Engineering, **2013, 5, 357-362**
http://dx.doi.org/10.4236/eng.2013.510B072 Published Online October 2013 (http://www.scirp.org/journal/eng)

A Restricted, Adaptive Threshold Segmentation Approach for Processing High-Speed Image Sequences of the Glottis

Mathew Blanco, Xin Chen, Yuling Yan*
Department of Bioengineering, Santa Clara University, Santa Clara, USA
Email: *yyan1@scu.edu

Received May 2013

ABSTRACT

In this paper, we propose a restricted, adaptive threshold approach for the segmentation of images of the glottis acquired from high speed video-endoscopy (HSV). The approach involves first, identifying a region of interest (ROI) that encloses the vocal-fold motion extent for each image frame as estimated by the different image sequences. This procedure is then followed by threshold segmentation restricted within the identified ROI for each image frame of the original image sequences, or referred to as sub-image sequences. The threshold value is adapted for each sub-image frame and determined by respective minimum gray-scale value that typically corresponds to a spatial location within the glottis. The proposed approach is practical and highly efficient for segmenting a vast amount of image frames since simple threshold method is adapted. Results obtained from the segmentation of representative clinical image sequences are presented to verify the proposed method.

Keywords: Segmentation; Glottis; Vocal Fold Motion; Difference Image; Adaptive Threshold

1. Introduction

Laryngeal imaging based analysis of vocal fold motion has been proved valuable for both diagnosing voice disorders and understanding the mechanism of voice production. High speed digital imaging (HSDI), or high speed video-endoscopy (HSV), has now become a clinical reality for imaging the vibrating vocal folds. The HSDI systems record images of the vibrating vocal folds at a typical rate of 2000 frames/sec, which is fast enough to resolve a specific, sustained phonatory vocal fold vibration. In the literature [1-9], glottal area waveform (GAW), along with other spatiotemporal waveforms of the glottis, has been successfully used to analyze the vocal fold vibration which may correlate with voice condition. The credibility of the analysis strongly depends on an accurate extraction of the GAW from images of the glottis. In order to obtain the GAW, the glottis, or the vocal fold opening region, needs to be segmented and the area calculated on a frame by frame basis. Clearly, it is crucial for us to develop effective and highly efficient segmentation algorithms for this purpose.

Image segmentation is fundamental to the field of image understanding and computer vision [10-13] and to establish an efficient segmentation algorithm is still challenging because of lacking in a universal segmentation algorithm for all image segmentation tasks.

The purpose of image segmentation is to divide an image into regions that are meaningful to some higher level processes. In this research, the meaningful region is the glottis, the air space between the pair of vocal folds. In the literature some algorithms for glottis segmentation have been reported, which include region growing algorithm [5,14,15] and active contour algorithm [16-20]. However, there are some limitations in these approaches, making them impractical for applications in the analysis of HSV image data sets. The region growing algorithm depends much on selection of the seed point that requires prior knowledge about the location of glottis [10]. On the other hand the active contour algorithm is extremely time consuming and susceptible to noises [11].

In a clinical setting, the HSV system is capable of capturing images of the vibrating vocal folds at a rate of at least 2000 frames per second. During an examination, a patient is instructed to phonate a sustained vowel phonation with a typical recording time of 2 seconds. In other words, each HSV recording contains 4000 image frames that need to be processed for further analysis and interpretation of the vocal fold dynamic behaviors [4]. As a result, it is essential to develop effective and efficient methods to segment the glottis rapidly and accurately. Since the time duration for each HSV recording is short, it is reasonable to assume that tremors of the hand of the clinician and of subject's neck and head are negligible. Additionally, following assumptions should hold:

*Corresponding author.

- The illumination is constant during the recording,
- The camera position is fixed during the recording.

While the motion of the vocal folds causes changes in the gray level in some region, the gray level intensity within other (motionless) regions remains almost unchanged. In order to successfully segment the glottis by threshold method, it is necessary to achieve well behaved histogram distributions. Since the motionless region is not of interest, it should first be removed. For this purpose, motion cue is used to obtain a sub-image, in which the size is adaptive to the glottis opening/closure status. As a result, the size of each sub-image varies so as to only contain a minimal but complete region of interest. In this way, the original image data is greatly reduced to facilitate faster segmentation and thus the simplest threshold method can be more efficiently and successfully adapted to segment the glottis.

In this work, we propose a two-step segmentation scheme based on the vocal fold motion analysis and adaptive thresholding as detailed in the following Method section.

2. Method

In this paper, the adaptive thresholding segmentation approach is based on an evaluation of the motion using difference image at corresponding spatial locations in the image sequence that highlights the region enclosing the vocal-fold motion extent. In addition, the images are segmented by adaptive thresholding, which is obtained in a restricted region of the original image, or termed sub-image. The threshold value varies for each image and is determined based on the grayscale minimum pixel in the sub-images, which typically corresponds to a location within the glottis.

We designed the following scheme for the segmentation task as illustrated in **Figure 1**:

1) Manually select an image frame from a HSDI recording where the vocal fold opening region is the smallest, as the reference image (RI).

2) Obtain the binary difference image (DI) based on the RI.

3) Use the median filter to eliminate the isolated points labeled one in the DI.

4) Obtain the sub-image which has a variable size for each image frame based on the DI.

5) Select the threshold value based on the lowest pixel value in each sub image frame and segment the sub-image.

2.1. Introduction to Image Segmentation and Motion Analysis

As illustrated in **Figure 2**, each image from a laryngeal image recording should be segmented into two regions: the vocal fold opening region (glottis), which is the object, and the remaining region, which is considered as the background. In general, the image segmentation techniques can be categorized into three classes [11]: 1) characteristic feature thresholding or clustering; 2) edge detection; and 3) region exaction. Among them, thresholding method is the simplest and most efficient.

Thresholding is the transformation of an input image $f(i, j)$ (a gray level image) to an output (segmented) image $g(i, j)$ (binary image),

$$g(i,j)=\begin{cases}1 & for \quad f(i,j)\geq T\\ 0 & for \quad f(i,j)<T\end{cases} \tag{1}$$

where T is the threshold value, $g(i, j)=1$ for image elements of objects; and $g(i, j)=0$ for image elements of the background (or vice versa). From Equation (1), it is clear that correct threshold selection is crucial for successful segmentation.

Motion is a powerful cue used by humans and many animals to exact objects of interest from a background of irrelevant detail [21]. Their applications of the motion cue in segmentation can be in both spatial and frequency domains. In this work, we exploit the basic spatial techniques since our applications focus on motion analysis in the spatial domain.

2.2. Glottis Area Segmentation

The different image is typically obtained by motion analysis in the spatial domain as defined by a binary image:

$$d(i,j)=\begin{cases}0 & if \quad \left|f_1(i,j)-f_2(i,j)\right|\leq \varepsilon\\ 1 & otherwise\end{cases} \tag{2}$$

where, $d(i, j)=1$ represents image areas enclosing motion, while $d(i, j)=0$ represents image areas with no or

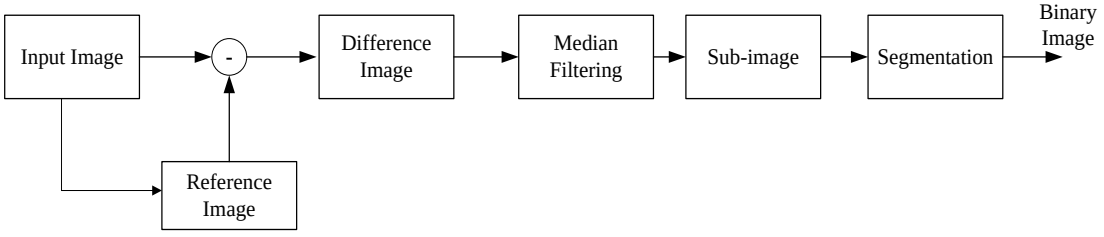
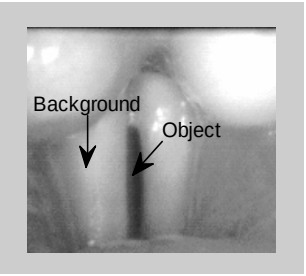
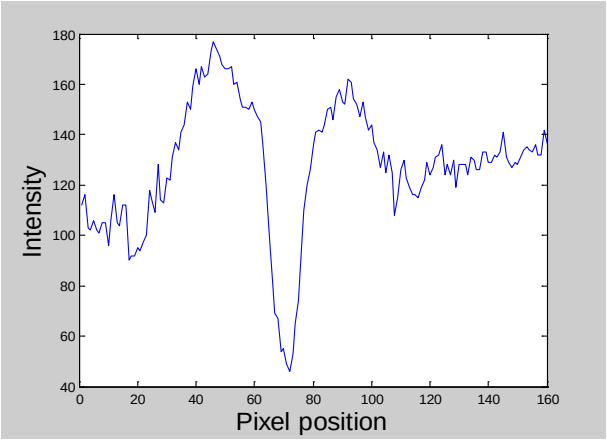


Figure 1. The scheme for the two-step segmentation.



(a)



(b)

Figure 2. (a) An image frame from the HSDI recording, and (b) the grey-level intensity profile along the mid-line of the vocal fold.

little motion. f_1 and f_2 are two consecutive gray level image frames within the original image sequences, and ε is a small positive number.

Here, we define the difference image (DI), a binary image, slightly differently as described below:

$$DI(x,y,t)=\begin{cases}1 \; if \; \left| f(x,y,t)-RI(x,y)\right| > T_1 \\ 0 \qquad otherwise\end{cases} \tag{3}$$

where T_1 is a positive constant. The optimal value of T_1 is determined based on experimenting with different datasets. The parameter *t* refers to the corresponding image frame at the recording time of *t*. Similarly, $DI(x,y,t)=1$ represents the vocal fold motion enclosure in and image frame at time *t*, and $DI(x,y,t)=0$ represents the background area within an image frame at time *t*. $RI(x,y)$ is the selected reference image frame that is used to compare with any input image. As mentioned earlier, an image frame having minimum glottis area is manually selected as the RI.

In each frame of the DI sequences, there might be pixels that are far from the glottis, mislabeled as '1'. The main reasons for this mislabeling are as follows:

1) Illumination is not constant during the image recording;

2) Vocal folds are not rigid. As a result, some regions near the vocal folds undergo moderate motion as the vocal folds vibrate.

In order to accurately obtain the sub-image and ensure it encloses entire region of the glottis, we apply a median filter to the DI for noise removal.

Median filtering is a non-linear smoothing method that is widely used to reduce the blurring of the edges [10]. This smoothing technique has been shown effective in eliminating spike noises. The key operation in the median filtering involves replacing the brightness of an individual pixel in the image by the median of the brightness values at several pixels in its neighborhood. The use of the median value can therefore reduce the effect of individual noise spike and smooth the image.

In the sub-image sequences, each image frame ideally contains a minimal region representing entire enclosure of the vocal fold motion extent. After the median filtering operation, the binary DI sequences are constructed and based on which we can determine the ROI that will be used for subsequent restricted, adaptive threshold segmentation processes applied to the sub-image sequences.

Further, we propose to use a variable threshold value for segmenting each sub-image, since it is prior knowledge that the darkest pixel point with minimum gray level intensity should be within the glottis, and in principle all pixels within the glottis should have lower values compared to areas outside the glottis in the sub-image. We thus obtain the threshold value based on the grayscale minimum value.

The algorithm is designed as follows,

1) Find the grayscale minimum (L) of each sub-image frame,

2) Obtain the threshold value $T_2 = \mathrm{L} + \mathrm{c}_2$,

3) Repeat above steps frame by frame.

Where, c_2 is a constant, the determination of c_2 is described in the following section.

After segmenting the sub-image sequences using the respective threshold values, we will obtain a binary segmented image sequences.

2.3. Parameters Determination

In this work, we use Matlab as a platform to conduct all analyses. In the proposed segmentation method, we need to determine the following parameters:

1) Size of the median filter convolution mask, [*m*,*n*],

2) Threshold value T_1, and constant c_2.

Different parameters can lead to different segmentation results. The method used for determining these parameters is based on trial and error. The parameters used in following analyses are T_1 = 0.10, c_2 = 0.15, and [*m*,*n*] is selected as [4,4].

3. Discussion and Conclusion

3.1. Discussion

Among threshold selection methods from gray-level histograms, Otsu method is widely used in many applications [22]. It is a nonparametric and unsupervised method for automatic threshold selection and image segmentation. An optimal threshold is selected by the discriminate criterion, namely, so as to maximize the separability of the resultant classes in gray levels. **Figure 3** shows an example of using Otsu method to segment the glottis from two representative HSDI frames (upper row). The segmentation results are shown in the lower row of **Figure 3**. It is clearly visualized that our method generated better segmentation results than those from Otsu method as shown in the middle row of **Figure 3**.

In **Figure 4**, the selected ROI, or the sub-image area, is shown for three consecutive original image frames (#10, 11, and 12). The size for each sub-image is shown to vary with the extent of the vocal fold motion, and each sub-image region encloses the entire glottis area.

In **Figure 5**, the left column displays four original frames within the obtained DI sequences. The right column shows the same frames after median filtering where

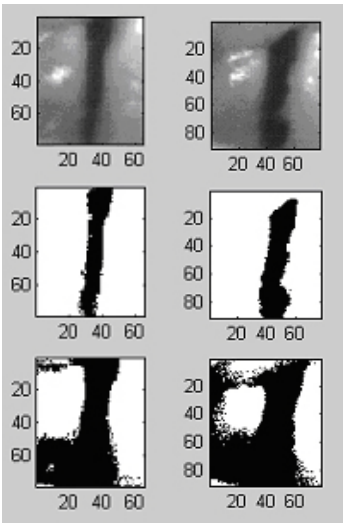


Figure 3. Comparison of the results of segmentation; the upper row shows two input images, the middle row shows the segmented images using our two-step approach, and the lower row shows the segmented images using Otsu method.

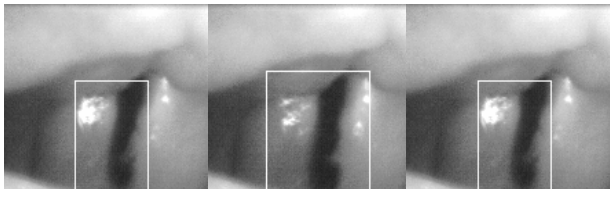


Figure 4. Sub-image frames showing the defined rectangular ROI.

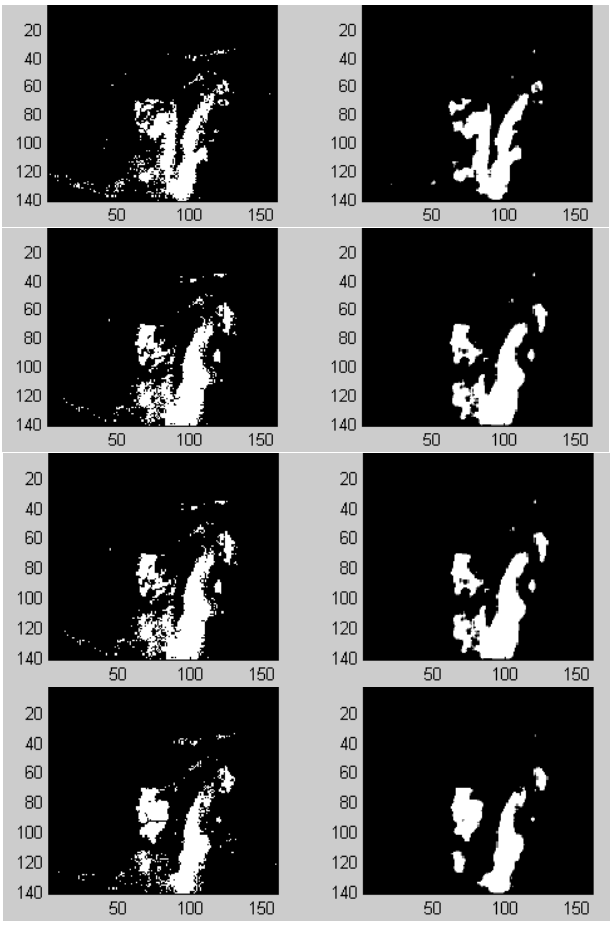


Figure 5. The left column shows four difference images; and the right column shows the results after applying a 4×4 median filter.

all pixels mislabeled "1" were effectively removed by the median filter. Finally, a series of segmentation results are shown in **Figure 6**, where both the sub-image region (rectangular ROI) and the accurately delineated glottis contour are outlined.

A comparison between the results of segmentation obtained from randomly selected three consecutive HSDI frames using Otsu and our method is shown in **Figure 7**. The top row shows the segmentation results obtained in the full image frame by Otsu method, and the lower row shows the results obtained from our method. It is clear that our first step to obtain the sub-image is critical for achieving robust and accurate segmentation results.

3.2. Conclusion

We developed a new approach for restricted, adaptive segmentation of images of the glottis that are acquired from the HSV system. By defining a sub-image set based on vocal fold motion cue, the subsequent threshold process is efficiently restricted to a ROI so that the effects of background are minimized, leading to a robust

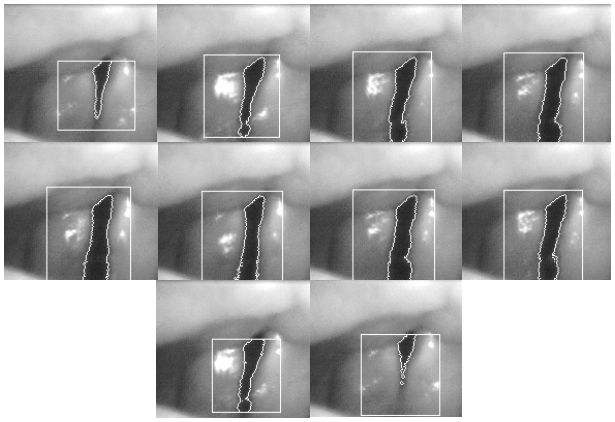


Figure 6. Serial segmentation results: the rectangle marks the defined ROI within which a restricted thresholding is performed to delineate the glottis (outlined).

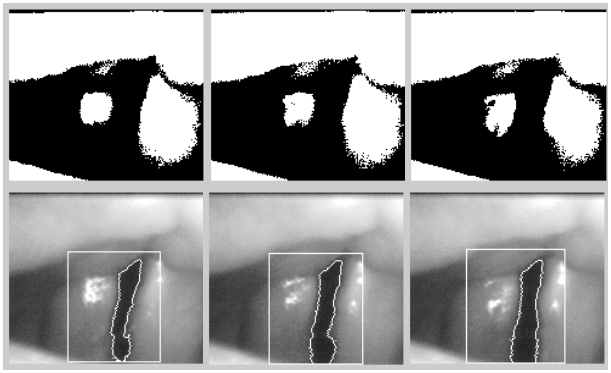


Figure 7. Results of segmentation from direct thresholding (top row) and from our algorithm (lower row).

and accurate segmentation outcome. From the segmentation results obtained from several clinical HSDI data sets using the proposed method, we can conclude that our method is effective and practical for applications in clinical settings.

REFERENCES

[1] R. Timke, H. von Leden and P. Moore, "Laryngeal Vibrations: Measurements of the Glottic Wave. Part I: The Normal Vibratory Cycle," *AMA Archives Otolaryngology*, Vol. 68, 1958, pp. 1-19. http://dx.doi.org/10.1001/archotol.1958.00730020005001

[2] J. Booth and D. Childers, "Automated Analysis of Ultra High-Speed Laryngeal Films," *IEEE Transactions on Biomedical Engineering*, Vol. 26, 1979, pp. 185-192. http://dx.doi.org/10.1109/TBME.1979.326556

[3] J. Noordzij and P. Woo, "Glottal Area Waveform Analysis of Benign Vocal Fold Lesions before and after Surgery," *Annals of Otology, Rhinology, and Laryngology*, Vol. 109, 2000, pp. 441-446.

[4] Y. Yan, K. Ahmad, M. Kunduk and D. Bless, "Analysis of Vocal Fold Vibrations from High-Speed Laryngeal Images Using a Hilbert Transform-Based Methodology," *Journal of Voice*, Vol. 2, 2005, pp. 161-175. http://dx.doi.org/10.1016/j.jvoice.2004.04.006

[5] X. Chen, D. Bless and Y. Yan. "A Segmentation Scheme Based on Rayleigh Distribution Model for Extracting Glottal Waveform from High-speed Laryngeal Images," 27*th Annual International Conference of the Engineering in Medicine and Biology Society*, Shanghai, 17-18 January 2005, pp. 6269-6272.

[6] Y. Yan, D. Bless and X. Chen, "Biomedical Image Analysis in High-speed Laryngeal Imaging of Voice Production," 27*th Annual International Conference of the Engineering in Medicine and Biology Society*, Shanghai, 17-18 January 2005, pp. 7684-7687.

[7] K. Ahmad, Y. Yan and D. Bless, "Vocal-Fold Vibratory Characteristics in Normal Female Speakers from High-speed Digital Imaging," *Journal of Voice*, Vol. 26, No. 2, 2012, pp. 239-253. http://dx.doi.org/10.1016/j.jvoice.2011.02.001

[8] K. Ahmad, Y. Yan and D. Bless, "Vocal Fold Vibratory Characteristics of Healthy Geriatric Females—Analysis of High-Speed Digital Images," *Journal of Voice*, Vol. 26, No. 6, 2012, pp. 751-759. http://dx.doi.org/10.1016/j.jvoice.2011.12.002

[9] Y. Yan and K. Izdebski, "Integrated Spatio-Temporal Analysis of High-Speed Laryngeal Imaging and Abnormal Vocal Functions—Their Role and Applications in the Study of Normal and Abnormal Vocal Functions," In: G. Demenko, Ed., *Speech and Language Technology*, Poznan, 2012.

[10] M. Sonka, V. Hlavac and R. Boyle, "Image Processing, Analysis and Machine Vision," 3rd Edition, Thomson Books/Cole, Toronto, 2008, pp. 74-77.

[11] K. Fu and J. Mui, "A Survey on Image Segmentation," *Pattern Recognition*, Vol. 13, No.1, 1981, pp. 3-16. http://dx.doi.org/10.1016/0031-3203(81)90028-5

[12] M. Atkins and B. Mackiewich, "Fully Automatic Segmentation of the Brain in MRI," *IEEE Transactions on Medical Imaging*, Vol. 17, No. 1, 1998, pp. 98-107. http://dx.doi.org/10.1109/42.668699

[13] J. Duncan and N. Ayache, "Medical Image Analysis: Progress Over Two Decades and the Challenges Ahead," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, 2000, pp. 85-106.

[14] Y. Yan, X. Chen, and D. Bless, "Automatic Tracing of Vocal-Fold Motion from High-Speed Digital Images," *IEEE Transactions on Medical Imaging*, Vol. 53, No. 7, 2006, pp. 1394-1400. http://dx.doi.org/10.1109/TBME.2006.873751

[15] J. Lohscheller, H. Toy, F. Rosanowski, U. Eysholdt and M. Döllinger, "Clinically Evaluated Procedure for the Reconstruction of Vocal Fold Vibrations from Endoscopic Digital High-Speed Videos," *Medical Image Analysis*, Vol. 11, No. 4, 2007, pp. 400-413. http://dx.doi.org/10.1016/j.media.2007.04.005

[16] B. Marendic, N. Galatsanos and D. Bless, "A New Active Contour Algorithm for Tracking Vibrating Vocal Folds," *IEEE International Conference on Image Processing*, 2001, pp. 397-400.

[17] J. Lohscheller, M. Döllinger, M. Schuster, R. Schwarz, U. Eysholdt and U. Hoppe, "Quantitative Investigation of the Vibration Pattern of the Substitute Voice Generator," *IEEE Transactions on Biomedical Engineering*, Vol. 51, No. 8, 2004, pp. 1394-1400. http://dx.doi.org/10.1109/TBME.2004.827938

[18] Y. Yan, G. Du, C. Zhu and G. Marriott. "Snake Based Automatic Tracing of Vocal-fold Motion from High-Speed Digital Imaging," 2012 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Kyoto, 25-30 March 2012, pp. 593-596.

[19] S. Karakozoglou, N. Henrich, C. D'Alessandro and Y. Stylianou, "Automatic Glottal Segmentation Using Local-Based Active Contours and Application to Glottovibrography," *Speech Communication*, Vol. 54, No. 5, 2012, pp. 641-654. http://dx.doi.org/10.1016/j.specom.2011.07.010

[20] 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, 2012, pp. 192-201. http://dx.doi.org/10.1016/j.bspc.2011.02.007

[21] J. Rong, J. Coatrieux and R. Collorec, "Combining Motion Estimation and Segmentation in Digital Subtracted Angiograms Analysis," *IEEE Sixth Multidimensional SignalProcessing Workshop*, Piscataway, 1989.

[22] N. Otsu, "Threshold Selection Method from Gray-Level Histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 9, 1979, pp. 62-66. http://dx.doi.org/10.1109/TSMC.1979.4310076

