Segmentation of the glottal space from laryngeal images using the watershed transform

Víctor Osma-Ruiz , Juan I. Godino-Llorente, Nicolás Sáenz-Lechón, Rubén Fraile

Dpt. of Ingeniería de Circuitos y Sistemas, Universidad Politécnica de Madrid, Ctra. Valencia, Km. 7, 28031 Madrid, Spain

Abstract

The present work describes a new method for the automatic detection of the glottal space from laryngeal images obtained either with high speed or with conventional video cameras attached to a laryngoscope. The detection is based on the combination of several relevant techniques in the field of digital image processing. The image is segmented with a watershed transform followed by a region merging, while the final decision is taken using a simple linear predictor. This scheme has successfully segmented the glottal space in all the test images used.

The method presented can be considered a generalist approach for the segmentation of the glottal space because, in contrast with other methods found in literature, this approach does not need either initialization or finding strict environmental conditions extracted from the images to be processed. Therefore, the main advantage is that the user does not have to outline the region of interest with a mouse click. In any case, some a priori knowledge about the glottal space is needed, but this a priori knowledge can be considered weak compared to the environmental conditions fixed in former works.

Keywords: Image segmentation; Watershed; Region merging; Just noticeable difference; Kymography

1. Introduction

Nowadays, our current life style has increased the number of individuals affected by vocal fold pathologies, so it is becoming increasingly necessary to find accurate and quick methods for diagnosis in this field. The most widespread method used for the diagnosis and evaluation of these pathologies is the direct observation of the vocal folds with an endoscope either in rest or during vibration. These observation techniques have contributed to the development of new tools to characterize the movement and the vibration pattern of the vocal folds with the aim of helping the medical doctors in the diagnosis.

There exist two main approaches to evaluate and register the movement of the vocal folds and the mucosal waveform : high speed video recordings, and low speed recordings illuminated with a stroboscopic light (in the following, stroboscopic recordings or videostroboscopy).

The high speed video recording allows the acquisition of stills of the vocal folds during phonation with a frame rate over 2000 pictures per second illuminating with a continuous light source. The high speed video recoding ensures an accurate registering of the whole vibratory cycle, because the frequency of vibration of the vocal folds is between 50–150 Hz (for males) and 200–300 Hz (for females). The main drawback of this technique is the price of the equipment.

The videostroboscopy uses low speed recordings with not more than 25 or 50 frames per second. According to the Nyquist theorem , this frame rate is clearly insufficient to register the vibratory cycle in detail. A low cost solution is to undersample the recording using a stroboscopic light that produces beams 0.1 ms long with a frequency slightly lower than the fundamental frequency of the vocal folds movement. This technique simulates the vibratory cycle and allows the visualization of the vocal folds movement with a virtual frequency that is clearly lower than the true frequency of vibration.

Despite its advantages, the videostroboscopy presents several problems inherent to the technique itself: some of the stills taken could be fuzzy and incorrectly illuminated and besides, the stills could not be taken at the right instant .

On the other hand, there exist some artefacts which influence the quality of the recordings in both of the techniques mentioned, such as the rotation of the camera, the side movements of the laryngoscope, and the movements of the patient during the recording. These factors delocalize the position of the vocal folds and glottal space within the frame, and their effect is more pronounced in the stroboscopic recordings because in order to register a complete cycle of vibration a longer recording interval is required (several seconds long instead of the typical maximum of 1 s at high speed).

Another source of variability is given by the degree of illumination that depends on the equipment used. In this sense, it is possible to differentiate an inter-video variability (i.e., the illumination depends on the light source and the recording equipment), and an intra-video variability (i.e., each photogram has a different degree of illumination due to the stroboscopic effects).

In order to illustrate the difficulties addressed in this paper and the effect of the factors mentioned, Fig. 1 shows two characteristic images of the larynx obtained with stroboscopic illumination. The position and orientation of the glottal space is different depending on the frame. Moreover, the images can be fuzzy (Fig. 1b) and show regions with a poor illumination, especially in the corners of the images, and black elongated areas introduced by the recording equipment, as well as variations in the luminance of the glottal space.

Taking all of this into account, the detection of the glottal space (or glottis [1]) in laryngeal images is not always an easy task. Moreover, it is a fundamental operation that has to be addressed in advance in order to calculate a large amount of parameters that quantify the phonation process: the glottal area waveform , the ratio vibratory amplitudes, the ratio vibratory periods, the ratio opening and close phase , fundamental frequency, open quotient, closed quotient, speed quotient, time periodicity index, amplitude periodicity index, and phase symmetry index . It is furthermore a prior step for the segmentation of the vocal folds .

There are many research works in existing literature that address the problem of the automatic detection of the glottal space as an initial step to analyze the phonation process: kimograms , vibration profiles , glottal space area time-evolution diagrams .

Research carried out in this field made it necessary to develop new digital image processing techniques orientated towards the automatic segmentation of the glottis. The techniques used before range from those based on the classical and simplest image processing operations (thresholding, filtering, morphological operations, etc) , to the active contours (snakes) , the balloon models , and the region growing techniques . The main problem that these techniques have is that the segmentation process strongly depends on the starting point (initialization). Furthermore, they are very sensitive to noise.

Palm developed variations of the snakes based method to improve behaviour against noise, as well as to obtain some degree of independence with the initialization procedure. However, this work needs a strict stop signal while looking for the glottis as a dark object centred on the image, something which does not always occur in most stroboscopic images due to the previously mentioned problems with illumination and movements.

, the initialization point is searched for by means of an advanced thresholding technique based on the histogram of the images. This method provides very good results with high speed video recordings, but its performance is lower with stroboscopic images because using this kind of illumination makes it difficult to separate the glottis from other dark areas (with a similar grey level) by means of a thresholding.

In ref. [14], the initialization is carried out analyzing the differences between two consecutive photograms to detect the movement areas (called Motion Energy Images). As in ref. , it is not easy to extrapolate the results to videostroboscopic images where the movement not only depends on the vibration of the vocal folds but also depends on the patient and camera movements mentioned earlier.

Other recent segmentation approach, such as fuzzy connectedness needs to define seeds to split the regions of the image, introducing some kind of user dependence.

The new method proposed in this paper does not need any kind of initialization. The method is based on the perceptual characteristics of the human eye to distinguish between different grey levels. Thus, the only a priori knowledge that this

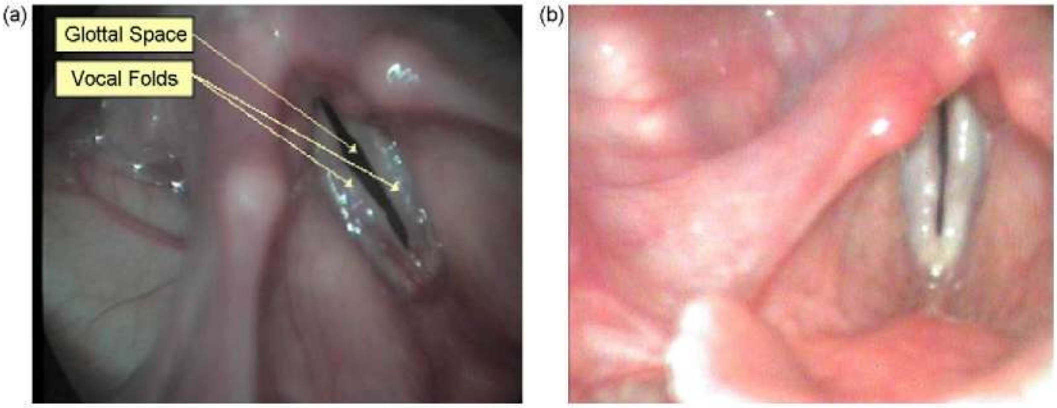

Fig. 1. Two different images of the larynx during phonation.

method integrates is easily applicable to any image of the larynx, because it is based on the characteristics that the glottis must always match in order to be recognized by a human observer. The technique used throughout this paper is called in existing literature Just Noticeable Difference (JND) . This technique integrates perceptual aspects, being a useful tool for the detection of the glottis in every image with independence of the means that were used to record them (videostroboscopy or high speed video).

The paper is organized as follows: Section 2 establishes the basic concepts and tools used in the process by means of a short review; Section 3 describes in detail the proposed method for the segmentation; Section 4 analyzes the results with a set of video images; and Section 5 enumerates the conclusions reached in view of the results.

2. Review of the digital image processing tools

2.1. Watershed transform

The watershed transform is one of the most valued tools in the field of digital image segmentation [19]. One of the main advantages of this technique lies in the fact that the result is a set of well delimited areas, so if we consider that these areas represent the searched objects, we will obtain an accurate edge detection defined by a set of connected pixels.

The concept of watersheds comes from the field of topography, referring to the division of a landscape in several basins or water catchment areas. From this point of view, we can consider the image as a topographic surface where each pixel is a point situated at some altitude as a function of its grey level .

The watershed transform simulates the rain over the surface associated with the image. 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. This process is repeated for all the points on the surface, so in the end every point will be assigned to a minimum and the surface will be divided into its catchment basins. An efficient implementation of this method is shown .

The goal is that each catchment basin matches an object in the image. Nevertheless, the result of the watershed transform is usually disappointing, due to the fact that thousands of catchment basins arise where only a few were expected . This problem is called oversegmentation and is due mainly to noise in the image.

A good solution to oversegmentation is to pre-process the initial image to reduce the noise. A widespread technique is the thresholding of the gradient image. Due to the fact that the gradient image has its maximum just over the edges of the objects present in the image, it is more logical to apply the transformation to the image gradient: before the watershed transform the gradient image is thresholded to remove the insignificant edges that appear due to noise . However, this pre-processing does not solve the problem completely, so a post-processing would be required to reach a better solution.

2.2. JND based region merging

The best solution to solve oversegmentation is to post-process the resulting image after the watershed transform to merge the catchment basins, following various criteria as described in existing literature . In general terms, all these methods are based on a continuous iteration over the watershed transform. Each iteration calculates the neighbour catchment basins that might be joined with a lower cost and merges them. The process ends either when there are as many catchment basins as wished (ideally one per object), or when the merging function cost exceeds an established threshold. The differences of the methods that might be used lies on the definition of the merging cost function. In this paper, the region merging that has been used is theoretically defined , where the merging cost function is calculated according the JND of the different grey levels of the image. The JND represents the sensibility of the human visual system to the changes of luminance, because it is well known that it is not able to differentiate certain changes of luminance.

, an expression (Eq. (1)) of the visibility threshold T is given (i.e., the threshold below which the eye is not able to detect the changes of luminance) as a function of the different grey levels of the image I (defining 0 like black and 255 like white). Fig. 2 shows graphically the relationship established in the Eq. (1). In view of the plot, a great insensibility of the human vision system can be observed against the changes of the grey level in the dark regions.

$$T(x,y) = \begin{cases} 17\cdot\left(1-\sqrt{\dfrac{I(x,y)}{127}}\right)+3 & \text{if } I(x,y)\le 127 \\[2ex] \dfrac{3}{128}\cdot(I(x,y)-127)+3 & \text{Otherwise} \end{cases} \tag{1}$$

2.3. Object parameterization: region invariant moments

The region moments are a set of parameters that enable the objects of an image to be described based on their characteristics (shape, texture, homogeneity, etc). There exist several methods to calculate the moments of an object. The ones used in this work have a statistical ground.

The moments integrated in this work are the classical invariant described by Gonzalez and Woods in ref. [28]. The binary invariant moments, that have been adapted from them, follow the

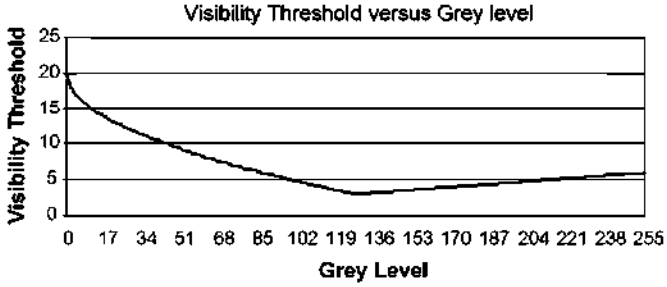

Fig. 2. Visibility threshold of the human visual system as a function of the grey level in the image.

same expressions, but consider that all pixels of the object have value 1. For this reason they become descriptors exclusively of the form of the object.

2.4. Linear prediction

Linear prediction is a simple technique of classification that recognizes and assigns new elements within previously defined classes by means of a supervised training .

Each of the target elements to be classified is characterized by an Nth order vector of parameters (e.g., the 7 invariant moments and the 7 binary invariant moments) that are selected with proven discriminative capabilities, so that they allow the different classes in study to be distinguished. A discriminant mathematical model is generated from a set of parameters belonging to known target elements. Thus, every unknown element is contrasted with the model and assigned to the most likely class as a function of the score given by the model.

An example of the linear prediction model is the Fisher discriminant analysis that classifies elements into two classes by means of a linear function (Eq. (2)):

$$D = \sum_{i=0}^{N} u_i X_i \tag{2}$$

where X_i represents the different parameters of the Nth dimensional feature vector, whereas u_i are the coefficients calculated by means of supervised training and defines the discriminant function that better separates the elements of the target class. The Fisher linear function is a projection of the feature space into a new subspace of dimension one, searching for the best separability. Thus, the classification is carried out by means of a linear combination of the feature space that provides a single score, and further establishing an optimum decision threshold over such score , that will be called the classification threshold.

Fig. 3 shows an example of the histogram given by the Fisher discriminant function in a two-class problem. The abscissa axis represents the scores given by the Fisher discriminant for all the elements under study, and the ordinate axis represents the number of cases that fall into each small interval of scores.

By referring to the histograms of Fig. 3 it can be established which is the best classification threshold that separates both classes, that do not necessarily have to be the average of the centroids of both classes. If the classification threshold is right shifted to the average of the centroids, the decision is more restrictive to classify elements belonging to class 1 (so, many class 1 elements will be labelled as class 0), with the advantage that only a few class 0 elements will be assigned to class 1 (target class).

Fixing the position of the classification threshold depends on the decision cost that can be assumed.

In any case, the error rate for each individual class (E_0 and E_1) and the total error (E_T) can be calculated as shown in Eq. (3), where N_i represents the number of elements belonging to class i, and $N_{i\to j}$ is the number of elements that belongs to class i but are assigned to class j.

$$E_0 = \frac{N_{0\to 1}}{N_0} \quad E_1 = \frac{N_{1\to 0}}{N_1} \quad E_\mathrm{T} = \frac{N_{0\to 1} + N_{1\to 0}}{N_0 + N_1} \tag{3}$$

For the classification task, there exist other techniques that are more powerful than linear prediction, such as Artificial Neural Networks (ANN), Support Vector Machines (SVM), Gaussian Mixture Models (GMM), etc . However, these techniques have not been considered in this work because the number of training patterns needed to get an accurate model exceeds the available images.

3. Methodology

The method that has been followed for the segmentation of the glottal space is graphically depicted in Fig. 4.

3.1. Watershed transform of the gradient image

The first step is to convert the original image (RGB) into a grey scale image by means of a transformation to the YIQ model

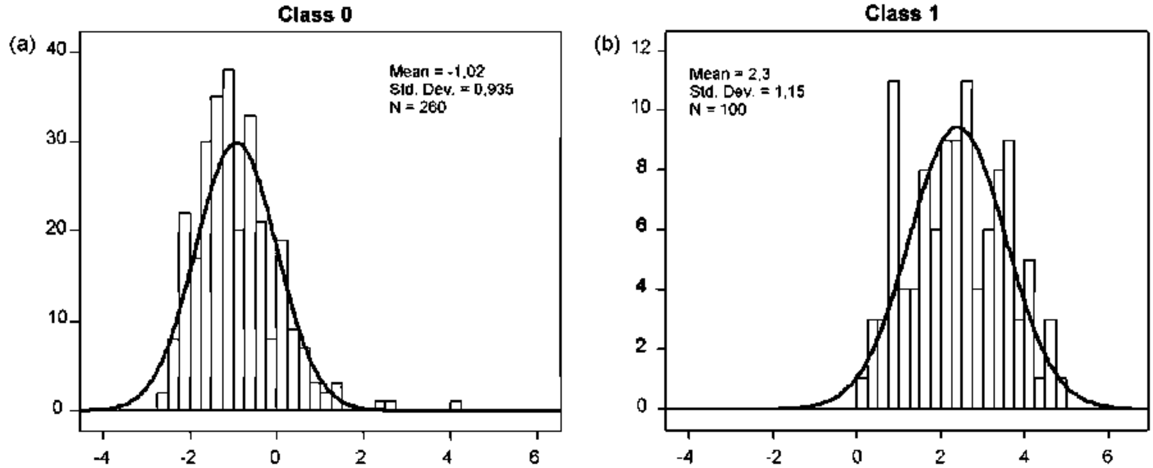

Fig. 3. Example of the histograms of the scores given by the Fisher discriminant for a two-class problem. (a) Scores for the element belonging to class 0 (non-target class); (b) scores belonging to class 1 (target class). Note that both histograms overlap.

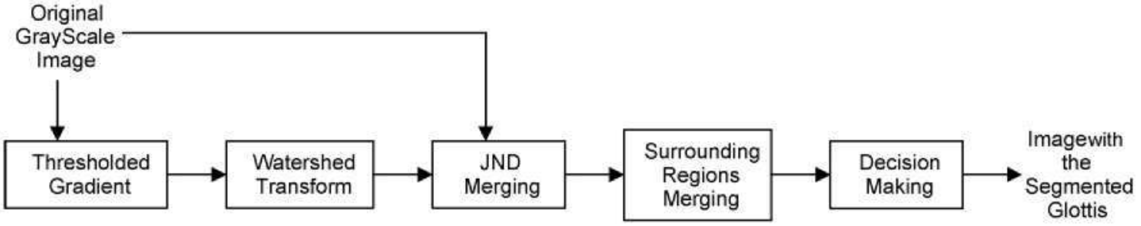

Fig. 4. Scheme that represents the steps followed for the segmentation of the glottal space.

. The luminance component (Y) is chosen and its gradient is calculated. A thresholding with a value of 2 is applied to the gradient image (i.e., those pixels of the gradient image with a grey level below 2 are assigned to 0, so they are converted into minima that can only belong to the internal part of any catchment basin). After the thresholding, the watershed transform is applied to the resulting image. This simple thresholding reduces by 20% the initial number of catchment basins, removing those that appeared due to the noise present in the image.

The threshold applied to the gradient image has been chosen to avoid removing significant edges of the image (although some catchment basins due to noise have been kept).

3.2. *JND based merging*

The second step is a region merging based on the JND. The merging cost function used to merge the basins is calculated according to the Eq. (4),

$$F_c = [|mR_1 - mR_2| - \text{MinJND}(mR_1, mR_2) + 255] \cdot \frac{\text{MinArea}(R_1, R_2)}{\text{LimitArea}} \qquad (4)$$

where mR_i represents the average value of the grey level in each basin, R_i, and LimitArea represents a limit value for the area (number of pixels that belong to a basin).

The goal of the first factor of the merging cost function (Eq. (4)), is to allow the merging of the basins when its result is below 255. This is because under this merging threshold, the human vision system considers that the average grey level of the basins is the same, so it is not able to discriminate between them.

In the second factor of the merging cost function (Eq. (4)), the limit area is empirically established as 0.5% of the total area of the image. This allows the very small regions to decrease their merging cost function. Despite this, in principle they could be different to the human eye. This fact is not a problem because, due to the oversegmentation introduced by the watershed transform, at the beginning of the merging process the basins will have a very small area. So there could exist small regions that belong to big areas with similar grey levels whose difference is below the human eye sensibility threshold. This modification of the merging cost function allows the regions to merge with their most similar neighbours. Thus, the segmentation process becomes isolated from the problems generated by the noise and the poor illumination of the stroboscopic images. The results are significantly improved, but there is still certain dependence of the merging threshold with the glottis area, that will be analyzed in the Section 4.

The JND function used in Eq. (4) follows basically the idea pointed out in Eq. (1), but some changes have been introduced in order to improve the results. The visibility threshold is 255 to the grey levels over 90% of the maximum value in the image histogram. This is to say that all the bright regions of the image are not distinguishable among them. Under this condition, the number of objects segmented after the merging process is drastically reduced. In any case the glottis is perfectly detected since it is a dark region (but not the darkest region as supposed in former works).

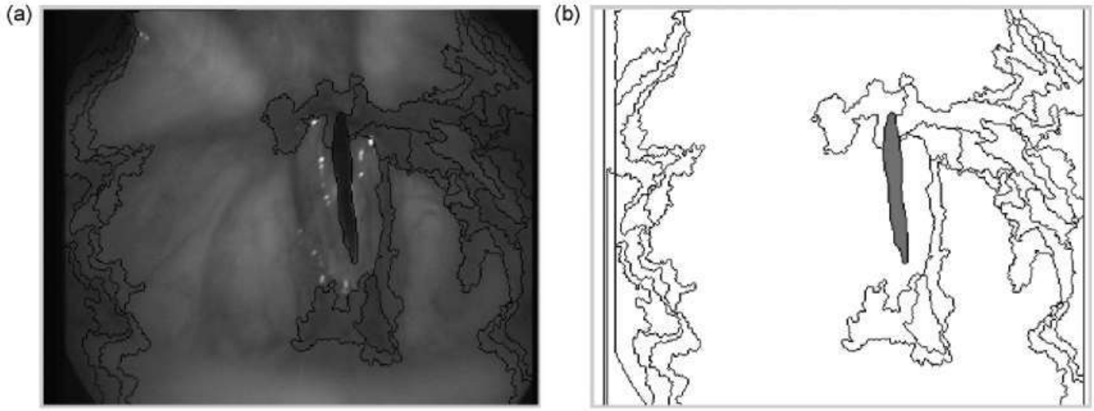

Fig. 5. An example of segmentation of the glottis after the first merging process (second step). (a) Original grey scale image with the region boundaries superimposed; (b) region boundaries.

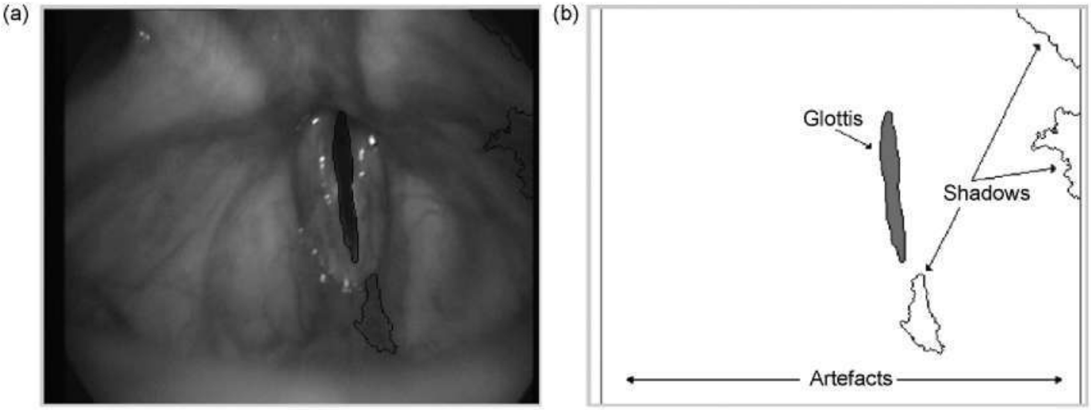

Fig. 6. An example of segmentation after the second merging process (third step). (a) Original grey scale image with the region boundaries superimposed; (b) region boundaries.

Fig. 5 shows the image segmented after the step 2 of the algorithm. The segmented image shows the glottis perfectly delimited as well as other regions with a homogeneous grey level to the human eye. In the example, the bright regions of the image have been merged together as if they were a single object.

3.3. Surrounding regions merging

The third step consists of another merging process, now attempting to merge all the neighbours that surround a region with a lower grey level than all of them. Now the goal is to reduce the number of segmented objects, by merging regions that cannot correspond to the glottis (note that from a human observer's point of view, the glottis should always be a dark object surrounded by a lighter area).

Hence, the process in this step consists of checking all the basins of the image in order to merge all that fulfills the aforementioned condition.

The segmentation obtained after the third step is similar to that represented in Fig. 6. The number of objects present in the image (between 5 and 12) has been substantially reduced. In this picture, it is distinguishable the back of the image, several shadows, and the glottis itself.

3.4. Decision making

The last step is a classification process to detect the glottis among the rest of the objects present in the image. The artefacts in both sides of the image and the back are easy to remove because their characteristics are quite different to those of the glottis: the back is easily removed keeping in mind its large area; and the dark regions at both sides are easily filtered according their low grey level (that is below 10) because the grey level of the glottis is typically between 30 and 70.

In order to distinguish the shadows and the glottis, a linear predictor has been used as described previously. The 88% of the available images (98 photograms taken from 13 videos out of 15) were used to train the predictor, totalling 263 shadow objects and 98 glottis objects. The rest of the images (13 photograms from the two remaining videos) were left aside for a subsequent validation. The decision to preserve the photograms from two videos was taken in order to validate the performance of the process under inter and intra-video illumination variations. The videos chosen are also representative of two different glottis sizes.

The parameters used for discrimination are the 7 invariant moments and the 7 binary invariant moments . The scores

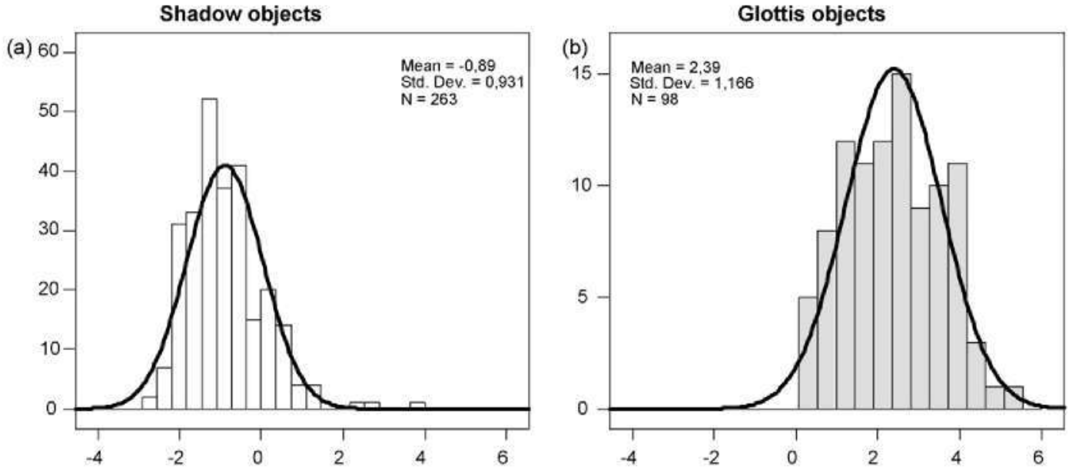

Fig. 7. Histogram of the scores given by the Fisher discriminant. Both histograms are overlapped. (a) Shadow regions; (b) Glottis.

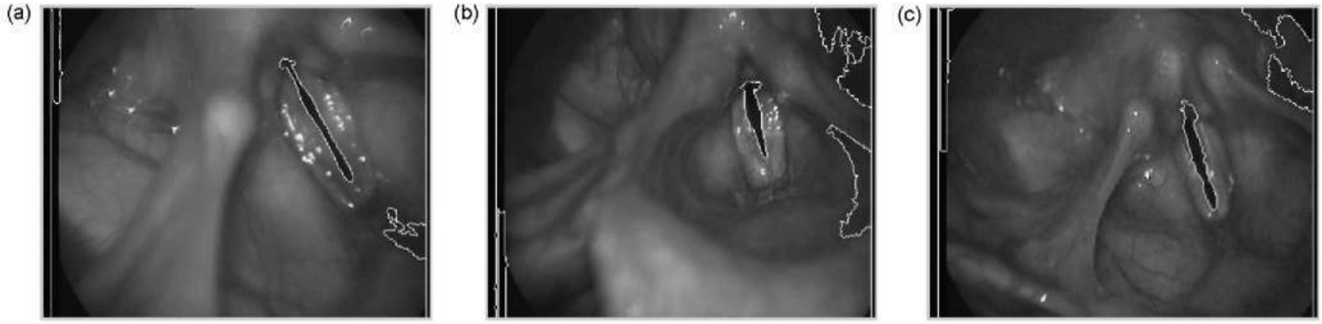

Fig. 8. Some mistakes reported by the algorithm after the first merging step. The objects that would disappear after the second merging step have been removed, in order to distinguish the segmentation of the glottis more easily.

given by the Fisher discriminant are represented in the histograms of Fig. 7. As expected, the histogram that represents the scores of the glottis overlaps with the one of the other regions. The classification threshold was fixed to 0 in order to ensure that all the objects corresponding to the glottis were rightly detected (the cost of rejecting the true glottis has been considered higher than the cost of considering a shadow as a candidate to be the glottis). Using this threshold, some shadows (non-target class) were interpreted by the classifier as a glottis region (target class). The error at this stage using such non-optimum threshold is 12.46%.

After the prediction stage the glottis is well detected and in some cases one or two further shadows are also detected as candidates. Thus, a new rule is applied to remove these shadows from the candidate list: if the remaining regions have a likelihood (score) to be glottis that is three times the likelihood to be glottis of the other regions, it is considered glottis; if there exists any region within this range, the selection of the glottis is based on the largest mean depth of the region (the mean depth being the difference of the mean gradient values of the edges and the minimum gradient inside the region). With this modification, the whole decision making system yields a correct classification rate of 98.98% with the training data.

The classifier has been developed using the software package SPPS 12.05 for Windows.

4. Analysis of the results

The methodology described in the previous sections has been tested with 111 images, taken from 15 videos recorded by the ENT service of the Gregorio Marañón Hospital in Madrid, with videostroboscopic equipment made by two different manufacturers. All the images used present the vocal folds opened.

The first region merging (second step of the whole process) accurately segments the glottis in all the images, but in some cases minor mistakes appear, such as those presented in Fig. 8. These mistakes appear in 5% of the images analyzed. The merging cost threshold was 255 in 75% of the images, and has been slightly changed to adjust the result in 12% of the images. The rest of the images (13% remaining) corresponded to two videos that had a much bigger glottis due the placement of the camera focus, so the merging cost threshold had to be increased to a level of 350 and 550.

After the second merging process the glottis is well segmented in all the images, and the number of regions detected was drastically decreased. In this stage, there is a great variability, but at least 70% of the non-desired objects were removed (in some images 90 was reached).

Lastly, after the decision making step, the glottis is appropriately detected in 98.98% of the training examples and for all the validation cases.

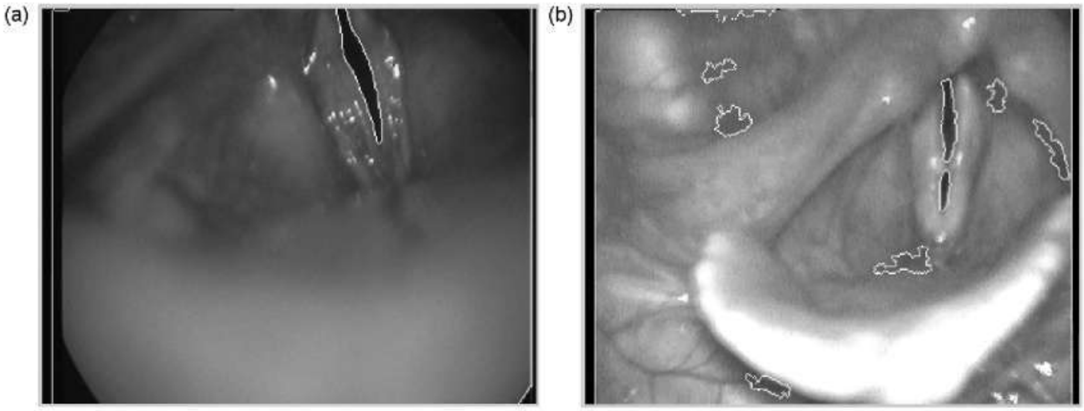

Fig. 9. Two examples of segmentation of the glottis. (a) The glottis is cut due to a movement of the camera; (b) the glottis is divided into two regions due to the presence of bilateral nodules.

The algorithm has gone well with different conditions of illumination and it is able to detect the glottis even when it is partially cut due to camera or patient movements (Fig. 9a). Moreover, the results are good even when the shape of the glottis is distorted due to the presence of some pathology. This is shown in Fig. 9b, where the glottis was divided in two segments due to the presence of bilateral nodules. In this case, the merging threshold had to be adjusted due to the change of the area and the decision stage must allow the preservation of two objects.

Regarding the efficiency of the system, the segmentation process can be considered very fast, being on average around 3.52 s per image (with a Pentium IV-3GHz with 1 GB of RAM). Although, a quick algorithm has been used for this purpose [22], the bottleneck is the calculation of the watershed transform and the merging. All the algorithms were developed in C++ language.

5. Conclusions

In this paper a new method for the automatic segmentation of the glottal space has been presented providing very good results even with stroboscopic images with a poor illumination. The method has proved to be very robust under different inter and intra-video illumination conditions.

The efficiency of the algorithm is based on perceptual criteria given by the JND. These criteria are based on the inability of the human eye to discriminate grey levels that are similar. Furthermore, the algorithm integrates the information that characterizes the glottis as a human observer sees it: the algorithm searches for an elongated shape, with a homogeneous grey level, and surrounded by a brighter region, avoiding conditions such as: the glottis is the darkest object in the image, or the glottis is centred in the image. This fact makes this method more generic than previous works found in the literature, providing accurate results even with images extracted from stroboscopic videos with artefacts introduced during the recording process (darker regions, non-centred glottis or cut in the edge due to a camera or patient movement).

The method has been trained with 98 images extracted from 13 videos and has been validated with 13 images extracted from 2 different videos. The glottis was accurately segmented in 98.98% of the images used for training, and in all the images used for the validation. In 75% of the images that have been used, the merging cost threshold had not been modified.

This presented method does not require any kind of initialization. This implies that the results of this method could also be used as seeds for other glottis segmentation approaches that need them.

The solution presented is very promising; however this algorithm has to be tested with a larger database in order to ensure its generalization capabilities.

Acknowledgements

This research was carried out under grants: TIC2003-08956-C02-01 and TEC2006-12887-C02 from the *Ministry of Science and Technology* of Spain.

References

Baken RJ, Orlikoff RF. Clinical measurement of speech and voice. 2 ed. Singular; 2000.

Svec JG, Schutte HK. Videokymography: high-speed line scanning of vocal fold vibration. Journal of Voice 1996;10(2):201–5.

Shannon CE. Communication in the presence of noise. Proceedings of the Institute of Radio Engineers 1949;37(1):10–21.

Yan Y, Bless D, Chen X. Biomedical image analysis in high-speed laryngeal imaging of voice production. Proceedings of the IEEE Engineering in Medicine and Biology 2005;1:7684–7.

Yan Y, Chen X, Bless D. Automatic tracing of vocal-fold motion from high-speed digital images. IEEE Transactions on Biomedical Engineering 2006;53(7):1394–400.

Manfredi C, Bocchi L, Bianchi S, Migali N, Cantarella G. Objective vocal fold vibration assessment from videokymographic images. Biomedical Signal Processing and Control 2006;1(2):129–36.

Qiu Q, Schutte HK, Gu L, Yu Q. An automatic method to quantify the vibration properties of human vocal folds via videokymography. Folia Phoniatrica et Logopaedica 2003;55:128–36.

Palm C, Lehmann TM, Bredno J, Neuschaefer-Rube C, Klajman S, Spitzer K. Automated analysis of stroboscopic image sequences by vibration profiles. In: Schutte HK, editor. Proceedings of the 5th International Workshop on Advances in Quantitative Laryngology, Voice and Speech Research. Groningen, Netherlands, 2001.

Wittenberg T, Moser M, Tigges M, Eysholdt U. Recording, processing, and analysis of digital high-speed sequences in glottography. Machine Vision and Applications 1995;8:399–404.

Sung MW, Kim KH, Koh TY, Kwon TY, Mo JH, Choi SH, et al. Videostrobokymography: a new method for the quantitative analysis of vocal fold vibration. The Laringoscope 1999;109(11):1859–63.

Sáenz-Lechón N, Osma-Ruiz VJ, Godino-Llorente JI. Kymogram synthesis from pre-recorded low speed video data. Proceedings of IEEE EMBS/BMES 2002;1:1088–9.

Eysholdt U, Tigges M, Wittenberg T, Pröschel U. Direct evaluation of high-speed recordings of vocal fold vibrations. Folia Phoniatrica et Logopaedica 1996;48:163–70.

Marendic B, Galatsanos N, Bless D. New active contour algorithm for tracking vibrating vocal folds. Proceedings of the IEEE International Conference on Computer Vision 2001;1:397–400.

Friedl S., Wittenberg T. Automatic segmentation of vocal folds using active shape models. In: Schade G, editor. Proceedings of the 6th International Conference on Advances in Quantitative Laryngology, Voice and Speech Research. Hamburg, IRB Verlag, 2003.

Bobick A, Davis J. The recognition of human movement using temporal templates. IEEE Transactions on Pattern Analysis and Machine Intelligence 2001;23(3):257–67.

Udupa JK, Saha PK. Fuzzy connectedness and image segmentation. Proceedings of the IEEE 2003;91(10):1649–69.

Ciesielski KC, Udupa JK, Saha PK, Zhuge Y. Iterative relative fuzzy connectedness for multiple objects with multiple seeds. Computer Vision and Image Understanding 2007;107:160–82.

Yang XK, Ling WS, Lu ZK, Ong EP, Yao SS. Just noticeable distortion model and its applications in video coding. Signal Processing: Image Communication 2005;20:662–80.

Bleau A, Leon LJ. Watershed-based segmentation and region merging. Computer Vision and Image Understanding 2000;77(3):317–70.

Gonzalez RC, Woods RE, Eddins SL. Segmentation using the watershed transform. In: Gonzalez RC, Woods RE, Eddins SL, editors. Digital image processing using MATLAB. NJ, USA: Pearson Prentice Hall; 2004. p. 417–25.

Bleau A, De Guise J, LeBlanc R. A new set of fast algorithms for mathematical morphology: I-Idempotent geodesic transforms. CVGIP: Image Understanding 1992;56(2):178–209.

Osma-Ruiz VJ, Godino-Llorente JI, Sáenz-Lechón N, Gómez-Vilda P. An improved watershed algorithm based on efficient computation of shortest paths. Pattern Recognition 2007;40:1078–90.

Hernandez SE, Barner KE. Joint region merging criteria for watershed-based image segmentation. Proceedings of IEEE ICIP 2000 2000;2:108–11.

Shen DF, Huang MT. A watershed-based image segmentation using JND property. Proceedings of IEEE ICASSP 2003 2003;3:377–80.

Haris K, Efstratiadis SN, Maglaveras N, Katsaggelos AK. Hybrid image segmentation using watersheds and fast region merging. IEEE Transactions on Image Processing 1998;7(12):1684–99.

Patino L. Fuzzy relations applied to minimize over segmentation in watershed algorithms. Pattern Recognition Letters 2005;26(6):819–28.

Bueno G, Musse O, Heitz F, Armspach JP. Three-dimensional segmentation of anatomical structures in MR images on large databases. Magnetic Resonance Imaging 2001;19(1):73–88.

Gonzalez RC, Woods RE. Digital image processing. Addison-Wesley; 1992.

Duda RO, Hart PE, Stork DG. Pattern classification. 2 ed. Wiley-Interscience; 2001.

Johnson RA, Wichern D. Applied multivariate statistical analysis. 4 ed. Prentice-Hall; 1998.

Víctor Osma-Ruiz was born in Cuenca, Spain, in 1974. He received the MSc in communications engineering from the Universidad Politécnica de Madrid, Spain, in 2002. From 1999, he has been with the Circuits and Systems Department at the Universidad Politécnica de Madrid as assistant professor. His main research interest is biomedical signal processing.

Juan Ignacio Godino-Llorente was born in Madrid, Spain. He received the MSc in communications engineering in 1996 and the PhD degree with honors in Computer Science from the Universidad Politécnica de Madrid, Spain, in 2002. Currently he is an associate professor at the Universidad Politécnica de Madrid. Also, he is the head of the Circuits and Systems Engineering Department, which belongs to the same University. His main research areas are in the field of biomedical signal processing, with applications to voice disorders and ECG signal processing.

Nicolás Saénz-Lechón was born in Barcelona, Spain, in 1972. He received the MSc in communications engineering from the Universidad Politécnica de Madrid, Spain, in 2002. He is a PhD student in the Circuits and Systems Department at the Universidad Politécnica de Madrid. His main research interest is biomedical signal processing.

Rubén Fraile was born in Vizcaya, Spain, in 1972. He received the MSc in 1995 and the PhD degree with honors in Communications Engineering in 2000, all from the Universidad Politécnica de Valencia, Spain. He is assistant professor at the Circuits and Systems Engineering Department at the Universidad Politécnica de Madrid.