Objective measurements to evaluate glottal space segmentation from laryngeal images*

J.M. Gutiérrez-Arriola, *Member, IEEE*, V. Osma-Ruiz, *Member, IEEE*, N. Sáenz-Lechón , *Member, IEEE*, J.I. Godino-Llorente, *Senior Member, IEEE*, R. Fraile, J.D. Arias-Londoño , *Member, IEEE*

Abstract— **Objective evaluation of the results of medical image segmentation is a known problem. Applied to the task of automatically detecting the glottal area from laryngeal images, this paper proposes a new objective measurement to evaluate the quality of a segmentation algorithm by comparing with the results given by a human expert. The new figure of merit is called Area Index, and its effectiveness is compared with one of the most used figures of merit found in the literature: the Pratt Index. Results over 110 laryngeal images presented high correlations between both indexes, demonstrating that the proposed measure is comparable to the Pratt Index and it is a good indicator of the segmentation quality.**

I. INTRODUCTION

When dealing with the problem of automatically determining the shapes of objects and structures in medical images, it is essential to assess the results of the segmentation. This task is usually performed by an expert who decides if the detected contour is correct or wrong. In spite of being tedious and time consuming, the expert cannot provide a measure of the degree of error that allows a grading of the process without having to visualize the segmented image.

Thus, it would be very convenient to compare the automatic contour with the segmentation outlined by an expert and to obtain a figure of merit to objectively quantify the quality of the automatic segmentation. Several objective measurements have been proposed [1][2][3][4][5][6]. Some of them compare the edge of the objects, being Pratt Index one of the most used in this context [1][2]. Other measurements aim to obtain information on the segmentation of the whole image where several regions are present. In [4] three measures are defined by means of the overlapping area matrix to measure the level of fragmentation of the segmented image.

Surveys found in the literature state that there is no universal measurement to assess the quality of the

segmentation output, and these measurements depend on the aspect to evaluate, so the use of several measurements is common before extracting conclusions.

On the other hand, detecting the glottal area (i.e. the gap between the abducted vocal folds) in laryngeal images is a basic task in the assessment of the phonatory function. In previous works described in [7][8], a watershed-based segmentation system was developed to detect the glottis in stroboscopic images. Although the glottis is correctly located in all the experiments, small errors can be observed delineating the glottal contour in 25% of the segmented images [9], as shown in Fig.1. The output of the segmentation system is the area of the glottis. Thus comparing both the boundary of the object and the area would be interesting.

The paper is organized as follows: section II describes the Pratt Index and the proposed Area Index; next, section III describes the segmentation technique used to evaluate the proposed index; section IV summarized the results; and, finally the last section is dedicated to the conclusions.

II. PRATT INDEX AND AREA INDEX FIGURES OF MERIT

A. Pratt Index figure of merit (F_P)

The figure of merit introduced by Pratt [1] has been chosen for edge comparison. It is defined as:

$$F_P = \frac{1}{\max\{N_I, N_A\}} \sum_{i=1}^{N_A} \frac{1}{1 + \alpha d_i^2} \tag{1}$$

where:

- N_I is the number of ideal edge points, that is, the number of points of the contour drawn up by the expert

- N_A is the number of the actual edge points. The number of points of the contour detected by the automatic segmentation system

- α is a scaling factor chosen to be 1/9 as in [1]

- d_i is the Euclidean distance from the *i*-th pixel of the automatically segmented contour to the nearest point (according to the Euclidean distance) in the edge segmented by the expert.

Pratt figure of merit varies between 0 and 1, indicating from a complete mismatch of the contours to a perfect coincidence, respectively. The known drawbacks are [2][3]:

Figure 1. Gray-scale laryngeal image. Segmentation is outlined in white. A missmatch in the delineation of the glottal area contour can be observed at the top part of the glottis

- This criterion is not symmetrical.

- It does not express shape errors.

- For a perfect segmented contour, Pratt figure of merit would not be 1 if the automatic system considers 4-connectivity and the expert 8-connectivity.

- It depends on the scaling factor α and the matching of points to evaluate d_i.

B. Area Index figure of merit (F_A)

A new figure of merit has been defined to compare two segmented objects based on their overlapped area. Several definitions are needed in order to obtain the Area Index figure of merit (F_A) as shown in Fig. 2 and described next:

- Common area (A_C) is the number of pixels that belong both to the object segmented by the expert and to the object extracted by the automatic system.

- External error area (A_E) is the number of pixels that belong only to the object segmented by the expert and are not present in the automatically segmented object.
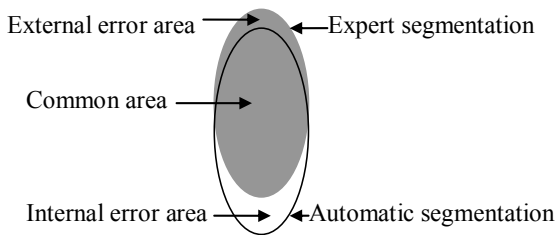

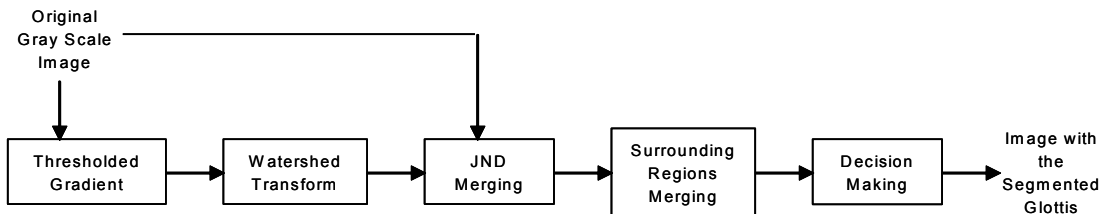Figure 2. Area definitions to calculate the Area Index

- Internal error area (A_I) is the number of pixels that belong only to the automatically segmented object and are not present in the object segmented by the expert.

Given these definitions, the total number of pixels can be calculated as:

$$N = A_C + A_E + A_I \tag{2}$$

And the Area Index figure of merit is calculated as:

$$A_{CN} = \frac{A_C}{N}; \quad A_{EN} = \frac{A_E}{N}; \quad A_{IN} = \frac{A_I}{N}$$

$$F_A = \frac{A_{CN}}{1 + A_{EN} \cdot A_{IN}} \tag{3}$$

F_A varies from 0, meaning that there is no common area, to 1 when both manual and automatic segmentations are identical. This criterion is symmetrical and does not need any extra-parameter definition.

I. AUTOMATIC SEGMENTATION SYSTEM

The method previously described by the authors in [7] performs glottis segmentation in laryngeal images following the scheme presented in Fig. 3. The four main steps that comprise the method are described in the following paragraphs:

Watershed transform of the gradient image [8]: the first step is to convert the original image (RGB) into a gray scale image by means of a transformation to the YIQ model. The luminance component (Y) is chosen and its gradient is calculated. A threshold with a value of 2 is applied to the gradient image (i.e. those pixels of the gradient image with a gray level below 2 are assigned to 0), removing those edges that appeared due to the noise present in the image. After thresholding, the watershed transform is applied to the resulting image, achieving the first region determination.

JND based merging: one of the drawbacks of the system is that Watershed transform is very sensitive to noise, causing the image to be divided into multiple regions where there are only a few (and the delimitation of only one is the goal of the system). The preprocessing mentioned above partially alleviates the problem but doesn't solve it. It is necessary to apply a subsequent merging to ensure the union of homogeneous regions. In this sense, the system presented in Fig. 3 introduces this block to merge the regions that are homogeneous to the human eye following on the basis of JND (Just Noticeable Difference).

Figure 3. Automatic segmentation system

5397

Surrounding regions merging: the third step consists of another merging process, now attempting to merge all the neighbors that surround a region with a lower gray level than all of them. Now the goal is to reduce the number of segmented objects by merging regions that can not 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).

Decision making: the last step is a classification process to detect the glottis among the rest of the objects present in the image. For this purpose, a linear predictor trained with the 7 invariant moments of the different objects is used.

II. EXPERIMENTS AND RESULTS

A. Synthetic images

Experiments with artificial images have been carried out to evaluate the correlation between Pratt Index and Area Index figures of merit. A square, a circle and a triangle are shifted horizontally, vertically and diagonally and then they are compared to their shifted version calculating both figures of merit, Fig. 4. Results are shown in Fig. 5. Both figures of merit are highly correlated and permit to evaluate the quality of the segmentation.

B. Laryngeal images

The data used to evaluate the results were taken from 30 stroboscopic videos recorded at the Gregorio Marañón Hospital in Madrid. Each frame is an image with 720x576 pixels. All patients have some organic pathology that affects the vocal folds. 110 images were randomly chosen and extracted from the 30 sequences. One expert and the automatic system segment the glottis in each image. The expert had no references about the results obtained by the automatic system.

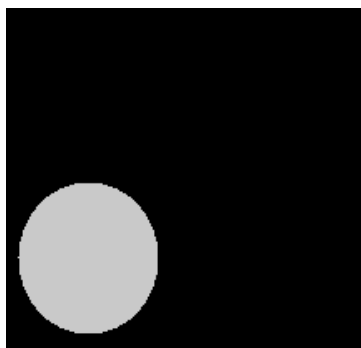Pratt and Area Indexes were calculated. Results are shown in Fig. 6 and 7.

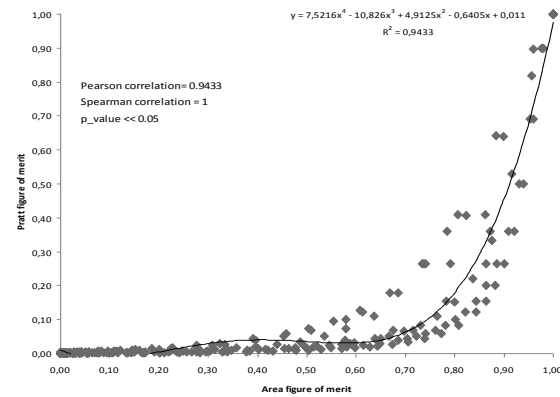Figure 4. A circle synthetic image. The circle is shifted and the compared to the original image.
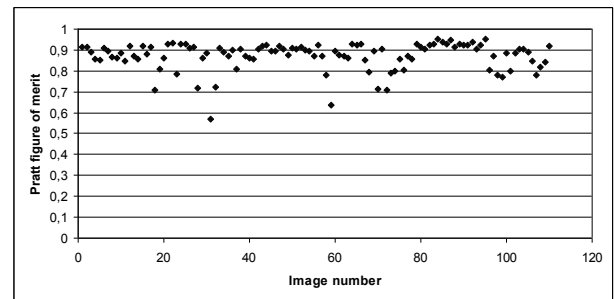

Figure 5. Correlation of both figures of merit

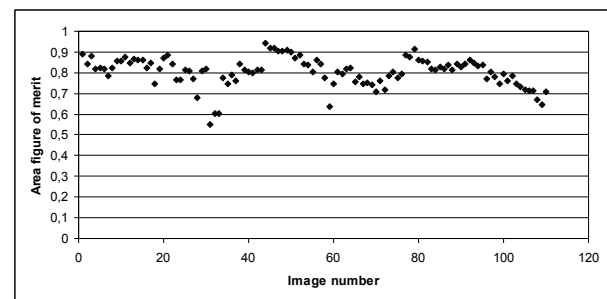Figure 6. Pratt figure of merit for each image.
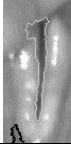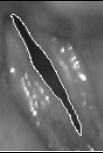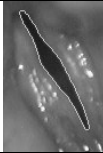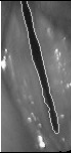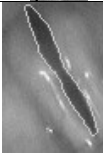
Figure 7. Area Index figure of merit for each image

Both figures of merit are over 0.7 for most of the images indicating that the manual and automatic segmentation are quite similar.

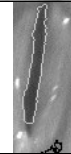Some examples can be seen in Table I. All images are a zoom of the gray-scale laryngeal image, segmentation is outlined in white. In examples 3 and 4 the segmentation results are indistinguishable and both indexes approach 1. If the error is a mismatch in the shape both indexes decrease in the same proportion. This is shown in examples 1, 2 and 5. However if the error is in the size of the segmented glottis, that is, the contour is similar but the automatically segmented glottis is bigger than that segmented by the expert, then the Pratt figure of merit remains high but the area index significantly decreases, as shown in examples 6, 7 and 8 of Table I.

5398

TABLE I.   COMPARISON OF RESULTS

Image example	Expert segmentation	Automatic segmentation	F_P	F_A
1			0.79	0.76
2			0.72	0.60
3			0.95	0.85
4			0.96	0.91
5			0.57	0.55
6			0.92	0.76
7			0.89	0.65
8			0.90	0.60

validate automatic segmentation systems. Results indicate that Area Index figure of merit is able to detect area variations while Pratt Index remains high if there is no great change in the perimeter.

REFERENCES

[1] W.K. Pratt, *Digital Image Processing. Fourth edition.* Wiley. 2007
[2] S. Chabrier, H. Laurent, C. Rosenberger, Y.J. Zhang. "Supervised evaluation of synthetic and real contour segmentation results" proc of 14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, 2006.
[3] N.L. Fernández-García, R. Medina-Carnicer, A. Carmona-Poyato, F.J. Madrid-Cuevas, M. Prieto-Villegas. "Characterization of empirical discrepancy evaluation measures" Pattern Recognition Letters, vol.25, pp. 35-47, 2004.
[4] A. Ortiz, G. Oliver. "On the use of overlapping area matrix for image segmentation evaluation: A survey and new performance measures". Pattern Recognition Letters, vol.27, pp. 1916-1926, 2006.
[5] Y.J. Zhang. "A survey on evaluation methods for image segmentation". Pattern Recognition, vol. 29, No. 8, pp. 1335-1346, 1996.
[6] Y.J. Zhang. "A review of recent evaluation methods for image segmentation". Sixth International Symposium on Signal Processing and its Applications, vol. 1, pp. 148-151, Kuala Lumpur, Malaysia, 2001.
[7] V.J. Osma-Ruiz, J.I. Godino-Llorente, N. Sáenz-Lechón, 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
[8] V.J. Osma-Ruiz, J.I. Godino-Llorente, N. Sáenz-Lechón, P. Gómez-Vilda, "An improved watershed algorithm based on efficient computation of shortest paths," Pattern Recognition, vol. 40, no. 3, pp. 1078-1090, 2007.
[9] V. Osma-Ruiz, J.M. Gutiérrez-Arriola, J.I. Godino-Llorente, N. Sáenz-Lechón, R. Fraile, J.D. Arias-Londoño. "Advanced Preprocessing of larynx images to improve the segmentation of glottal area". Proceedings of MAVEBA, Florence, Italy, 2009.

III. CONCLUSIONS

Two figures of merit have been calculated to evaluate and compare the glottal area segmentation performed by an expert to that obtained from an automatic segmentation system: the Pratt Index, and the proposed Area Index. The proposed figure of merit is a generic measurement that can be used to evaluate the goodness of image segmentation algorithms. It compares two segmented objects based on their overlapped area.

In view of the comparison, the preliminary results show that the proposed index is able to objectively evaluate the quality of the segmented contour, being a candidate to compare automatic and manual segmentations and to