Automatic Evaluation of Uterine Cervix Segmentations

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

ABSTRACT

In this work we focus on the generation of reliable ground truth data for a large medical repository of digital cervicographic images (cervigrams) collected by the National Cancer Institute (NCI). This work is part of an ongoing effort conducted by NCI together with the National Library of Medicine (NLM) at the National Institutes of Health (NIH) to develop a web-based database of the digitized cervix images in order to study the evolution of lesions related to cervical cancer. As part of this effort, NCI has gathered twenty experts to manually segment a set of 933 cervigrams into regions of medical and anatomical interest. This process yields a set of images with multi-expert segmentations. The objectives of the current work are: 1) generate multi-expert ground truth and assess the difficulty of segmenting an image, 2) analyze observer variability in the multi-expert data, and 3) utilize the multi-expert ground truth to evaluate automatic segmentation algorithms. The work is based on STAPLE (Simultaneous Truth and Performance Level Estimation), which is a well known method to generate ground truth segmentation maps from multiple experts' observations. We have analyzed both intra- and inter-expert variability within the segmentation data. We propose novel measures of "segmentation complexity" by which we can automatically identify cervigrams that were found difficult to segment by the experts, based on their inter-observer variability. Finally, the results are used to assess our own automated algorithm for cervix boundary detection.

Keywords: Observer Performance Evaluation, Technology Assessment,Uterine cervix, Cervical cancer, Image segmentation and indexing, Medical image archives, STAPLE, F-measure

1. INTRODUCTION

The purpose of this work is to generate multi-expert ground truth segmentation data for a unique medical repository of digital cervicographic images, as well as to define a "segmentation complexity" measure and quantify the intra- and inter-expert variability. The multi-expert ground truth is very important for evaluating the complexity of the segmentation problem at hand and can be of great help in developing tools for automatic content analysis of cervigrams. When building an automated system for segmenting medical images, one encounters various questions regarding the ground truth and the assessment measurements which should be used. This problem is very common in the medical imaging field and is addressed here for cervigrams for the first time. The database we use, generated by NCI, contains a collection of 933 cervigrams each segmented by up to twenty medical experts. Most of the images were marked by a small number of experts with a focus on two important regions - the cervix boundaries and the acetowhite region (see Figure 1). The cervix boundary defines the region of medical and anatomical interest within the cervigram. The acetowhite region is a white-appearing epithelium that is visible for a short period of time following the application of acetic acid. Some acetowhite regions correlate with uterine cervix cancer progression, and thus are of clinical significance.

Address all correspondence to H. Greenspan, E-mail:hayit@eng.tau.ac.il

Medical Imaging 2007: Image Perception, Observer Performance, and Technology Assessment,
edited by Yulei Jiang, Berkman Sahiner, Proc. of SPIE Vol. 6515
65151J, (2007) · 1605-7422/07/$18 · doi: 10.1117/12.708889

Each image within the database possesses a different number of manual markings varying from one to twenty. As Figure 1 illustrates, for some images, a strong variability exists between the experts' segmentations of the relevant regions. There are "simple" images, in which most of the experts agree on the tissue boundaries (Figure 1(c),(g)) and "complex" images, where the experts have substantially differing markings which vary in size and location (Figure 1(a),(e)).

In this work, the multi-expert ground truth is estimated by applying STAPLE (Simultaneous Truth and Performance Level Estimation)[1] to the experts' segmentations. STAPLE is a well known method that generates ground truth segmentation maps from the observations of multiple experts and measures the performance levels of each of the experts. STAPLE has been used in the literature in varying applications, such as generating ground truth maps for MR brain images,[2] constructing a brain MRI atlas for two-year-old children,[?] and combining two-class maps to obtain a complete segmentation of a brain tissue.[3] It has also been used for object recognition.[4]

In the current work we use the STAPLE output to quantify the intra- and inter-expert variability within the cervigrams database. A "segmentation complexity" is defined and the results of an automated algorithm for cervix boundary detection are quantitatively evaluated. The STAPLE algorithm is briefly described in section 2. Computational measures extracted from STAPLE's output are presented in section 3. Experimental results using the cervigram archive are described in section 4.

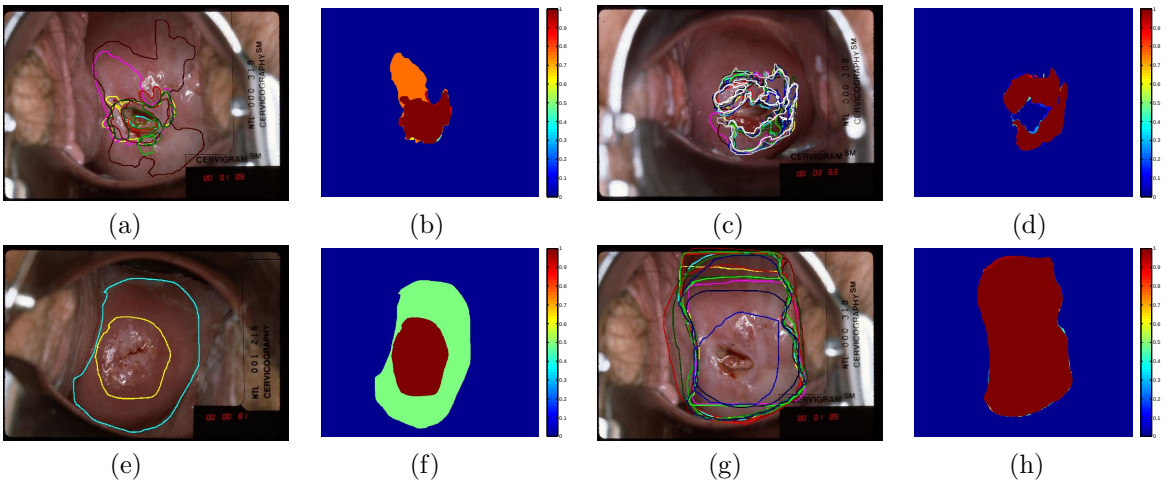


Figure 1. (a)(c)(e)(g) Examples of manually segmented cervigrams (different color per expert); (b)(d)(f)(h) Corresponding multi-expert ground truth, generated by STAPLE. Pixel probabilities are color-coded from blue (low) to red (high); (a),(c) acetowhite region; (e),(g) cervix boundary.

2. THE STAPLE ALGORITHM

The STAPLE algorithm[1] takes a collection of image segmentations as input. It then simultaneously computes: (1) a probabilistic estimate of the true segmentation and (2) a measure of the performance level represented by each segmentation. The algorithm is formulated as an instance of the expectation-maximization (EM) algorithm.[?] The performance levels, or quality achieved by each expert, are represented by 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 is computed as:

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

where f(T_i) is the prior probability for tissue 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:

$$a_i^{(k)} \equiv f(T_i = 1) \prod_j f(D_{ij}|T_i = 1, p_j^{(k)}, q_j^{(k)}) = f(T_i = 1) \prod_{j:D_{ij}=1} p_j^{(k)} \prod_{j:D_{ij}=0} (1 - p_j)^{(k)}, \tag{2}$$

$$b_i^{(k)} \equiv f(T_i = 0) \prod_j f(D_{ij}|T_i = 0, p_j^{(k)}, q_j^{(k)}) = f(T_i = 0) \prod_{j:D_{ij}=0} q_j^{(k)} \prod_{j:D_{ij}=1} (1 - q_j)^{(k)}, \tag{3}$$

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:

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

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

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

The sensitivity estimator, p_j, can be interpreted as the ratio of the j_{th} 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 the amount of agreement between the experts. This map is regarded as the ground truth segmentation generated by STAPLE. Figure 1 shows examples of computed multi-expert ground truths that correspond to the expert markings for both the acetowhite region (b, d) and the cervix boundary (f, h). Pixel probabilities are color-coded from blue (low probability) to red (high probability). The intersection of all experts markings is colored red as can be expected.

3. COMPUTATIONAL MEASURES

In this section we introduce a set of computational measures and analysis schemes, that utilize the STAPLE output to quantify the "segmentation complexity" (Section 3.1), quantify the intra- and inter-expert variability (Section 3.2) and quantitatively evaluate the results of an automated algorithm for cervix boundary detection (Section 3.3).

3.1. Measuring Segmentation-Complexity

The multi-expert ground truth generated by STAPLE is used to quantify the segmentation complexity of a single image. Inhomogeneity of pixel probabilities within the multi-expert ground truth reflects the amount of disagreement among the experts and, thus, the complexity of the segmentation problem: a lower amount of inhomogeneity is correlated with a simpler segmentation task. Different descriptors that reflect this inhomogeneity can be extracted from the multi-expert ground truth and used for the complexity evaluation. Following is a description of the set of descriptors that we investigated. In the definitions below, the multi-expert probabilistic ground truth is treated as a gray-level image I, with 256 levels of agreement among the experts. Pixel intensity is denoted by x_i. The descriptors are computed for the object area only ($x_i > 0$).

1. **Entropy:**

$$entropy(I) = -\sum_i p_i * log(p_i) \tag{6}$$

where p_i is the probability of gray level i, as defined by a histogram with 256 bins. An intuitive understanding of entropy relates to the amount of uncertainty in the segmentation: an image with only one gray level, as is the case when all the experts agree on the segmentation, has no uncertainty in it and its entropy is zero. Disagreement between the experts generates additional gray levels within the segmentation map this leads to broader distributions and to higher entropy values.

2. **Standard deviation (STD):**

$$std(I) = (\frac{1}{n-1}\sum_{i=1}^{n}(x_i - x_m)^2)^{\frac{1}{2}}, \tag{7}$$

where x_m is the average gray-level of the segmentation and n is the number of pixels in the image. The standard deviation reflects the distribution of the gray levels within the segmentation map around their mean. A low distribution of gray levels, reflected by a low standard deviation value, is correlated with a strong agreement between the experts and vise versa.

3. **Entropy or Standard Deviation Scaled by Mean -** While the entropy and the standard deviation represent the distribution of values, they don't represent their magnitude. This fact is especially important for our application, where two different images, that possess two different levels of agreement in their segmentation, can get the same entropy or standard deviation values. The distribution of gray-levels can be narrow but around low values of agreement, thus using the entropy or the standard deviation will generate low values and the actual strong disagreement that exist in the segmentation won't be reflected. A normalized set of measures, in which the entropy and the standard deviation are scaled by the inverse of the mean gray-value of the distribution, is used in this work in order to cope with such cases. We call these descriptors "entropy scaled by mean" (ESM) and "standard deviation scaled by mean" (SSM), respectively, and define them as follows:

$$ESM(I) = 100 * \frac{entropy(I)}{mean(I)^2}; \qquad SSM(I) = 100 * \frac{std(I)}{mean(I)^2}, \tag{8}$$

The dynamic range of the mean values is narrow and therefore to enable a better differentiation between the images we square the mean gray value.

4. **Relative Area of Disagreement -** an additional descriptor suggested in this work is the relative area of disagreement (RAD) defined as the ratio between the disagreement area, where the multi-expert ground truth possess varying gray-levels, and the object agreement area, where the ground truth gray-level is the highest (255). It represents the significance of the disagreement area. The relative area of disagreement will not distinguish between images in which it has the same size but a different distribution of gray-levels.

The complexity evaluation process itself is performed using two methods: In the first method a complexity threshold is learned from a training set of segmentations for each of the descriptors. The complexity of a new segmentation can be evaluated by using this threshold. In the second method the entropy and the mean of the segmentations within the training set are regarded as points in a two dimensional feature space and the images are clustered into groups with different complexity levels using these features. The complexity of a new image is next rated using a classification procedure.

Using the two methods described, we can quantify the complexity of a single image and define it as easy or difficult to segment. We can also analyze the variability of the segmentation complexity across all of the images within the database, thus evaluating the complexity of the database itself. Finally, we can characterize the complexity of segmenting different regions within the cervix.

3.2. Evaluating the Performance of a Single Expert

We use the STAPLE output of sensitivity and specificity to quantify the quality of a single expert's segmentation across multiple images. The sensitivity of an expert for a single image can be interpreted as the ratio of the expert's true positive detection to the total amount of object believed to be in the data by the multi-expert ground truth. High values of sensitivity indicate that the expert correctly identified most of the region agreed on by the different experts. The specificity can be interpreted similarly with respect to the background. High values of specificity indicate that the expert identified most of the background agreed on by the multi-expert ground truth. The mean performance levels of each expert across all the images measure that expert's tendency to mark specific areas, as compared to the other experts.

3.3. Evaluating Automatic Segmentation Results

Given a new segmentation map, we wish to evaluate it as compared 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 among the results of different algorithms. The analysis is based on the performance levels of the automated algorithm, which are evaluated in the current work using the following methods:

1. Computation of the sensitivity and specificity performance levels with respect to the multi-expert ground truth using Equation (5) (as suggested by Warfield[1]).
2. Considering the algorithm as an additional "expert" and using its segmentation results along with the manual markings of the experts during the learning phase. The sensitivity and specificity performance levels for the algorithm are then estimated as part of the STAPLE's output.

Having computed the sensitivity and specificity parameters for a specific image, a comparison can be made to the parameters of the human experts for that image, examining each of the parameters separately. In this work we suggest a more accurate analysis, by observing the two parameters simultaneously. The sensitivity and specificity of the algorithm and of the different experts are considered as points in a two dimensional feature space and their dispersion is examined. A better segmentation, as compared to the other experts or algorithms, is a one with higher sensitivity and specificity values. Cases where one parameter has a high value, while the other parameter has a very low value, can be easily detected. Such cases indicate on a segmentation that doesn't agree with the multi-expert ground truth. In order to computationally evaluate the segmentation quality, while reflecting the nature of this dispersion, we use the *F-measure*,[5] which is often used in information retrieval to combine recall and precision measures (i.e.[6]). The *F-measure* is defined as:

$$F = \frac{pq}{\alpha p + (1-\alpha)q}, \tag{9}$$

where $\alpha = 0.5$ in the current case. It captures the trade off between sensitivity, p, and specificity, q, as their weighted harmonic mean. The higher the value of the *F-measure* the more accurate is the segmentation at hand, as compared to the other experts (algorithms).

4. EXPERIMENT AND RESULTS

In this section we present experimental results using the suggested computational measures and analysis schemes. The STAPLE algorithm was applied to the 933 manually segmented cervigrams within the NCI database in order to evaluate the segmentation quality of two marked regions: the acetowhite region and the cervix boundaries. We used the STAPLE implementation available through the ITK toolkit*.

*http://www.itk.org/

4.1. Evaluation of the Segmentation-Complexity Measure

Trying to assess the relative effectiveness of the different descriptors to differentiate between simple and complex segmentations, we first investigate the correlation between the descriptor values and the level of agreement among the experts. The descriptors were computed for each of the images using the multi-expert ground truth generated by the STAPLE algorithm.

Figure 2 shows examples of the multi-expert ground truth segmentation for both the cervix boundary and the acetowhite region. The corresponding descriptors, computed for each of these examples, are listed. We make the following observations:

- The **Entropy** measures the scatter of the gray levels, without factoring in the gray-level magnitude. This results in cases such as Figure 2(g), where the entropy is very low but the disagreement between the two experts is very high.
- The **Standard deviation (STD)** has a similar deficiency: it accounts for the gray-level distribution around the mean, but not the mean itself. Like the entropy descriptor, the standard deviation will fail in cases such as Figure 2(g).
- The **Mean** descriptor presents the average gray-level of the segmentation map. A correlation can be detected between high mean values and a high degree of agreement between the experts but the mean value alone lacks the ability to differentiate between cases with similar mean and different distributions. The mean value is therefore used only to define the ESM and the SSM descriptors (in order to improve the differentiation ability of the entropy and the STD descriptors, respectively).
- The **Relative Area of Disagreement (RAD)** reflects the maximum amount of disagreement among the experts with respect to the entire disagreement area, but does not take into account the distribution within this area. Hence, a single bad segmentation can significantly affect its value. Figures 2(g) and (h) are two examples that reflect the importance of the gray levels distribution. Visually inspecting these examples, the disagreement between the experts in (h) is large compared to the one in (g), as there are only two segmentations in (g) and one of them can be regarded, for the sake of argument, as a bad segmentation. This fact is not reflected by the RAD descriptor, which is higher for (g). The other descriptors, that conceder the gray levels distribution, indicate on a stronger disagreement in (h), as expected.
- The **Entropy Scaled by Mean (ESM)**, which we propose, successfully differentiates among the different levels of agreement in all of the presented examples. From these examples it is evident that higher values of ESM indicate a stronger disagreement among the experts. Using both the entropy and the mean appears to more clearly capture the distribution of the expert disagreement, as compared to using entroy alone. The **STD Scaled by Mean (SSM)** attained similar results, with higher values of SSM corresponding to greater expert disagreement.

The results of this experiment suggest that the ESM and the SSM are good descriptors for segmentation complexity.

Next, we investigate the segmentation complexity using the two-dimensional feature space of mean and entropy. Low entropy and high mean values are correlated with easier segmentations, where expert agreement is strong. Figure 3 shows the distribution of the entropy and the mean descriptors for 100 ground-truth segmentations. Each of the 100 images was marked by two experts. The experiment was conducted for segmentation of (a) the cervix boundary and (b) the acetowhite region.

We can detect three main groups within the distribution of the acetowhite segmentations in Figure 3(b): group A - includes the low-entropy-low-mean segmentations; group B - includes the low-entropy-high-mean segmentations and group C -includes the mid-entropy-mid-mean segmentations. Group A corresponds to segmentations with strong disagreement among the experts, group B, to segmentations with strong agreement, and group C, to segmentations with an intermediate level of disagreement among the experts. Figure 4 shows segmentation examples drawn from each of these groups. The top row shows examples from group A, the middle row from group B and the bottom row from group C. The distinct clustering patterns for the three groups is clearly visible.

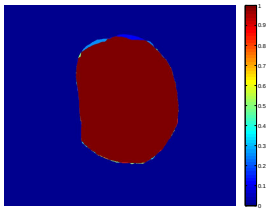
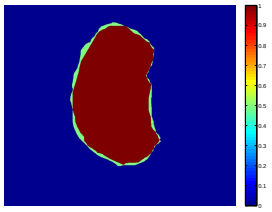
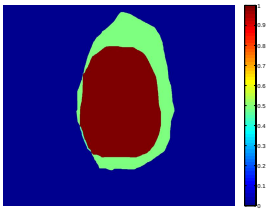
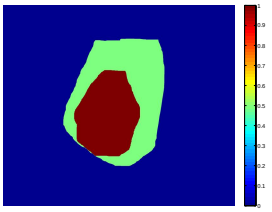
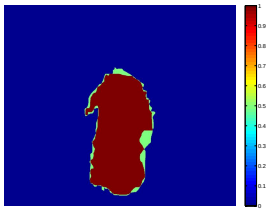
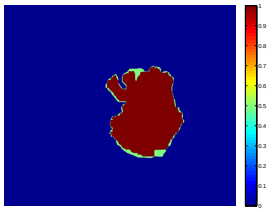
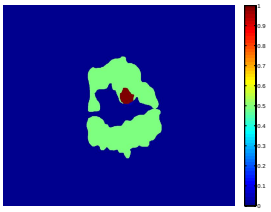
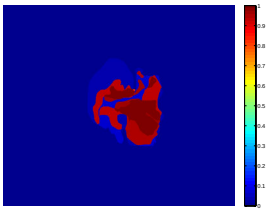
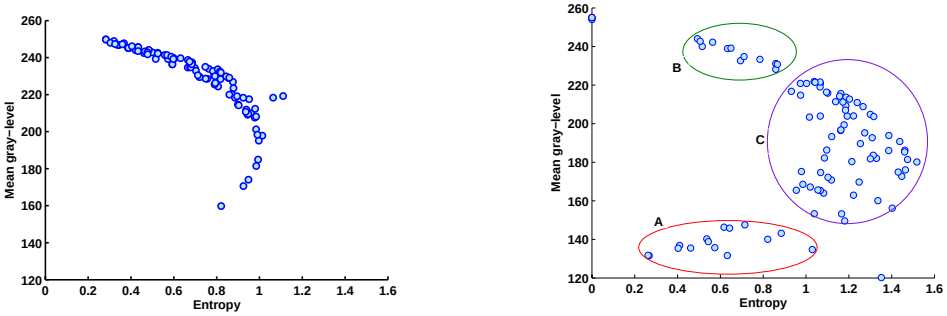
			
(a)	(b)	(c)	(d)
Entropy=0.419	Entropy=0.385	Entropy=0.96	Entropy=0.985
STD=28.353	STD=34.034	STD=61.16	STD=63.28
Mean=250	Mean=245	Mean=209	Mean=181
RAD=0.052	RAD=0.075	RAD=0.351	RAD=0.575
SSM=0.045	SSM=0.0566	SSM=0.939	SSM=0.192
ESM=0.0007	ESM=0.0006	ESM=0.002	ESM=0.003
			
(e)	(f)	(g)	(h)
Entropy=0.564	Entropy=0.493	Entropy=0.267	Entropy=2.746
STD=38.63	STD=36.04	STD=25.6	STD=115.47
Mean=242	Mean=244	Mean=131	Mean=147
RAD=0.099	RAD=0.085	RAD=0.961	RAD=0.846
SSM=0.066	SSM=0.06	SSM=0.148	SSM=0.529
ESM=0.0009	ESM=0.0008	ESM=0.002	ESM=0.013

Figure 2. Examples for multiple-expert ground truth data for the cervix boundary (top row) and the acetowhite region (bottom row). (a),(b),(e),(f): examples of agreement among experts; (c),(d),(g),(h): examples of disagreement among experts; Corresponding complexity descriptors are listed below each example.



(a) (b)

Figure 3. Entropy versus mean results for 100 images marked by two experts. (a) cervix boundary segmentation; (b) acetowhite region segmentation.

We compared these results with the well-known Dice and Sensitivity similarity measures, which are often used when only *two* segmentations are available, considering one of them as the reference or the ground truth. Given two binary segmentation maps, S and R, the Dice measure is defined as: $Dice = \frac{S \cap R}{S \cup R}$ and the Sensitivity as: $Sensitivity = \frac{S \cap R}{R}$, where R denotes the reference or ground truth segmentation. High values of Dice and Sensitivity indicate strong agreement between the two segmentation maps. The Dice and Sensitivity measures computed for the examples in Figure 4 are listed under the corresponding segmentation maps. Sensitivity was computed as the mean of two results, in which a different expert is considered as the ground truth. Images from group A have the lowest Dice and Sensitivity values, group B have the highest values and group C has intermediate values. These results are strongly correlated with the classes observed in the entropy-mean feature space.

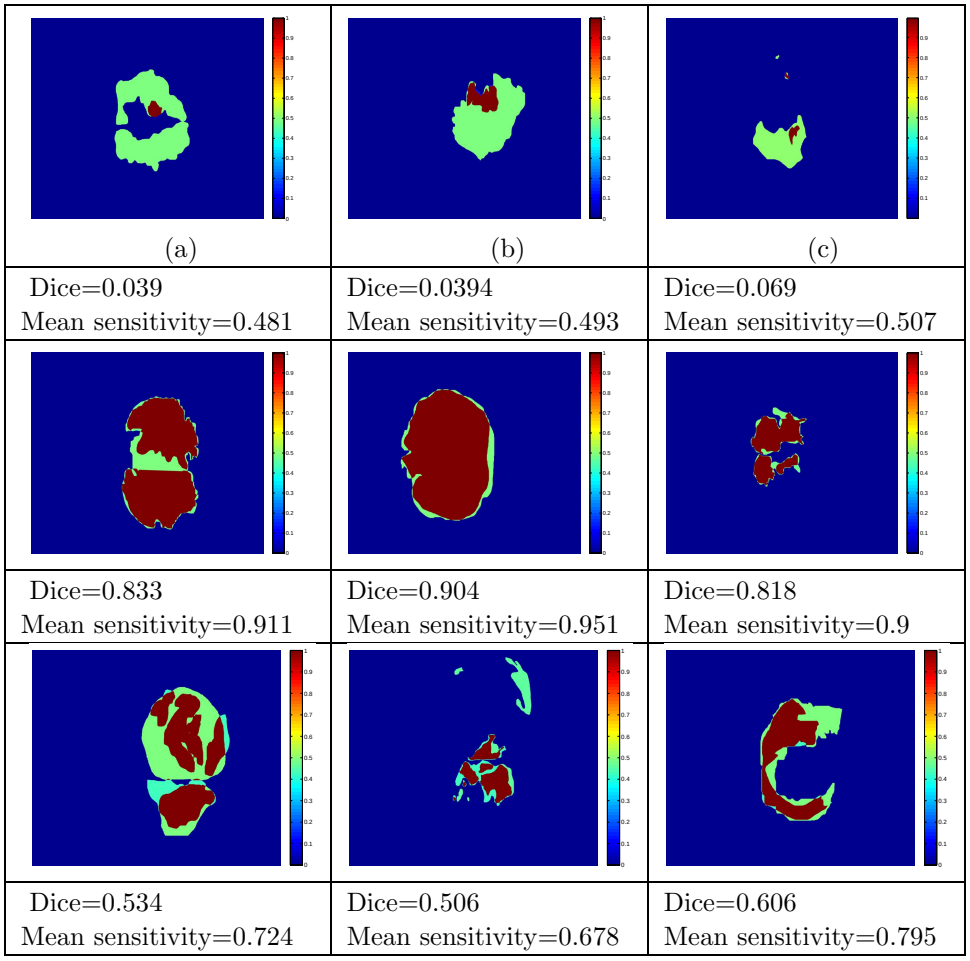
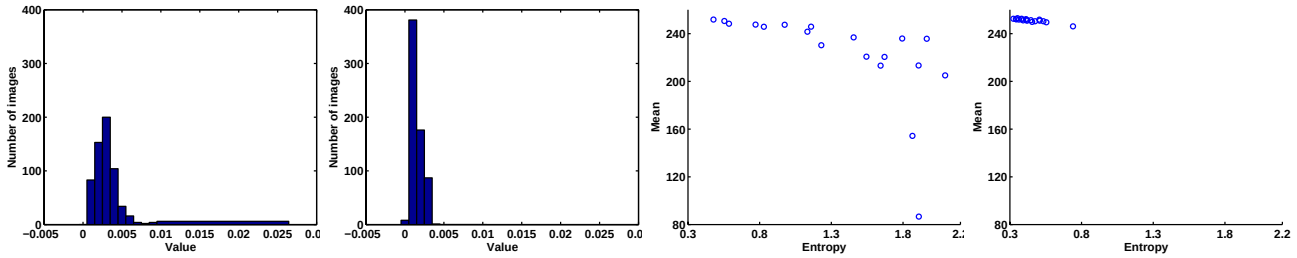


Figure 4. Example segmentations from the three groups of the entropy-mean feature space observed for the acetowhite region. Top row: group A - strong disagreement; Middle row: group B - strong agreement; bottom row: group C - some disagreement.

Finally, we compared the complexity of the two segmentation tasks, i.e., the complexity of segmenting the acetowhite regions, as compared to the complexity of segmenting the cervix boundary. Figure 5 displays the distribution of the ESM descriptor over the entire database for the acetowhite segmentation (a) and for the cervix boundary segmentation (b). 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, which makes the cervix boundary segmentation task the easier one. Figure 5(c) and (d) show the distribution of images that were segmented by

more than ten observers in the entropy-mean feature space . The cervix distribution (d) is narrow and has low entropy and high mean values, reflecting the strong agreement between the experts. The acetowhite distribution (c) is more spread out indicating on a larger disagreement and on a more complex segmentation task.



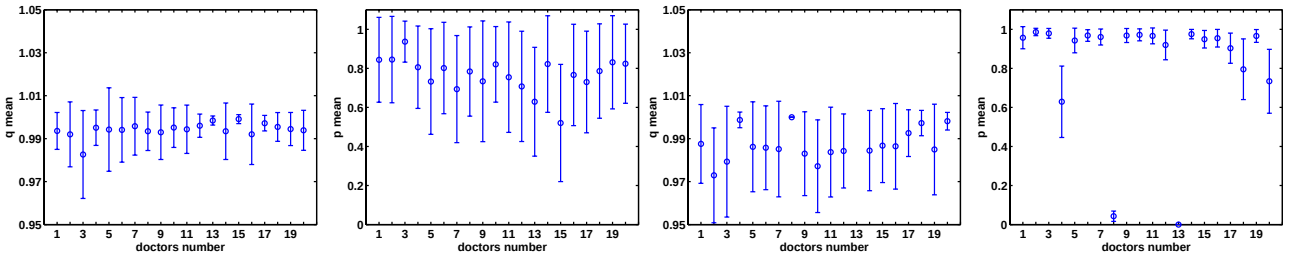
(a) (b) (c) (d)

Figure 5. (a)(b)-ESM distribution over 933 images: (a) for acetowhite ($mean = 0.003$); (b) for cervix boundary (mean = 0.001). (c)(d)-Distribution in the entropy-mean feature space for images with more than ten experts markings: (c) achetowhite region (16 images); (d) cervix boundary (20 images).

4.2. Evaluation of Single Expert Performance

Collecting all the images segmented by the same expert, we evaluate in the current experiment the expert's performance. STAPLE was initially used to compute the sensitivity and specificity performance levels for each of the images segmented by the expert. Their mean values were then computed across all the images segmented by that expert. Figure 6 shows the error bars (mean and standard deviation values) for the specificity and sensitivity computed for each expert (x-axis) for the tasks of acetowhite segmentation (a),(b) and cervix boundary segmentation (c),(d).

The sensitivity distribution has a larger variability than the specificity distribution because the relative disagreement area of the background is smaller than the relative disagreement area of the object (note the different scaling of the Y-axis in both cases). The sensitivity distribution of the cervix segmentation has a more uniform distribution as compared to the sensitivity distribution of the acetowhite segmentation (Figure 6(b) and (d)). This reflects the larger disagreement that is present between the experts in the case of acetowhite segmentation.



(a) (b) (c) (d)

Figure 6. Error bars for specificity and sensitivity computed per expert (X-axis). (a),(b) specificity and sensitivity, respectively, for the acetowhite segmentation; (c),(d) specificity and sensitivity, respectively, for the cervix boundary segmentation.

We may see segmentation characteristics of individual observers in these results. For example expert ♯15 in Figure 6(b) has the lowest mean sensitivity value, which reflects a poor agreement between his results and

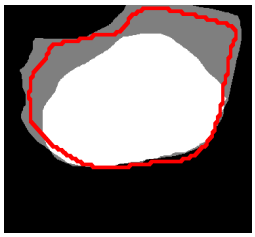
the multi-expert ground truth. Expert ♯3 has the highest mean sensitivity value meaning large portions of his markings agree with the ground truth. From Figure 6(a) we can see that expert ♯3 has also the lowest specificity value, this indicates that although this expert marked most of the acetowhite region as agreed on by the others, he also marked the largest portions from the background. Combining this two results we can conclude that this expert prefers to stay on the "safe side" and mark larger acetowhite regions at the expense of detecting some false positives.

A note should be made on the validity of the above analysis. In general, the experts in this experiment marked different images, and each expert a different number of images. In order to make a precise and fair comparison among them, the same images should be segmented by all, and those segmentations used for the evaluation. Having generated such a database the performance levels output by STAPLE algorithm can be used in an analysis similar to the above.

4.3. Evaluation of Automatic Cervix Boundary Segmentation

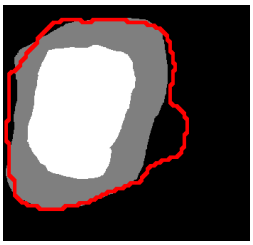
We used the STAPLE sensitivity and specificity parameters to evaluate the performance of an automated algorithm for segmentation of the cervix boundary. The algorithm tested is based on a new active contour functional that incorporates a local convexity feature.[7] The algorithm was devised especially for the purpose of segmenting the cervix boundaries. In this algorithm the standard edge indicators, based on image gradients, are replaced by edge indicators that are based on the cervix convexity. This is motivated by the fact that most of the cervix boundaries are outlined by folds of skin that resemble narrow valleys and are distinctively concave.

The sensitivity and specificity were initially computed with respect to the multi-expert ground truth, using Equation (5) and were treated independently. Figure 7 shows two segmentation results, where the contour of the automatic algorithm (marked by a red line) is superimposed on the multi-expert ground-truth (the segmentations of only two experts are available in these cases). The corresponding sensitivity, p, and specificity, q, values are given both for the experts and for the algorithm. In Figure 7(a), the sensitivity and specificity values of the algorithm fall between the experts' parameters. In this case, we may reasonably consider the algorithm to be another expert. In the second segmentation example, shown in Figure 7(b), the algorithm performance parameters fall between the experts' sensitivity values, but outside the range of the specificity values. Our interpretation is that the algorithm has detected the object better than one of the experts, but marked more background as compared to both of them.



Observer	q	p
1^{st} expert	0.945	0.998
2^{nd} expert	0.999	0.775
Automated algorithm	0.959	0.914

(a)



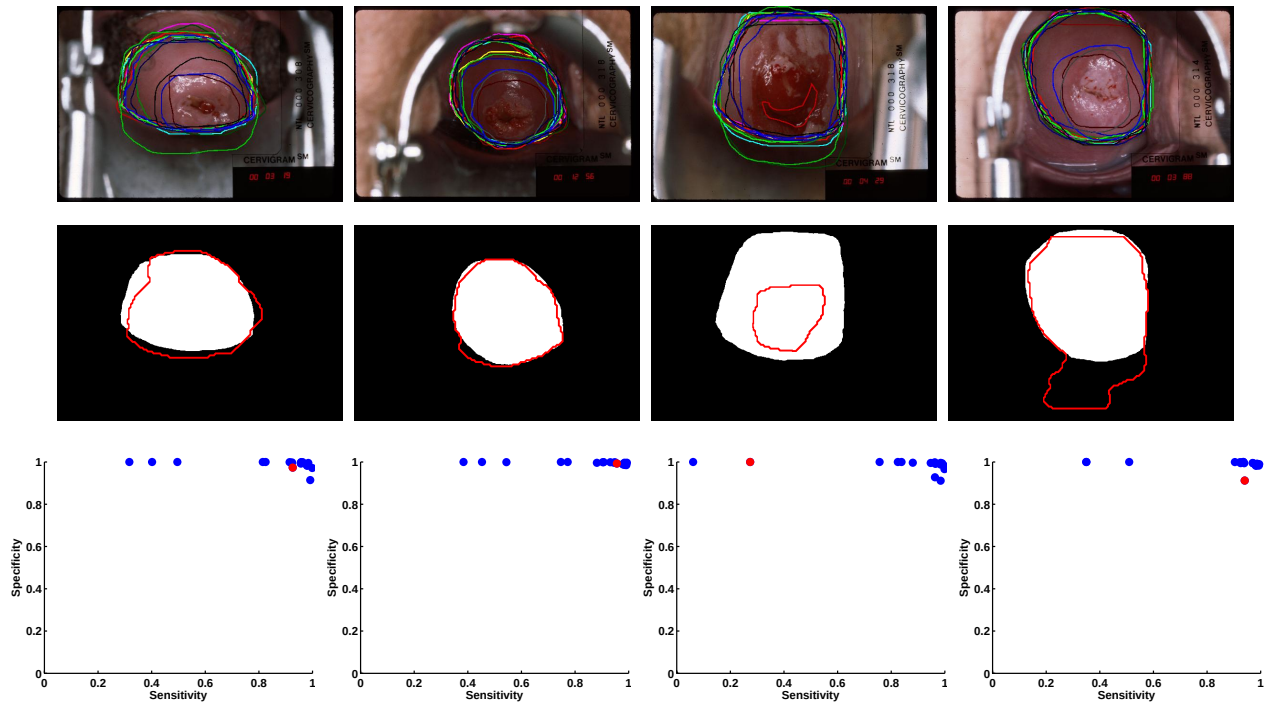
Observer	q	p
1^{st} expert	1	0.623
2^{nd} expert	0.937	1
Automated algorithm	0.908	0.959

(b)

Figure 7. Segmentation results of an automated algorithm (marked by a red line) imposed on the multi-expert ground truth. Sensitivity and specificity results for the experts and the automated algorithm are listed.

We next used the algorithm results as one of the input segmentations to the STAPLE and regarded it as an additional observer (expert). This kind of analysis was applied for images with more than ten segmentations; thus, the algorithm's result did not strongly influence the multi-expert ground truth and the performance levels. The sensitivity and specificity values assigned to each of the experts (including the algorithm) were plotted in a two-dimensional feature space and the corresponding *F-measure* was calculated. The best segmentation is the one with the highest *F-measure*. The results were ordered according to the increasing *F-measure* values and the position of the automated algorithm within this *F-measure* list was inspected. A few examples of the

sensitivity-specificity feature space are shown in Figure 8 (bottom row). Each example includes the multi-expert markings color-coded by expert on the original cervigram (top-row) and the multi-expert ground truth with the automated algorithm boundaries shown in red (middle-row). The *F-measure*, calculated for the algorithm performance and its location in the ordered list, are presented under corresponding results.



F-measure = 0.95 (12/18) (a)

F-measure = 0.97 (12/18) (b)

F-measure =0.43 (16/17) (c)

F-measure = 0.93 (15/18) (d)

Figure 8. Examples for evaluation of automatic segmentation results. Top-row - the experts marking on the original image (different colors for different experts). Middle row - multi-expert ground truth with the automated segmentation results imposed in red. Bottom row - distribution of the experts and the algorithm segmentation results in sensitivity (X-axis) and specificity (Y-axis) feature space. Experts marked in blue, the automated algorithm is marked in red. The *F-measure* result for the automated algorithm and its location in the ordered *F-measure* list, are listed under corresponding results.

The algorithm's performance parameters in Figure 8(a) are close to those of most of the experts, reflecting strong agreement between the algorithm and the experts. It has a better sensitivity value than five experts and a better specificity value than one expert. This fact is also reflected by the position of the algorithm in the ordered *F-measure* list. The result implies that the algorithm has detected the object region with high accuracy. Its sensitivity has a value of approximately 0.9, which implies 90 percent detection of the multi-expert ground truth. The specificity value is lower than most of the experts; only one expert have a lower value, meaning only one expert marked more background. The example in Figure 8(c) shows a smaller object marked by the algorithm as compared to the multi-expert ground truth. This fact is correlated with its low sensitivity value (approximately 0.25). Only one expert marked a smaller object area out of the multi-expert ground truth. Note that because the automated algorithm marked the object entirely within the multi-expert ground truth its specificity is unity (1), which illustrates that this parameter can be misinterpreted if used independently of sensitivity. Using the *F-measure* the algorithm is in position (16/17), which captures well the segmentation quality. In all of the presented examples the algorithm's *F-measure* values are within the range of the experts values.

5. CONCLUSIONS

This work concentrates on creating a reliable ground truth for the cervigram segmentation task. In addition, we have defined a quantitative measure of segmentation complexity by using STAPLE output and the ESM and SSM descriptors. Our work suggests that of the alternatives investigated, these two descriptors best represent disagreement among experts. We have proposed an additional method to define and evaluate segmentation complexity, using a two-dimensional entropy-mean feature space. We have also used (1) ESM and (2) entropy-mean feature space to characterize the complexity of segmenting acetowhite lesions versus segmenting cervix boundaries. In all the methods tested we concluded that the acetowhite segmentation is the more complex, and that the amount of disagreement among experts is large. This can be explained by the fact that, while the cervix is a single connected region, the acetowhite tissue may consist of multiple regions distributed across the cervix and, in addition, the acetowhite tissue is visually more difficult to detect and has less well-defined boundaries.

The task of automated cervical 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 propose can be used to classify a database into "simple" and "complex" images, which may aid evaluation of performance of automated algorithms per complexity group.

Evaluating an algorithm performance with respect to a multi-expert ground truth for a *small* number of experts can be accomplished by comparing the mean sensitivity and specificity values of the algorithm (computed using Equation 5) over a large set of images. These results can be used to compare among different algorithms in an objective manner. When a *large* number of experts is present, the two-dimensional sensitivity-specificity feature space can be used, along with the *F-measures*. The *F-measure* was shown to capture well the correlation between the two parameters and as such it provides an objective measure to evaluate the performance of both the experts and different algorithms.

ACKNOWLEDGMENTS

We would like to thank Dr. Simon K. Warfield for his support with the STAPLE implementation software. 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. S. K. Warfield, H. K. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation," *IEEE Transactions on Medical Imaging* **23**(7), pp. 903–921, 2004.
2. M. B. Cuadra, L. Cammoun, T. Butz, O. Cuisenaire, and J.-P. Thiran, "Comparison and validation of tissue modelization and statistical classification methods in T1-weighted MR brain images," *IEEE Transactions On Medical Imaging* **24**(12), pp. 1548–1565, 2005.
3. H. Li, T. Liu, G. Young, L. Guo, and S. T. C. Wong, "Brain tissue segmentation based on DWI/DTI data," in *Proc. of IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 57–60, 2006.
4. F. Mattern, T. Rohlfing, and J. Denzler, "Adaptive performance-based classifier combination for generic object recognition," in *Proc. of International Fall Workshop Vision, Modeling and Visualization (VMV)*, pp. 139–146, 2005.
5. C. J. van Rijsbergen, *Information Retrieval*, Butterworths, London, 1979.
6. D. R. Martin, C. C. Fowlkes, and J. Malik, "Learning to detect natural image boundaries using local brightness, color, and texture cues," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **26**(5), pp. 530– 549, 2004.
7. G. Zimmerman, S. Gordon, and H. Greensand, "Automatic landmark detection in uterine cervix images for indexing in a content-retrieval system," in *Proc. of IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1348–1351, 2006.