SNAKE BASED AUTOMATIC TRACING OF VOCAL-FOLD MOTION FROM HIGH-SPEED DIGITAL IMAGES

Yuling Yan[1], Gan Du[1], Chi Zhu[2] and Gerard Marriott[3]*

1. Santa Clara University, Santa Clara, CA, 95053, USA
2. Maebashi Institute of Technology, Gunma, Japan
3. University of California, Berkeley, CA, 94720, USA

ABSTRACT

High-speed digital imaging (HSDI) of the larynx provides important information on the vocal fold vibrations that are closely associated with voice condition. We present an active contour (snake)-based algorithm for the automatic delineation of the glottis within image sequences captured from an HSDI system. The algorithm has three steps: first, a rough segmentation is performed by global thresholding, and followed by the detection of an ellipse-shaped region that approximates the glottal geometry, secondly, parameters of the ellipse are estimated using the principal component analysis (PCA) method and thirdly, the snake-method is applied using the estimated ellipse as an initial contour. The performance of the proposed approach is demonstrated through the use of clinical samples of the HSDI recordings obtained from subjects having both normal and pathological voice conditions. Finally the proposed method is compared with existing snake-based methods in terms of efficiency and segmentation accuracy.

Index Terms—High-speed laryngeal imaging, vocal fold vibration, active contour, snake

1. INTRODUCTION

Characteristic analysis of the vocal fold vibration provides an important component for understanding the mechanism of phonation and for voice assessment. The vocal folds oscillate at a frequency of 100Hz-400Hz during normal phonation - the emerging HSDI systems that are capable of recording images of the glottis at a typical rate of 2000 frame/second is fast enough to resolve the actual vibrations of the vocal folds [1-3]. Our previous study showed that the area of the glottis extracted from laryngeal image sequences can provide valuable information on the properties of the vocal fold vibration that correlates with the voice quality and health condition [2, 3].

The glottal area waveform (GAW) is a plot of the area of the vocal fold opening, or glottis, as a function of time. The performance of the subsequent GAW analysis depends on the accuracy of the GAW extraction, which requires both effective and highly efficient method for image segmentation from the vast amount of HSDI data. In the past, researchers have attempted to develop algorithms for the vocal fold edge detection using stroboscopic or HSDI data [3-9]. Manual approaches for image segmentation are frequently employed to extract glottis on a frame-by-frame basis - this operation may achieve accurate results but at an expense of time. Automatic methods developed for the glottal edge detection include region-growing [3] and active-contour (snake) methods [6-9]. The region-growing algorithm however, depends on an accurate selection of a seed-pixel point, while the existing active-contour methods are sensitive to noise and moreover, the requirement for large iteration places a severe limit to their ability to process vast image frames within a reasonable time frame. Since active contour methods have been shown to perform well in many medical imaging processing applications by generating satisfactory object boundary in edge detection [10], we elected to direct our approach towards a modification and adaptation of the active contour-based algorithm for glottal edge detection. As we mentioned earlier, there are two critical practical issues that need to be addressed for the application of the active-contour method, one is the noise sensitivity, and the other is the expensive computational cost. To this end, the proposed approach will address these critical issues and provide a practical solution to meet the stated challenges.

*Direct all correspondence to this author: yyan1@scu.edu

2. METHODS

The selection of an initial curve in the active contour method can significantly affect the segmentation results - in most cases, the geometry of the glottis can be approximated using an ellipse and so we begin by obtaining a rough estimation of the initial curve through global threshold segmentation. Next, an ellipse-like region is detected and the ellipse parameters are estimated using the principal component analysis (PCA) method. The ellipse is then drawn as an initial curve. Finally the snake is applied to this process. The number of iterations is determined by the ellipse fitting error - the greater the fitting error, the longer the iteration time is required to achieve desired accuracy. For the processing of subsequent image frames, the region of interest is searched based on the current result, which helps to limit the influence of background noises. We will test the proposed method by using HSDI data captured from subjects with normal and pathological voice conditions.

2.1. Procedures for the HSDI recording

A Kay-Pentax (Lincoln, Park, NJ) high-speed imaging system was used to record the laryngeal images. This system acquires images at a rate of 2000 frames per second with a spatial resolution of 160 × 140 pixels. Several subjects having normal and abnormal voice productions were instructed to produce a sustained vowel /i/ phonation. The recording time for each HSDI record is usually within 2 seconds.

2.2. Global thresholding

Prior to the application of the snake method, we use the thresholding to convert initial grey-scale image into a binary image. Since the region within the glottis has low intensity, a fixed fraction of the maximum intensity in each frame is used as the threshold value for a rough segmentation. As illustrated in Fig. 1 (b), each image frame can be segmented into two regions: the glottis (object), and the remainder (background).

2.3. Segmentation with snake

The active contour model introduced by Kass [11] is an automatic method for image segmentation. The contour, or snake, is represented as a curve, which is guided by external forces that pull the snake towards certain features such as lines and defined edges. Assume the curve is defined as $v(s) = (x(s), y(s)), s \in [0,1]$, it will move through the spatial domain of the image to minimize the following energy function:

$$E = \int_s \frac{1}{2}(a|v_s|^2 + b|v_{ss}|^2) + E_{ext}(v(s))ds \qquad (1)$$

Where, v_s and v_{ss} represent the first and second derivatives of v with respect to s, and a is a measure of the snake's tension and b is a measure of the snake's rigidity. E_{ext} is the external energy acting on the snake that is determined by the image gradient. The minimization procedure is an iterative technique using sparse matrix methods and the minimization criterion is based on differential calculus [11]. After initializing a curve close to the object boundary, the snake starts deforming to fit the local minima so as to move toward the desired object boundary and finally settles on it.

The initial contour setting plays an important role in the active contour method. Here, we detect an accurate location for the first frame (see Fig. 1). Since the strong background with similar intensity to the object often appears in the lower half of the frame (anterior vocal fold), the projection (integral of the intensity profile) along vertical axis (Y-axis) within upper half (Fig. 1 (c)) and lower half (Fig. 1 (d)) of the frame are calculated and displayed respectively, this is followed by multiplication of the two profiles (Fig. 1 (e)) from which the peak point can be detected. To obtain the horizontal (left-right) opening of the vocal fold, each point from the peak point (on the projection curve) to the left and to the right side is examined respectively, and the first point from each side having a value of less than 1% of the peak value is found and taken as the left and right boundary point respectively. Similarly, the projection along X-axis (Fig. 1(f)) is obtained and the vertical (up-bottom) opening of the vocal fold can be obtained; in this case, the secondary peak (Fig. 1(f)) can be easily excluded since it is very close to the boundary of the image frame and therefore carries no useful information to the process. Based on the procedures described above, the ellipse curve can be drawn and it then serves as an initial contour (Fig. 1 (g)). In particular, the pixel with an X-coordinate and Y-coordinate respectively corresponding to the peak position in Fig. 1 (e) and (f) is taken as the center of the ellipse, while the length of the major axis ($axis\_major$) of the ellipse is determined by the vertical opening and the minor axis ($axis\_minor$) by the horizontal opening of the vocal fold.

For the processing of subsequent image frames, the ROI is selected based on the previous frame (as shown in Fig. 2). Assume an image size of *H* by *W* pixels, and that the left, right, up and bottom range of previous frame is described by *leftx, rightx, upy* and *bottomy* respectively, the ROI of current frame is then set as a rectangular window with its position defined by four corner coordinates:
[max(*leftx-r1*,1), max(*upy-r2*,1)], [max(*leftx-r1*,1), min(*bottom y+r2*,W)],[min(*rightx+r1*,H),max(*upy-r2*,1)],and [min(*rightx +r1*,H), min(*bottomy+r2*,W)], where *r1* and *r2* are selected based on the maximum range of the vocal fold movement.

Let $f(x,y)$ denote the binary image within the ROI, the barycentre coordinates (*mx, my*) of $f(x,y)$ is taken as the center of the initial ellipse and it can be obtained as follows:

$$mx = Int(\frac{1}{Nb}\sum_{x=1}^{M}\sum_{y=1}^{N} xf(x,y))$$
$$my = Int(\frac{1}{Nb}\sum_{x=1}^{M}\sum_{y=1}^{N} yf(x,y)) \qquad (2)$$

Where, *Nb* is the number of pixels whose value is equal to 1 within the ROI of the binary image, *M* and *N* are the width and length of the ROI respectively. Int(*x*) stands for the nearest integer of *x*. Next, PCA method is applied to $f(x,y)$ and the eigenvalues and eigenvectors are then obtained. The parameters $axis\_major$ and $axis\_minor$ of the ellipse can be estimated as:

$$axis\_major = 4 \times \sqrt{eval1}$$
$$axis\_minor = 4 \times \sqrt{eval2} \qquad (3)$$

Where, $eval1$ and $eval2$ are the maximum and second maximum eigenvalue respectively.

To determine the number of iterations, a relative fitting error, *fiterr,* is defined and computed as follows:

$$fiterr = \frac{|cover_{ellipse} - vocalfold_{rough}|}{vocalfold_{rough}} \times 100\% \qquad (4)$$

Where, $cover_{ellipse}$ is the area that initial ellipse covers, and $vocalfold_{rough}$ is the rough vocal fold area obtained via global threshold within the ROI.

The iteration number, *itn,* is determined adaptively:

$$itn = \begin{cases} 25 & if\ fiterr \le 0.11 \\ 75 & if\ 0.11 < fiterr \le 0.15 \\ 90 & if\ fiterr > 0.15 \end{cases} \qquad (5)$$

3. RESULTS

The proposed method was applied to HSDI recordings from both subjects with normal and pathological voice conditions. The global threshold value was set as 60% the maximum intensity value, and *r1* and *r2* were set equally to 10 (pixels). The segmentation results for both normal and pathological ceases are obtained and subsequently the GAW and the bilateral vocal fold displacements are obtained (Figs. 3&4).

To evaluate the performance of the proposed method, we applied the snake-based method described in [6] for comparisons. In this method, the initialization is performed by user interaction for the beginning frame, and for subsequent frames, the snake position found in the preceding frame is used as the initial position for current frame. The results of analysis are shown in Fig. 5. Evidently this method does not work well for the HSDI data, this is likely due to the fast vocal-fold dynamics that generates a difference between the neighboring frames that is too large to be traced by the snake.

Next we applied a second snake-based method introduced by [7] to the same image data for comparisons. In this method,

each image frame is processed independently. We showed that the method worked well only in some cases where little background is present and with a sufficiently large number of iterations. However, since this method does not take into account the shape and the position of the glottis, it is sensitive to background and often fails to achieve reasonable segmentation accuracy in the presence of background even with a large number of iteration (Fig. 6).

Finally we compared the efficiency of each method used in these comparative analyses. With less significant background, the method in [7] requires an average of 500 iterations or approximately 0.14 seconds to process one image frame. Similarly, the method in [6] required 175 iterations or 0.11 seconds to process one image frame. In contrast, our method requires at most 90 iterations, or approximately 0.02 seconds for the same task. Combined previous tracing results and the comparisons, evidently our method outperforms both existing snake-based methods in terms of segmentation accuracy and computational efficiency.

4. CONCLUSION

We introduced an improved snake-based method for effectively and efficiently tracing the glottis from image sequences generated by the HSDI modality. The advantages of the new method over the existing methods include; first, it utilizes information from the previous image frame to provide a rough estimate of the location of the object (glottis) thus to avoid adverse influence of the background. Second, the initial snake points are set by an ellipse curve with ellipse parameters estimated by the PCA method. This process allows one to determine an initial contour that approximates the actual glottis. Moreover, the computational efficiency of this approach is further improved by adaptive determination of iteration number based on the ellipse-fitting error.

In summary, the proposed approach provides significant improvement with regard to computational cost and enables highly efficient frame-by-frame processing of the vast amount of image data that is typical of the HSDI recordings.

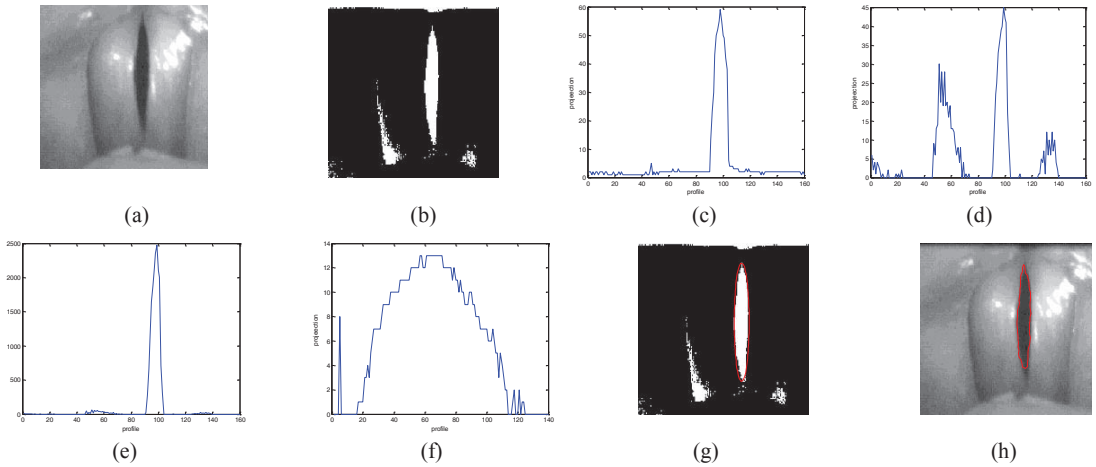


Fig.1. (a), original image frame; (b), binary image after thresholding; (c), y-axis projection within upper-half frame; (d), y-axis projection within lower-half frame; (e), multiplication of (c) & (d); (f), projection along x-axis; (g) & (h), initial and final glottis contours

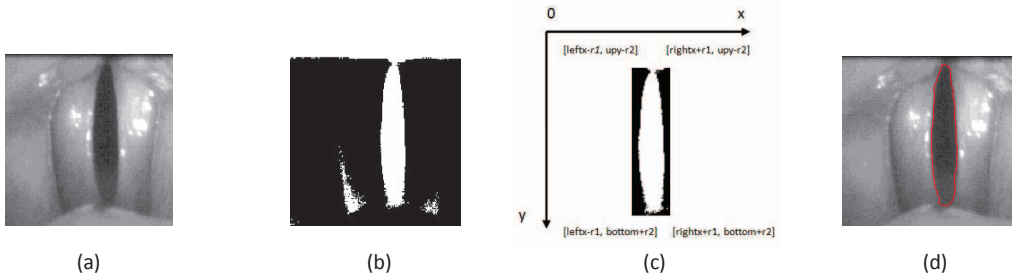


Fig. 2. Segmentation results and glottis delineation obtained from frame 2: (a) original image; (b) binary image after thresholding; (c) defined ROI ; (d) delineated glottis contour.

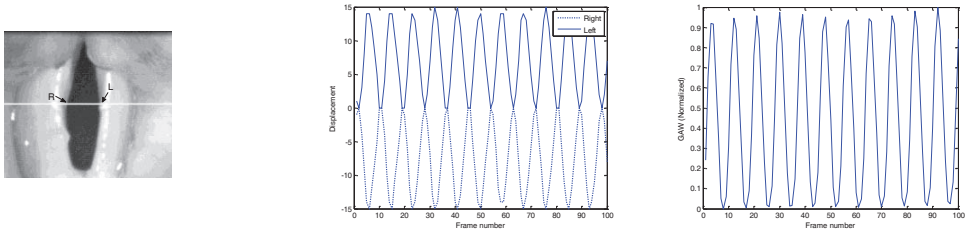
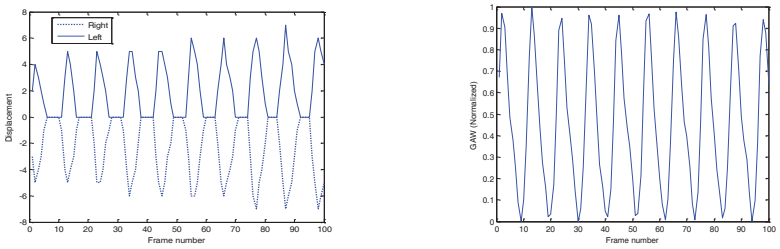


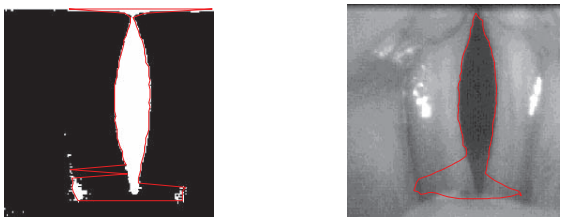
Fig. 3 (a) One image frame of the HSDI recording (from normal voicing) showing the location where the vocal fold displacements are calculated, (b) displacement of the left (solid line) and right vocal fold (dotted line) respectively over 11 vibratory cycles, (c) GAW plot



(a) (b)

Fig. 4 Results of analysis obtained from the pathological voice sample: (a) displacement of the left (solid line) and right (dotted line) vocal fold respectively over 10 vibratory cycles, (b) GAW plot over 10 vibratory cycles.

Fig. 5 Segmentation results from three selected image frames obtained using the snake-based method as described in ref.[6].



(a) (b)

Fig. 6 Segmentation results obtained from the same image (first frame in Fig. 5) using method described in ref. [7]: (a) Initial vocal fold contour; b) final glottis contour after 500 iterations

5. REFERENCES

[1] H. Larsson, S. Hertegard, P. A. L Indestad, B. Hammarberg, "Vocal fold vibrations: high-speed imaging, kymography and acoustic analysis," *Laryngoscope,* vol. 100, pp. 2117-2122, 2000.

[2] Y. Yan, K. Ahmad, M. Kunduk and D. Bless, "Analysis of vocal-fold vibrations from high-speed laryngeal images using Hilbert transform based methodology," *Journal of Voice*, vol. 19, no. 2, pp. 161-175, 2005.

[3] Y. Yan, X. Chen and D. bless, "Automatic tracing of vocal fold motion from high-speed digital images," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 7, pp. 1394-1399, 2006.

[4] Y. Zhang and J. J. Jiang, "Spatiotemporal chaos in excised larynx vibrations," *Phys Rev E*. vol. 72, pp. 35201-35204, 2005.

[5] C. Tao, Y. Zhang, J. J. Jiang, "Extracting physiologically relevant parameters of vocal folds from high-speed video image series," *IEEE Trans. Biomed. Eng.*, vol. 54, no. 5, pp. 794-801, 2007.

[6] A. K. Saadah, N. P. Galatsanos, D. Bless and C. A. Ramos, "Deformation analysis of the vocal folds form videostroboscopic image sequences of the larynx," *J. Acoust. Soc. Am*, vol. 103, pp. 3627-3639, 1998".

[7] C. Manfredi, L. Bocchi, S. Bianchi, N. Migali and G. Cantarella , "Objective vocal fold vibration assessment from videokymographic images," *Biomedical signal processing and control,* vol. 1, pp. 129-136, 2006.

[8] 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 Trans on Medical Imaging,* vol. 27, No. 3, pp300-309, 2008.

[9] D. D. Mehta, D. D. Deliyski, T. F. Quatieri, 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.

[10] A. Yezzi, P. Olver and A. Tannenbaum, " A geometric snake model for segmentation of medical imagery," *IEEE Trans. Med. Imaging,* vol. 16, no. 2, pp. 199-209, 1997.

[11] M. Kass, A. Witkin and D. Terzopoulos, "Snake:active contour models, " *Int. J. Comput. Vis.* vol. 1, no. 4, pp. 321-331, 1988.