INTERSPEECH 2013

Automatic Glottal Tracking From High-Speed Digital Images Using a Continuous Normalized Cross Correlation

Gustavo Andrade-Miranda[1], *Juan Ignacio Godino-Llorente*[1]

[1]Circuits and Systems Engineering Department, Universidad Politécnica de Madrid, Ctra. de Valencia, km. 7, 28031 Madrid, Spain

gxandrade@ics.upm.es, igodino@ics.upm.es

Abstract

The present work describes a new approach for the automatic tracking of the glottal area from high-speed digital images of the larynx. This approach involves three processes: Firstly, the frame with the maximal glottal opening is found automatically, by detecting the frame for which the sum of pixel intensities is minimum. Secondly, a segmentation algorithm based on active contours is used to detect the glottal space and build an initial template. Finally, using the normalized cross correlation the surface that represents the best matching between the initial template and the next frame is obtained. The surface area corresponds, at the same time, to the glottal space and to the new template. The process is repeated for each frame in order to obtain the new template and the new best cross correlation surface. This process is done iteratively until the last frame of the sequence has been reached. The performance, effectiveness and validation of the approach is demonstrated even in high-speed recordings in which the images present an inappropriate closure of the vocal folds.

Index Terms: glottal gap, keyframe, normalized cross correlation, HSDI, matching, template

1. Introduction

Biomedical images play an important role for a precise, fast and reliable diagnosis of the vocal folds vibration. In this sense, laryngeal images provide visual cues about the vibratory patterns that commonly the acoustic measurement cannot provide. These images are recorded using videoendoscopic techniques. There are two basic videoendoscopic procedures that are used to capture the vibratory movement of the vocal folds [1]: slow motion stroboscopy (SMS) and high speed digital imaging (HSDI). The HSDI systems record images of the larynx at a typical rate of 2000 frames/second, while the rate obtained in slow motion is only around of 25 or 50 frames per second. HDSI illuminates using a continuous light whereas SMS uses a stroboscopic lamp to show the movement of the vocal folds. A clear advantage of the HDSI with respect to SMS is that the stills are not fuzzy and incorrectly illuminated. However, both methods present camera rotations, side movements of the laryngoscope, and movements of the patient, causing a delocalization of the vocal folds and the glottal gap that complicates the application of automatic image processing techniques.

On the other hand, an accurate detection of the glottal gap and its tracking along time is required to objectively characterize the vibratory patterns of the vocal folds. This is usually carried out synthesizing different representations such as Glottal Vibration Profiles (GVP) [2], Glottal Area Waveforms (GAW) [3], Kymograms[4], and extracting some important measurements such as: the ratio of vibratory amplitude, ratio of periods of vibration, etc [5]. It is known that these parameters are correlated with voice quality and health condition [6], and help the specialist to evaluate the phonation in an objective way.

Currently, the previous task of identifying the glottal gap is usually carried out using semi automatic methods. In this context, and with the exponential growth of computer power and the constant improvement of the algorithms used for image processing, the hard task of automatically segmenting the glottal space has achieved a dramatic advancement. However, many of the techniques found in the literature still have weaknesses that make them impractical in a clinical environment, in which the automatization and reliability are fundamental. The most common techniques reported in the literature to detect the glottal space are based on histogram [7], region growing methods [8], watershed [9] and active contour delineation methods[10].

Histogram based algorithms are simple thresholding methods in which the threshold is determined based on the histogram of each image. The glottis has darker intensity levels than the vocal fold tissues, so a threshold can be used to differentiate between the glottis and the vocal folds. However, the laryngeal images often have low contrast and with heterogeneous profiles. Hence, selecting a global threshold may result in an erroneous delimitation of the glottal gap during the closure phase of the vocal folds, since, at this moment, the glottis and the background intensity are practically the same.

The region growing method checks the neighbour pixels of an initial group of points called seeds, and determines if the neighbours of the seeds should be added to the region. One of the problems of this method is that it requires a solid criterion and edges relatively well delimited in order to converge towards the glottal space; another drawback occurs when the vocal folds are almost closed; in these cases the glottis is sometimes divided in two or three parts, and the algorithm fails rejecting some of these segments (i.e. they are not considered belonging to the glottal gap).

The watershed simulates a rain over the image where each pixel represent an altitude as a function of its grey level. The drops that fall over a point will flow along the path of steepest descent until reaching a minimum. Such a point is labelled as belonging to the reception basin associated with this minimum. The result of the watershed produces thousand of catchment basis (oversegmentation), for that reason a pre-processing step is required to reduce the noise. However, in most of the laryngeal images the watershed does not eliminate all the objects that do not belong to the glottis [9], for that reason a post-processing step is required, causing loss of glottal information (i.e when the glottis is divided in two or three parts).

Lastly, the active contour algorithms or snakes are popular in medical images since they can be coupled appropriately to non-rigid and amorphous contours. Despite all the benefits provided by the active contours their effectiveness in processing thousands of high speed images is limited due to high computation times. In addition, during the glottal closure phase the snake has a high probability to converge to an erroneous point.

The main problem with the aforementioned algorithms is that they do not take into account the temporal dimension of the problem, so each frame is treated individually leaving aside the information obtained from the previous frames, so new techniques referring to the object tracking problem should be explored.

The literature is plenty of algorithms to track objects along time. They can be divided in two groups: target representation and localization, and filtering and data association. The first is mostly a bottom-up process and typically the computational complexity for these algorithms is low. In this group we find algorithms such as: blob tracking, kernel-based tracking, contour tracking, visual feature matching [11]. The second one is mostly a top-down process and commonly uses estimation or filtering as a core component for instead, the Kalman Filter, extended Kalman Filter, unscented Kalman Filter, point mass filter and particle filter. The aforementioned methods are combined with Bayes theorem for tracking a single target that is moving at approximately constant speed (non maneuvering). To track a maneuvering object They use the generalized pseudo-Bayesian Filter, interacting multiple model filter and many others [12].

Therefore, new methods are needed to accurately and efficiently overcome the limitations exposed, paying special attention to the tracking of the vocal folds along time. This paper is not intended to completely solve the problem of tracking the glottal space, but rather provide a new tool that could be used to refine existing techniques or to initialize more elaborated segmentation algorithms, for instance, active contours or watershed. The method proposed uses for the first glottis detection, the temporal information computed from a sequence of frames to reduce the number of false detections (Keyframes), and iteratively updated the information in each frame based on an adaptive cross correlation. Among the advantages of the method are: high degree of adaptability to existing techniques, easy implementation, and the possibility to distinguish the vocal folds from the glottis.

The rest of the work is organized as follows: Section 2, briefly introduces some basic concepts. Section 3 develops the methodology implemented for the glottis tracing. Section 4 evaluates the results obtained using the new approach and finally in section 5 presents some conclusions.

2. Correlation

The use of the correlation for object identification is an idea that goes back to the beginning of image processing and computer vision. It computes the similarity among an image $I(x, y)$ and a given template $T(x, y)$ [13]. Correlation is translation-invariant but requires that both template and image have the same scale and orientation. In laryngeal sequences the scale variation can be understood as the change of the vocal folds opening during phonation. Many approaches have been suggested to solve the scale and orientation changes. Some of them perform the correlation after transforming the image and the template into a domain with different invariances [14] [15]. However, the use of transformations increase the computational load, rendering it infeasible in real-time implementations or when the amount of data to be processed is significant.

The highest peak in the standard correlation (SC) surface indicates the location of the top-left corner of the best match in the search image. The main problem with SC is that it is highly sensitive to illumination conditions, since the highest peak always appears at the brightest spot in the image. There are several variations from the SC, some of them developed in the transformed domain. One of the most robust approaches is the phase correlation (PC). The phase correlation normalizes to unit magnitude the transform coefficient before computing the correlation in the frequency domain. Thus, the correlation is based only on the phase information and is insensitive to changes in the intensity of the image. This method has an interesting property: it attenuates all the other elements in the correlation surface that do not match with the best position, but at the cost of being more sensitive to noise than SC. Furthermore, the value of the peak is highly dependent on the scene content. Therefore, it is very difficult to set a single threshold, which is needed to compare the peak value for template updating and other later stages of the tracking algorithm.

In order to handle the limitations of SC and PC, researchers have proposed the normalized correlation (NC). This correlation has two salient features: it is less sensitive to varying illumination conditions than SC, and its values are normalized within the range $[0, 1]$. However, its counterpart in the frequency domain does not exist, so it is computationally more intensive than SC or PC.

Finally, the literature reports the normalized correlation coefficient (NCC), which is the most common correlation used for object localization. It is more stable to differences in the illumination, and its values are normalized within the range $[-1, 1]$. It can be expressed as:

$$c(m,n) = \frac{\sum_{x,y}[I(x,y) - \bar{I}_{m,n}][T(x-m,y-n) - \bar{T}]}{\sqrt{\sum_{x,y}[I(x,y) - \bar{I}_{m,n}]^2 \sum_{x,y}[T(x-m,y-n) - \bar{T}]^2}} \quad (1)$$

where $\bar{I}_{m,n}$ and $\bar{T}$ are the mean intensity values of the image and the template, respectively. However the NCC has two disadvantages. Firstly, it requires non constant intensity values of I or T; otherwise, the correlation value will be indeterminate. Secondly, its implementation in the spatial-domain is computationally intensive. However, there is an efficient method to compute it using FFT and the concept of summed-area table (SAT) [16]. Despite the age of the NCC algorithm and the existence of more recent techniques that address its various shortcomings, it is probably fair to say that a suitable replacement has not been universally recognized. One of the reasons is that NCC makes few requirements on the image sequence. Additionally it can be used to provide simple feature tracking or it can be easily combined with more elaborated techniques.

For this work the NCC has been selected due to its easy implementation and because it provides valuable information about the glottis and vocal folds contour. To overcome the scaling limitation due to the opening and closing of the vocal folds, the hypothesis is that a good approach would be an adaptive strategy updating the template after processing each frame. For the cases in which the NCC produce an indeterminate correlation, a history of the previous templates is used.

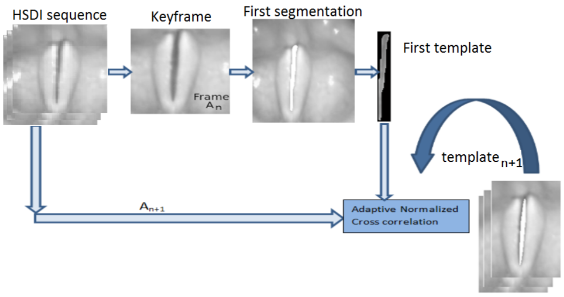

Figure 1: *Outline of the proposed algorithm.*

3. Method

The scheme of the proposed method is presented in Figure 1. The function of each block is detailed next:

3.1. Keyframe detection and first frame segmentation

In the first step, the algorithm finds useful information about the glottal cycles, using the frames of the sequence with the maximal glottal opening. These frames can be detected as the ones for which the sum of pixel intensities is minimum, since in those moments the glottis is open and the amount of dark pixels is maximum. Such frames are called keyframes. The keyframes represent the maximal open states of the glottal cycle within the sequence under consideration. The same approach has also been used in [17] for finding the frames with maximal glottal area.

$$keyframes(p) = \underset{i=1..k}{\operatorname{argmin}} \left(\sum_x \sum_y I_i(x, y) \right) \qquad (2)$$

The next step is choosing one of the p keyframes as initial point of the algorithm. In our case, to simplify the sequence traversed, the first frame is selected from the p keyframes. Finally, it is necessary to obtain the first glottis segmentation and propagate it through the video sequence. To achieve this, we use an automatic segmentation based on snakes using the algorithm in [18]. The figure 2 shows the first key frame and the initial template.

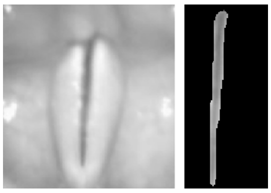

Figure 2: *First keyframe and first template.*

The templates are composed by superimposing the glottis area obtained in the previous frame with a black background. The black background act like an edge enhancer in order to highlight the glottis contour.

However, the region of interest covers only a part of the entire image. For localization and computational reasons, there is no need to process the entire image. So, a dynamic searching window can be defined, based only on the information extracted for the templates. The term dynamic is because the window is able to change its position according to the new position of the template.

This dynamic window is built take into account the information from the previous position of the glottis in x and y by creating an boundary box that enclose the result obtained. The dimensions of the searching window are variable and indicate the tolerance to abrupt movements; for instance, when the searching area increases, the motion tolerance and the false matches increase too. On the other hand, when the searching window decreases its area, the precision of the tracking decreases as well, and the probability to lose the target (glottis) increases. In our case the dimension was setting in order to obtain a complete visualization of the objects of interest.

A searching window reduces the number of peaks generated by the cross correlation, in this manner we have avoided the use of tedious transformations in the frequency domain used usually to eliminate the non representative peaks [19]. This ability increase the tolerance to the involuntary movements of the camera. Figure 3 shows a clear example of a frame with its respective searching window.

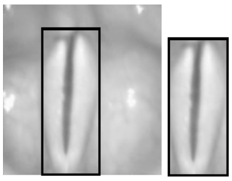

Figure 3: *Next frame with its respective search window.*

3.2. Adaptive Normalized Cross-Correlation

One way of making the method more robust to scaling and rotation is to use templates at different scales and orientations. An Adaptive Correlation (AC) is therefore our choice for the task of tracking the glottal space variation. The normalized cross correlation matrix is in the range of $[-1, 1]$. The values closer to 1 indicates more similarities between the template and the searching window, otherwise if the values are closer to -1 means that the template is uncorrelated with the searching window.

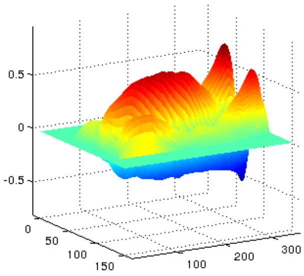

Figure 4: *Normalized Cross Correlation surface.*

The result of the correlation (Figure 4) is an enhanced version of the original image, in which the information concerning to the glottal space and the vocal folds are distinguished with clearness. The glottis in figure 5 is the red region surrounded by the vocal folds (dark blue). Therefore using only techniques based on thresholding or gradients, it is possible to assess them separately and get information such as the position, shape and area.

This information is updated iteratively by including the new template and the new searching window. This procedure is repeated until the last frame is reached. Additionally, a buffer

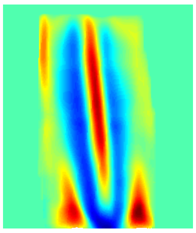

Figure 5: *Top view of Normalized Cross Correlation Surface.*

of the previous templates is required during each new iteration. The reason is that during the last cross correlation of the glottal cycle, the glottis passes from a partially open state to fully closed, or the opposite, so the likelihood of mismatching is high. At this point the new template is fixed to a zero matrix causing an indeterminate cross correlation. The buffer (previous frame) is then replaced with the zero matrix, and the procedure is repeated each time that the template only has zeros values.

4. Results

The methodology described in the previous section has been tested with HSDI sequences taken from the database recording by Dr. Erkki Bianco and Gilles Degottex. The resolution of the videos is 256x256 pixels and the sampling rate is 4000 frames/seconds. Figure 6 show three different frames and its respectively normalized cross correlation surface. In these images is easy to see that the use of a searching window eliminates a huge number of erroneous peaks. However, there are some small peaks that remain in the top and bottom corners of the cross correlation surface, but they can be easily eliminated using thresholding or a gradient criterion. The most significant peaks represent the vocal folds and the glottis. The correlation clearly shows the great differences between them: while one tends to be more positive (glottis), the other tends to be more negative (vocal folds); this behaviour remains constant for all the frames of the video sequence. Another important feature occurs in the second image, in which the glottis is partially closed. In this case the problem is not only referred to the variation of the glottis scale but also to a shape variation. The Adaptive Correlation fits small variations in shape between consecutive frames. Finally, the last frame shows one of the most demanding scenarios during glottis detection, in which the glottis is divided in two or more sections. In that scenario the proposed algorithm has a great performance being able to distinguish perfectly the regions that belong to the glottis from the vocal folds.

The algorithm presented was tested as a initialization for the algorithm in [20]. The final segmentation in each frame was compared with a semiautomatic segmentation [21] using the Pratt Index [22]. This algorithm calculates a figure of merit that measures the similarity between boundaries, where 1 indicates that the two edges are equal and 0 that there is not similarity. The Figure 7 summarizes the results obtained from 2 HSDI sequences in which the quality of the segmentation is analized using a 5-point scale directly linked with the Pratt Index.

5. Discussions and Conclusions

An adaptive correlation method has been developed. It selectively updates the correlation template in response to glottis changes in an HSDI laryngeal sequence. The template is updated continuously and a previous frame is used only when the matching is the zero matrix. The template update enables the correlation to track the glottis adjusting to the varying scale, while the dynamic searching window solve the problems introduced by abrupt changes of the camera position. Normalized cross correlation is a simple but effective method to be used as a similarity measure, being invariant to linear brightness and contrast variations. Its easy hardware implementation makes it useful for real-time applications.

The motivation of this paper is to explore techniques that even being traditional for video tracking have not been considered previously in the state of art for the detection and tracking of the glottal gap. The experimentation has shown that its use provides valuable information, not only to detect and track the glottal space but also the vocal folds.

The results obtained are very promising, even more if we consider combining them with other techniques; however this algorithm need to be tested in different conditions to ensure its generalization capabilities.

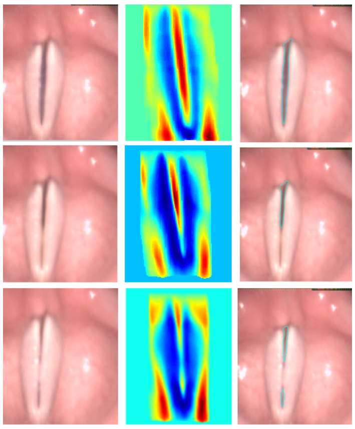

Figure 6: *Results; first column correspond to the original image, the second column is the result of apply the adaptive correlation and the third column is the final delineation of the glottis using [20]. Images courtesy of E.Bianco & G.Degottex-IRCAM.*

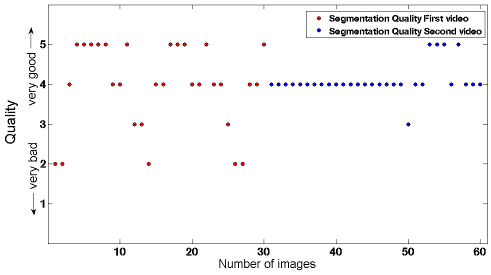

Figure 7: *Segmentation subjective assessment of 60 frames on a 5-point scale.*

6. Acknowledgements

This research was carried out under grants: TEC2009-14123-C04 and TEC2012-38630-C04-01 from the Spanish Ministry of Education.

The authors would also thank to Dr. Erkki Bianco and Gilles Degottex for the database provided.

7. References

[1] R. Baken and R. F. Orlikoff, *Clinical Measurement of Speech and Voice*, 2nd ed. United States of America: Singular Publishing Group, 2000.

[2] C. Palm, T. Lehmann, S. N.-R. Bredno, Klajman, and K. Spitzer, "Automated analysis of stroboscopic image sequences by vibration profile diagrams," in *5th International Workshop on Advances in Quantitative Laryngology, Voice and Speech Research*, 2001.

[3] P. Woo, "Quantification of videostrobolaryngoscopic findings-measurements of the normal glottal cycle," *The Laryngoscope*, vol. 106, no. S79, pp. 1–27, 1996.

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

[5] V. Osma, "Contribución al procesado digital de imágenes para la caracterización de patologías laríngeas," Ph.D. dissertation, Universidad Politécnica de Madrid, 2010.

[6] 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, pp. 239–253, 2012.

[7] 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," *Speech, Language and Hearing Research*, vol. 54, no. 1, pp. 47 – 54, 2011.

[8] J. Lohscheller, H. Toy, F. Rosanowski, U. Eysholdt, and M. Dollinger, "Clinically evaluated procedure for the reconstruction of vocal fold vibrations from endoscopic digital high-speed videos," *Medical Image Analysis*, vol. 11, no. 4, pp. 400 – 413, 2007.

[9] V. Osma-Ruíz, 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*, vol. 32, no. 3, pp. 193 – 201, 2008.

[10] B. Marendic, N. Galatsanos, and D. Bless, "New active contour algorithm for tracking vibrating vocal folds," in *Image Processing, 2001. Proceedings. 2001 International Conference on*, vol. 1, 2001, pp. 397–400.

[11] E. Maggio and A. Cavallaro, *Video Tracking: Theory and Practice*. Wiley, 2011.

[12] D. Subhash Challa, Mark R. Morelande and R. J. Evans, *Fundamentals of object tracking*. Cambridge, 2011.

[13] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. John Wiley & Sons Inc, 1973.

[14] R. A. Messner and H. H. Szu, "An image processing architecture for real time generation of scale and rotation invariant patterns," *Computer Vision, Graphics, and Image Processing*, vol. 31, no. 1, pp. 50 – 66, 1985.

[15] D. Casasent and D. Psaltis, "Position, rotation, and scale invariant optical correlation," *Appl. Opt.*, vol. 15, no. 7, pp. 1795–1799, Jul 1976.

[16] F. C. Crow, "Summed-area tables for texture mapping," in *SIGGRAPH '84: Proceedings of the 11th annual conference on Computer graphics and interactive techniques*. New York, NY, USA: ACM, 1984, pp. 207–212.

[17] S.-Z. Karakozoglou, N. Henrich, C. dAlessandro, and Y. Stylianou, "Automatic glottal segmentation using local-based active contours and application to glottovibrography," *Speech Communication*, vol. 54, no. 5, pp. 641 –654, 2012.

[18] G. Andrade, N. Sáenz-Lechón, V. Osma-Ruíz, and J. I. Godino-Llorente, "A new approach for the glottis segmentation using snakes," in *6th International Joint Conference on Biomedical Engineering Systems and Technologies*. INSTICC, 2013, p. 34.

[19] A. Mendez, E. Ismaili Alaoui, B. Garcia, E. Ibn-Elhaj, and I. Ruiz, "Glottal space segmentation from motion estimation and gabor filtering," in *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*, 2009, pp. 5756–5759.

[20] S. Lankton and A. Tannenbaum, "A localizing region-based active contours," *IEEE Trans. on Image Processing*, pp. 2029–2039, 2008.

[21] V. Vezhnevets and V. Konouchine, ""growcut" - interactive multi-label nd image segmentation by cellular automata," in *international conference on Computer Graphics and Vision*.

[22] I. E. Abdou and W. K. Pratt, "Quantitative design and evaluation of enhancement/thresholding edge detectors," *Proceedings of The IEEE*, vol. 67, pp. 753–763, 1979.