Computerized Medical Imaging and Graphics xxx (2009) xxx–xxx

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

journal homepage: www.elsevier.com/locate/compmedimag

Evaluation of uterine cervix segmentations using ground truth from multiple experts

Shiri Gordon[a], Shelly Lotenberg[a], Rodney Long[b], Sameer Antani[b], Jose Jeronimo[c], Hayit Greenspan[a,*]

[a] Tel Aviv University, Tel-Aviv 69978, Israel
[b] National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA
[c] National Cancer Institute, National Institutes of Health, Bethesda, MD 20852, USA

ARTICLE INFO

Article history:
Received 29 January 2008
Received in revised form 8 November 2008
Accepted 2 December 2008

Keywords:
Image segmentation
Evaluation of segmentation
Multi-expert ground truth
Segmentation complexity
Cervical cancer
Uterine cervix images

ABSTRACT

This work is focused on the generation and utilization of a reliable ground truth (GT) segmentation for a large medical repository of digital cervicographic images (cervigrams) collected by the National Cancer Institute (NCI). NCI invited twenty experts to manually segment a set of 939 cervigrams into regions of medical and anatomical interest. Based on this unique data, the objectives of the current work are to: (1) Automatically generate a multi-expert GT segmentation map; (2) Use the GT map to automatically assess the complexity of a given segmentation task; (3) Use the GT map to evaluate the performance of an automated segmentation algorithm.

The multi-expert GT map is generated via the STAPLE (Simultaneous Truth and Performance Level Estimation) algorithm, which is a well-known method to generate a GT segmentation from multiple observations. A new measure of segmentation complexity, which relies on the inter-observer variability within the GT map, is defined. This measure is used to identify images that were found difficult to segment by the experts and to compare the complexity of different segmentation tasks. An accuracy measure, which evaluates the performance of automated segmentation algorithms is presented. Two algorithms for cervix boundary detection are compared using the proposed accuracy measure. The measure is shown to reflect the actual segmentation quality achieved by the algorithms.

The methods and conclusions presented in this work are general and can be applied to different images and segmentation tasks. Here they are applied to the cervigram database including a thorough analysis of the available data.

1. Introduction

Cervicography is a photographic method for cervical cancer screening that permits archive and study of cervical cancer. The method uses visual testing based on color change of cervix tissues when exposed to 5% acetic acid. This helps to detect abnormal cells that turn white (acetowhite) following the application of 5% acetic acid [9]. In this method the uterine cervix is photographed with a special 35 mm camera with a ring flash, used to provide enhanced illumination of the target region. The resulting image is termed a cervigram. The National Cancer Institute (NCI) has collected a large database of cervigrams as part of an ongoing effort for investigating the role of HPV in the development of cervical cancer and its intraepithelial precursor lesions in women [20]. This database contains a subset of 939 cervigrams that were each segmented by up to twenty medical experts [11]. The segmentation was performed using the Boundary Marking Tool software, developed by the National Library of Medicine (NLM) and NCI [10].

Two clinically important regions were marked by the experts within each image: the cervix boundary and the acetowhite region. The cervix boundary defines the region of medical and anatomical interest within the cervigram. The acetowhite region is the white-appearing epithelium, following the application of 5% acetic acid. The experts were blinded to clinical patient information, such as cytology and HPV status. Examples of manual markings, varying per image from one to twenty, can be seen in Fig. 1. As we consider the segmented images it is evident that several key issues need to be addressed in multiple-expert scenarios: What is the ground truth? Is it the intersection of the markings or their union? Is one expert better than the other? Was the segmentation task a difficult one? As Fig. 1 illustrates, there are "simple" cases, in which most of the experts agree on the tissue boundaries (Fig. 1(c and g)) and more "complex" cases, where the experts have substantially differing markings which vary in size and location (Fig. 1(a

* Corresponding author. Tel.: +972 3 6405839; fax: +972 3 6407939.
E-mail address: hayit@eng.tau.ac.il (H. Greenspan).

0895-6111/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compmedimag.2008.12.002

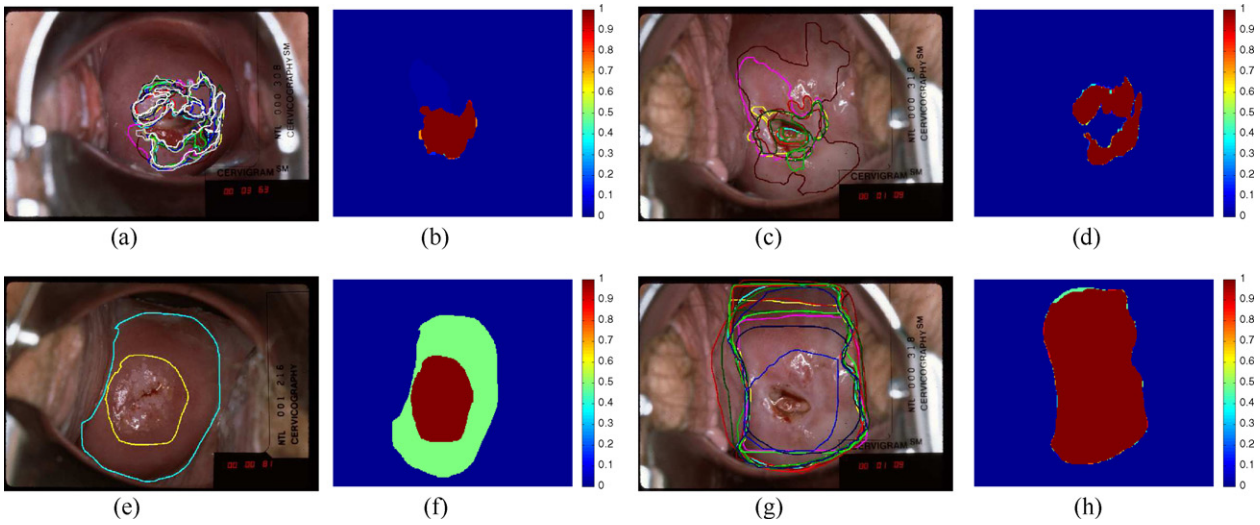

Fig. 1. (a, c, e and g) Examples of manually segmented cervigrams (a different color per expert). (a and c) Acetowhite region; (e and g) cervix boundary. (b, d, f and h) Corresponding multi-expert ground truth, generated by STAPLE. Pixel probabilities are color-coded from blue (low) to red (high).

and e)). How can this level of agreement between the experts be quantified? When building an automated system for cervigram segmentation and analysis, how should we quantify the performance of an automated segmentation algorithm as compared to the markings of multiple experts? And what are the assessment measures that should be used?

Quantitative evaluation and validation of medical image analysis is a well-known challenge. Several measures for the comparison of automated segmentation results to manual segmentation of a *single expert* have become a standard in the field [6,7,21]. Several works attempt to handle the above-listed questions for the case of multiple expert data [2,14,17,23–25]. These works focus on generating an *average* segmentation map using contour-based or area (volume)-based metrics. For example, a simple majority voting rule to generate the multi-expert ground truth map is presented in [25] and others. A shortcoming of this method is that it lacks a strategy for determining the number of experts that should agree before the structure is considered to be present. It treats each expert equally without regard to potential variability in quality of segmentation and does not admit use of *a priori* information about the structure being segmented. A well-known algorithm, that copes with these issues is the STAPLE (Simultaneous Truth and Performance Level Estimation) algorithm [23], which takes a collection of binary segmentations and computes simultaneously a probabilistic estimate of the true segmentation and the performance levels of each input segmentation. The performance parameters are computed using the area-based metrics of sensitivity and specificity. When no "multi-expert ground truth" is necessary, an additional solution is to use the Williams index [26] in order to evaluate a given segmentation against the joint agreement among several experts. This method has been shown to give similar results to the STAPLE-based analysis [17]. STAPLE has been used in the literature in varying application domains, such as generating ground truth maps for Magnetic Resonance Images (MRI) of the brain [3,19], 3D medical structures [1] and open curves of vascular structures [12]. It was used for constructing a brain MRI atlas for two-year-old children [13], and in combining two-class maps to obtain a complete segmentation of a brain tissue [15]. It has also been used for object recognition [18].

In the current work the STAPLE algorithm is applied to the cervigram images for the first time. It is used in order to combine the different expert markings and to generate a single ground truth map. A new segmentation-complexity measure is defined based on the multi-expert ground truth map. The performance parameters of sensitivity and specificity are the common assessment measures used to evaluate segmentation quality in the STAPLE literature (some additional methods are presented and discussed in [28]). These measures are known to possess incommensurate magnitudes, as they represent percentages from different populations of pixels [1]. The current work addresses this difficulty and demonstrates its effect on the STAPLE performance. It then defines an accuracy measure (similar to ref. [1]) to evaluate the results of automatic segmentation algorithms as compared to the multi-expert ground truth map. The accuracy measure is used for the evaluation of two automated algorithms for cervix boundary detection. The focus of the work is on the cervigram database, but the methods proposed are general.

The paper is organized as follows: The STAPLE algorithm is described in Section 2. Its sensitivity to the size of different populations is also discussed. Methods for performance analysis based on the STAPLE output are presented in Section 3. Experimental results on the cervigram database are described in Section 4. A discussion concludes the work in Section 5. This work extends and elaborates on an earlier work presented by the authors [16].

2. The STAPLE algorithm

The STAPLE algorithm [23] takes a collection of binary image segmentations as an input. The object pixels within these segmentations are marked as one and the background pixels as zero. The algorithm simultaneously computes: (1) a probabilistic estimate of the true segmentation and (2) a measure of the performance level represented by each input segmentation (expert). The algorithm is formulated as an instance of the expectation-maximization (EM) algorithm [4]. The performance levels, or quality achieved by each expert, are represented by the sensitivity and specificity parameters. The *sensitivity* (p_j) of expert j represents the "true positive fraction": $p_j = Pr(D_{ij} = 1|T_i = 1)$. The *specificity* ($q_j$) of expert j represents the "true negative fraction": $q_j = Pr(D_{ij} = 0|T_i = 0)$, where D_{ij} is the decision made by expert j for pixel i (1 meaning: present in the expert's segmentation and 0, absent) and T_i is the hidden true segmentation for pixel i.

The EM algorithm estimates the performance level parameters (p, q) while maximizing the complete data log likelihood function. It iterates as follows: In the E-step the unobserved true segmentation

Please cite this article in press as: Gordon S, et al. Evaluation of uterine cervix segmentations using ground truth from multiple experts. Comput Med Imaging Graph (2009), doi:10.1016/j.compmedimag.2008.12.002

is computed as:

$$f(T_i|D_i, p^{(k-1)}, q^{(k-1)}) = \frac{\prod_j f(D_{ij}|T_i, p_j^{(k-1)}, q_j^{(k-1)}) f(T_i)}{\sum_{T_i'} \prod_j f(D_{ij}|T_i', p_j^{(k-1)}, q_j^{(k-1)}) f(T_i')}, \quad (1)$$

where $f(T_i)$ is the prior probability for pixel i and k is the iteration step. Considering a binary segmentation, factoring over all the experts and using the definitions for p_j and q_j, the following formulas are derived:

$$a_i^{(k)} \equiv f(T_i = 1) \prod_j f(D_{ij}|T_i = 1, p_j^{(k)}, q_j^{(k)})$$

$$= f(T_i = 1) \prod_{j:D_{ij}=1} p_j^{(k)} \prod_{j:D_{ij}=0} (1 - p_j)^{(k)}, \quad (2)$$

$$b_i^{(k)} \equiv f(T_i = 0) \prod_j f(D_{ij}|T_i = 0, p_j^{(k)}, q_j^{(k)})$$

$$= f(T_i = 0) \prod_{j:D_{ij}=0} q_j^{(k)} \prod_{j:D_{ij}=1} (1 - q_j)^{(k)}, \quad (3)$$

where $j : D_{ij} = 1$ denotes the set of indexes for which the decision of the rater at pixel i has the value 1. Using these formulas, a compact expression for the conditional probability of the true segmentation at each pixel, W_i, is defined:

$$W_i^{(k-1)} \equiv f(T_i = 1|D_i, p^{(k-1)}, q^{(k-1)}) = \frac{a_i^{(k-1)}}{a_i^{(k-1)} + b_i^{(k-1)}}. \quad (4)$$

The experts performance level parameters are estimated in the M-step using the following equations:

$$p_j^{(k)} = \frac{\sum_{i:D_{ij}=1} W_i^{(k-1)}}{\sum_i W_i^{(k-1)}}; \quad q_j^{(k)} = \frac{\sum_{i:D_{ij}=0} (1 - W_i^{(k-1)})}{\sum_i (1 - W_i^{(k-1)})}. \quad (5)$$

The *sensitivity estimator*, p_j, can be interpreted as the ratio of the jth expert true positive detections to the total amount of the structure $T_i = 1$, where in both cases each pixel is weighted by W_i: the strength of belief in $T_i = 1$. Similarly, the *specificity estimator*, q_j, can be interpreted as an estimator for the specificity given a degree of belief in the underlying $T_i = 0$ state.

The unobserved true segmentation computed in the E-step is a probability map where each pixel is assigned the probability of being part of the segmented object according to (1) the amount of agreement among the experts and (2) the performance levels of the experts. This map is regarded as the "multi-expert ground truth segmentation" generated by STAPLE. Fig. 1 shows examples of multi-expert ground truth maps that correspond to the expert markings for both the acetowhite region (b and d) and the cervix boundary (f and h). Pixel probabilities are color-coded from blue (low probability—zero) to red (high probability—one). The intersection of all experts' markings is colored red (with the highest probability value) as expected.[1]

In the current task of cervigram segmentation, the object area, which is the cervigram region, is relatively small as compared to the area of the background. The amount of pixels for which the experts disagree when marking the object, is even smaller as compared to the background (i.e. Fig. 1(f and h)). In such cases the range of sensitivity values and the range of specificity values, computed per image for the different expert markings, are incommensurate. The specificity values that are computed with respect to the background area (Eq. (5)) obtain much higher values with a much narrower range. This is due to the fact that most of the background pixels are marked correctly by the experts, while the differences between their markings are very small with respect to the overall background area. This behavior has a major influence on the estimated ground truth segmentation: The b_i values (Eq. (3)) are small compared to the a_i values and the resulting W_i (Eq. (4)) values are higher than expected. A similar range for the two performance measures (sensitivity and specificity) can be obtained only when the object and the defined background are of the same size. Such a case seldom happens in real-life segmentations.

We propose the following procedure to obtain more comparable performance measures: The union of the different expert markings is considered to be the object area. The background area is modified to include the same amount of pixels as in the defined object. These pixels are equally distributed around the object. The idea of modifying the background area prior to the manual or automatic segmentation process, was previously suggested in order to improve the segmentation results and the STAPLE-based validation [28]. Fig. 2 illustrates the influence of the background modification on the output of the STAPLE algorithm. Fig. 2(a) presents five overlapping segmentation masks of ellipses with different orientations and sizes (each mask is color coded differently and marked by a different number), the rest of the image is considered to be the original background area. Fig. 2(d) presents the suggested background modification. The union of the masks is colored red. The modified background, colored green, is equally distributed around this region with the same amount of pixels. The rest of the pixels within the image, colored blue, are masked out and ignored throughout the rest of the computations. Fig. 2(b and c) are the ground truth segmentation maps generated by the STAPLE algorithm when using the original background area and the modified background, respectively. Fig. 2(e and f) are corresponding histograms, reflecting the distribution of the probability values within each map. Note that the y axis is truncated at the value of 0.1, to allow for a better observation of the distribution values within the object area. The original ground truth map, (b), generated by applying the STAPLE algorithm to the entire image, possesses higher probability values as compared to the map generated with the modified background (c). As a result, a larger area is considered to be an object with high probability. This result does not reflect the actual overlap that exists between the experts. It is expected for example, that the region with the highest probability will be the region marked by the largest amount of experts: the intersection of masks 1, 2, 4, 5. This is not the case in map (b), where an additional region, the intersection of masks 1, 2, 3, possesses the same value. The ground truth map generated with the modified background, (c), has a broader range of probability values and reflects the overlap between the different expert markings more accurately.

Table 1 presents the sensitivity and specificity values computed for each of the masks for the original background case (p, q) and for

Table 1
STAPLE simulation: sensitivity and specificity values computed for each of the segmentation masks of Fig. 2 using the original background (p, q) and the modified background (p_m, q_m).

	I_1	I_2	I_3	I_4	I_5
p	1.00	0.48	0.44	0.31	0.24
q	0.94	0.99	0.98	1.00	1.00
p_m	1.00	0.67	0.26	0.46	0.36
q_m	0.68	0.97	0.86	1.00	1.00

[1] A colored version of this paper is available online.

Please cite this article in press as: Gordon S, et al. Evaluation of uterine cervix segmentations using ground truth from multiple experts. Comput Med Imaging Graph (2009), doi:10.1016/j.compmedimag.2008.12.002

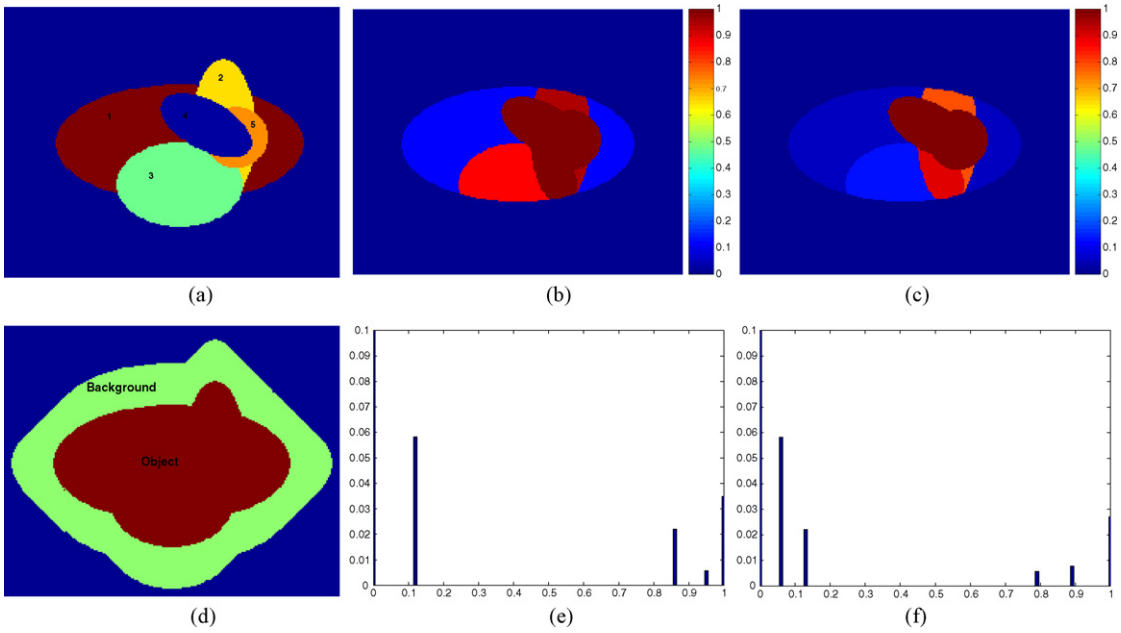

Fig. 2. STAPLE simulation. (a) Input segmentation masks imposed on the same image (each binary mask is color coded differently); (b) multi-expert ground truth generated by STAPLE using the original background; (c) multi-expert ground truth generated by STAPLE using the modified background. (d) background modification process: object area (red), modified background area (green) the rest of the pixels (blue) are ignored; (e and f) histograms of the ground truth maps, presented under corresponding results.

the modified background case (p_m, q_m). Without the background modification the specificity values are higher and with a narrower range (0.94:1 as compared to 0.68:1). Note the significant change in the specificity value of expert 1 (who has the lowest specificity). In addition, the sensitivity values do not accurately reflect the level of agreement between the experts. For example, with the original background, the sensitivity of expert 3 is higher than the sensitivity of experts 4 and 5. This is in contradiction to the fact that the markings of experts 4 and 5 possess a better overlap with the high-probability region of the multi-expert ground truth. When the modified background is used, expert 3 attains the lowest sensitivity value, as expected.

3. Performance analysis based on multi-expert ground truth maps

The multi-expert ground truth (GT) map generated by STAPLE is a fuzzy probability map, that includes the probability per pixel to belong to the object region within the image. In this section we propose to use the map for quantifying segmentation complexity (Section 3.1) and for quantitatively evaluating the results of an automated segmentation algorithm (Section 3.2).

3.1. Measuring segmentation complexity

We propose to compute a set of measures, or "descriptors", to represent the complexity of a given segmentation task. The underlying assumption is that the variability of pixel probabilities within the GT map indicates the level of disagreement between the human experts. This, in turn, is an indication for the increased complexity of the segmentation task. The probability values within the GT map are linearly stretched prior to the descriptors computation, to enable a comparison of the values across the images. Highest probability value is set to one and lowest value to zero. The probability value of a pixel i within the GT map, I, is denoted by W_i. The descriptors are computed only for the object area, as defined by the STAPLE output ($W_i > 0$).

We propose the following descriptors:

1. **Entropy:**
   Entropy is a well known measure of distribution homogeneity. It is computed here using the histogram representation of the distribution within the GT map:

$$\text{entropy}(I) = -\sum_{i=1}^{N} h_i \times \log(h_i) \tag{6}$$

   where h_i is the value of bin i of the histogram ($\sum h_i = 1$) and $N = 100$. An intuitive understanding of entropy relates to the amount of uncertainty in the segmentation: A GT map that contains a single probability value, as is the case in complete overlap of the expert segmentations, has entropy value of zero (no uncertainty). Disagreement between the experts generates additional probability values within the GT map, which leads to a broader distribution and to a higher entropy value.
2. **Standard deviation (STD):**
   The width of a distribution can be measured using its standard deviation:

$$\text{STD}(I) = \left(\frac{1}{n-1} \sum_{i=1}^{n} (W_i - W_m)^2 \right)^{1/2}, \tag{7}$$

   where n is the number of pixels considered to be an object and W_m is the mean of their probability values. A low STD value is associated with a narrow distribution of probabilities within the GT map and vise versa.
3. **Entropy or Standard Deviation scaled by Mean:**
   While the entropy and the standard deviation represent the spread of the distribution well, they do not represent the probability values themselves. This information is important when measuring segmentation complexity. A narrow distribution may be located in the high-probability range, or in a low-probability range. The first case corresponds to GT maps with large areas of strong agreement and the second case corresponds to GT maps with large areas of strong disagreement between the experts. (Examples and discussion will be provided in Section 4). In order to cope with such cases, we propose a *normalized* set of

Please cite this article in press as: Gordon S, et al. Evaluation of uterine cervix segmentations using ground truth from multiple experts. Comput Med Imaging Graph (2009), doi:10.1016/j.compmedimag.2008.12.002

descriptors. In this set the entropy and the standard deviation are scaled by the square of the distribution's mean. These descriptors are termed "entropy scaled by mean" (ESM) and "standard deviation scaled by mean" (SSM), respectively and are defined as:

$$\mathrm{ESM}(I) = \frac{\mathrm{entropy}(I)}{\mathrm{mean}(I)^2}; \quad \mathrm{SSM}(I) = \frac{\mathrm{std}(I)}{\mathrm{mean}(I)^2}. \tag{8}$$

We now have four types of descriptors. We classify a given segmentation task as "simple" or "complex", relative to a selected set of these descriptors. We use either a thresholding or a clustering approach to carry out this classification. A threshold for the complexity can be learned from a training set of segmentations, for the ESM (SSM) descriptor, following which each new segmentation task can be categorized using the selected threshold. For the clustering scheme, we use a 2D feature space of the entropy and the mean descriptors. Training data is used to cluster the space into varying complexity levels. Based on the 2D clustering of the complexity feature space, each new image input to the system can be categorized as less or more complex (depending on its own GT map descriptor set). It is also possible to analyze the variability of the segmentation complexity across the images within the database. In addition, the complexity of segmenting different regions within the cervix can be compared, thus distinguishing between easy and difficult segmentation tasks.

3.2. Evaluating automatic segmentation results

Given a new segmentation map, created independently of STAPLE, it may be desirable to compare it quantitatively to the STAPLE multi-expert ground truth. Such an analysis can be used to assess the performance of an automated segmentation algorithm and to compare the results of different algorithms. This analysis can be made using the following methods:

1. Computation of the sensitivity and specificity performance levels of the new segmentation as compared to the multi-expert ground truth, using Eq. (5)[23].
2. Computation of the *accuracy* [5] of the new segmentation as compared to the multi-expert ground truth. The *accuracy* of a given classifier, defined on a binary set of samples with positive, P, and negative, N, labels, is computed as the total correct fraction: $((TP + TN)/(P + N))$, where TP and TN are the amount of true positives and true negatives detected by the classifier. In the current case, the multi-expert ground truth is a set of real numbers, W, which define the probability for a positive label; denote the new segmentation by D. Then we compute accuracy as [1]:

$$\mathrm{accuracy} = \frac{\sum_{i:D_i=1} W_i + \sum_{i:D_i=0} (1 - W_i)}{N}, \tag{9}$$

where D and W are treated as vectors, and N is the number of samples being considered. Higher *accuracy* values indicate better correspondence to the ground truth.

In the first method, the sensitivity and specificity parameters computed for a specific image are compared to the parameters attained by the human experts for that image. Each of the parameters is evaluated separately. This method has two main drawbacks: First, as the number of experts increases, it is more complicated to rank their results and compare them to the results of the algorithm. A single measure is more appropriate in that case. Second, as demonstrated in Section 2, the range of the sensitivity and the specificity values depends on the relative size of the object and background within the image, thus care should be taken when combining them into a single measure. In addition, these measures can be used to compare results within a single image (of the same data [1]), but not across the images in the database, as the size of the objects varies considerably. A statistical evaluation of segmentation algorithm results is applicable only when the comparison is performed between different algorithms and on the same data.

In earlier work [16], the *F-measure* was suggested in order to combine the sensitivity, p, and specificity, q, into a single value. The *F-measure* is defined as the weighted harmonic mean of the two parameters [22]:

$$F = \frac{pq}{\alpha p + (1 - \alpha) q}, \quad \alpha = 0.5. \tag{10}$$

Being dependent on sensitivity and specificity, the *F-measure* is strongly affected by the relative object/background area. This measure also assumes a similar range of sensitivity and specificity values within a single image (by setting $\alpha = 0.5$), which is seldom true in the current case. Other measures that combine sensitivity and specificity into a single measure suffer from the same faults (including the shortest distance from the (0, 1) corner, used in ROC analysis [5], and the mean predictive value (*PV*) [23], that reduces to $((p + q)/2)$ in the binary case).

The *accuracy*, used in the current work, accounts for the amount of accurately detected labels as compared to the image size and not the size of the different regions within it. The modification of the background area, in order to balance the object/background proportions, makes this measure even less sensitive to their relative size, as compared to the other options.

Table 2 presents the *F-measure* (*F*), mean predictive value (*PV*), and the *accuracy* (*acc*) results computed for the different segmentations, $I_1, \ldots, I_5$, within the simulation of Fig. 2. These results illustrate the benefits of using the *accuracy* measure. The values are computed using both the original and the modified backgrounds. A higher value indicates a more accurate segmentation, as compared to the STAPLE-generated, multi-expert ground truth. In order to identify the segmentation that is most similar to the multi expert ground truth, the different values are sorted in decreasing order, and the different segmentations are ranked accordingly. The desired ranking according to our perceptual understanding may be: I_2, I_4, I_5, I_3, I_1, where I_2 is the most similar segmentation to the ground truth.

The following observations can be made: (1) The *accuracy* measure obtains a similar ranking of segmentations for the original background (*acc*) and for the modified background (acc_m). This observation indicates reduced sensitivity of the *accuracy* measure to the relative size of the object and the background; (2) The ranking of the *accuracy* measure appears to correspond with our perceptual understanding; and, (3) The range of the *accuracy* values is increased when using the modified background. This generates a

Table 2
STAPLE simulation.

	I_1	I_2	I_3	I_4	I_5	Rank results
F	0.97	0.65	0.60	0.47	0.39	I_1, I_2, I_3, I_4, I_5
F_m	0.81	0.79	0.40	0.63	0.53	I_1, I_2, I_4, I_5, I_3
PV	0.97	0.74	0.71	0.66	0.62	I_1, I_2, I_3, I_4, I_5
PV_m	0.84	0.82	0.56	0.73	0.68	I_1, I_2, I_4, I_5, I_3
acc	0.945	0.959	0.947	0.954	0.950	I_2, I_4, I_5, I_3, I_1
acc_m	0.73	0.92	0.77	0.91	0.90	I_2, I_4, I_5, I_3, I_1

F-measure, mean predictive value and *accuracy* results computed for each of the segmentation masks of Fig. 2 with the original background (F, PV, acc) and with the modified background (F_m, PV_m, acc_m). Corresponding ranking order of segmentations, from most to least similar, is included (left to right).

Please cite this article in press as: Gordon S, et al. Evaluation of uterine cervix segmentations using ground truth from multiple experts. Comput Med Imaging Graph (2009), doi:10.1016/j.compmedimag.2008.12.002

better distinction between the quality of the different segmentations within a single image and provides additional support for the background modification suggested in Section 2.

The above observations are not exhibited in the case of the *F*-measure and the *PV*. The different range of values obtained for the sensitivity and the specificity measures with the original background and with the modified background, affects the values and corresponding ranking of the *F-measure* and the *PV*. This indicates that these measures are more sensitive to the relative proportions of the object and background. An additional observation relates to the ranking order itself, where I_1 is wrongly ranked as most similar to the GT. This misplacement occurs because a similar range of sensitivity and specificity is assumed in the computation. The range of the specificity values, improved by the modified background (Table 1), is still narrower than that of the sensitivity.

4. Experiments and Results

A set of experiments was conducted in order to evaluate the proposed computational measures and analysis schemes. The database used contains a set of 932 manually segmented cervigrams out of the 939 cervigrams of the NCI database [11]. Seven additional

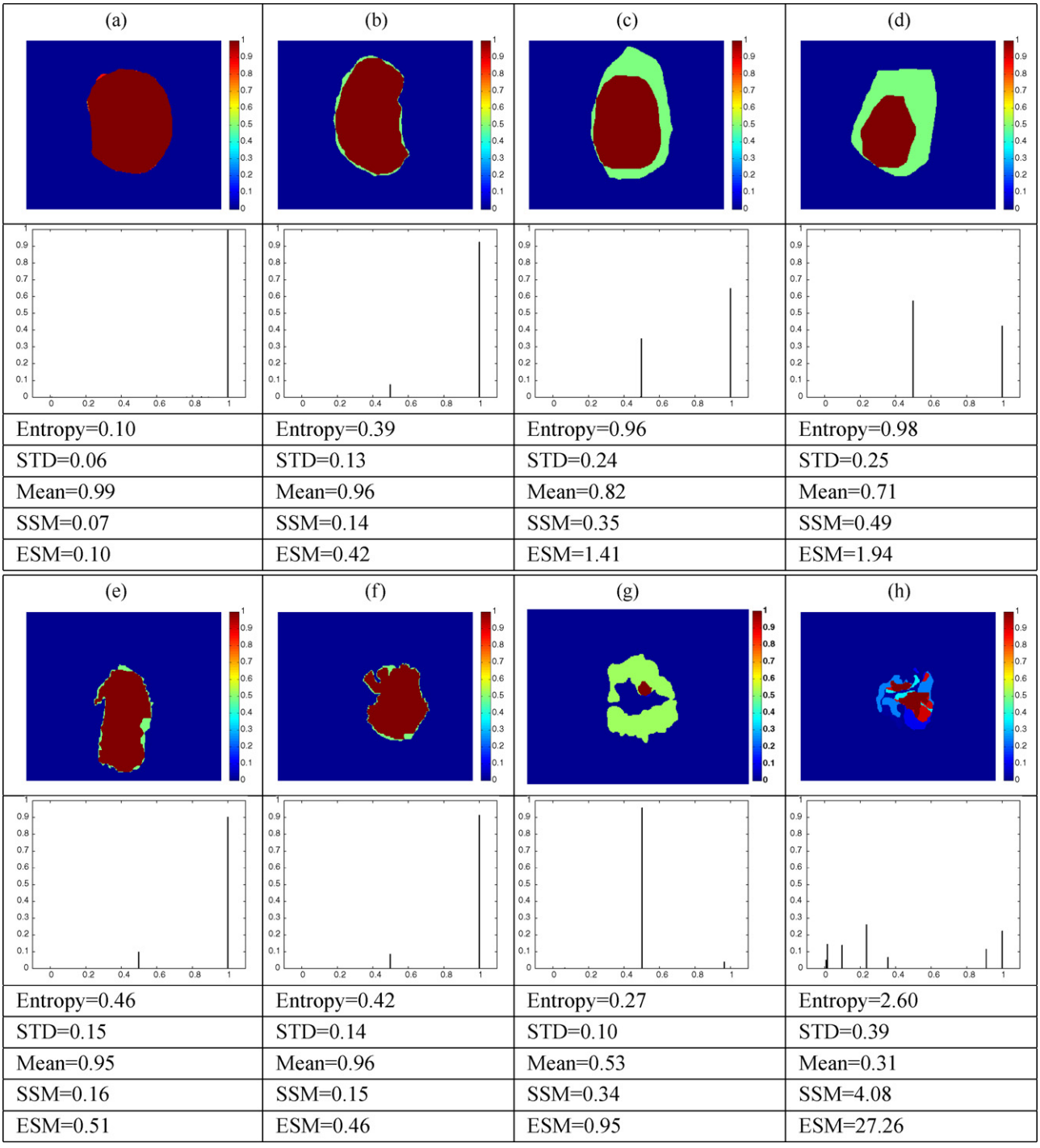

Fig. 3. Examples for multi-expert ground truth data for the cervix boundary (top row) and the acetowhite region (bottom row). (a, b, e and f) Examples of agreement among experts; (c, d, g and h) examples of disagreement among experts. Corresponding histograms and complexity descriptors are presented under each example.

Please cite this article in press as: Gordon S, et al. Evaluation of uterine cervix segmentations using ground truth from multiple experts. Comput Med Imaging Graph (2009), doi:10.1016/j.compmedimag.2008.12.002

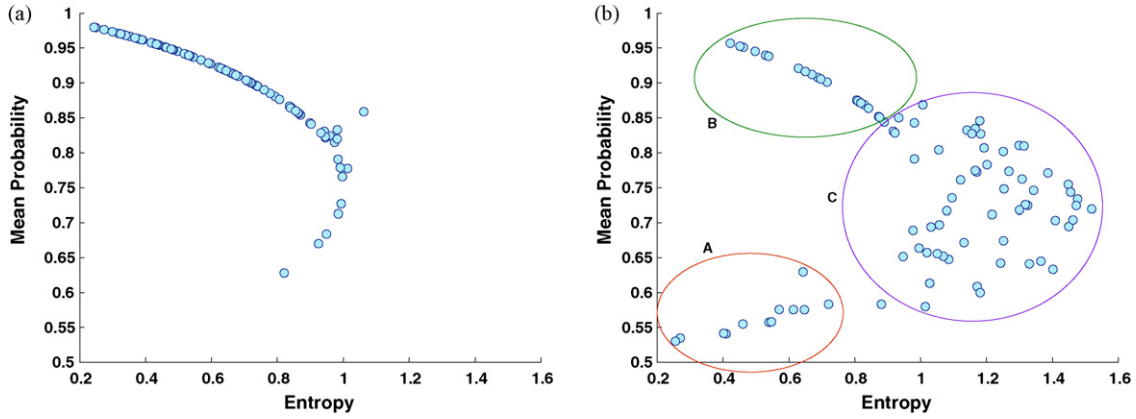

Fig. 4. Scatter in entropy-mean feature space of 100 images marked by two experts. (a) Cervix boundary segmentation; (b) Acetowhite region segmentation.

cervigrams were discarded since their segmentations exhibited no overlap whatsoever between the human experts. The NCI database was divided into two main groups. One group contains 20 cervigrams that were marked by twenty medical experts. The second group contains the remaining 912 cervigrams, with each marked by two experts out of the twenty medical experts. The markings of two regions are examined: the acetowhite region and the cervix boundaries. A version of the STAPLE algorithm that considers the modified background (Section 2) was used throughout the analysis to generate the multi-expert ground truth.[2]

4.1. Evaluation of the segmentation complexity descriptors

We start by examining the correlation between the proposed segmentation complexity descriptors (Section 3.1) and the level of agreement among the experts, as given by the multi-expert ground truth generated by STAPLE. Fig. 3 shows examples of the multi-expert ground truth segmentation for both the cervix boundary and the acetowhite region. Histograms of the probability values within each map are presented under corresponding examples (for the object region only). The segmentation complexity descriptors, computed for each of these examples, are listed. The following observations can be made:

- The *Entropy* measures the distribution of the probability values within the map, without taking into account their magnitudes. This results in cases such as the one presented in Fig. 3(g), where the entropy is very low but the disagreement between the two experts is clearly visible.
- The *Standard deviation (STD)* has a similar deficiency: it accounts for the distribution of the probability values around their mean, but does not consider the mean value itself. Like the entropy, the standard deviation will fail (i.e., take on low values, even though the disagreement between the experts is high) in cases such as the one presented in Fig. 3(g).
- The *Mean* descriptor is the average value of the probabilities within the ground truth map. A correlation can be detected between high mean values and strong agreement between the experts. The mean value, however, lacks the ability to differentiate between cases with similar mean and different distributions.
- The *Entropy Scaled by Mean (ESM)* combines the benefits of both the entropy and the mean descriptors and successfully differentiates among the different levels of agreement in all of the presented examples. Low values of ESM are correlated with high levels of agreement. The *STD Scaled by Mean (SSM)* attains similar results.

From these observations it is evident that high level of agreement between the experts is captured well by the ESM or the SSM descriptors, where both the distribution and the mean of the probability values within the ground truth segmentation are considered. It is important to note that the ESM/SSM values are strongly influenced by the number of expert markings, since a larger number of experts may produce a wider range of probability values within the multi-expert ground truth. A more reliable comparison would be between images that were marked by the same number of experts.

In a second experiment, we evaluate segmentation complexity by clustering in the two-dimensional feature space of entropy and mean. In this feature space, low entropy and high mean values are correlated with easier cases, where expert agreement is high. Fig. 4 presents a scatter plot of the entropy and the mean descriptors computed for 100 expert segmentations. These segmentations were randomly selected out of the 912 cervigrams that were marked by two experts. The experiment was conducted for segmentation of (a) the cervix boundary and (b) the acetowhite region.

The distribution of the cervix boundary segmentations, Fig. 4(a), is mainly concentrated within the low-entropy-high-mean region of the feature space. This indicates a strong agreement among the experts within most of the cervigrams. In the distribution of the acetowhite segmentations, Fig. 4(b), three main groups can be detected: Group A includes the low-entropy-low-mean segmentations. Group B includes the low-entropy-high-mean segmentations and Group C includes the high-entropy-mid-mean segmentations. This may be interpreted as follows: Group A corresponds to segmentations with strong disagreement among the experts. Group B corresponds to segmentations with strong agreement, and Group C corresponds to segmentations with an intermediate level of disagreement among the experts. Fig. 5 shows ground truth segmentation examples for each of these groups along with their ESM values. From top to bottom, each row shows examples for groups A, B and C, respectively. The distinction between the maps in the entropy-mean feature space is highly correlated with the level of agreement between the experts, visually detected in the maps within each group (where the red color corresponds to regions of strong agreement). The ESM descriptor attains the lowest values within the images of group B, where the level of agreement between the experts is high, as expected. The distinction between groups A and C is less evident when using the ESM descriptor.

[2] The original STAPLE algorithm is available via the ITK toolkit (http://www.itk.org/).

Please cite this article in press as: Gordon S, et al. Evaluation of uterine cervix segmentations using ground truth from multiple experts. Comput Med Imaging Graph (2009), doi:10.1016/j.compmedimag.2008.12.002

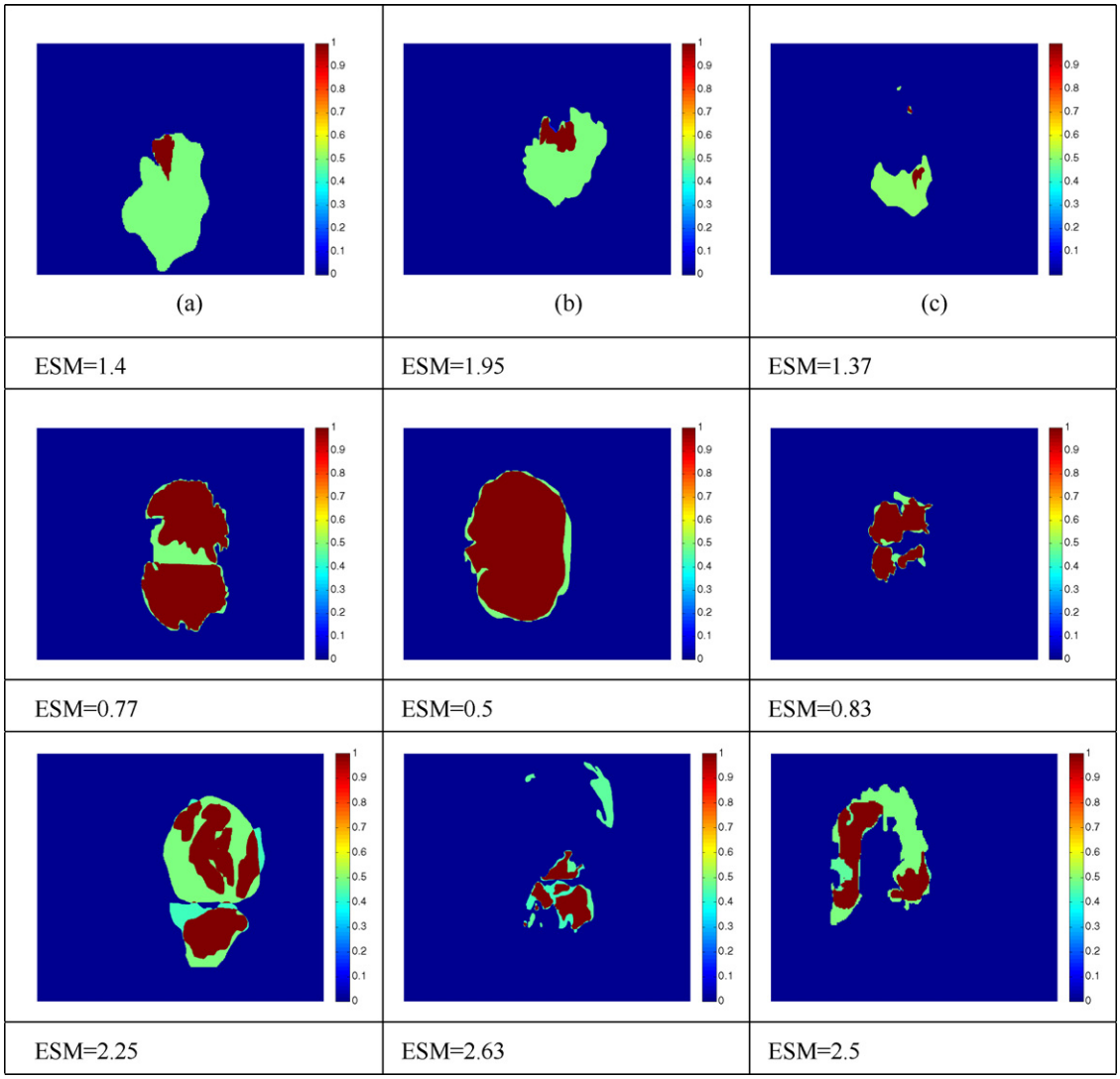

Fig. 5. Example segmentations for the three groups observed in the entropy-mean feature space of the acetowhite region (Fig. 4(b)). Top row: group A—low-entropy-low-mean; middle row: group B—low-entropy-high-mean; bottom row: group C—high-entropy-mid-mean.

The images within the three groups were presented to a medical expert who was asked to describe the visual appearance of the acetowhite regions within them. According to the expert most of the acetowhite regions within group A are not visibly clear (described as pale, with diffused and weak acetowhitening). This may explain the poor level of agreement between the experts detected in these images. The distinction between the images within groups B and C is less evident, but most of the acetowhite regions within these groups are described to be very clear and well delimited.

4.2. A comparison between the complexity of the acetowhite and cervix boundary segmentation tasks

The complexity descriptors can be used to assess the segmentation complexity of different images as well as of different segmentation tasks. In the following experiment, the complexity of segmenting the acetowhite region is compared to the complexity of segmenting the cervix boundary. Fig. 6 presents the distributions of the ESM descriptor for the cervix boundary segmentation (a), and for the acetowhite segmentation (b). The distributions were computed for images that were marked by two experts. The cervix boundary segmentation has a narrower distribution with a lower mean value. This indicates strong agreement among the experts in most of the cases, and suggests that the cervix boundary segmentation task is the easier task. Fig. 6(c and d) show scatter plots in entropy-mean feature space for images that were segmented by more than ten observers. The cervix boundary segmentation results, (c), are concentrated in the low-entropy-high-mean region. This reflects the strong agreement between the experts in all cases. The scatter of the acetowhite segmentation results in (d) is more spread out, thus indicating a larger disagreement across the different cases and, correspondingly, a more complex segmentation task. Similar results can be detected in the scatters presented in Fig. 4, in which the results of 100 cervigrams are presented.

4.3. Evaluation of automatic cervix boundary segmentation

We compared two algorithms for cervix boundary detection, using the accuracy measure defined in Section 3.2. The first algorithm (algorithm I) detects an initial coarse region of interest (ROI) located around the cervix region [8]. The second algorithm (algorithm II) is based on a new active contour functional that incorporates a local convexity feature and was devised especially for

Please cite this article in press as: Gordon S, et al. Evaluation of uterine cervix segmentations using ground truth from multiple experts. Comput Med Imaging Graph (2009), doi:10.1016/j.compmedimag.2008.12.002

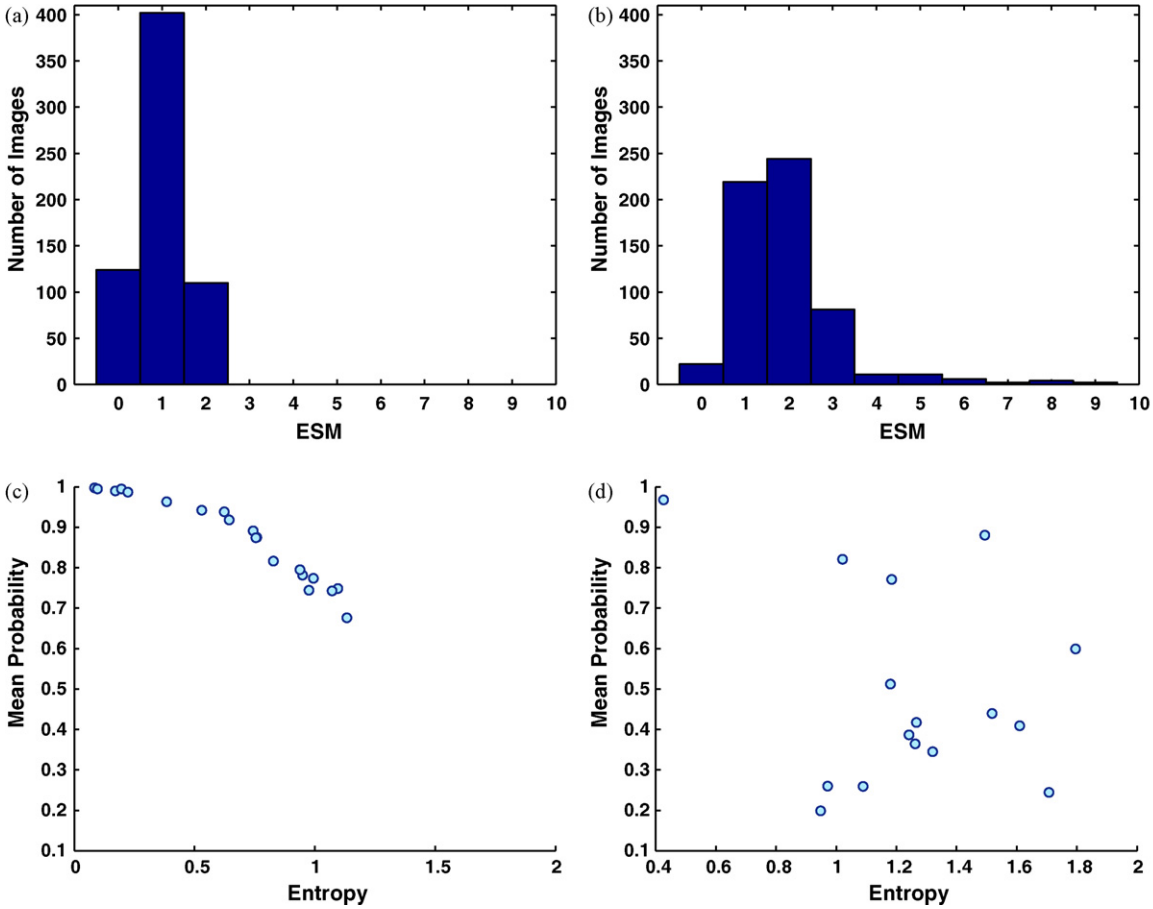

Fig. 6. Distribution of images that were marked by two experts: (a) cervix boundary, mean = 0.93 (636 images); (b) acetowhite region, mean = 1.93 (602 images). Scatters in the entropy-mean feature space for images with more than ten experts markings: (c) cervix boundary (20 images); (d) acetowhite region (16 images).

the purpose of cervix boundary detection [27]. Algorithm I is used to initialize algorithm II and is therefore expected to have inferior results.

A comparison was conducted among the following assessment measures: sensitivity (p), specificity (q) and *accuracy*. All measures were computed using the multi-expert ground truth generated by STAPLE. The experiment was conducted on 636 images in which the cervix boundary was marked by two experts. Table 3 presents the mean and the standard deviation values attained for each of these measures for the two segmentation algorithms tested. Algorithm I attains very high sensitivity values but they are correlated with very low specificity values. These results indicate that the cervix region is always located within the detected ROI but that large portions of the background are also included. Algorithm II, which was designed with the goal of obtaining a more accurate delineation of the cervix boundary, significantly reduces the amount of falsely detected regions. This comes at the expense of missing some of the cervix region pixels. The *accuracy* measure, which combines the detection quality of the cervix region and the background into a single measure, favors algorithm II as expected.

Table 3
Evaluation of two algorithms for cervix boundary detection using different performance measures.

	Algorithm I	Algorithm II
p	0.98 (0.03)	0.87 (0.1)
q	0.38 (0.22)	0.75 (0.19)
accuracy	0.65 (0.12)	0.8 (0.09)

Mean and standard deviation results for 636 images are presented (mean(std)).

Fig. 7 demonstrates the benefits of the *accuracy* over the *F-measure*, and the mean predictive value (*PV*), when more than two expert segmentations are available. In this example the three quality measures are used to assess the performance of the experts themselves, as compared to the STAPLE-generated, multi-expert ground truth. 10 expert segmentations are used in the comparison. The original cervigram with the different markings is presented in (a), and the ground truth segmentation generated by STAPLE is presented in (b). The *F*-measure, *PV* and *accuracy*, computed for each of the available expert segmentations, are listed on top of corresponding maps in (c, d and e), respectively. The maps are sorted in decreasing order of similarity to the STAPLE ground truth, beginning with the most similar map. The sorted maps in (c) and (d) demonstrate the tendency of the *F-measure* and the *PV* to favor larger segmentations. Note the second and third most similar maps in (c): These maps correspond to regions that were marked by single experts and are certainly not within the region of highest probability in the GT map. The sensitivity in these cases is high, as the maps include most of the ground truth region. The specificity is not low enough when compared to the other cases, due to its narrow range of values. Both the *F-measure* and the *PV* are strongly affected by the sensitivity, which leads to erroneous results. The sorted maps in (e), where the *accuracy* measure is used, have a ranking closer to our intuition. The most similar maps in this case correspond to the region of high probability within the GT map, where the level of agreement between the experts is high.

Please cite this article in press as: Gordon S, et al. Evaluation of uterine cervix segmentations using ground truth from multiple experts. Comput Med Imaging Graph (2009), doi:10.1016/j.compmedimag.2008.12.002

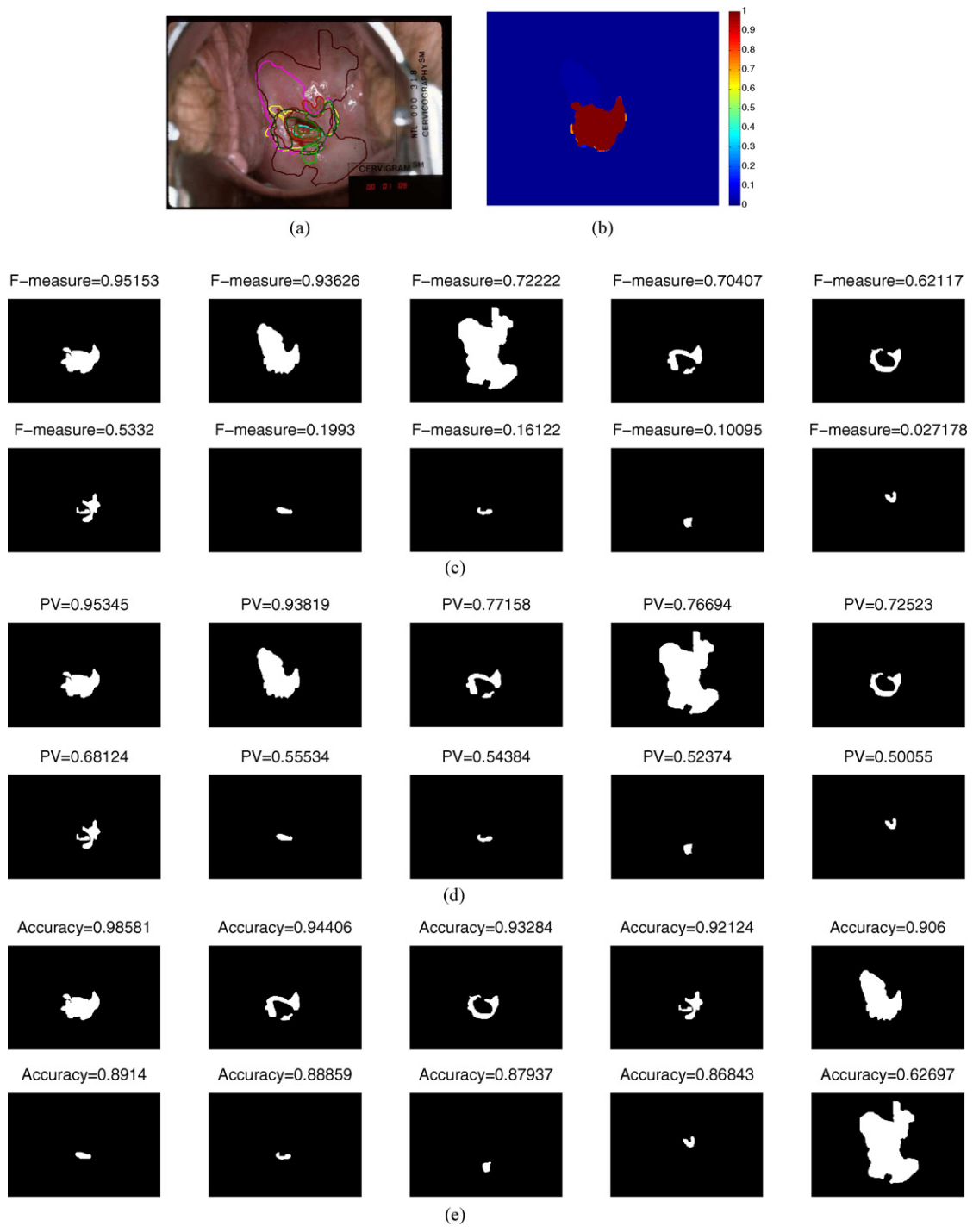

Fig. 7. (a) Markings of 10 experts imposed on original cervigram. (b) Multi-expert ground truth generated by STAPLE. (c, d, e) Sorted segmentation maps according to *F-measure*, *PV* and *a* ccuracy, respectively. The maps are sorted in decreasing order from most to least similar to the ground truth segmentation.

5. Conclusions

This work focuses on generating reliable multi-expert ground truth for the cervigram segmentation task. In addition several descriptors based on the output of the STAPLE algorithm are discussed, including the ESM and SSM descriptors that measure segmentation complexity of a single image, and the *accuracy* measure that evaluates the performance of automated segmentation algorithms as compared to the markings of multiple experts.

Our results demonstrate the correlation between (1) the ESM and SSM descriptors and (2) high levels of agreement among experts. We have demonstrated the superiority of these descriptors over alternatives such as simple entropy, standard deviation, and mean probability measures. The evaluation of segmentation complexity in the entropy-mean feature space was shown to be more accurate than the ESM descriptor, when trying to distinguish between different types of disagreement among the experts. The ability of the *accuracy* measure to evaluate the results of automated segmentation algorithms was demonstrated, and this measure was shown to provide a reliable evaluation that factors in the detection quality of the object as well as that of the background, without being too sensitive to their relative sizes. The ESM and the entropy-mean

Please cite this article in press as: Gordon S, et al. Evaluation of uterine cervix segmentations using ground truth from multiple experts. Comput Med Imaging Graph (2009), doi:10.1016/j.compmedimag.2008.12.002

feature space were used to characterize the complexity of segmenting acetowhite lesions versus segmenting cervix boundaries. In all of the presented experiments, the acetowhite segmentation was shown to be a more complex segmentation task, with a larger amount of disagreement among experts. This result can be explained by the fact that the acetowhite tissue may consist of multiple regions distributed across the cervix, the tissue is visually more difficult to detect, and it has less well-defined boundaries. The cervix region, on the other hand, is a single connected region that is clearly visible within the cervigram.

The task of automatic uterine cervix image analysis is in its preliminary stages. Detection and segmentation of cervigram tissues is very challenging due to the large diversity of the cervigram images within the database and the different artifacts present in the cervigrams. Tuning algorithms to the segmentation characteristics of a single expert would be unsatisfactory, due to the large multi-expert variability that exists. The complexity definition that we have proposed can be used in future tasks to classify a database into "simple" and "complex" images. This may aid in the performance evaluation and analysis (per complexity group) of automated segmentation algorithms being developed.

We also conducted an initial qualitative comparison between the visual appearance of the acetowhite lesions (described by a single expert) and the segmentation complexity specified by the clusters in the entropy-mean feature space. In this comparison a correlation was detected between lesions that are difficult to detect and images that are complex to segment (where the agreement between the experts was poor). In future work we plan a more thorough analysis of the correlation between segmentation complexity and other medical findings available in the NCI database. Finally, the focus of this paper is the cervigrams database, but the methods discussed here are general and can be applied to a variety of medical image archives and application domains.

Acknowledgement

We would like to thank Dr. Simon K. Warfield for his notes and support with the STAPLE implementation software. We also would like to acknowledge the medical expert contributions to this work from the National Institutes of Health/American Society for Colposcopy and Cervical Pathology (NIH-ASCCP) Research Group. This research was supported by the Intramural Research Program of the U.S. National Institutes of Health, National Library of Medicine, and Lister Hill National Center for Biomedical Communications.

References

[1] Cates JE, Lefohn AE, Whitaker RT. GIST: an interactive, GPU-based level set segmentation tool for 3D medical images. Medical Image Analysis 2004;8(3):217–31.
[2] Chalana V, Kim Y. A methodology for evaluation of boundary detection algorithms on medical images. IEEE Transactions on Medical Imaging 1997;16(5):642–52.
[3] Cuadra MB, Cammoun L, Butz T, Cuisenaire O, Thiran J-P. Comparison and validation of tissue modelization and statistical classification methods in T1-weighted MR brain images. IEEE Transactions on Medical Imaging 2005;24(12):1548–65.
[4] Dampster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B 1977;39(1):1–38.
[5] Fawcett T. ROC Graphs: notes and practical considerations for researchers. Technical Report HPL-2003-4, HP Labs, Palo Alto, CA; 2004.
[6] Fenster A, Chiu B. Evaluation of segmentation algorithms for medical imaging. In: Proceedings of Engineering in Medicine and Biology Society, IEEE-EMBS 2005. 2005. p. 7186–9.
[7] Gerig G, Jomier M, Chakos M. Valmet: A new validation tool for assessing and improving 3D object segmentation. Lecture Notes in Computer Science 2001;2208:516–28.
[8] Gordon S, Zimmerman G, Long R, Antani S, Jeronimo J, Greenspan H. Content analysis of uterine cervix images: initial steps towards content based indexing and retrieval of cervigrams. Proceedings of SPIE Medical Imaging 2006;6144:1549–56.
[9] Jeronimo J, Castle PE, Herrero R, Burk RD, Schiffman M. HPV testing and visual inspection for cervical cancer screening in resource-poor regions. International Journal of Gynecology and Obstetrics 2003;83:311–3.
[10] Jeronimo J, Long LR, Neve L, Bopf M, Antani S, Schiffman M. Digital tools for collecting data from cervigrams for research and training in colposcopy. Journal of Lower Genital Tract Disease 2006;10(1):16–25.
[11] Jeronimo J, Massad LS, Schiffman M, for the NIH/ASCCP Research Group. Visual appearance of the uterine cervix: correlation with human papillomavirus detection and type. American Journal of Obstetrics and Gynecology 2007;197(1):47.e1–8.
[12] Jomier J, LeDigarcher V, Aylward SR. Comparison of vessel segmentations using staple. In: Proceedings of MICCAI. 2005. p. 523–30.
[13] Josh S, Davis B, Homier M, Gehrig G. Unbiased diffeomorphic atlas construction for computational anatomy. NeuroImage 2004;23:151–60.
[14] Ladak H, Wang Y, Downey D, Fenster A. Testing and optimization of a semi-automatic prostate boundary segmentation algorithm using virtual operators. Medical Physics 2003;30(7):1637–47.
[15] Li H, Liu T, Young G, Guo L, Wong STC. Brain tissue segmentation based on DWI/DTI data. In: Proceedings of IEEE International Symposium on Biomedical Imaging (ISBI). 2006. p. 57–60.
[16] Lotenberg S, Gordon S, Long R, Antani S, Jeronimo J, Greenspan H. Automatic evaluation of uterine cervix segmentations. Proceedings of SPIE Medical Imaging 2007.
[17] Martin-Fernandez M, Bouix S, Ungar L, McCarely RW, Shenton ME. Two methods for validating brain tissue classifiers. In: Proceedings of MICCAI. 2005. p. 512–22.
[18] Mattern F, Rohlfing T, Denzler J. Adaptive performance-based classifier combination for generic object recognition. In: Proceedings of International Fall Workshop Vision, Modeling and Visualization (VMV). 2005. p. 139–46.
[19] Rohlfing T, Russakoff DB, Maurer Jr CR. Extraction and application of expert priors to combine multiple segmentations of human brain tissue. In: Proceedings of MICCAI. 2003. p. 578–85.
[20] Schiffman M, Adrianza ME. ACUS-LSIL triage study: design, methods and characteristics of trial participants. Acta Cytologica 2000;44(5):726–42.
[21] Udupa JK, LeBlanc VR, Zhuge Y, Imielinska C, Schmidt H, Currie LM, et al. A framework for evaluating image segmentation algorithms. Comuterized Medical Imaging and Graphics 2006;30:75–87.
[22] van Rijsbergen CJ. Information Retrieval. London: Butterworths; 1979.
[23] Warfield SK, Zou HK, Wells WM. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Transactions on Medical Imaging 2004;23(7):903–21.
[24] Warfield SK, Zou HK, Wells WM. Validation of image segmentation by estimating rater bias and variance. In: Proceedings of MICCAI. 2006. p. 839–47.
[25] Warfield S, Dengler J, Zaers J, Guttmann C, Wells W, Ettinger G, et al. Automatic identification of gray matter structures from MRI to improve the segmentation of white matter lesions. Journal of Image Guided Surgery 1995;1:326–38.
[26] Williams GW. Comparing the joint agreement of several raters with another rater. Biometrics 1976;32:619–27.
[27] Zimmerman G, Gordon S, Greenspan H. Automatic landmark detection in uterine cervix images for indexing in a content-retrieval system. In: Proceedings of IEEE International Symposium on Biomedical Imaging (ISBI). 2006. p. 1348–51.
[28] Zou KH, wells WM, Kikinis R, Warfield SK. Three validation metrics for automated probabilistic image segmentation of brain tumors. Statistics in Medicine 2004;23:1259–82.

Shiri Gordon was born in Tel-Aviv, Israel in 1971. She received a B.Sc. degree in mechanical engineering and a M.Sc. degree in electrical engineering both from Tel-Aviv University, Israel, in 1995 and 2002, respectively. She was a Ph.D. student at the Biomedical Engineering Department of the Faculty of Engineering, Tel-Aviv University, Israel, working with Dr. Hayit Greenspan. She is currently with Superfish Ltd. Her research interests include medical image processing and analysis, content-based image retrieval, statistical image modeling and segmentation, machine learning and information theory.

Shelly Lotenberg was born in Tel-Aviv, Israel in 1977. She received a B.A. degree in Computer Science in the Interdisciplinary center, Herzelia, Israel in 2002. She is currently a M.Sc. student at the Biomedical Engineering Department of the Faculty of Engineering, Tel-Aviv University, Israel, working with Dr. Hayit Greenspan. Her research interests include medical image processing and analysis, content-based image retrieval, statistical image modeling and segmentation.

Rodney Long was born in Crosbyton, TX, USA, in 1950. He received the M.A. degree in mathematics from the University of Texas in 1976 and M.A. in applied mathematics from the University of Maryland in 1985. Since 1990, he has been an electronics engineer for the Communications Engineering Branch at the National Library of Medicine. Prior to his current job, he worked for 14 years in industry as a software developer and as a systems engineer. His research interests are in telecommunications, image processing, and scientific/biomedical databases.

Sameer Antani earned his B.E. (Computer) degree from the University of Pune, India, in 1994, and his M.E. and Ph.D. degrees in Computer Science and Engineering from the Pennsylvania State University, USA, in 1998 and 2001, respectively. He is a Staff

Scientist with the Lister Hill National Center for Biomedical Communications, an intramural R&D division of the National Library of. His research interests are in image and text data management for large biomedical multimedia archives. Dr. Antani is a member of the IEEE, the IEEE Computer Society, and SPIE. He serves on the steering committee for IEEE Symposium for Computer Based Medical Systems (CBMS).

Jose Jeronimo was born in Peru in 1963. He received a medical degree from the Federico Villarreal University of Lima, Peru. From 2001 to January 2008 he worked at the Hormonal and Reproductive Epidemiological Branch at the National Cancer Institute (USA); and since February 2008 he works at the Program for Appropriate Technology in Health (PATH) in Seattle, USA. His research interests focus on cervical cancer prevention, early diagnosis and treatment.

Hayit Greenspan is an Associate Professor at the Biomedical Engineering Department at Tel-Aviv University, Israel. She received her B.S. and M.S. degrees in Electrical Engineering from the Technion- Israel Institute of Technology, in 1986 and 1989, respectively, and the Ph.D. degree in Electrical Engineering from CALTECH—California Institute of Technology, in 1994. She was a Postdoc with the Computer Science Division at U.C. Berkeley from 1995 to 1997. In 1997 she joined Tel-Aviv University and founded the Biomedical Image Processing lab, which she heads. Hayit is interested in analysis of biological and medical image data, content-based image and video search and retrieval, statistical image modeling and segmentation. Recent research areas include MRI brain analysis, Echo-Doppler analysis, uterine cervix cancer detection and medical image annotation and retrieval in large scale X-ray archives.

Please cite this article in press as: Gordon S, et al. Evaluation of uterine cervix segmentations using ground truth from multiple experts. Comput Med Imaging Graph (2009), doi:10.1016/j.compmedimag.2008.12.002