Using Cluster Analysis for Image Processing in High Speed Video Laryngoscopy

Ettler Tomáš
Department of Computer Science and Engineering
University of West Bohemia
Pilsen, Czech Republic
E-mail: thritton@kiv.zcu.cz

Nový Pavel
Department of Computer Science and Engineering
University of West Bohemia
Pilsen, Czech Republic
E-mail: novyp@kiv.zcu.cz

Abstract – This paper summarizes findings related to the problematics of glottis detection in video sequences obtained by medical examination of vocal cords by high speed videolaryngoscopy (HSV). The glottis detection is based on cluster analysis method K-means which complements the existing set of segmentation methods used in detection algorithms. This method has been tested on a large corpus of HSV sequences from clinical practice on ENT department.

Keywords-vocal cords; glottis; laryngoscopy; image segmentation; cluster analysis; K-means;

I. INTRODUCTION

Vocal cords examination by High Speed Video laryngoscopy (HSV) became standard method used in the field of otorhinolaryngology and phoniatry. The video recording of the glottis allows to observe the real movement of the vocal cords in a slow, time-spaced form. Since watching the video alone by the doctor may not always be sufficient to evaluate the vocal cords kinematics, a number of supporting tools were developed to analyze the vocal cords behavior by various methods. For this purpose, methods of analysis and processing of individual vocal cord images (sequence of frames are used as a two-dimensional signal, i.e. a 2D signal distributed over time). To analyze the behavior of the vocal cords, it is therefore necessary to define parameters that characterize the temporal changes of the vocal cords during one or more periods. Based on these parameters, the criteria used to assess the quality of the vocal fold kinematics are further determined [1], [2].

The basic task of processing images in the HSV sequence is to detect the glottis on each frame, which is the key to assessing the kinematics of the vocal cords.

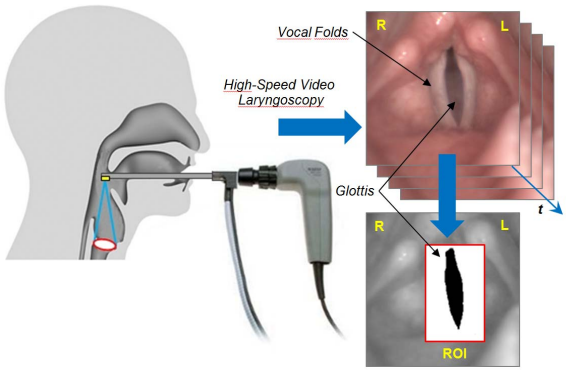
Currently, commercial software is available for HSV analysis, which typically allows manual or semi-automatic glottis detection, manual vocal cords axis determination, calculating and displaying pixel value changes in time, calculating glottis area size and creating a videokymogram at any position on the frame. They also allow visualization and sorting of individual images to enable visually determine the opening and closing moment of vocal cords and calculate the Open Quotient (OQ) or to capture the change in brightness at predefined points on the vocal cords during one or more periods. When evaluating some results, which meet the necessary prerequisites, Fourier analysis (FFT) is used to obtain another parameter set.

This paper deals with the description of segmentation method that we use for vocal glottis detection. It is a modified method of cluster analysis, specifically K–means method with static and adaptive parameters.

II. HIGH SPEED VIDEOLARYNGOSCOPY AND GLOTTIS DETECTION

High-speed laryngoscopy is an optical indirect laryngoscopic examination technique. This is a rigid endoscopic method supplemented by a high-speed camera with a frame rate of at least 1000 fps and a minimum resolution of 256x256 pixels, Fig. 1. The device is able to capture and store the real movement of the vocal cords during their entire oscillation in the form of a video sequence which can be slowed down. This allows detailed analysis of the vocal folds and the shape of the glottis during the opening and closing phases of the vocal cords.

Figure 1. Scheme of high speed videolaryngoscopy and the glottis detection [15].



The HSV recordings we process are captured by the HSV HRES ENDOCAM 55621 system[1], with 4000 fps and 256x256 pixels. The video recording is taken for a few seconds while phoning the vocal "i:" simultaneously with acoustic signal. Sometimes both recordings, video / audio, are supplemented with an electroglottographic EGG record.

[1] Richard Wolf GmbH, http://www.richardwolf.com .

ISBN 978-80-261-0892-4, © University of West Bohemia, 2020

A. Video recording quality

The success of the glottis detection is determined by the quality of the video recording. Because it is an examination with a rigid laryngoscope with the request of phonation of the vocal "i:" and the short time available for obtaining the video, the quality of the resulting video can vary.

Factors that mostly affect video quality:

- Angle of camera – projection plane (optical system) should be parallel to plane of vocal cords, Fig. 1.
- Image blur – it leads to inaccurate detection of the glottis boundary.
- Over lighting – this causes false contours and areas in the image that can lead to false region of interest and glottis detection.
- Poor lighting – Loss of visual information, image has low contrast and contains noise.
- Overlap – Some anatomical structures of the larynx may overlap the glottis if the HSV camera is misaligned or rotated.
- Camera movement – the camera moves due to the tremor of the hand. But the dynamics of camera position change is usually slow relative to the capture rate. Therefore, this movement is neglected in image processing, registration method is applied or the method can compensate that.
- Noise – it affects processing and the accuracy especially when the illumination of anatomical structures in the image is low.
- Presence of body fluids – a mucus overlapping some anatomical parts may cause false reflections and false anatomical structures may be detected.
- Size of the vocal cords in the image – at a limited camera resolution of 256x256 pixels, the small size of the vocal cords (large distance of the HSV camera from the vocal cords) causes inaccuracy in the glottis detection and consequently in parameter calculations.

B. Glottis detection

Currently during examination, the glottis is usually detected manually or semi-automatically with user input. Semi-automatic methods are mostly included in proprietary commercial software, but the processing methods are usually not described in detail. The most commonly published approaches to glottis segmentation are:

- Automatic thresholding [3], [4], [5].
- Watershed segmentation [6], [7], [8].
- Region growing [9], [7].
- Gabor filtering [10], [11].
- Active contour [12], [13], [14], [23].

However, applying these methods separately is not very successful when processing complicated images with lower quality. Therefore, the combined methods are used by many authors [3], [5], [16], [17], [22].

In cooperation with the ENT Clinic of the University Hospital Pilsen, we use two methods for automatic glottis detection in our applications, the *Max-Min-Thresholding* method and the *cluster analysis* method.

The *Max-Min-Thresholding* [18], [19] method consists of a sequence of steps and includes several processes that are applied to both individual frames and the entire video sequence. The method is the result of testing a number of approaches and their modifications in the field of point transformations, filtration, automatic thresholding and construction of continuous areas in the image. After exhausting the possibilities of this method we started to use another approach, *cluster analysis*, specifically MacQueen algorithm, K-means method applied on image pixels.

III. Glottis Segmentation by Cluster Analysis Method

Based on our experience, we analyzed the brightness and color composition of HSV recordings images. The glottis area differs in color from its surroundings, most notably in the red component of the RGB color model. We therefore consider individual pixels as separate objects with parameters derived from the character of the image data and the pixel position. Thus, pixels can be divided into classes with similar properties by means of cluster analysis [20].

In general, cluster analysis represents procedures for grouping objects into more or less homogeneous groups based on their similarity. The most commonly used cluster analysis method is the *K-means* method [21]. According to individual parameters, objects are classified into k classes ($CLASS_j$) with the least difference in parameters within the class or with the greatest difference in parameters between classes. Individual classified objects are placed in m-dimensional space, where m is the number of the parameters. Thus, the m-dimensional space coordinates are indexes to the parameter values and each parameter can have a different weight. The distances between objects are then calculated in this space to look for clusters of objects with similar properties. As the input of the method, centers ($Center_j$) are specified (which can be some selected objects or newly created objects with specified parameters) to reach the division of objects into classes accurately and quickly.

The first step of the method is the initial division of the input data xx_i, $i = 1$ to n, where n is the number of objects, into classes $CLASS_j$ according to the selected criterion. The criterion (1) of the smallest distance of the object relative to each primary center $Center_j$ is used, where $j = 1$ to k, k is the number of $CLASS_j$ classes.

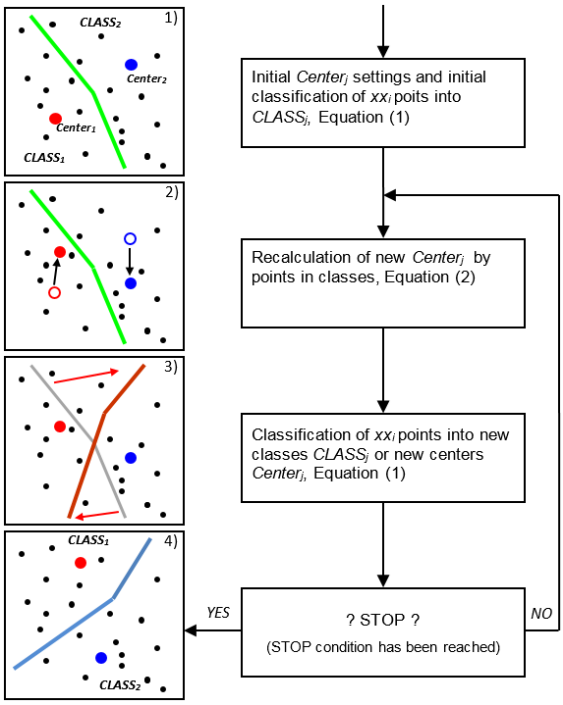
$$CLASS_j = \underset{j=1\ldots k}{arg\ min} \left\| xx_i - Center_j \right\| \quad (1)$$

After division into classes, in the second step, new centers $Center_j$ are computed for each class, e.g. their center points. The new centers are determined according to xx_i objects belonging to $CLASS_j$ (2), where n_j is the number of xx_i objects in $CLASS_j$.

$$Center_j = \frac{1}{n_j} \sum_{i \in CLASS_j} xx_i \quad (2)$$

This method can be repeated until the state is steady, when the division of the points into the classes no longer changes or the classification can be terminated after a specified number of cycles, Fig. 2.

Figure 2. Visualization of the K-means cluster analysis algorithm.



A. *Cluster analysis with static parameter weights*

Considering pixels as objects, there are several parameters that can be selected for classification together with their weights. Parameters defined for cluster analysis:

- R – red component.
- G – green component.
- B – blue component.
- X – x coordinate in the frame.
- Y – y coordinate in the frame.
- RmB – difference between R and B component.
- C – distance from the center of the frame.
- LR_{Diff} – difference of red component value between neighboring left and right pixel (introduced later).

The optimal weight setting is a very complex heuristic task where the values of the individual components were determined experimentally, TABLE I.

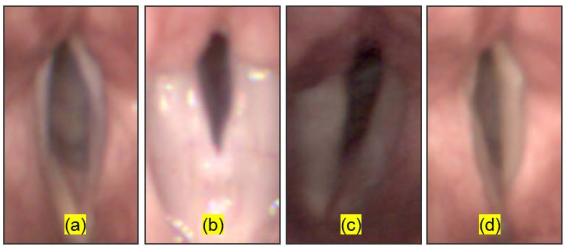
Before applying the cluster analysis method it is necessary to determine the number of classes, it means number of primary centers. Two classes should be sufficient to detect points belonging to the glottis and its surroundings, but in low-contrast images, the two classes may be insufficient.

TABLE I. DESCRIPTION AND THEIR STATIC WEIGHTS OF CLUSTER ANALYSIS PARAMETERS

Parameter description with the initial and result weight settings *coeff₀* and *coeff₁*			
Parameter	***Value range***	***coeff₀***	***coeff₁***
R	Component value (0 to 255)	1,00	1,00
G	Component value (0 to 255)	0,50	0,10
B	Component value (0 to 255)	0,50	0,10
X	Coordinate value (0 to 255)	1,00	1,00
Y	Coordinate value (0 to 255)	0,50	0,10
RmB	Value difference (-255 to 255)	1,00	1,00
C	Distance in pixels (0 to 181)	1,00	1,00
LR_{Diff}	Value difference (-255 to 255)	-	-

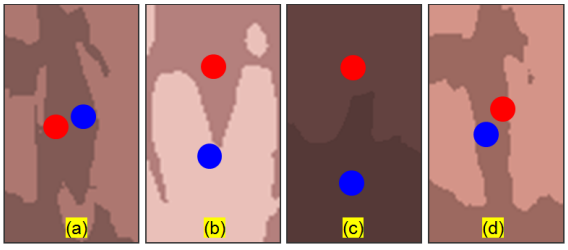
Four images with localized ROI (Region of interest) area of the vocal cords were chosen as an example for illustrating the method application, Fig. 3.

Figure 3. Images with localized vocal cords.



The visualization of the results shows the image after classification, Fig. 4, where the area of points belonging to one class is represented by the average color of the points belonging to the class. The highlighted color point is the center point of the class, located at the position of average coordinates of all points in the class.

Figure 4. Classification result after the test with 2 classes, $j = 2$, using weights $coeff_0$.



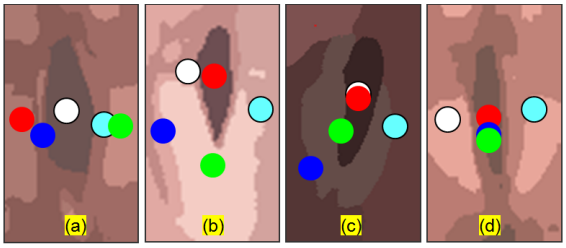
In further tests, the number of classes was increased to j = 3, 4, 5 and various parameter weight

settings were used. In the end the weight settings have stabilized at $coeff_1$ from TABLE I.

In the last test, changes in the weight of the parameters and the selected number of classes improved the result and the glottis was well detectable not only in all tested images, Fig. 5, but also in other images in our corpus.

The calculated centers of the most open vocal cords are then used for all frames in the video sequence to detect glottis in each frame.

Figure 5. Classification result after the after the test with 5 classes, $j = 5$, using weights $coeff_1$.



B. *Cluster analysis with adaptive parameter weight*

In the cluster analysis method with static parameters, each parameter weight has been experimentally set and these weights are used to calculate distances between individual objects (pixels in the image). This static setting is further applied to the other analyzed HSV sequences. However, since the characteristics of each recording may vary, a modified method has been proposed based on the adaptive weights for each video recording calculated from the input image data. For every weight it is necessary to define relation depending on the selected input parameter(s). The set of the input parameters (from the image) was set:

- X_{max} – ROI width of the input image.
- Y_{max} – ROI height of the input image.
- R_{max} – Maximum R value in the selected ROI.
- R_{min} – Minimum R value in the selected ROI.
- R_{sum} – Sum of red component values of all point in the ROI.

We assume that for videos with narrow ROI (width is much lower than height), the weight of the X parameter should be greater, while in the wider ROI region, the weight of the X parameter decreases. This should prevent poor detection at the right/left edges of the vocal cleft due to the large distance from the center. The same applies to the R parameter where images with higher contrast needs higher weight of R parameter than low contract images to avoid wrong classification of the most dark or light points in the glottis area,

In order to find these suitable relationships and their settings, a learning mechanism with a teacher was created, where, according to the knowledge of the correct result, several possible settings were tested, from which the best one was subjectively selected. Learning was done on several sample videos to achieve the most general settings. This procedure helped to find the best relationship for calculating parameter weights.

The condition for calculating the relationships is a simple calculation which should have complexity at most n (where n is the number of pixels in the ROI), e.g. to find the maximum and minimum of the color component. In this way we have reached the resulting relationships, (2), (3), (4), (5), (6), (7), and resulting for weights, TABLE II.

$$W_R = 2 * MIN\left(5, \frac{110}{R_{max} - R_{min}}\right) \quad (3)$$

$$W_{RmB} = \frac{W_R}{2} \quad (4)$$

$$W_X = \frac{120}{X_{max}} \quad (5)$$

$$W_C = \frac{W_R}{2} + W_X \quad (6)$$

$$W_{LRDiff} = \frac{R_{max} - R_{min}}{200} \quad (7)$$

The weights for parameters G, B and Y were decreased to zero for low effect in the result to speed up the segmentation process.

TABLE II. DESCRIPTION AND THE EXAMPLE OF ADAPTIVE WEIGHTS OF CLUSTER ANALYSIS PARAMETERS

Parameter description with the initial and result weight settings $coeff_0$ and $coeff_1$ for specific image			
Parameters	***Weight of parameter***	***$coeff_0$***	***$coeff_1$***
W_R	R	3,00	1,29
W_G	G	0,00	0,00
W_B	B	0,00	0,00
W_X	X	1,00	0,47
W_Y	Y	0,20	0,00
W_{RmB}	RmB	1,00	0,64
W_C	C	2,00	1,12
W_{LRDiff}	LDDiff	3,00	1,12

IV. RESULTS

The implemented method of *cluster analysis with static weight of parameters* increases the success rate of glottis detection when comparing the results with the method *Max-Min-Thresholding* [18], [19].

For a comprehensive testing of glottis segmentation methods, we used a data corpus containing 549 HSV recordings, TABLE III. For this comparative analysis, we selected 130 video sequences from this set of videos that were identified as problematic in terms of glottis detection (e.g. containing low contrast, blur, noise). These were records where the *Max-Min-Thresholding* failed or achieved different results compared to the *Cluster analysis with static parameter weights* method more often. For 419 out of 549 videos, both methods

achieved comparable results. The accuracy of glottis detection has always been assessed in cooperation with an otorhinolaryngologist.

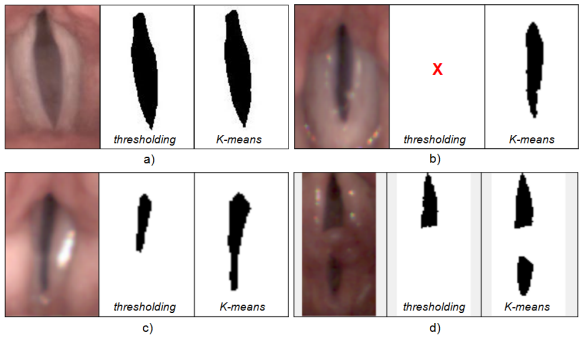
Glottis was detected correctly in 97 cases using the *Cluster analysis with static parameter weights* method, and using the *Max-Min-Thresholding* method in 42 cases. The result of the *Max-Min-Thresholding* method was more accurate than the *cluster analysis* method in only 4 cases of the glottis detection.

TABLE III. DATA CORPUS STRUCTURE USED FOR ALGORITHM TESTING OF GLOTTIS DETECTION

HSV data corpus	**video recordings**		
	total sum	***men***	***women***
diagnosis	549	190	359
cystis vocal	7	0	7
vocal polyp	31	13	18
chordectomy	5	0	5
papillom	13	13	0
vocal nodules	16	6	10
carcinoma	8	8	0
granuloma	2	2	0
Reinke's edema	18	1	17
reccurent laryngeal n. paresis	104	32	72
tonsillectomy	64	10	54
hemangioma	3	0	3
thyroid glan	23	2	21
healthy vocal cords	133	34	99
dg. is not determined	122	69	53

The success of glottis detection by *Max-Min-Thresholding* and *cluster analysis* methods for the purpose of this paper is presented in individual casuistries, Fig. 6. Casuistries contain an example of correct detection of glottis in healthy vocal cords, Fig. 6a, complete failure of the *Max-Min-Thresholding* method for vocal cords with nodule on the left side, Fig. 6b, and partial failure of the *Max-Min-Thresholding* method in case of left-sided paresis of the reversible nerve, Fig. 6c, and the polyp on the left side, Fig. 6d.

Figure 6. Comparation of glottis detection using *Thresholding method* and *Cluster analysis* (static weights) method on selected.



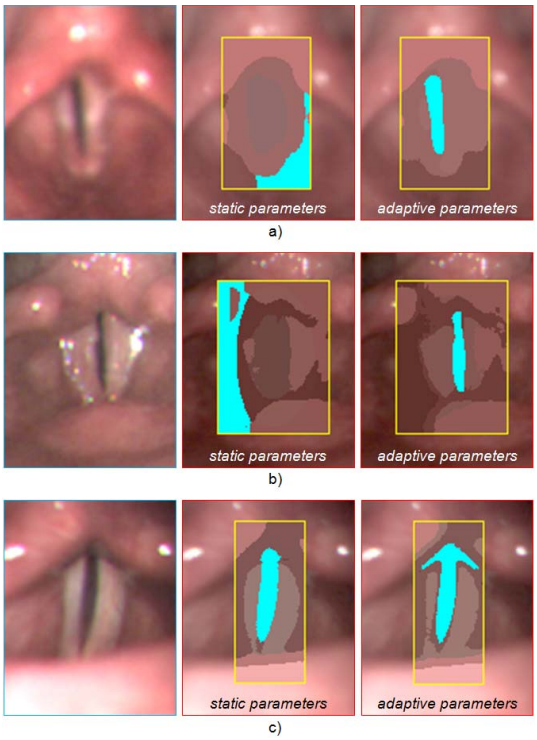
In all cases, the failure of the thresholding method occurs due to the brightness and color conditions in the images, not overlapping anatomical structures, camera or vocal cords movement, blur or angle of capture.

After comparison of the cluster analysis with static parameter weights, TABLE I, and adaptive parameter weights, TABLE II, there was slight improvement in success rate of correct glottis detection. From 100 random test cases, better result for adaptive weights was seen in 5 cases (recognition failed for static weights and was successful using adaptive weights, e.g. blur image on Fig. 7a or non-uniform lighting Fig. 7b), in 1 case the recognition was better using static weights, Fig. 7c.

The differences are caused by different image characteristics where the cluster analysis with adaptive parameter weights is able to react on input data and select more suitable settings.

The specific reasons why the methods works in one case and fail in another are very complex and are individual for each case. In the example on Fig. 7c, it probably failed because of too high weight for R component parameter and low weight for X parameter. It is probably impossible to tune the parameter weights to have 100% success rate.

Figure 7. Causistry of results using cluster analysis with static and adaptive parameter weights.



V. CONCLUSIONS

Experience with solving the problem of glottis segmentation in HSV video sequences shows that published methods of automatic thresholding, Watershed segmentation, Region growing, Active contours, Gabor filtering and others, have comparable results for a large class of video sequences with similar image characteristics.

The success of the methods differs in special complicated cases, where a very diverse range of methods has a significant influence on the success and accuracy of glottis detection.

Presented methods are used in own software for computing many parameters of detected glottis area and evaluating vocal cords movement for possible early diagnosis of potential issues during the examination on the ENT department.

The methods are tested on many HSV recordings from the vocal cods examinations. The data corpus is still growing and current number of usable HSV recordings is 692. All the results are evaluated to refine and tune the presented methods.

ACKNOWLEDGMENT

This work was supported by the Grant No. SGS-2019-018 Data and Software Engineering for Advanced Applications.

REFERENCES

[1] I. R. Titze, Principles of Voice Production, 2nd ed., National Center of Voice and Speech: Iowa City, IA, USA, 2000, ISBN: 0-87414-122-2, pp. 87-183.

[2] J. Švec, Studium mechanicko-akustických vlastností zdroje lidského hlasu. [Studies on the mechanic-acoustic properties of the human voice. Thesis. In Czech] Olomouc, Palacký University, Faculty of Natural Sciences, Department of Experimental Physics, 1996.

[3] F. H. Ng, "Automatic thresholding for defect detection," Pattern Recognition Letters, 27(14): pp. 1644-1649, 2006. https://doi.org/10.1016/j.patrec.2006.03.009.

[4] N. Otsu, "A threshold selection method from gray-level histograms," IEEE Transactions on Systems, Man, and Cybernetics, vol. 9, no. 1, pp. 62–66, 1979.

[5] J. Kittler and J. Illingworth, "Minimum error thresholding," Pattern Recognition, vol. 16, no. 1, pp. 41-47, 1986.

[6] A. Bleau and L. Leon, "Watershed-based segmentation and region merging," Computer Vision and Image Understanding, 77(3): pp. 317-370, 2000.

[7] V. Osma-Ruiz, J. I. Godino-Llorente, N. Sáenz-Lechón and R. Fraile, "Segmentation of the glottal space from laryngeal images using the watershed transform," Computerized Medical Imaging and Graphics, 32(3), pp. 193-201, 2008.

[8] G. Andrade-Miranda, J. I. Godino-Llorente, L. Moro-Velázquez and J. A. Gómez García, "An automatic method to detect and track the glottal gap from high speed videoendoscopic images," BioMedical Engineering OnLine, 14:100 DOI 10.1186/s12938-015-0096-3. 2015.

[9] B. Peng and L. Zhang, "Automatic Image Segmentation by Dynamic Region Merging," IEEE Transactions on Image Processing, vol. 12, no. 12, pp. 3592-3605, 2011.

[10] A. Méndez, E. M. Ismaili Alaoui, B. García, E. Ibn-Elhaj and J. Ruiz, "Glottal Space Segmentation from Motion Estimation and Gabor Filtering, 31st Annual International Conference of the IEEE EMBS, Minneapolis, USA, pp. 5756-5759, 2009.

[11] Ch. Palm, D. Keysers, T. Lehmann and K. Spitzer, "Gabor Filtering of Complex Hue/Saturation Images for Color Texture Classification," In Proceedings of 5th Joint Conference on Information Science (JCIS 2000), Atlantic City, USA, pp. 45-49, 2000.

[12] M. Kass, A. Witkin and D. Terzopoulos, "Snakes: Active contour models." In First International Conference on Computer Vision, pp. 259–268, London, 1987.

[13] S. Allin, J. Galeotti, G. Stetten and S. H. Dailey, "Enhanced Snake Based Segmentation of Vocal Folds," In 2nd IEEE International Symposium on Biomedical Imaging: Macro to Nano. IEEE; 2004.

[14] F. Schenk, M. Urschler, C. Aigner, I. Roesner, P. Aichinger and H. Bischof, "Automatic glottis segmentation from laryngeal high-speed videos using 3D active contours," *Proceedings of the 18th Conference on Medical Image Understanding and Analysis,* pp. 111-116, The British Machine Vision Association, 2014.

[15] P. Schlegel, M. Sammler, M. Kunduk, M. Döllinger, Ch. Bohr and A. Schützenberger, "Influence of Analyzed Sequence Length on Parameters in Laryngeal High-Speed Videoendoscopy," Applied Sciences, December 2018,vol. 8, issue 12, https://doi.org/10.3390/app8122666.

[16] M. Blanco, X. Chen and Y. Yan, "A Restricted, Adaptive Threshold Segmentation Approach for Processing High-Speed Image Sequences of the Glottis," Scientific Research, Engineering, 5, pp. 357-362, Published Online, October 2013. (http://www.scirp.org/journal/eng).

[17] J. Demeyer, T. Dubuisson, B. Gosselin and M. Remacle, "Glottis Segmentation with a High-Speed Glottography: a Fully Automatic Method," 3rd Advanced Voice Function Assessment International Workshop, Madrid, 2009.

[18] J. Pešta, J. Slípka, M. Vohlídková, T. Ettler, P. Nový and F. Vávra, "Vocal Cord Kinematics-New Evaluation Parameters, "Otorinolaryngologie a foniatrie, 65, c. 2, pp. 88-96, Praha, 2016.

[19] T. Ettler, "Detection and Evalution of Glottis in High Speed Video Recording," Professional work for the state doctoral exam; University of West Bohemia; Pilsen, 2017.

[20] S. Tatiraju and A. Mehta, "Image Segmentation using k-means clustering," EM and Normalized Cuts, Technical report, Department of EECS, 2008.

[21] J. B. MacQueen, "Some Methods for classification and Analysis of Multivariate Observations," Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, 1. University of California Press. pp. 281–297. MR 0214227. Zbl 0214.46201. Retrieved 2009-04-07. 1967.

[22] D. Aghlmandi and K. Faez, "Automatic Segmentation of Glottal Space from Video Images Based on Mathematical Morphology and the Hough Transform," International Journal of Electrical and Computer Engineering (IJECE), vol.2, no.2, pp. 223-230, 2012, ISSN: 2088-8708.

[23] F. Schenk, P. Aichinger, I. Roesner and M. Urschler, "Automatic high-speed video glottis segmentation using salient regions and 3D geodesic active contours," Annals of the BMVA, vol. 2015, no. 1, 2015, pp. 1–15.

[24] J. C. Russ, The Image Processing Handbook, Fourth Edition, CRC Press, 2002, ISBN 0-8493-2532-3.

[25] D. D. Mehta, D. D. Deliyski, T. F. Quatieri and R.E. Hillman, "Automated Measurement of Vocal Fold Vibratory Asymmetry From High-Speed Videoendoscopy Recordings," Journal of Speech, Language, and Hearing Research, vol. 54. pp. 47-54, 2011.