting FMOs.

approach with the original method [1]. We perform preprocessing of the dataset in such a way that the size and color of the foreground resembles the foreground used for training.

Since our ultimate goal is tracking in real-world sports videos, we also include YouTube sports videos for performance evaluation. However, they are not annotated and so we provide only visual assessment.

Network	mAP	mAR	F1
Faster-RCNN with ResNet-50	33.2	15.5	21.2
ENet	36.5	52.7	41.2

TABLE I: Performance comparison between Enet segmentation and modified Faster-RCNN network as measured on bounding boxes. Metrics used in the table are standard Pascal mean Average Precision @0.5 and mean Average Recall @0.5

B. Network architecture

To decide the main direction of our research, we tested two current machine learning approaches: Semantic Segmentation and Object Detection.

After running performance and metric assessment tests of several segmentation frameworks, we opted for U-Net architecture called ENet [14] consisting of inception blocks proposed in [15]. The initial choice of this network design was based on the inference speed and performance on our benchmark dataset.

The choice of Object Detection network was based on study published in [16] revealing the Faster-RCNN [17] network with ResNet-50 [18] feature extractor backbone as a well balanced framework in terms of speed and accuracy.

ENet provide binary segmentation masks and RCNN bounding boxes. Both networks were adjusted to facilitate 15-channel input images to be able to process 5-frame sequences, initialized with publicly available weights pre-trained on ImageNet [19] and trained using our synthetic data generator. The performance of both networks was evaluated on the FMO dataset using mean Average Precision and mean Average Recall with Intersection over Union (IoU) threshold set to 0.5. See the performance comparison Table I. To compare both approaches, results of ENet were converted to bounding boxes by calculating axis-aligned rectangles circumscribing connected components in the segmentation masks.

Since ENet outperforms RCNN, we decided to base our approach on semantic segmentation.

The basic idea behind the FMO trace segmentation is training the network to recognize prolonged objects with no apparent texture, typically of white color to resemble most common sports balls. This represented in our opinion the majority of the problematic sports videos.

Single image segmentation, which is the standard input scenario for most of the segmentation methods, performs poorly in detecting FMOs and produces a large number of false positives. The overall measured Precision, Recall and F1 score (as defined in Section IV-B) was 4.3, 4.3 and 3.6, respectively. This is expected, as the proposed network learns to recognize

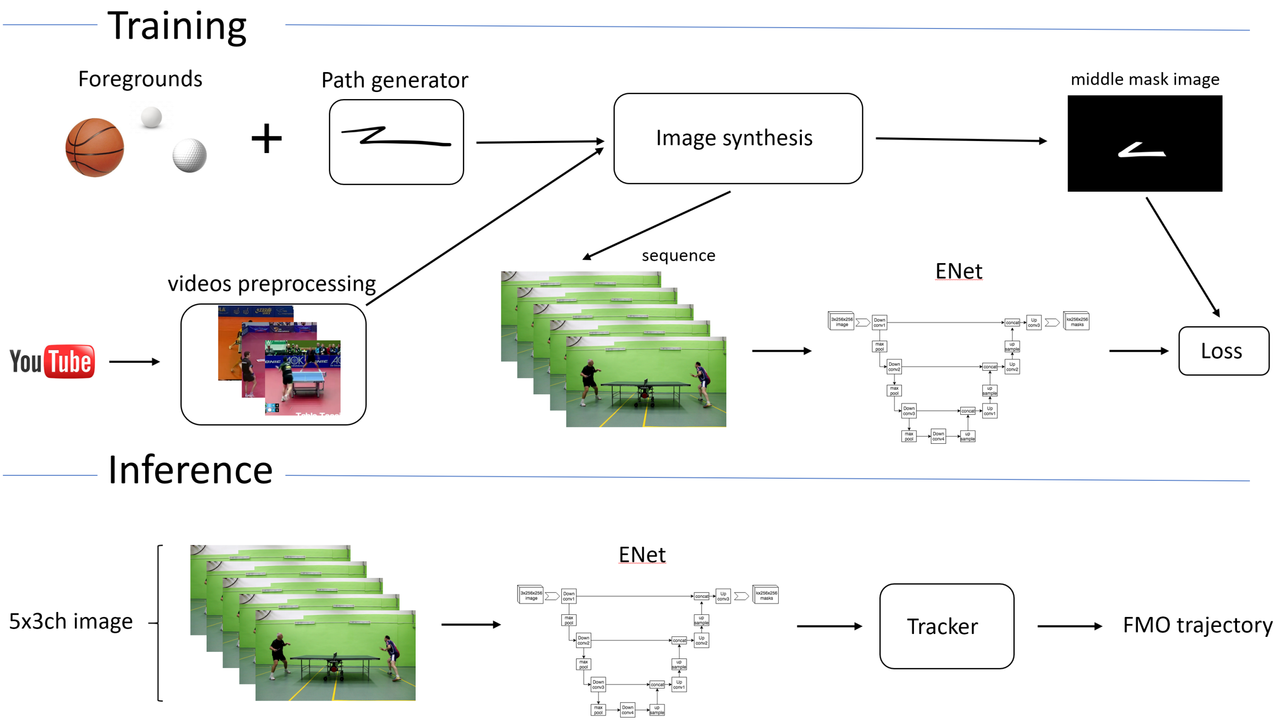


Fig. 3: Processing pipeline: During the training phase (top section) the sequences are dynamically synthesized using preprocessed video sequences, foregrounds, and path generator. Next, the frames are concatenated and input the network as a 15-channel image. During the inference phase (bottom part) the sequences are segmented by the network and Kalman based tracker is used for path prediction.

bright smears and therefore falsely segment any bright spots or lines in the image. In other cases, the model learns to ignore white lines, if they are in scene backgrounds, and does not detect FMOs at all. To overcome these limitations, we propose to use a sequence of consecutive frames as a network input. The idea is that image sequences improve trace consistency in time. We tested several multi-frame approaches, namely three and five frames either concatenated along color channels or as a full 4D input to the 4D network. Even though this approach is mathematically equivalent to the channel concatenation, it can provide faster learning and less false positives. The best results were achieved by using five consecutive video frames concatenated along color channels, i.e. the input to the network is a single 15-channel image; see Figure 3.

In our experiments, the original ENet architecture produced segmentation images with insufficient segment border precision. To address this problem, we have replaced the standard max-pooling with max-pooling-with-argmax and used the argmax values in corresponding upsampling unpooling layers. From this modification, more detailed segmentations were obtained.

The images used for training are synthetic FMO sequences based on real-world sports background images. Because every deep network is only as good as the dataset used for training, we have created a tool for generating synthetic sequences. This approach has proven to be very effective as the system is able to successfully segment fast-moving objects in real-world images, even though the network has never seen any during the training.

The majority of the state-of-the-art deep learning methods heavily depends on re-using the learned parameters from their successful predecessors. In our case, transfer learning led to worse performance. We hypothesize that this is due to the specificity of our task, which cannot exploit learned convolution kernels from other problems based on the extraction of texture features.

C. *Dataset generator*

Due to the nonexistence of a large annotated FMO dataset for training, we propose our own FMO sequence generator that obeys Newton's laws of motion. First, we collected YouTube sports videos which we used as a background. To eliminate any false fast-moving object from the videos, we have generated sequences of median images. Every frame of such a sequence was calculated as a median of 5 consecutive frames. Next, we created a foreground generator based upon selected ball images from a variety of sports. Finally, we designed a physically plausible generator of trajectories, including random bounces or occlusions; see examples in Figure 5.

In the core of the image synthesizer is a random motion path generator that takes into account a fully simulated camera (including CCD size resolution and aperture properties) as well as motion of the simulated object in space. The generator

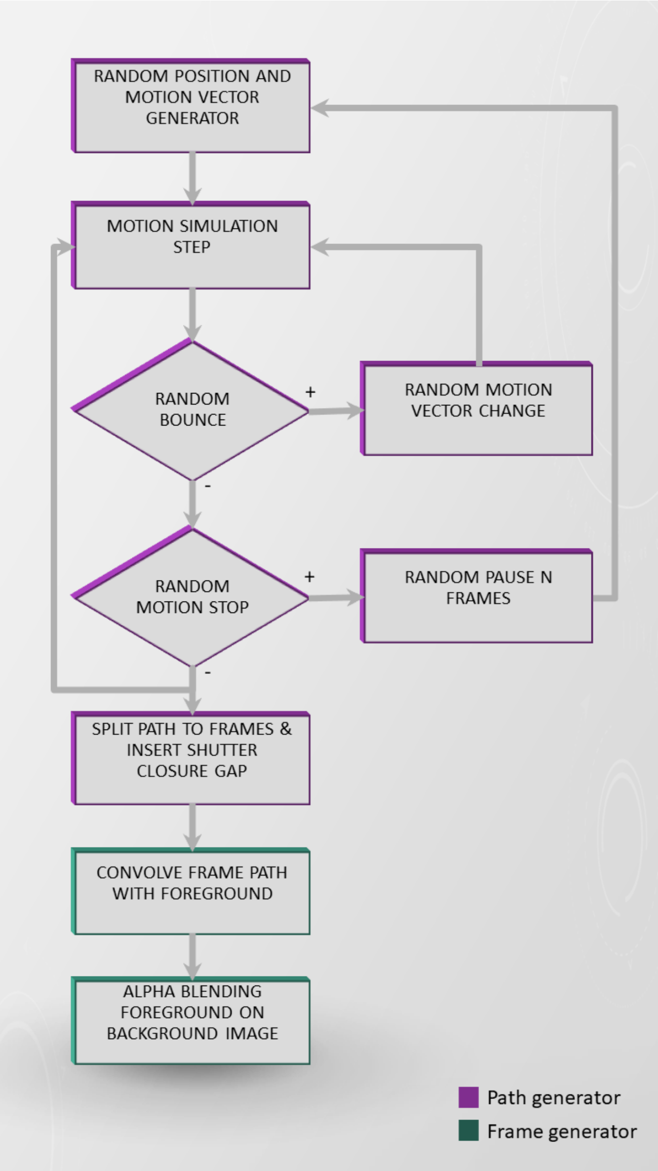

RANDOM POSITION AND MOTION VECTOR GENERATOR
MOTION SIMULATION STEP
RANDOM BOUNCE
+
RANDOM MOTION VECTOR CHANGE
-
RANDOM MOTION STOP
+
RANDOM PAUSE N FRAMES
-
SPLIT PATH TO FRAMES & INSERT SHUTTER CLOSURE GAP
CONVOLVE FRAME PATH WITH FOREGROUND
ALPHA BLENDING FOREGROUND ON BACKGROUND IMAGE
🟪 Path generator
🟩 Frame generator

Fig. 4: Generator pipeline of synthetic images. The motion is generated in sub-frame steps, which are consequently split to 5 frames with shutter closure gap emulation. The trajectory for a given frame is convolved with the foreground to form the blurred motion and the resulting foreground image is alpha-blended with the background according to (1).

begins from a random initial speed vector and then iterates in time simulating the motion. For better plausibility, the gravitational acceleration is taken into account too. Sudden velocity changes (e.g. hit from a racket), bounces (from wall, ground or table) occlusions and sudden motion stops are simulated as well. The processing pipeline scheme is depicted in Figure 4.

The generated trajectory is then convolved with the foreground to create the motion trace and finally inserted as a weighted sum into the sequence of background images using the following formula.

$$I_t(x) = [P_t * b_f F](x) + (1 - [P_t * M])B(x), \qquad (1)$$

where P_t is the path PSF normalized to sum to 1, $F(X)$ is the random foreground image, b_f is the overexposure brightness factor (described in next paragraph), $M(x)$ is the foreground indicator function and $B(x)$ is the background image. The used foreground image is created as a random selection of real-world white ball images that are tinted in random bright color and resized to a pre-defined range of foreground sizes.

Another aspect that had been taken into account is fast-moving object overexposure. This is due to the 'HDR' effect of the moving object. The overall brightness of the object in one frame can, and often is, brighter than the maximum brightness point in the rest of the image. Typically what every camera has to solve is the conversion of high brightness range of the world to the quantized 255 brightness values. This is done by several techniques that are out of the scope of this article. This conversion usually includes some form of clipping of the brightness levels which are too high to optimize overall image brightness balance. In a typical image without any FMO the overexposed parts of the image are clipped to the maximum allowed brightness. But, in the case of a fast-moving object, the true brightness of the object when stopped is an integration of its brightness along the object trajectory. In other words, the overall brightness of the object is spread out along the object path so it does not exceed the maximum pixel brightness of any point in the image. Therefore, it is often the case, that the true brightness of the object, when aggregated along the path, exceeds the maximum brightness of the image, especially with the white ball. If this effect would not have be taken into the consideration, the rendered object would seem very dim in the resulting image. This led us to set the factor of absolute brightness of the foreground between 0.8 - 1.4 of the maximum brightness.

As the ground truth mask image used in the training phase, we use the foreground path mask corresponding to the middle frame of the sequence. It is calculated again as a foreground mask convolved with the trajectory corresponding to the middle frame ($[P_3 * M]$); see Figure 3 for illustration.

D. Tracking

On top of the segmentation pipeline, we have implemented a simple tracker. The tracker is responsible for final object trajectory estimation. First, we select the blob which most likely represents the tracking object. This is achieved by simply selecting the largest connected component in the segmentation image. For sequences containing many false positives, more sophisticated logic should be applied. We used a weighted composition of two measures: connected component size and shape. Since we are looking for a prolonged object, we use second central moments of the connected components to estimate the prolongation.

Sequences of the bounding box positions are used by the tracker to extrapolate the object trajectory. For frames with

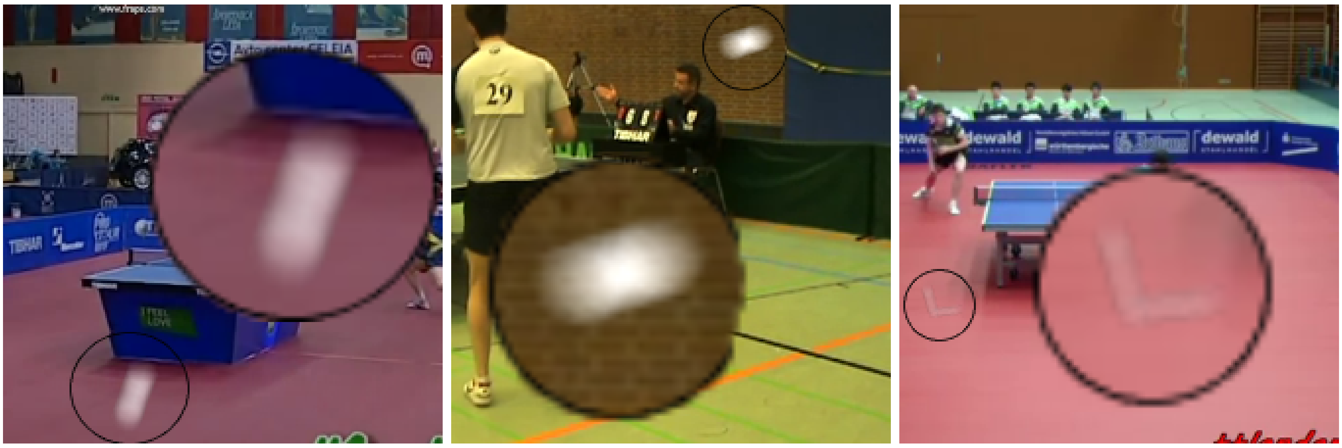


Fig. 5: Results of FMO synthetic data generator. The rightmost image shows example of small emulated bounce.

missing or too small blobs, we utilize a Kalman filter to estimate the missing trajectory or predicting trajectories in cases the object is lost or occluded. The output of the tracker is a sequence of coordinates representing the estimated object trajectory.

IV. Experiments

In this section we present performance of the proposed method and compare it to the original work [1]. We focus our attention to real-world applications with both inference speed and accuracy for small ball-like object detection.

A. Training

Initially, the network was trained on synthetic data with a wide range of foreground parameters. We used the modified ENet described in III using Adam optimizer, learning rate set to 0.01 and exponential learning rate decay; weight decay set to $2e-4$; average cross-entropy loss function; 200k iterations.

During the second stage of the training, the model was fine-tuned using the same architecture on a narrow size range of the synthetic foregrounds. The initial learning rate was lowered to 0.001 and optimizer switched to SGD. The training session ran for 50k iterations.

B. Evaluation

The proposed method was evaluated on the FMO dataset [1], where it achieved comparable or better results than the previously published method.

The performance criteria correspond to evaluation statistics in the original paper. These are precision TP/(TP + FP), recall TP/(TP + FN) and F1-score 2TP/(2TP + FN + FP), where TP, FP, FN is the number of true positives, false positive and false negatives, respectively. A true positive detection has an intersection over union (IoU) with the ground truth polygon greater than 0.5 and an area larger than other detections. The second condition ensures that multiple detections of the same object generate only one TP. False negatives are FMOs in the ground truth with no associated FP detection.

The results for both the original method and the proposed approach are listed in Table II. We can conclude, that overall mean F1-score is slightly better for our method, as well as mean recall. We avoid significant under-sizing of the resulted segmentation of the FMO trace, which causes high precision values over small recall value. Therefore, we argue that our approach results are more balanced in terms of precision and recall performance metrics.

The performance of the method reflects the purpose of our algorithm. It performs well on sequences with small ball-shaped objects moving fast relative to their size (ping-pong, softball, tennis, and squash); see Figure 6. Poor performance was recorded on sequences with foregrounds different from balls (like darts or archery) and on sequences with low background-foreground contrast (darts window and blue ball). The method performs poorly on data with a larger size-to-velocity ratio (frisbee and volleyball). Even though these sequences are part of the FMO dataset, foregrounds on these sequences are larger and are not moving faster than their diameter, as per FMO definition in Section I.

Our approach is advantageous in the fact that the network can be easily fine-tuned with image synthesizer setup for another sequence type, such as particular background (i.e. tennis tournament), particular foreground (i.e. yellow ball), foregrounds of different sizes, etc. For comparison, we have re-trained the network to detect foregrounds of bigger size and slower motions. The results are in the most right section of Table II. The segmentation network stopped to be sensitive to smaller foregrounds, such as ping pong, squash, or tennis, and starts to perform in cases with larger foregrounds, like frisbee or volleyball.

C. Computational time

Another benefit of the ENet neural network is the short inference time. The state-of-the-art approaches [2], [3], [4] are based on foreground de-blurring and therefore are inherently slow. In [4] authors state that the mean time is 4 seconds per frame. Our method is capable of near real-time execution while using a widely available graphics card, such as NVIDIA

	original work				Learning-based I			Learning-based II		
	n	Pr.	Rec.	F1	Pr.	Rec.	F1	Pr.	Rec.	F1
volleyball	50	100	45.5	62.5	0	0	0	33.3	42.9	37.5
volleyball passing	66	21.8	10.4	14.1	**20**	**16.2**	**17.9**	**85**	**98.1**	**91.1**
darts	75	100	26.5	41.7	37	**62.5**	**46.5**	33.3	**100**	**50**
darts window	50	25	50	33.3	**33.3**	33.3	33.3	**33.3**	33.3	33.3
softball	96	66.7	15.4	25	**83.3**	**83.3**	**83.3**	54.5	**66.7**	**60**
archery	119	0	0	0	**25**	**20**	**22.2**	**18.8**	**100**	**31.6**
tennis serve side	68	100	58.8	74.1	66.7	**76.9**	71.4	35.3	**85.7**	50
tennis serve back	156	28.6	5.9	9.8	**35.3**	**69.2**	**46.8**	26.4	**70**	**38.4**
tennis court	128	0	0	0	**33.3**	**40.8**	**36.7**	**25.5**	**58**	**35.5**
hockey	350	100	16.1	27.7	24.1	**86.7**	**37.7**	20	**91.7**	**32.8**
squash	250	0	0	0	**26**	**84.4**	**39.7**	**21.6**	**75.9**	**33.6**
frisbee	100	100	100	100	0	0	0	94.7	94.7	94.7
blue ball	53	100	52.4	68.8	40	26.7	32	58.3	**43.8**	50
ping pong tampere	120	100	88.7	94	58.6	66.7	62.4	0	0	0
ping pong side	445	12.1	7.3	9.1	**45.4**	**79.1**	**57.7**	0	0	0
ping pong top	350	92.6	87.8	90.1	56	**98.9**	71.5	0	0	0
Average per frame	2476	53.7	31	35.5	38.3	**68.5**	**47.2**	21.7	**49.7**	27.8

TABLE II: Performance of the original CVPR2017 method [1] in comparison to the proposed method (method I - trained for smaller foregrounds; method II - trained for bigger foregrounds). The results suggest the better overall performance of the trace segmentation in overall F1 performance score for method I.

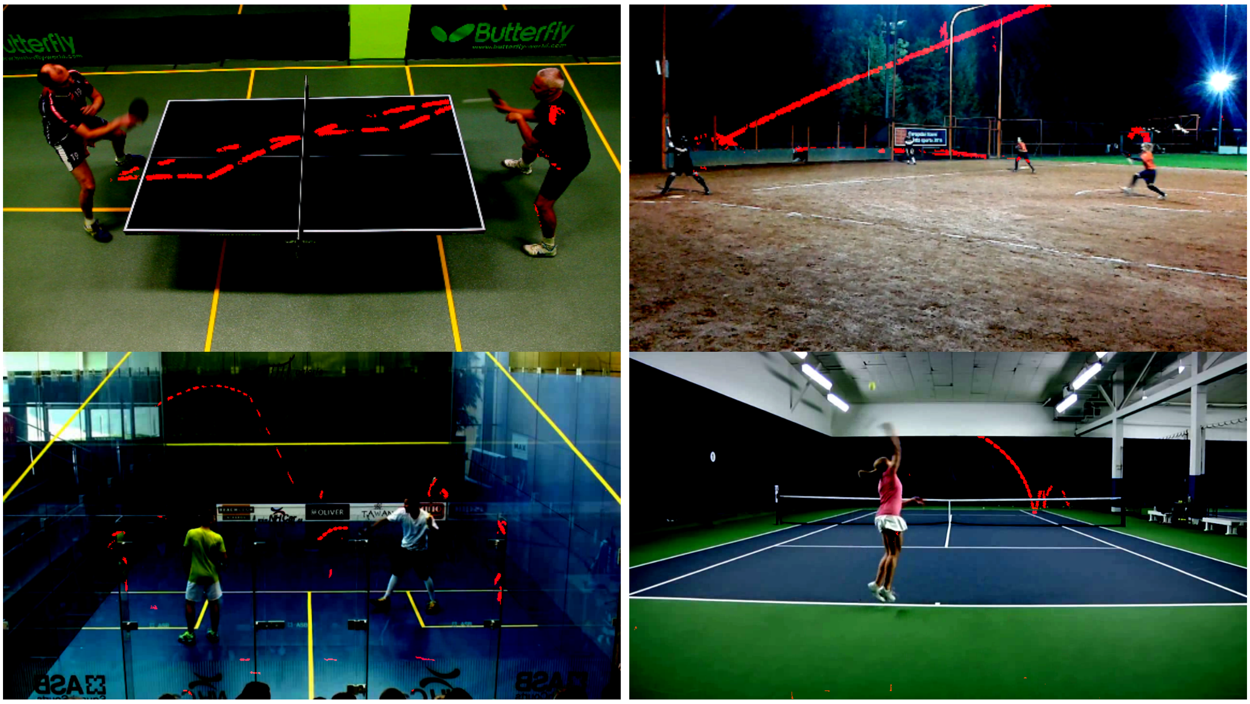


Fig. 6: Examples of segmentation results on FMO dataset. Notice false positives caused by racket or players movements or glass reflections.

video resolution	average fps
864 x 1536	2
576 x 1024	4.7
430 x 768	8.6
324 x 576	11.8
216 x 384	23.1

TABLE III: Some examples of video inference times achieved using NVidia Tesla X GPU.

GeForce 2080ti or similar. For more details refer to Table III, where we summarize mean frame evaluation times for NVidia Tesla P100 GPU with different image resolutions.

D. YouTube sport videos

We have created a tool that automatically downloaded more than 900 000 YouTube sports videos to create a base of our synthetic data generator backgrounds. Over 1800 of these sequences contain ping-pong matches, which we used for visual assessment of our framework. Although we measured our performance on the FMO dataset, we also aim for good performance on real wold sequences. Examples of ping-pong sequence evaluation can be seen in Figure 7.

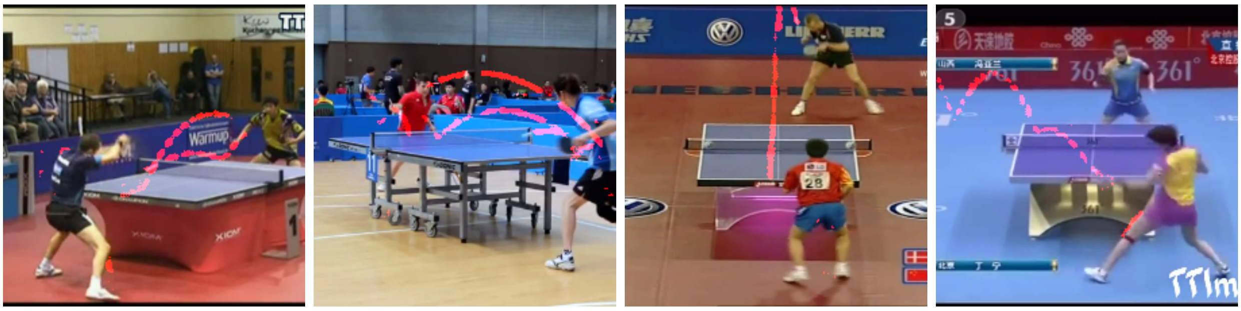


Fig. 7: Images show evaluation examples of YouTube real-world ping pong sequences.

V. Conclusion

We have implemented a learning-based method that performs near real-time detection and tracking of real-world fast moving objects. The proposed approach overcomes limitations of previous methods in this field, namely the long computation time and difficulty to detect small and very fast objects. We have introduced a synthetic physically plausible fast moving object sequence generator, which we use for network training. The simplicity of adapting the generator to another type of foreground followed by network fine-tuning allows us to detect foregrounds of different sizes and colors.

In the future work, we would like to focus on optimizing the processing pipeline with respect to speed in order to achieve true real-time performance in high-resolution videos and automatically track all kinds of sports balls in video streams. This can be further utilized in various applications such as instantaneous ball speed detection, ball misses, or detection of balls out of bounds.

Acknowledgment

This work was supported by Czech Science Foundation grant GA18-05360S and by the Praemium Academiae awarded by the Czech Academy of Sciences. We thank Prof. Jiri Matas for valuable comments and suggestions.

References

[1] D. Rozumnyi, J. Kotera, F. Sroubek, L. Novotny, and J. Matas, "The world of fast moving objects," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5203–5211.

[2] D. Rozumnyi, J. Kotera, F. Šroubek, and J. Matas, "Non-causal tracking by deblatting," in *German Conference on Pattern Recognition*. Springer, 2019, pp. 122–135.

[3] J. Kotera and F. Šroubek, "Motion estimation and deblurring of fast moving objects," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 2860–2864.

[4] J. Kotera, D. Rozumnyi, F. Sroubek, and J. Matas, "Intra-frame object tracking by deblatting," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[5] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *ECCV*, 2018.

[6] S. Hare, S. Golodetz, A. Saffari, V. Vineet, M.-M. Cheng, S. L. Hicks, and P. H. Torr, "Struck: Structured output tracking with kernels," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 10, pp. 2096–2109, 2015.

[7] J. Zhang, S. Ma, and S. Sclaroff, "Meem: robust tracking via multiple experts using entropy minimization," in *European conference on computer vision*. Springer, 2014, pp. 188–203.

[8] K.-W. Chen and Y.-P. Hung, "Multi-cue integration for multi-camera tracking," in *2010 20th International Conference on Pattern Recognition*. IEEE, 2010, pp. 145–148.

[9] S. Gladh, M. Danelljan, F. S. Khan, and M. Felsberg, "Deep motion features for visual tracking," in *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2016, pp. 1243–1248.

[10] H. Liu, Q. Hu, B. Li, and Y. Guo, "Long-term object tracking with instance specific proposals," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 1628–1633.

[11] J. Fu, J. Liu, Y. Wang, J. Zhou, C. Wang, and H. Lu, "Stacked deconvolutional network for semantic segmentation," *IEEE Transactions on Image Processing*, 2019.

[12] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[13] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 603–612.

[14] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," *arXiv preprint arXiv:1606.02147*, 2016.

[15] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[16] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama *et al.*, "Speed/accuracy trade-offs for modern convolutional object detectors," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7310–7311.

[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 91–99.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

Automatic Estimation of Mucosal Waves Lateral Peak Sharpness – Modern Approach

Aleš Zita [a]***, Šimon Greško*** [a]***, Adam Novozámský*** [a]***, Michal Šorel*** [a]***, Barbara Zitová*** [a]***, Jan G. Švec*** [b]***, Jitka Vydrová*** [c]
[a] ***The Czech Academy of Sciences, Institute of Information Theory and Automation, Prague, Czech Republic***
[b] ***Palacký University, Faculty of Sciences, Department of Experimental Physics, Voice Research Lab, Olomouc, Czech Republic***
[c] ***Voice Centre Prague, Medical Healthcom, Ltd., Prague, Czech Republic***

Abstract

Videokymographic (VKG) images of the human larynx are often used for automatic vibratory feature extraction for diagnostic purposes. One of the most challenging parameters to evaluate is the presence of mucosal waves and their lateral peaks' sharpness. Although these features can be clinically helpful and give an insight into the health and pliability of vocal fold mucosa, the identification and visual estimation of the sharpness can be challenging for human examiners and even more so for an automatic process. This work aims to create and validate a new method that can automatically quantify the lateral peak sharpness from the VKG images using a convolutional neural network.

Introduction

Videokymography is one of the fast-growing fields of vocal cords vibration visualization techniques. The method uses a line scanner camera to visualize a vibratory pattern of the larynx and its neighboring tissue (Figure 1). Vertically stacked scanned lines create a spatial-temporal videokymographic image (see Figure 2). Physicians use this visualization to evaluate the vibration characteristics of the vocal folds for diagnostic purposes, often with the help of an automatic software tool capable of extracting the essential characteristics and features [1]. The line scanner camera operates with a frequency of 7200 fps and typically produces 25 VKG images every second [2, 3]. Due to a large amount of data, VKG images are suitable for computer processing.

A phase difference between the movement of the lower and upper vocal cord edges causes the mucosal wave on the medial surface of the vocal cords. In the VKG images, a significant phase difference appears as a double contour during the glottis closure phase and as sharp lateral oscillation peaks (refer to Figure 3). If the phase difference between the upper and lower vocal cord margins is relatively small, it will show up as a restricted glottal wave and rounded lateral peaks in the VKG images. The occurrence and the shape of mucosal waves on the vibrating vocal cords are crucial indicators of larynx health conditions.

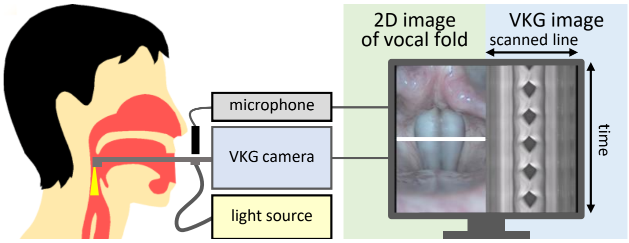


Figure 1. *Videokymography examination of the patient. The whole videokymographic frame comprises a 2D space image of the vocal fold (left side) and the temporal image of the scanned middle line highlighted in white in the 2D image (right side).*

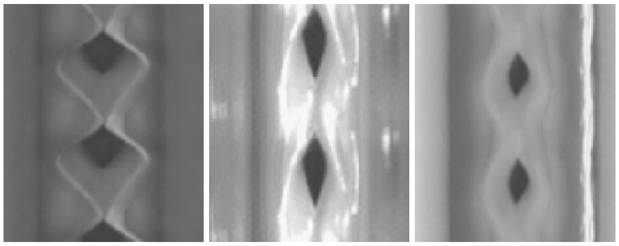


Figure 2. *Videokymographic images, where the vertical axis represents the temporal and the horizontal axis denotes the spatial domain. Here we see three different types of sharpness from left to right - sharp, rather rounded, and rounded.*

The performance of deep learning systems increased significantly in the last few years. In some areas, the machine learning approach exceeds the actual human experts. The main goal of this study is to verify the usability of deep learning for mucosal wave presence detection and the sharpness evaluation of waves' lateral peaks, one of the most complicated tasks in videokymographic image analysis.

In lateral peak sharpness assessment, manual ratings of the same images can vary between the examining experts; even single human professional evaluations may differ when repeated. Several influences cause these inconsistencies in rating, such as different levels of experience, length of practice, or such a trifle as the order of individual images. The combination of all these can bias the final assessment.

Related Work

Several researchers focused on the mucosal wave properties automatic estimation from the Videokymographic images [4, 5, 6], but due to its complexity, only a few have addressed the wave lateral peak sharpness.

Jiang et al. in 2000 [7] used an indirect method of peak sharpness estimation by quantifying the vertical phase difference using a sinusoidal model approximation. Although the method is correct in theory, the more complex shapes of the glottal contour are hard to process or interpret.

Yamauchi et al. [8] choose a different approach. They defined a *Lateral Peak Index* as an angle formed by two lines be-

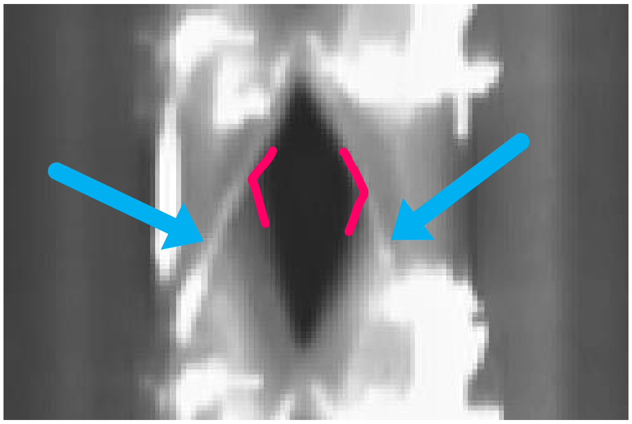


Figure 3. Mucosal wave as viewed on the VKG image. The blue arrows denotes the mucosal wave and the red marking shows the lateral peak

tween the start of the open phase and the lateral peak; and between the lateral peak and the end of the open phase. They quantified the sharpness of the lateral peak using the defined index. The drawbacks of this approach are that the index is sensitive to unrelated factors and discounts the changes of curvature of the vocal fold waveform that influence peak sharpness.

In [9], the researchers proposed four quotients that are good indicators of lateral peak sharpness. A set of proposed quotients was automatically calculated from the glottal contour line using the VKG Analyzer tool [1]. Four of the derived quotients had the best correspondence with the visual ratings of human experts, namely, PQ_{95}, PQ_{80}, OTQ_{95}, and OTQ_{80}. They are the variants of the *Plateau Quotients*, defined as the proportion of time during which the vocal fold displacement exceeds $R\%$ of vibration amplitude within the open phase (denoted as PQ_R) and the *Open Time Percentage Quotients*, defined as the proportion of time during which the vocal fold displacement exceeds a chosen percentage (R) of the vibration amplitude within a period (denoted as OTQ_R).

All the publications mentioned above attempted to estimate the sharpness of lateral peaks from VKG images using conventional image processing and mathematical approaches. To our knowledge, no other teams pursued this topic using a machine learning system.

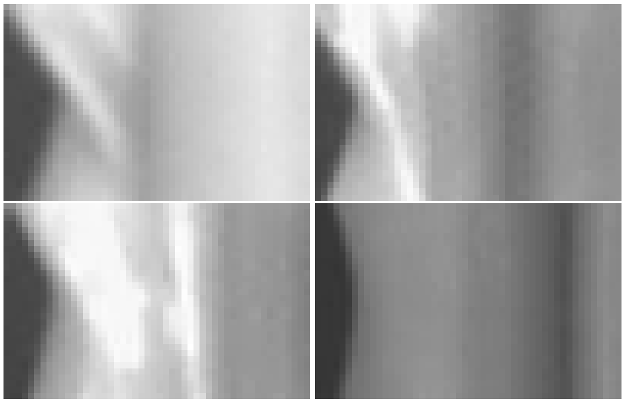


Figure 4. Cropped parts of videokymograms for sharpness classification; from left to right - sharp, rather sharp, rather rounded, and rounded.

Methodology

The proposed automatic mucosal wave lateral sharpness estimation method is defined as a deep learning classification task using a convolutional neural network (CNN). The technique uses fine-tuning of a pre-trained CNN architecture, trained on a set of data from a given application domain, in our case, videokymograms. The final system works with two same neural networks with different trained weights. The first one classifies the lateral peaks of a videokymogram into one of four classes of sharpness [*sharp, rather sharp, rather rounded, rounded*]; see the examples in Figure 4. The second one estimates the mucosal wave length into one of four ranges [*0-25%, 25-50%, 50-75%, 75-100%*].

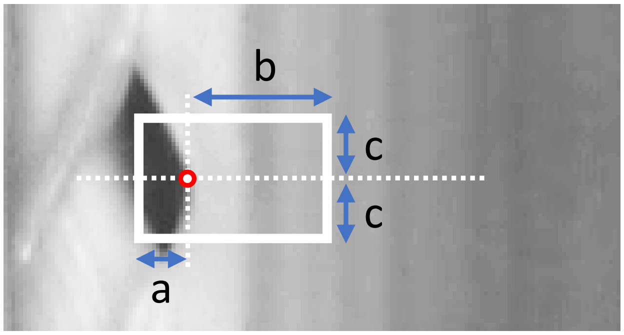


Figure 5. Cropping of the neighborhood of the lateral peak.

After trying and testing several different approaches, the best results were achieved using the following pipeline: We took advantage of the fact that videokymograms typically contain several visually almost identical vocal cord opening and closing cycles. Using a VKG Analyzer tool [1], we segment the vocal folds and detect significant points (lateral peaks, opening, closing). That gives us the beginning and end of each cycle. Within each cycle, we cropped the local neighborhood following these coordinates $[x_l - a : x_l + b, y_l - c + 1 : y_l + c]$; where $[x_l, y_l]$ represents the lateral point of a particular cycle. The graphic representation is in Figure 5.

A neural network subsequently classifies these cropped parts one after the other. In this way, we detect the relevant features separately for each cycle. The resulting overall value for the whole videokymogram is the most frequent (mode) for both the right and left sides. Additionally, the agreement of the results on individual cycles within a videokymogram determines the reliability of the estimation.

Parameters a, b, and c need to be set according to the size of one cycle in the image (in pixels). Gender, length of the throat, or the frequency of the vocal cords, is one of the main factors which cause these differences. We can normalize the whole image or adapt the size of the cropping. In our case, we selected images with similar sizes of cycles from different examinations of patients. Therefore we could set up the values globally as a=10, b=40, and c=16. No other normalization of size was needed.

We used *MobileNetV2*[10] with the *Adam*[11] optimizer algorithm as the backbone of our algorithm for its advantage of few parameters and few operations, which leads to easy implementation, fast inference, and not demanding hardware. The same neural network architecture and dataset were used to automatically estimate the lateral peak's sharpness and mucosal waves' presence. We trained the system on the left and right halves of the

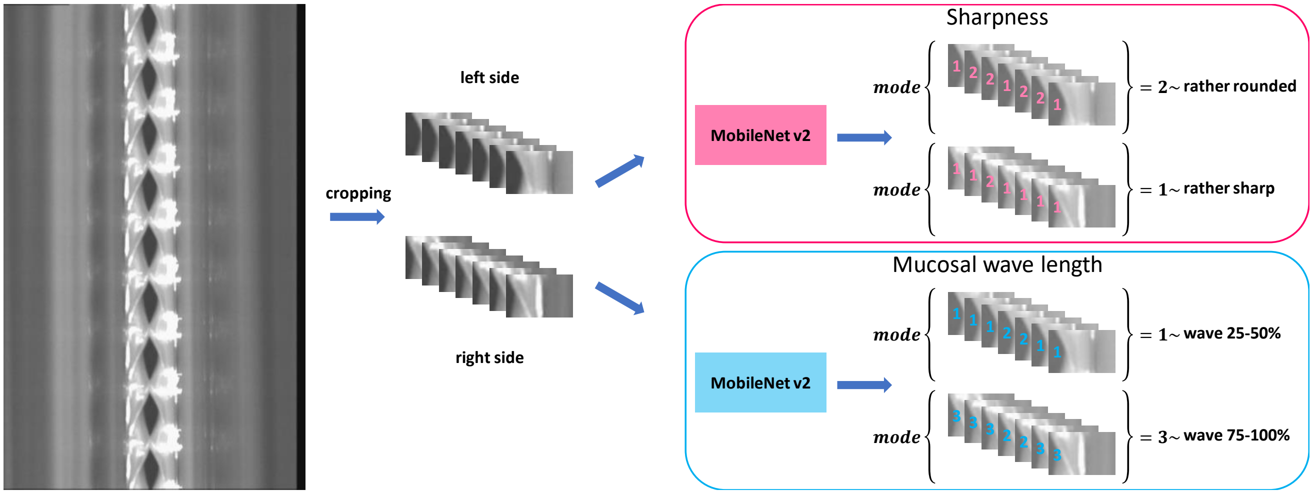


Figure 6. *Schematic representation of our method. The first step is a specific pre-processing operation (image enhancement and normalization). Then we cut individual complete cycles from the left and right sides of the kymogram. The cropped data is sent to CNN for sharpness and mucosal wave length classification in the next step. The final decision is to find the most common class on both sides.*

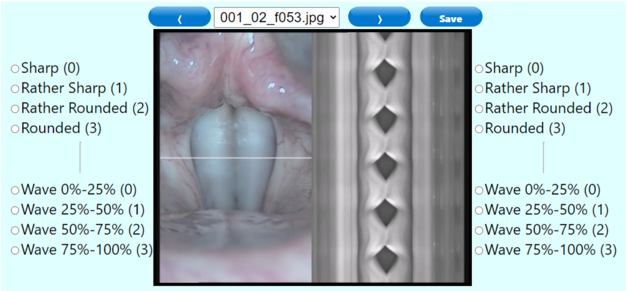
videokymogram together, with the left side first rotated around the vocal tract axis to the same position as the right side. Figure 6 shows the whole pipeline.

Dataset

All data used in this study are from examinations performed on patients of different ages, gender, and health conditions at the *Voice and Hearing Centre, Medical Healthcom, Ltd, Prague*. All VKG images come from the second generation VKG camera (Kymocam, CYMO, b.v. Groningen, the Netherlands) with a combination of different connected laryngoscopes, objective adapters, or light sources. We used two vocal cord specialists from the same department to evaluate the images.

The robustness of the CNN-based approach strongly depends on the quality of the training set used. Therefore, we have focused on data collection and subsequent annotation using the proposed web annotation tool we created for this task, see Figure 7. Using this tool, we could randomly display individual images to experts and store their sharpness ratings. For control, we showed each image several times in random order. In the resulting database, we only included images on which the experts agreed.

After this evaluation, our database consists of 319 expert-rated VKG images from clinical records. The images were processed and analyzed by VKG Analyzer software [1], and particular cycles were extracted and saved for subsequent sharpness analysis. In this way, a database of 3695 cropped parts with a size of 32x50 pixels was created, which was then divided into training



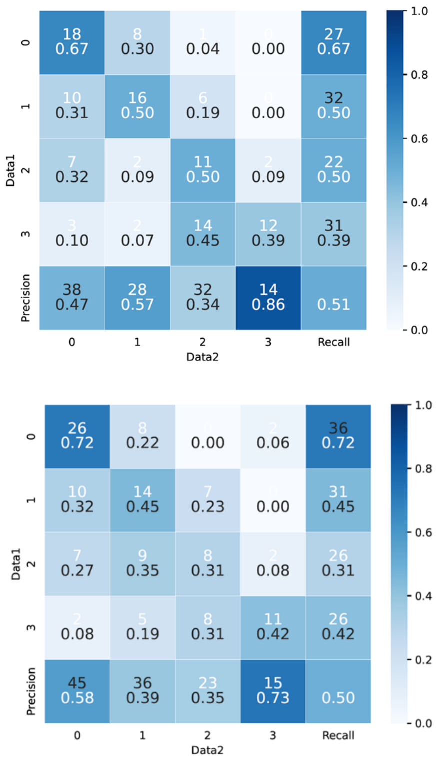
○Sharp (0)
○Rather Sharp (1)
○Rather Rounded (2)
○Rounded (3)

○Wave 0%-25% (0)
○Wave 25%-50% (1)
○Wave 50%-75% (2)
○Wave 75%-100% (3)

○Sharp (0)
○Rather Sharp (1)
○Rather Rounded (2)
○Rounded (3)

○Wave 0%-25% (0)
○Wave 25%-50% (1)
○Wave 50%-75% (2)
○Wave 75%-100% (3)

Figure 7. *A screenshot from the proposed tool used for the training dataset manual annotations. Annotators evaluated the sharpness of left and right lateral peaks and the level of mucosal wave length (percentage denotes the distance to the neighboring tissue, see Figure 3 for illustration of 100% wave).*



Data1 \ Data2	0	1	2	3	Recall
0	18 0.67	8 0.30	1 0.04	0 0.00	27 0.67
1	10 0.31	16 0.50	6 0.19	0 0.00	32 0.50
2	7 0.32	2 0.09	11 0.50	2 0.09	22 0.50
3	3 0.10	2 0.07	14 0.45	12 0.39	31 0.39
Precision	38 0.47	28 0.57	32 0.34	14 0.86	0.51

Data1 \ Data2	0	1	2	3	Recall
0	26 0.72	8 0.22	0 0.00	2 0.06	36 0.72
1	10 0.32	14 0.45	7 0.23	0 0.00	31 0.45
2	7 0.27	9 0.35	8 0.31	2 0.08	26 0.31
3	2 0.08	5 0.19	8 0.31	11 0.42	26 0.42
Precision	45 0.58	36 0.39	23 0.35	15 0.73	0.50

Figure 8. *The confusion matrices between professional physicians and the convolutional network in* ***wave presence detection****. The tables show the evaluated features' precision, recall, and accuracy. Professional physicians' ratings are on the vertical axes and the proposed method results on the horizontal axis. The upper table belongs to the left side of the vocal cord, the lower to the right.*

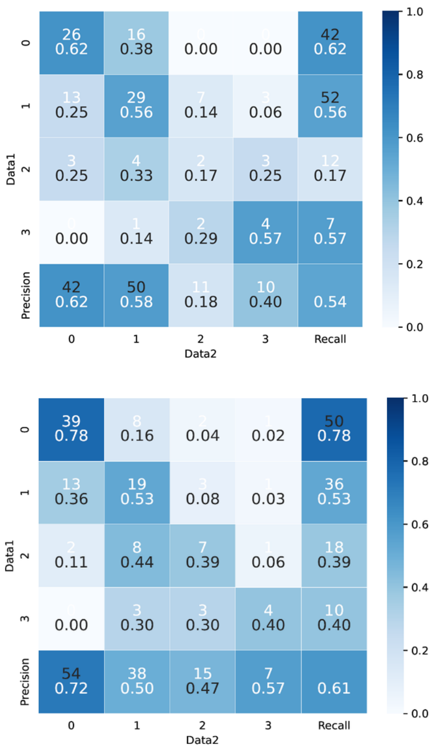


Figure 9. *The confusion matrices between professional physicians and the convolutional network in* ***lateral peak sharpness****. The tables show the evaluated features' precision, recall, and accuracy. Professional physicians' ratings are on the vertical axes and the proposed method results on the horizontal axis. The upper table belongs to the left side of the vocal cord, the lower to the right.*

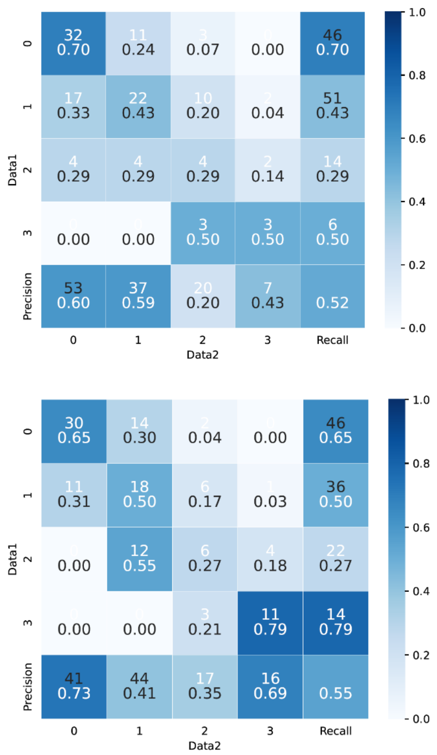


Figure 10. *The performance comparison of the junior evaluators' ratings vs. the professional physician. The table shows the professional (ground truth) on the vertical axis and the examining rater on the horizontal axis. The bottom right corner denotes the overall accuracy. The upper table belongs to the left side of the vocal cord, the lower to the right.*

and test parts in a ratio of 3:2. Examples of individual sharpness classes can be seen in Figure 4. Due to the vocal cords' symmetry, we could evaluate the left and right vocal cords simultaneously. The results of our method on the validation set in the form of confusion matrices can be seen in the following section of this document.

Results

The network performance results are presented in the form of confusion matrices. Figures 8 and 9 show both evaluated features' values for the left and right sides. The vertical axis corresponds to the ground truth value determined by the expert (values 0 to 3), while the horizontal axis corresponds to the values determined by the machine learning algorithm. The integer values in the contingency table correspond to the number of cases (combinations of expert and algorithm values). Numbers below the combination values are the same values normalized to the sum of one, row-wise. The rightmost column shows the recall values, and the bottom row shows the precision values. The bottom right corner then displays the overall percentage of exact match (accuracy). Then we wanted to compare the ratings of two junior evaluators with our professional physicians. The results of this comparison are shown in Figure 10; the interpretation is the same as in Figure 8.

Finally, the proposed machine learning algorithm achieves an accuracy of **0.54** and **0.61** for right and left lateral peak sharpness, which exceeds the match of success of junior evaluators (0.52 and 0.55). Similarly, compared to junior evaluators' values, it improves the rating for wave length, reaching an accuracy of **0.51** and **0.50** for the left and right sides. The system's performance will improve in the future through our continuous fine-tuning of the network with newly acquired data. We consider this to be the study's main result, as it will allow us to automate and objectify the estimation of an important parameter for evaluating the condition of the vocal cords.

Conclusion

We have developed a CNN-based tool for automatically estimating lateral peak sharpness and the mucosal wave length in videokymograms. The performance of this tool was evaluated on a small dataset, and the results indicate that the proposed method can accurately assess the sharpness of lateral peaks, and the level of accuracy is higher or comparable to that of non-specialists.

In addition, when trained on a more extensive and diverse dataset, we expect significant improvements which will be approaching accuracy achieving professional physicians. In that case, we could handle a more comprehensive range of videokymograms and accurately assess the sharpness of lateral peaks in a broader range of vocal conditions. Overall, this tool shows promise as a reliable and efficient method for evaluating the health and function of the vocal cords.

Acknowledgments

This work was partially supported by the Technology Agency of the Czech Republic, grant No. TH04010422, by the Czech Science Foundation, grant no. GA21-03921S, and by the *Praemium Academiae* awarded by the Czech Academy of Sciences.

References

[1] Aleš Zita, Adam Novozámský, Barbara Zitová, Michal Šorel, Christian T Herbst, Jitka Vydrová, and Jan G Švec, "Videokymogram analyzer tool: Human–computer comparison," *Biomedical Signal Processing and Control*, vol. 78, pp. 103878, 2022.

[2] Jan G. Švec and Harm K. Schutte, "Videokymography: High-speed line scanning of vocal fold vibration," *Journal of Voice*, vol. 10, pp. 201–205, 1996.

[3] J. G. Švec and F. Šram, "Videokymographic examination of voice," in *Handbook of Voice Assessments*, E. P. M. Ma and E. M. L. Yiu, Eds., pp. 129–146. San Diego, CA: Plural Publishing, 3rd edition, 2011.

[4] Adam Novozámský, Jiři Sedlář, Aleš Zita, Filip Šroubek, Jan Flussef, Jan G. Švec, Jitka Vydrová, and Barbara Zitová, "Image analysis of videokymographic data," in *2015 IEEE International Conference on Image Processing (ICIP)*, 2015, pp. 78–82.

[5] Jack J Jiang, Yu Zhang, Michael P Kelly, Erik T Bieging, and Matthew R Hoffman, "An automatic method to quantify mucosal waves via videokymography," 2008.

[6] S. Pravin Kumar and Jan G. Švec, "Kinematic model for simulating mucosal wave phenomena on vocal folds," *Biomedical Signal Processing and Control*, vol. 49, pp. 328–337, 3 2019.

[7] Jack J. Jiang, Ching I.B. Chang, Joseph R. Raviv, Sameer Gupta, Franklin M. Banzali, and David G. Hanson, "Quantitative study of mucosal wave via videokymography in canine larynges," *Laryngoscope*, vol. 110, pp. 1567–1573, 2000.

[8] Akihito Yamauchi, Hisayuki Yokonishi, Hiroshi Imagawa, Ken-Ichi Sakakibara, Takaharu Nito, Niro Tayama, and Tatsuya Yamasoba, "Quantitative analysis of digital videokymography: A preliminary study on age- and gender-related difference of vocal fold vibration in normal speakers," 2015.

[9] S.P. Kumar, K.V. Phadke, J. Vydrová, A. Novozámský, A. Zita, B. Zitová, and J.G. Švec, "Visual and automatic evaluation of vocal fold mucosal waves through sharpness of lateral peaks in high-speed videokymographic images," *Journal of Voice*, vol. 34, 2020.

[10] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[11] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," 2014.

Author Biography

Aleš Zita received his M.Sc. degree in informatics from the Faculty of Mathematics and Physics, Charles University, Prague, in 2013. He is pursuing his Ph.D. with the Institute of Information Theory and Automation Cooperating Institute of Charles University, Prague. His research interests include medical imaging, image segmentation, machine learning, and image forensics.