A Methodology for Evaluation of Boundary Detection Algorithms on Medical Images

Vikram Chalana and Yongmin Kim,* *Fellow, IEEE*

***Abstract*— Image segmentation is the partition of an image into a set of nonoverlapping regions whose union is the entire image. The image is decomposed into meaningful parts which are uniform with respect to certain characteristics, such as gray level or texture. In this paper, we propose a methodology for evaluating medical image segmentation algorithms wherein the only information available is boundaries outlined by multiple expert observers. In this case, the results of the segmentation algorithm can be evaluated against the multiple observers' outlines. We have derived statistics to enable us to find whether the computer-generated boundaries agree with the observers' hand-outlined boundaries as much as the different observers agree with each other. We illustrate the use of this methodology by evaluating image segmentation algorithms on two different applications in ultrasound imaging. In the first application, we attempt to find the epicardial and endocardial boundaries from cardiac ultrasound images, and in the second application, our goal is to find the fetal skull and abdomen boundaries from prenatal ultrasound images.**

***Index Terms*— Average polygon, boundary detection, evaluation, gold-standard, image segmentation, polygon metrics, validation.**

I. Introduction

RESEARCHERS in the area of medical image analysis have long sought to extract contours of different body organs and tissue types from medical images of various modalities. We believe that objective evaluation of these medical image segmentation algorithms on a large set of clinical data is one of the important steps toward establishing validity and clinical applicability of an algorithm. However, very few medical image segmentation researchers have carried out such an evaluation of their algorithms on a large number of clinical data sets, e.g., [1]–[6]. Many researchers have compared their algorithms on phantoms [7] or *in-vitro* studies [8], [9], which are typically idealistic representations of real data. Predicting performance on real data, based on such results, may be difficult. Even those who have evaluated their algorithms on real clinical data have used different criteria and different statistics [10]–[13], making it difficult to compare the performance of their algorithms against other algorithms. Thus, there is a compelling need for validation and comparison of medical image segmentation algorithms (as well as any other medical imaging algorithms) using standardized protocols. In this paper, we propose a methodology for evaluation and comparison of boundary detection algorithms for medical image segmentation.

Ideally, if the expected result were known, we would have the gold-standard segmentation and we may compare the segmentation output from the computer with this gold standard. This comparison would be done for a large number of clinical data sets to test the hypothesis that the segmentation output from the computer is not statistically different from the gold standard. The reasons that make such a medical image segmentation evaluation task difficult are as follows.

- Lack of a definitive gold standard. In medical image segmentation, typically, the only standards available for comparison are segmentations produced by expert observers. Such segmentations cannot be considered as gold standards because they may suffer from observer bias and inter- and intraobserver variability.
- Difficultly in defining a metric. A well-defined metric is needed to compare the computer-generated segmentation results to the segmentation results produced by expert observers. Such metrics are difficult to define for image segmentation because of the complex multidimensional nature of segmentation data.
- Lack of standardized statistical protocols. Summarizing the results and making conclusions about the algorithm performance require statistical analysis using standard protocols. Due to the lack of a gold standard and difficulty in defining metrics, defining such a standard statistical protocol is difficult.
- Tedious and time-consuming data collection. Collecting a large number of data sets with user-defined segmentation results is a very difficult task. For the expert observers, hand-segmenting the images is tedious and time consuming.

Other researchers have addressed some of the issues identified above for medical image segmentation evaluation, but no research group has addressed all the issues. For defining the evaluation metric for boundary detection, most researchers have used parameters derived from the boundaries, such as, area or perimeter of the boundaries for comparison [11], [13]. A few researchers have also used metrics based on distances between boundaries for their evaluation [3], [10]. DeGraaf *et al.* [14] used a metric based on the number of edit operations to perform on the segmentation results. Hammoude [15] used

Manuscript received October 28, 1996; revised May 8, 1997. This work was supported in part by a grant from Siemens Medical Systems, Ultrasound Group, Issaquah, WA. The Associate Editor responsible for coordinating the review of this paper and recommending its publication was N. Ayache. *Asterisk indicates corresponding author.*

V. Chalana is with the MathSoft Data Analysis Products Division, Seattle, WA 98109-3044 USA.

*Y. Kim is with the Department of Electrical Engineering, University of Seattle, Box 352500, Seattle, WA 98195–2500 USA (e-mail: kim@ee.washington.edu).

Publisher Item Identifier S 0278-0062(97)07586-1.

an evaluation method based on a pixel-by-pixel comparison of pixels enclosed by two different boundaries.

For addressing the problem of interobserver variability, Detmer *et al.* [10] have independently evaluated the interobserver variability of the hand-segmentation process, but they were not able to use these results in their statistical evaluation procedure. Most other researchers who have validated their segmentation techniques on relatively large clinical data sets have ignored the intra- or interobserver variability by comparing the computer-generated boundaries to only one observer's hand-outlined boundaries [11], [13].

In this paper, we address the first three problems identified above for a particular case of image segmentation—boundary detection of a single object from an image. First, we propose a metric to measure the distance between a computer-generated boundary and a hand-outlined boundary. Next, we propose a method based on averaging multiple expert observers' outlines to generate a gold-standard boundary. Next, we propose a few statistical methods for validating the computer-generated boundaries against boundaries outlined by expert observers. Finally, to provide concrete examples of the application of these techniques, we will consider its application in two different domains, both related to finding organ boundaries from ultrasound images. In the first application, we will consider a computer algorithm for detecting endocardial and epicardial contours on short-axis cardiac ultrasound images [16]. In the second application, we will consider an algorithm for detecting the boundaries of the fetal head and abdomen from prenatal ultrasound images that generates automatic measurements of diagnostically important parameters, such as the biparietal diameter (BPD), the head circumference (HC) and abdomen circumference (AC) [17].

We believe that comparing the result of the computer segmentation to only one observer's outline may not be sufficient, because a single observer's boundary may be subject to the observer's bias and intra- and interobserver variability. The methodology that we propose here compares the computer-generated boundaries to the multiple expert observers' boundaries to check whether the computer-generated boundaries differ from the manually outlined boundaries as much as the manually outlined boundaries differ from one another.

The methodology proposed in this paper will not only be useful for evaluation of individual image segmentation algorithms, but will also be applicable for comparison of different boundary detection algorithms (or the same algorithm with different parameters) for the same application. In much of the research in the field of medical image segmentation, performance of new algorithms is usually reported on data sets that are limited in size and scope, without systematic comparison to existing or commonly known algorithms, and without studies revealing the types of data on which proposed methods are effective, e.g., [1]–[6] and [10]–[13]. Establishment of standardized evaluation protocols along with a database of images for different applications may help alleviate some of these problems. The data collection problem is not addressed in this research; however, we believe that this problem can be solved by a multicenter research initiative to set up a large database of medical images of different modalities stored with the various hand-segmented images. This methodology may also prove useful for systematic selection of algorithms or the parameters of one algorithm for a particular application.

II. Methods

A. Error Metric

The first decision for image segmentation evaluation is to choose a parameter to be compared. Parameters derived from the boundaries, such as the area enclosed or the perimeter, may be compared or the boundaries themselves may be compared directly. The derived parameters are usually application dependent, and often the accuracy of measuring these parameters is the functional goal of image segmentation. In our cardiac boundary detection application, for example, the derived parameters that need to be compared are the areas enclosed by the end-diastolic (ED) and end-systolic (ES) boundaries. In our fetal head detection application, two parameters of clinical interest are the HC and the BPD of the skull. For comparing the derived parameters, we use the absolute difference between the computer-generated parameter value and the user-measured parameter value as our distance metric. We have found cases where these computer-generated parameters agreed well with the manually measured parameters, but the boundaries from which these measurements were derived did not agree as well. Hence, comparing the boundaries directly will provide a more stringent evaluation of the segmentation scheme.

We define a metric to measure the distance, $e(\mathcal{A}, \mathcal{B})$, between the two given curves, $\mathcal{A}$ and $\mathcal{B}$. When common biological landmarks are available on the two curves, establishing correspondence between the curves is straightforward. However, in the absence of landmarks, we first have to establish artificial correspondence between points on the two curves and then measure the distance between the corresponding points. If the two curves are represented as sets of points $\mathcal{A} = \{\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_m\}$ and $\mathcal{B} = \{\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_m\}$, where each $\mathbf{a}_i$ and $\mathbf{b}_i$ is an ordered pair of the x and y coordinates of a point on the curve, we define the distance to the closest point (DCP) for $\mathbf{a}_i$ to the curve $\mathcal{B}$ as

$$d(\mathbf{a}_i, \mathcal{B}) = \min_j \|\mathbf{b}_j - \mathbf{a}_i\|. \tag{1}$$

The Hausdorff distance between the two curves is defined as the maximum of the DCP's between the two curves [18]

$$e(\mathcal{A}, \mathcal{B}) = \max\left(\max_i \{d(\mathbf{a}_i, \mathcal{B})\}, \max_j \{d(\mathbf{b}_j, \mathcal{A})\}\right). \tag{2}$$

The closest point distance associates each point on both curves to a point on the other curve, and the Hausdorff distance finds the largest distance between the associated points. Fig. 1 shows two examples of the Hausdorff distance between two curves.

A distance between the two curves is a metric if it satisfies the following three properties:

1) $e(\mathcal{A}, \mathcal{A}) = 0$ and $e(\mathcal{A}, \mathcal{B}) \geq 0$;
2) $e(\mathcal{A}, \mathcal{B}) = e(\mathcal{B}, \mathcal{A})$;
3) $e(\mathcal{A}, \mathcal{C}) \leq e(\mathcal{A}, \mathcal{B}) + e(\mathcal{B}, \mathcal{C})$.

It is easy to see that the Hausdorff distance satisfies the first two properties. The third property, which is the triangle inequality, is also satisfied as shown in the Appendix.

Other metrics can also be defined to measure distances between boundaries. One such metric is the root mean squared (rms) radial distance between boundaries. This distance is defined by first choosing a common centroid of the two boundaries from which radial lines are drawn projecting outward. The intersection of these radial lines with the two curves define the corresponding points, and the metric is defined as the rms distance between all such points. Such a measure has been previously used for evaluation of ultrasound boundary detection techniques [3], [10]. The deficiency of this method is that the boundaries are assumed to be star shaped, i.e., each point of the boundary is visible from the centroid. Whereas this assumption is satisfied for simple-shaped boundaries, it fails for more complex-shaped boundaries. Another metric which can be used to measure the distance between two boundaries is to do a pixel-by-pixel comparison of those pixels enclosed by the two different boundaries. First, binary images are constructed for each boundary, wherein a pixel is nonzero if it is inside a boundary and zero if it is outside. Next, a pixel-wise XOR operation is performed on the two images and the average number of nonzero pixels in the resulting image defines a metric for the two boundaries. Hammoude [15] used this method for evaluating an ultrasound image segmentation method. The only shortcoming of this metric is that it is computationally intensive.

B. Averaging Curves

We now define a procedure to evaluate an average curve, given two or more curves. This procedure is based on establishing one-to-one correspondence between the points constituting two or more curves using a modification of the methods for shape registration described by Sampson *et al.* [19] and Besl and McKay [20]. In the absence of a gold-standard contour, the average of the multiple observers' curves can be used as a gold standard. Other researchers have used a similar averaging procedure to establish a gold-standard contour [21]. Their approach is based on a shape-based interpolation method which is very similar to our method proposed below, except that our method explicitly establishes correspondence between the curves to be averaged. Such point-wise correspondence has many advantages as will be discussed below.

Given a set of M curves $\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_M$, each with m equidistant points, we need to find the average curve $\mathcal{Y}$. We establish the initial single-point correspondence by choosing a point $\mathbf{x}_{11}$ at random on $\mathcal{X}_1$ and finding a point closest to $\mathbf{x}_{11}$ on each curve $\mathcal{X}_2, \mathcal{X}_3, \cdots, \mathcal{X}_M$. We denote these points by $\mathbf{x}_{21}, \mathbf{x}_{31}, \cdots, \mathbf{x}_{M1}$. For the remaining $m-1$ points on the curve, the correspondence is established sequentially, i.e., the point $\mathbf{x}_{12}$ on curve $\mathcal{X}_1$ corresponds to points $\mathbf{x}_{22}, \mathbf{x}_{32}, \cdots, \mathbf{x}_{M2}$ on curves $\mathcal{X}_2, \mathcal{X}_3, \cdots, \mathcal{X}_M$, respectively. A point on the average curve is given by the centroid of the M corresponding points

$$\mathbf{y}_i = \frac{1}{M} \sum_{j=1}^{M} \mathbf{x}_{ji} \tag{3}$$

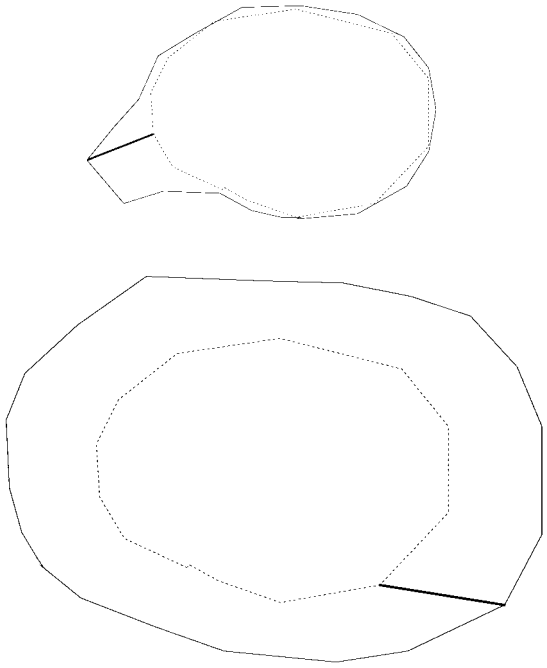


Fig. 1. Two examples illustrating the Hausdorff distance between the two curves. The Hausdorff distance is measured along the thick line. The Hausdorff distance in both examples is about 35 pixels, whereas the mean distance between corresponding points in the first example is about ten pixels and in the second example is about 30 pixels.

for each $i = 1, 2, \cdots, m$. Next, from each point on $\mathcal{Y}$, a normal to the curve at that point is drawn and the intersection of this normal with each of the M input curves is determined. These points of intersection define another set of correspondence between the M input curves. This new correspondence is averaged again to give a new average curve. The process is iterated until the average curve does not change any more. Typical averaging procedures converge in less than five iterations. An example of the averaging procedure is shown in Fig. 2. The normal to a point on a digital curve is computed by an efficient procedure based on computing the eigenvectors of a 2×2 scatter matrix described by Anderson and Bezdek [22].

This averaging procedure establishes correspondence between two curves. The average distance between the corresponding points may also be used to measure the difference between the two given curves. This distance is different from the Hausdorff distance defined in the last section. Whereas the Hausdorff distance measures the distance between points that differ the most, this distance finds the mean distance between the two curves. In Fig. 1, the Hausdorff distances for the two examples are approximately the same; however, the average distances in the two examples are very different. Although the average distance between two curves, computed this way, is not a valid metric (it does not satisfy the triangle inequality), it is still a useful distance measure that we will use.

Correspondence points between the two curves also allow us to analyze the regional difference between the two curves. For example, we can compute the signed distances between two curves to estimate the bias of an algorithm. Establishing

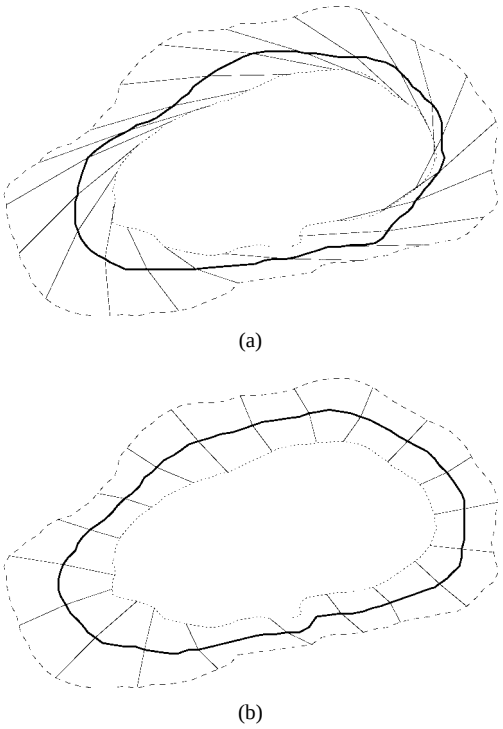


Fig. 2. Averaging of the two curves. (a) Shows the initial correspondence and the average curve (thick line) using this correspondence and (b) the final correspondence and the average curve (thick line) obtained after five iterations.

correspondence points also allows us to weigh different regions of the curves differently while computing the distance measure; e.g., some regions in the curve are detected more accurately and hence may be given larger weights.

C. Statistical Evaluation

For our statistical evaluation, we assume that multiple expert observers' hand-outlined boundaries are available for evaluating the computer-generated boundaries. Our methodology compares the computer-generated boundaries to the multiple expert observers' boundaries to check whether the computer-generated boundaries differ from the manually outlined boundaries as much as the manually outlined boundaries differ from one another.

We used two kinds of statistics to compare the computer-generated boundaries to the multiple hand-outlined boundaries. The first method is a modified version of Williams Index [23] which computes the ratio between the average computer-to-observer agreement and the average interobserver agreement. The second method computes the percentage of observations for which the computer-generated boundary lies within the interobserver range.

1) Williams Index: Williams [23] proposed a method to compare the agreement of an observer with the joint agreement of other observers; thus, helping to answer the question: "Does the individual observer agree with the set of observers as often as a member of the set agrees with another member of the set?" This index was originally defined only for nominal data. We have extended the Williams index to numeric multivariate data.

If $(n+1)$ observers numbered from 0 to n make measurements (or ratings) on N subjects (or images), this statistic aims to compare observer 0 with the reference group of n observers. In our case, the observers 1 through n are the expert observers who manually outline the boundaries on each of the N images and the observer 0 is the computer algorithm which automatically produces a boundary on each of the N images.

First, the proportion of agreements for each pair of observers j, j' is computed, where j and j' index the observers from 0 to n. Let us denote this proportion by $P_{j,j'}$. The average level of agreements between the observer 0 and the reference group of observers is computed by

$$P_0 = \frac{1}{n}\sum_{j=1}^{n} P_{0,j} \tag{4}$$

and the average level of agreements between the n reference observers by

$$P_n = \frac{2}{n(n-1)}\sum_{j}\sum_{j':j'\neq j} P_{j,j'}. \tag{5}$$

The index

$$I = \frac{P_0}{P_n} \tag{6}$$

is used for comparing the observer 0 to the group of n observers. If the upper limit of the confidence interval (CI) of this index is greater than the value one, we can conclude that the measurement data are consistent with the hypothesis that the individual observer agrees with the group at least as well as the group members agree with each other (i.e., the individual observer is a reliable member of the group).

To generalize the Williams index to numeric as well as to multivariate data, we redefine the proportion of agreements between two observers to be equal to the reciprocal of the average disagreements, $D_{j,j'}$, between the two observers j and j'

$$P_{j,j'} = \frac{1}{D_{j,j'}} \tag{7}$$

and the average disagreement between the two observers is defined

$$D_{j,j'} = \frac{1}{N}\sum_{i=1}^{N} e(\mathbf{x}_{ij}, \mathbf{x}_{ij'}) \tag{8}$$

where $\mathbf{x}_{ij}$ is a vector observation on subject i by an observer j, N is the number of subjects, and the function $e(\mathbf{x}, \mathbf{y})$ is a distance metric between two observations, $\mathbf{x}$ and $\mathbf{y}$.

Using this new definition of the proportion of agreements, we define the modified Williams index

$$I' = \frac{\frac{1}{n}\sum_{j=1}^{n}\frac{1}{D_{0,j}}}{\frac{2}{n(n-1)}\sum_{j}\sum_{j':j'\neq j}\frac{1}{D_{j,j'}}}. \tag{9}$$

With this modification, we can now compute an index similar to the Williams index for any kind of numeric multivariate data where we can define a distance metric $e(\cdot, \cdot)$.

The CI for this index is estimated using a jackknife nonparametric sampling technique [24]. This sampling procedure works by leaving out one of the N observations at a time and computing the Williams index for $N-1$ observations. Denote the data set with the ith observation removed by $\mathbf{X}_{(i)}$. We have N such data sets $\mathbf{X}_{(1)}$, $\mathbf{X}_{(2)}$, $\cdots$, $\mathbf{X}_{(N)}$ and N estimates of the modified Williams index, $I'_{(1)}$, $I'_{(2)}$, $\cdots$, $I'_{(N)}$. The jackknife estimate of the standard error in the computation of the modified Williams index is given by

$$se = \left\{ \frac{1}{N-1} \sum_{i=1}^{N} [I'_{(i)} - I'_{(\cdot)}]^2 \right\}^{1/2} \tag{10}$$

where

$$I'_{(\cdot)} = \frac{1}{N} \sum_{i=1}^{N} I'_{(i)}. \tag{11}$$

Thus, the 95% CI for the estimate of the modified Williams index is

$$I'_{(\cdot)} \pm z_{0.95}\, se \tag{12}$$

where $z_{0.95} = 1.96$ is the 95th percentile of the standard normal distribution.

This modified Williams index may be used both for evaluating the direct distances between the computer-generated and manually outlined boundaries and for evaluating the differences between the parameters derived from the boundaries, such as the area or the perimeter.

2) Percent Statistic: Our second statistical technique computes the percentage of cases for which the computer-generated boundary (or measurement) lies within the interobserver range. For parameters derived from the boundaries, this percentage can be easily computed. For each observation, we just have to test whether the computer-generated measurement is less than the largest and greater than the smallest observer-made measurement. For multidimensional measurements, such as the boundaries, establishing whether the computer-generated boundary lies within the interobserver range is not straightforward. The key question is how to define what it means to be inside the interobserver range. We define a computer-generated boundary to be within the interobserver range if it lies within a multidimensional convex polyhedron formed by the observer-outlined boundaries. Fig. 3 illustrates this concept in a simplified way. The manual outlines are points in a $2m$-dimensional Euclidean space which are represented in the figure as points in a two-dimensional space. If the convex hull of the points representing the manual outlines includes the point representing the computer-generated boundary, we say that the computer-generated boundary is within the interobserver range. There are other possible ways to establish whether the computer-generated boundary lies within the interobserver range. One method is to first establish correspondence points between all observer-outlined curves and the computer-generated curve. Next, we can count the number of points on the computer-generated curve lying between the corresponding observer-curve points. If a majority of the computer-curve points are within the observer-curves points, we can say that the computer-generated curve lies within the interobserver range.

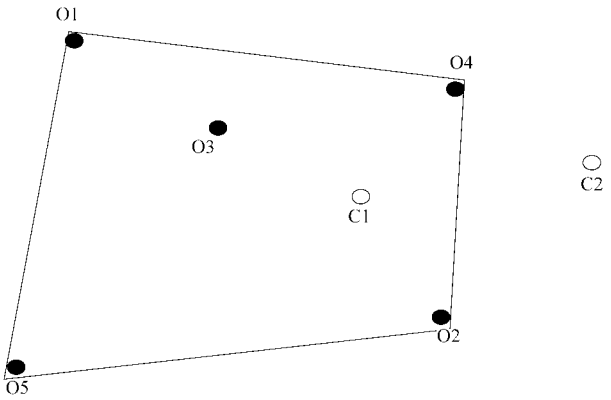


Fig. 3. Five observers' and two computer-generated boundaries (or measurements) collapsed onto a two-dimensional plane. The points O1, O2, O3, O4, and O5 represent the five observer-outlined boundaries and the points C1 and C2 represent the two computer-generated boundaries. The shaded area is the convex polygon which bounds all the observer-outlined boundaries. The computer-generated boundary C1 lies inside this convex polygon and is considered to be within the interobserver range, whereas C2 lies outside this range.

An approximate, but quick and easy, way to test whether a point $\mathcal{C}$ lies in the convex hull of a set of points $\mathcal{O}_1$, $\mathcal{O}_2$, $\cdots$, $\mathcal{O}_n$, is to just check whether

$$\max_i \{e(\mathcal{C}, \mathcal{O}_i)\} \leq \max_{i,j} \{e(\mathcal{O}_i, \mathcal{O}_j)\} \tag{13}$$

i.e., whether the maximum computer-to-observer distance is less than or equal to the maximum interobserver distance. Using such a test, we can compute the statistic of interest, the percentage of observations for which the computer-generated boundary lies within the interobserver range. Under the hypothesis that the $n+1$ observers produce boundaries which are samples from the same distribution, the expected probability that one observer's boundary lies outside the range of the other n observers' boundaries is $1/(n+1)$. Thus, under the hypothesis that the computer-generated boundaries and the observer-outlined boundaries are samples from the same distribution, the expected percent of times that the computer-generated boundaries lie within the interobserver range is $100n/(n+1)$. For three human observers, this expected percentage is 75%; for four observers, it is 80%; and for five observers, it is 83%. To test whether the data is consistent with the hypothesis, we compute the 95% CI of the percentage statistic and check whether it includes the expected value. If the data is not consistent with the hypothesis that the computer-generated boundaries and observer-outlined boundaries are samples from the same distribution, then the CI will not include the expected value.

Whereas the Williams index gives information about averages because it computes the ratio between the average computer-to-observer agreement and the average interobserver agreement, the percentage statistic gives information about corresponding relationships between the computer measurements and the observer measurements. The percent statistic is

useful because it tells us the number of times that the algorithm is successful, i.e., the number of times it produces boundaries which are within the interobserver range; however, the 95% CI test for the percentage statistic is a very stringent test.

3) Classical Techniques: If we are to compare the parameters derived from the boundaries only, some classical statistical techniques may be used as well. The most-commonly used technique is the linear regression. The parameters derived from the computer-generated boundaries are correlated with those derived from the hand-outlined boundaries.

Recently, Bland and Altman [25] have proposed the use of another technique to measure the agreement between two variables. Their method plots the pair-wise differences between the two variables versus the best estimate of their true value. In our case, the true value may be taken to be the average of the parameter values derived from the hand-outlined boundaries. This method tests for any proportional error or bias in the measurements.

Statistics on interobserver variability were measured by calculating the average of all absolute interobserver differences. To measure reliability, we calculated Cohen's kappa [26], which is a coefficient of agreement between two observers, corrected for agreement by chance. This coefficient was originally developed for nominal data, but we used the modifications described by Berry and Mielke [27] to evaluate reliability of multiple observers for numerical data.

D. Comparison of Algorithms

Often there is a need to compare the results of applying two or more different algorithms on the same data set. We propose the use of the average of the observer boundaries (the pseudo ground truth) as the basis for the comparison. The computer-generated boundaries from the different algorithms are compared to these ground-truth boundaries resulting in an error measurement for each algorithm and for each image. The algorithm which results in a smaller overall error is preferred over the other algorithms. The comparison of the errors is carried out using Friedman's two-way analysis of variance by ranks [28]. This is a nonparametric procedure, involving ranking of the errors due to the different algorithms for each data set. The null hypothesis is that each algorithm performs identically and, thus, the average rank for each algorithm over the entire data set is the same. The test statistic is

$$\chi^2_{k-1} = \frac{12}{Nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3N(k+1) \qquad (14)$$

where

- N number of data sets (or images);
- k number of algorithms;
- R_j sum of ranks for the algorithm j.

The statistic is compared to the χ^2 distribution of $k-1$ degrees of freedom to determine whether the difference in ranks can be attributed to chance. If the null hypothesis is rejected, i.e., a difference in the ranks is significant, the different algorithms are compared using a multiple comparison procedure described by Daniel [28]. The algorithms j and j' are considered significantly different at a level α if

$$|R_j - R_{j'}| \geq z_{\alpha/k(k-1)} \sqrt{\frac{Nk(k+1)}{6}} \qquad (15)$$

where $z_{\alpha/k(k-1)}$ is the z-value corresponding to the standard normal distribution.

E. Data Used

We illustrate the use of the methodology described above to validate the results of boundary detection in two different applications, both in ultrasound imaging. In both cases, the boundaries were detected using a computer algorithm based on active contour models [16], [17].

For the cardiac boundary detection application, we have short-axis echocardiogram sequences from 44 randomly selected patients collected during routine sonographic examination. Four experienced observers hand-outlined the endocardium (inner boundary) and the epicardium (outer boundary) of the left ventricle on the ED and ES images in each sequence. The computer algorithm generated the epicardium and endocardium for the entire sequence [16]. This algorithm requires the user to draw a rough initial boundary representing the end-diastolic epicardial boundary. We showed that the algorithm is very robust to variations in this initial boundary [16]. First, we compared the epicardial and endocardial boundaries directly to the hand-outlined boundaries using techniques described above. Next, we compared the derived parameters, the enclosed area within the epicardial and the endocardial boundaries.

For the fetal head and abdomen detection application, we have images of the fetal head and abdomen from 30 randomly selected patients going through routine sonographic examination. Four experienced observers outlined the fetal skull and abdomen and measured the BPD, HC, and AC on each image. The computer algorithm also generated the fetal skull and the abdomen boundary and the BPD, HC and AC measurements [17]. This algorithm requires the user to mark a point near the center of the organ of interest. We showed that the algorithm is very robust to variations in this initial point [17]. First, we compared the computer-generated boundary directly to the hand-outlined ones. Next, we compared the HC, BPD, and AC measurements derived from the skull boundary.

III. Results and Discussion

A. Cardiac Boundary Detection

Fig. 4 shows two sample short-axis cardiac images at end diastole along with the hand-outlined and computer-generated epicardial and endocardial boundaries. Table I shows the Hausdorff distance and the average distance by directly comparing the computer-generated boundaries to the four observers' hand-outlined boundaries averaged over the 44 data sets. Table I also shows the measured Williams index and the percentage statistic and their CI's.

Table II shows the computer-to-observer differences, the interobserver differences, Williams index, and the percent statistics for the epicardial and endocardial areas. It also shows

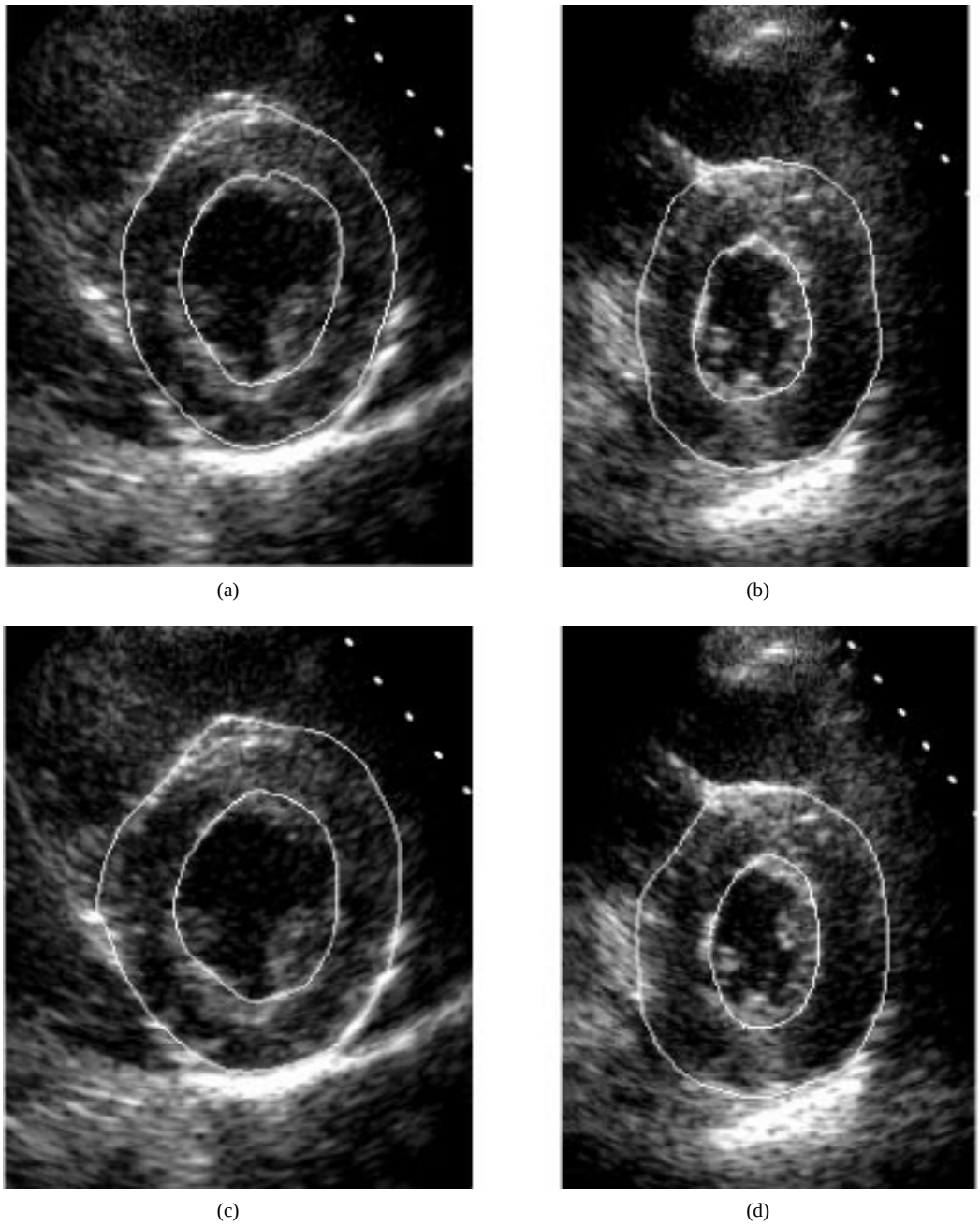


Fig. 4. (a) and (b) Two sample short-axis cardiac images with the hand-outlined epicardial and endocardial boundaries. (c) and (d) The same images with computer-generated epicardial and endocardial boundaries.

the correlation coefficient between the computer-generated measurements and the average observer measurements of these derived parameters.

From these results, we can clearly see why it is important to compare the computer's measurements to multiple observers' measurements. In Table I, we see that the mean computer-to-observer difference is almost the same for both the epicardial and endocardial boundaries; however, the Williams index and the percent statistic are very different for epicardium and endocardium. This is because the interobserver variability for outlining the epicardium is much larger than for outlining the endocardium. The Williams index for epicardium boundary detection is very close to one, indicating that the computer-generated boundaries differ from the hand-outlined boundaries as much as the hand-outlined boundaries differ from one another. For endocardial boundaries, this is not the case, however. Table I also shows that although the upper limit of the 95% Williams index CI is greater than the expected value of 1.0 for the epicardial boundaries in both the Hausdorff distance and the average distance, the upper limit of the 95% statistic CI for Hausdorff distance does not exceed its expected value of 80%.

From Table II, we can derive similar inferences. Here, the mean computer-to-observer difference is smaller for endocardial boundaries than for the epicardial boundaries. However, the upper limit of the Williams index CI does not exceed 1.0

TABLE I
DIRECT COMPARISON OF THE COMPUTER-GENERATED BOUNDARIES TO THE FOUR OBSERVERS' BOUNDARIES FOR CARDIAC BOUNDARY DETECTION. THE EXPECTED VALUE OF WILLIAMS INDEX (WI) IS 1.0 AND THE EXPECTED VALUE OF THE PERCENT STATISTIC IS 80%. COD = MEAN COMPUTER TO OBSERVER DIFFERENCE. IOD = MEAN INTEROBSERVER DIFFERENCE. P = PERCENT STATISTIC

	COD (mm)	IOD (mm)	WI	95% CI	P(%)	95% CI
Hausdorff Distance (Epi.)	8.79 ($\sigma = 3.33$)	8.31 ($\sigma = 2.62$)	0.95	(0.88, 1.01)	61.4	(52.7, 70.0)
Average Distance (Epi.)	3.53 ($\sigma = 1.33$)	3.79 ($\sigma = 1.53$)	1.07	(1.06, 1.08)	77.3	(69.8, 84.6)
Hausdorff Distance (Endo.)	8.94 ($\sigma = 2.93$)	6.79 ($\sigma = 2.11$)	0.75	(0.71, 0.81)	32.9	(24.7, 41.2)
Average Distance (Endo.)	3.88 ($\sigma = 1.53$)	2.67 ($\sigma = 0.88$)	0.69	(0.68, 0.70)	30.7	(22.5, 38.8)

TABLE II
COMPARISON OF THE COMPUTER-GENERATED EPICARDIAL AND ENDOCARDIAL AREAS TO THE FOUR OBSERVERS' MEASUREMENTS FOR CARDIAC BOUNDARY DETECTION. r = CORRELATION COEFFICIENT

	COD (cm^2)	IOD (cm^2)	WI	95% CI	P(%)	95% CI	r
Epi. area	3.26 ($\sigma = 3.23$)	4.27 ($\sigma = 4.54$)	1.30	(1.29, 1.32)	85.2	(78.9, 91.5)	0.95
Endo. area	2.43 ($\sigma = 1.71$)	1.45 ($\sigma = 1.52$)	0.59	(0.58, 0.60)	30.7	(22.5, 38.9)	0.91

for the endocardium as it does for the epicardium because the interobserver variability is much larger for epicardial areas. We can see that the correlation coefficient is fairly high for both the endocardial and epicardial areas.

We can clearly see that using only the statistics like the mean computer-to-observer difference and the correlation coefficient are not indicative of how well an algorithm performs because they do not establish a guideline for how good the statistics have to be. Measurement of the interobserver variability establishes such a guideline and can be used as a clinically useful standard to measure the performance of image segmentation algorithms.

We used a method described in the last section to compare the performance of different algorithms on the same task. We compared three different variations of the active contour algorithm for detection of the endocardial boundary by computing the boundary distances of the computer-generated boundaries to the pseudo ground-truth boundaries (established by averaging the four observers' boundaries). The three different variations of the algorithm represent the different preprocessing methods applied to the image before it is input to the active contour algorithm. In the first method, the image was prefiltered with a 5×5 Gaussian kernel while the image was filtered via a grayscale morphological opening operation with a 5-pixel-diameter disk in the second method. In the third method, no prefiltering was applied to the image.

We computed the average distance between the boundaries instead of the Hausdorff distance, and boxplots of these distances for the three algorithms are shown in Fig. 5. The Friedman's rank sum test indicated a significant difference between the performance of the three algorithms ($p < 0.001$). Multiple comparison showed that the third algorithm consistently outperformed the other two algorithms. The mean boundary distances over all 44 images for the three algorithms were 3.87, 4.58, and 3.61 mm, respectively. Thus, in this case, the third algorithm (with no prefiltering) was preferred over the others. We concluded that prefiltering the image, in this case, removed some low contrast information that is essential for accurate segmentation of the boundaries.

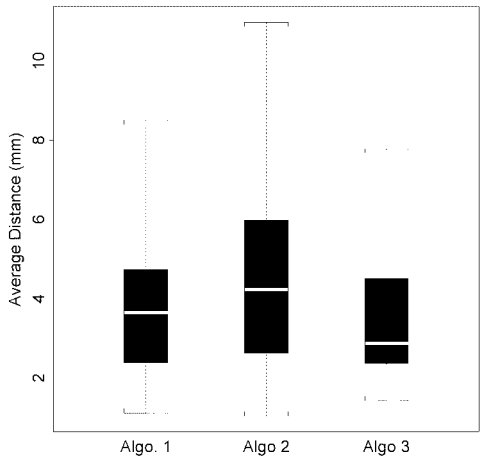
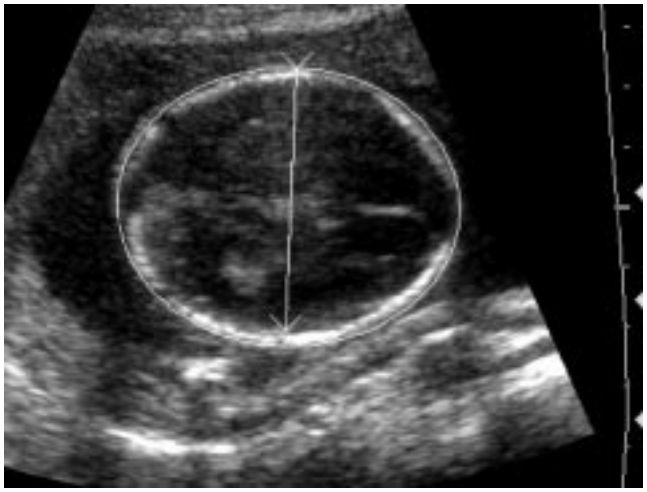


Fig. 5. Boxplots of the average boundary distances from the computer-generated boundary to the pseudo ground-truth boundaries for three different algorithms for detecting the endocardial boundary. The box represents the middle half of the data, the whiskers extend to the extreme values and the white line inside the box represents the median. Friedman's rank sum test showed a significant difference between the three algorithms. Multiple comparison showed that the third algorithm performed consistently better than the other two.

B. Fetal Size Measurements

Fig. 6 shows two of the 30 images and the automatically detected skull boundaries. The line used to measure the BPD is also shown. Table III shows the Hausdorff distance and the average distance by directly comparing the computer-generated boundaries to the four observers' hand-outlined boundaries. The upper limit of the 95% Williams index CI for



(a)

(b)

Fig. 6. Two images where the algorithm detected the skull boundary and measured the BPD and HC.

TABLE III
DIRECT COMPARISON OF THE COMPUTER-GENERATED BOUNDARIES TO THE FIVE OBSERVERS' BOUNDARIES FOR FETAL SKULL AND ABDOMEN DETECTION. CO = MEAN COMPUTER-TO-OBSERVER DISTANCE, IO = MEAN INTEROBSERVER DISTANCE, WI = WILLIAMS INDEX, CI = CONFIDENCE INTERVAL

	CO (mm)	IO (mm)	WI	95% CI
Fetal Head				
Hausdorff Distance	4.64 ($\sigma = 2.61$)	3.83 ($\sigma = 1.90$)	0.83	(0.70, 0.96)
Average Distance	2.09 ($\sigma = 0.95$)	1.92 ($\sigma = 0.82$)	0.92	(0.81, 1.03)
Fetal Abdomen				
Hausdorff Distance	8.88 ($\sigma = 6.25$)	5.48 ($\sigma = 5.22$)	0.61	(0.49, 0.73)
Average Distance	4.05 ($\sigma = 3.13$)	2.91 ($\sigma = 3.49$)	0.69	(0.57, 0.83)

TABLE IV
COMPARISON OF COMPUTER-GENERATED MEASUREMENTS TO THE GOLD-STANDARD (MEAN OF THE FOUR OBSERVERS' MEASUREMENTS) USING ABSOLUTE DIFFERENCES. r = CORRELATION COEFFICIENT

	CO (mm)	CO (%)	IO (mm)	IO (%)	r
BPD	0.71 ($\sigma = 0.61$)	1.19 ($\sigma = 0.85$)	0.83 ($\sigma = 0.66$)	1.33 ($\sigma = 0.82$)	0.999
HC	5.22 ($\sigma = 5.27$)	2.07 ($\sigma = 1.67$)	8.46 ($\sigma = 3.28$)	3.54 ($\sigma = 0.99$)	0.996
AC	12.6 ($\sigma = 9.48$)	6.35 ($\sigma = 5.26$)	11.62 ($\sigma = 10.6$)	5.65 ($\sigma = 6.53$)	0.974

TABLE V
WILLIAMS INDEX AND PERCENT STATISTIC FOR BPD, HC, AND AC MEASUREMENT. WI = WILLIAMS INDEX, P = PERCENT STATISTIC, CI = CONFIDENCE INTERVAL

	WI	95% CI	P	95% CI
BPD	1.07	(1.02,1.11)	48.5	(33.9, 63.1)
HC	1.12	(1.09,1.41)	66.7	(56.3, 83.1)
AC	0.82	(0.61,1.03)	51.4	(37.3, 65.5)

the average distance measure for the head is greater than the value 1.0, thus, indicating that the computer-generated boundaries agree as much with the observer-outlined boundaries as the observer-outlined boundaries agree with one another. However, the upper limit of the Hausdorff distance CI is less than one. As shown in Fig. 1, the Hausdorff distance and the average distance measure two different quantities. The fact that the Hausdorff distance between the computer-generated boundaries and the hand-outlined boundaries is larger than the average distance shows that even though the boundaries lie generally close to each other, there are outliers on the boundaries (as in Fig. 1) on some images. The choice of the measure to use depends on the application and type of errors allowed. In our case, the functional goal is to measure the head circumference and the biparietal diameter. Outliers on the boundaries do not affect the HC and the BPD much as will be seen from the following results.

Tables IV and V show the computer-to-observer differences, the interobserver differences, the Williams index, and the percentage statistic for BPD, HC, and AC. It also shows the correlation coefficient between the computer-generated measurements and the average observer measurements of BPD, HC, and AC. The upper limits of the Williams index CI for BPD, HC, and AC includes are all greater than 1.0. The computer-generated BPD measurements differ from the expert observers' measurements by 0.71 mm, whereas the AC measurements differ by 12.6 mm. Both measurements were found to be comparable to the interobserver differences as was illustrated in the computation of Williams index. Thus, the measured interobserver variability provides the answer to the question posed in the beginning; "How close to the observers' measurements do the computer measurements have to be?" Such an evaluation is not possible if only one observer's data are available.

As before, we compared three different variations of the active contour algorithm for detection of the fetal skull by computing the boundary distances of the computer-generated boundaries to the pseudo ground-truth boundaries. Fig. 7 shows boxplots of the average boundary distances for the three different algorithms. As is clear from the boxplots, the Friedman rank sum test did not indicate any significant difference between the performance of the three algorithms. The mean boundary distances over all 30 images for the three algorithms were 2.09, 1.98, and 2.06 mm, respectively. Thus, in this case, any of the three algorithms can be chosen without any significant difference in the performance. In

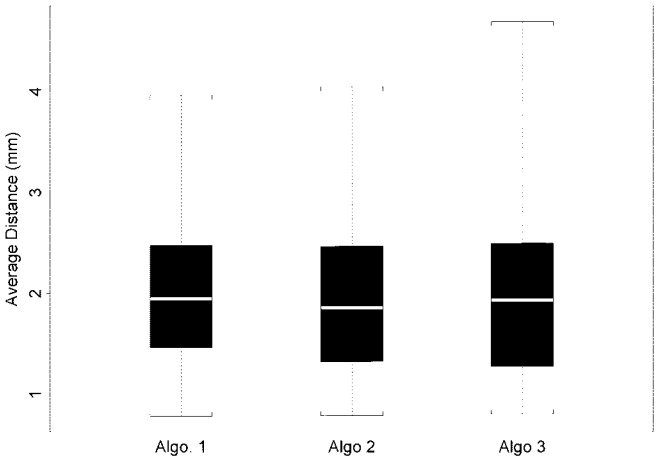


Fig. 7. Boxplots of the average boundary distances from the computer-generated boundary to the pseudo ground-truth boundaries for three different algorithms for detecting the fetal skull. Friedman's rank sum test showed no significant difference in performance between the algorithms.

this case, since the contrast in the image was high to begin with, prefiltering of the image did not adversely affect the segmentation performance.

IV. Conclusions

In this paper, we have proposed a protocol for evaluating medical image segmentation algorithms where the only information available is multiple observers' hand-outlined boundaries. We have applied this methodology and found it useful in evaluating image segmentation algorithms for two different ultrasound imaging applications. With this methodology of using multiple observers' outlines, we found new pieces of information about the performance of the algorithms which would not be possible with conventional evaluation techniques using only one observer. We have also developed a method for comparing the performance of two or more different algorithms. We believe that the objective and quantitative evaluation and comparison of various medical image segmentation algorithms using such a methodology on a standard large data set is an important step toward their acceptance and clinical use.

The segmentation evaluation methodology proposed in this paper has several limitations which need to be addressed in the future. One of the limitations is that this methodology does not take the bias of the individual observers into consideration. The method only considers the variance between observers. Without an independent ground truth, consistent observer bias is difficult to quantify. However, observer bias on individual images dependent on quantifiable measures of image quality or shading can be computed by considering statistical models. Future work may involve formulating such models, including building quantitative measures for image quality. Another limitation of our method is that it does not consider variability in the computer measurements. Typically, segmentation algorithms involve human input for initialization. Characterizing the algorithm performance with respect to the variation in this human input is also important. Udupa *et al.* [29] recently proposed methods to characterize this human input with respect to the reduction in interobserver variability and the reduction in overall time required for segmentation. Future work may involve attempts to integrate these ideas into the statistical framework proposed in this paper.

Appendix

In this appendix, we attempt to establish the triangle inequality for Hausdorff distances. Hausdorff distance is redefined here. If the two curves are represented as sets of points $\mathcal{A} = \{\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_m\}$, and $\mathcal{B} = \{\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_m\}$, where each $\mathbf{a}_i$ and $\mathbf{b}_i$ is an ordered pair of the x and y coordinates of a point on the curve, we define DCP for $\mathbf{a}_i$ to the curve $\mathcal{B}$ as

$$d(\mathbf{a}_i, \mathcal{B}) = \min_j \|\mathbf{b}_j - \mathbf{a}_i\|. \tag{16}$$

The Hausdorff distance between the two curves is defined as the maximum of the DCP's between the two curves [18]

$$e(\mathcal{A}, \mathcal{B}) = \max\left\{\max_i[d(\mathbf{a}_i, \mathcal{B})], \max_j[d(\mathbf{b}_j, \mathcal{A})]\right\}. \tag{17}$$

By the triangle inequality for Euclidean distances

$$\|\mathbf{a}_i - \mathbf{c}_j\| \leq \|\mathbf{a}_i - \mathbf{b}_k\| + \|\mathbf{b}_k - \mathbf{c}_j\| \tag{18}$$

for any point $\mathbf{b}_k$ on curve $\mathcal{B}$. By definition, $d(\mathbf{a}_i, \mathcal{C}) \leq \|\mathbf{a}_i - \mathbf{c}_j\|$ for all j; thus

$$d(\mathbf{a}_i, \mathcal{C}) \leq \|\mathbf{a}_i - \mathbf{b}_k\| + \|\mathbf{b}_k - \mathbf{c}_j\| \tag{19}$$

for any point $\mathbf{c}_j$ on curve $\mathcal{C}$ and $\mathbf{b}_k$ on curve $\mathcal{B}$. This equation can also be written as

$$d(\mathbf{a}_i, \mathcal{C}) \leq d(\mathbf{a}_i, \mathcal{B}) + d(\mathbf{b}_{k_i}, \mathcal{C}) \tag{20}$$

where $\mathbf{b}_{k_i}$ is the closest point to $\mathbf{a}_i$ on curve $\mathcal{B}$. Considering the maximum distances we can write

$$\max_i\{d(\mathbf{a}_i, \mathcal{C})\} \leq \max_i\{d(\mathbf{a}_i, \mathcal{B})\} + \max_i\{d(\mathbf{b}_{k_i}, \mathcal{C})\}. \tag{21}$$

Now, it can be seen by the definition of maximum distance that $\max_k\{d(\mathbf{b}_k, \mathcal{C})\} \geq \max_i\{d(\mathbf{b}_{k_i}, \mathcal{C})\}$; thus, (21) can be rewritten as

$$\max_i\{d(\mathbf{a}_i, \mathcal{C})\} \leq \max_i\{d(\mathbf{a}_i, \mathcal{B})\} + \max_k\{d(\mathbf{b}_k, \mathcal{C})\}. \tag{22}$$

Similarly, we can show that

$$\max_i\{d(\mathbf{c}_i, \mathcal{A})\} \leq \max_i\{d(\mathbf{c}_i, \mathcal{B})\} + \max_k\{d(\mathbf{b}_k, \mathcal{A})\}. \tag{23}$$

Combining (22) and (23), and rearranging terms, we get

$$\begin{aligned}\max\left(\max_i\{d(\mathbf{a}_i, \mathcal{C})\}, \max_i\{d(\mathbf{c}_i, \mathcal{A})\}\right) \\ < \max\left(\max_i\{d(\mathbf{a}_i, \mathcal{B})\}, \max_k\{d(\mathbf{b}_k, \mathcal{A})\}\right) \\ + \max\left(\max_k\{d(\mathbf{b}_k, \mathcal{C})\}, \max_i\{d(\mathbf{c}_i, \mathcal{B})\}\right)\end{aligned} \tag{24}$$

which is the triangle inequality for the Hausdorff distance.

ACKNOWLEDGMENT

The authors would like to thank Dr. P. Sampson of the Department of Statistics, Dr. D. Haynor of the Department of Radiology, and Dr. P. Detmer of Department of Surgery of the University of Washington for their valuable input.

REFERENCES

[1] L. H. Staib and J. S. Duncan, "Left ventricular analysis from cardiac images using deformable models," *IEEE Comput. in Cardiol., Mag.*, pp. 427–430, 1989.

[2] D. Adam, O. Hareuveni, and S. Sideman, "Semiautomated border tracking of cine echocardiographic ventricular images," *IEEE Trans. Med. Imag.*, vol. MI-6, pp. 266–271, 1987.

[3] N. Friedland and D. Adam, "Automatic ventricular cavity boundary detection from sequential ultrasound images using simulated annealing," *IEEE Trans. Med. Imag.*, vol. 8, pp. 344–353, 1989.

[4] J. Feng, W.-C. Lin, and C.-T. Chen, "Epicardial boundary detection using fuzzy reasoning," *IEEE Trans. Med. Imag.*, vol. 10, pp. 187–199, 1991.

[5] J. W. Klinger, C. L. Vaughan, T. D. Fraker, and L. T. Andrews, "Segmentation of echocardiographic images using mathematical morphology," *IEEE Trans. Biomed. Eng.*, vol. 35, pp. 925–934, 1988.

[6] I. L. Herlin and N. Ayache, "Feature extraction and analysis methods for sequences of ultrasound images," *Image and Vision Computing*, vol. 10, pp. 673–682, 1992.

[7] E. R. Wolfe, E. J. Delp, C. R. Meyer, F. L. Bookstein, and A. J. Buda, "Accuracy of automatically determined borders in digital two-dimensional echocardiography using a cardiac phantom," *IEEE Trans. Med. Imag.*, vol. MI-6, pp. 292–296, 1987.

[8] W. Zwehl, R. Levy, E. Garcia, R. Haendchen, W. Childs, S. Corday, S. Meerbaum, and E. Corday, "Validation of a computerized edge detection algorithm for quantitative two-dimensional echocardiography," *Circ.*, vol. 68, pp. 1127–1135, 1983.

[9] C. H. Chu, E. J. Delp, and A. J. Buda, "Detecting left ventricular endocardial and epicardial boundaries by digital two-dimensional echocardiography," *IEEE Trans. Med. Imag.*, vol. 7, pp. 81–90, 1988.

[10] P. R. Detmer, G. Bashein, and R. W. Martin, "Matched filter identification of left-ventricular endocardial borders in transesophageal echocardiograms," *IEEE Trans. Med. Imag.*, vol. 9, pp. 396–404, 1990.

[11] E. A. Geiser, D. A. Conetta, M. C. Limacher, V. O. Stockton, L. H. Olivier, and B. Jones, "A second-generation computer-based edge detection algorithm for short-axis two-dimensional echocardiographic images: Accuracy and improvement in interobserver variability," *J. Amer. Soc. Echocardiol.*, vol. 3, pp. 79–90, 1990.

[12] J. E. Perez, A. D. Waggoner, B. Barzilai, H. E. Melton, J. G. Miller, and B. E. Sobel, "On-line assessment of ventricular function by automatic boundary detection and ultrasonic backscatter imaging," *J. Amer. College Cardiol.*, vol. 19, pp. 313–320, 1992.

[13] B. F. Vandenberg, L. S. Ruth, P. Stuhlmuller, H. E. Melton, and D. J. Skorton, "Estimation of left ventricular cavity area with an on-line, semi-automated echocardiographic edge detection system," *Circ.*, vol. 86, pp. 159–166, 1992.

[14] C. deGraaf, A. Koster, K. Vincken, and M. Viergever, "A methodology for the validation of image segmentation algorithms," in *Proc. IEEE Symp. Computer-Based Medical Systems*, 1992, pp. 17–24.

[15] A. Hammoude, "Computer-assisted endocardial border identification from a sequence of two-dimensional echocardiographic images," Ph.D. thesis, Univ. Washington, Seattle, WA, 1988.

[16] V. Chalana, D. T. Linker, D. R. Haynor, and Y. Kim, "A multiple active contour model for cardiac boundary detection in echocardiographic sequences," *IEEE Trans. Med. Imag.*, vol. 15, pp. 290–298, 1996.

[17] V. Chalana, T. C. Winter, D. R. Cyr, D. R. Haynor, and Y. Kim, "Automatic fetal size measurements from ultrasound images," *Academic Radiol.*, vol. 3, pp. 628–635, 1996.

[18] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the Hausdorff distance," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 15, pp. 850–863, 1993.

[19] P. D. Sampson, F. L. Bookstein, F. H. Sheehan, and E. L. Bolson, "Eigenshape analysis of left ventricular outlines from contrast ventriculograms," in *Advances in Morphometrics, Proceedings of NATO Advanced Study Institute*, L. Marcus, M. Corti, A. Loy, G. Naylor, and D. Slice, Eds. New York: Plenum, 1995.

[20] P. Besl and N. McKay, "A method for registration of 3-D shapes," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 14, pp. 239–256, 1992.

[21] C. Bouma, W. Niessen, K. Zuiderveld, E. Gussenhoven, and M. Viergerver, "Evaluation of segmentation algorithms for intravascular ultrasound images," *Visualization and Biomed. Computing*, pp. 203–212, 1996.

[22] I. M. Anderson and J. C. Bezdek, "Curvature and tangential deflection of discrete arcs: A theory based on the commutator of scatter matrix pairs and its application to vertex detection in planar shape data," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-6, pp. 27–40, 1984.

[23] G. W. Williams, "Comparing the joint agreement of several raters with another rater," *Biometrics*, vol. 32, pp. 619–627, 1976.

[24] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap.* London, U.K.: Chapman and Hall, 1993.

[25] J. Bland and D. Altman, "Statistical methods for assessing the agreement between two methods of clinical measurement," *Lancet*, vol. 1, pp. 307–310, 1986.

[26] J. Cohen, "A coefficient of agreement for nominal scales," *Educational, Psychological Meas.*, vol. 20, pp. 37–46, 1960.

[27] K. J. Berry and P. W. Mielke, "A generalization of Cohen's kappa agreement measure to interval measurement and multiple raters," *Educational, Psychological Meas.*, vol. 48, pp. 921–933, 1988.

[28] W. W. Daniel, *Applied Nonparametric Statistics.* Boston, MA: Houghton Mifflin, 1978.

[29] J. K. Udupa, D. Odhner, J. Tian, G. Holland, and L. Axel, "Automatic clutter-free volume rendering for MR angiography using fuzzy connectedness," in *Proc. SPIE Conf. Medical Imaging*, 1997, vol. 3034, pp. 114-119.