Comments on: A Methodology for Evaluation of Boundary Detection Algorithms on Medical Images

Article *in* IEEE Transactions on Medical Imaging · May 2004

DOI: 10.1109/TMI.2004.826358 · Source: PubMed

3 authors:

Carlos Alberola-López
Universidad de Valladolid
240 PUBLICATIONS **2,833** CITATIONS
SEE PROFILE

Marcos Martin-Fernandez
Universidad de Valladolid
172 PUBLICATIONS **1,342** CITATIONS
SEE PROFILE

Juan Ruiz-Alzola
Universidad de Las Palmas de Gran Canaria
94 PUBLICATIONS **1,229** CITATIONS
SEE PROFILE

Comments on: A Methodology for Evaluation of Boundary Detection Algorithms on Medical Images

Carlos Alberola-López*, Marcos Martín-Fernández, and Juan Ruiz-Alzola

***Abstract*—In this paper we analyze a result previously published about a comparison between two statistical tests used for evaluation of boundary detection algorithms on medical images. We conclude that the statement made by Chalana and Kim (1997) about the performance of the *percentage test* has a weak theoretical foundation, and according to our results, is not correct. In addition, we propose a one-sided hypothesis test for which the acceptance region can be determined in advance, as opposed to the two-sided confidence intervals proposed in the original paper, which change according to the estimated quantity.**

***Index Terms*—Boundary detection, confidence interval, hypothesis test, percentage statistic, validation, William index.**

I. Introduction

In [1], Chalana and Kim propose a methodology for evaluation of automatic boundary detection algorithms in medical images. One of their most interesting contributions is the proposition of two statistical tests that check whether a boundary detection algorithm can be validated. The authors pose the validation problem so as to check whether *the computer-generated boundaries differ from the manually outlined boundaries as much as the manually outlined boundaries differ from one another*.

This general idea is implemented by means of two different statistical tests, the first one of which is a generalization, made by the authors, of the William index (WI) [4], and the second is what the authors call interchangeably *percent* or *percentage statistic* (PS). The PS, as indicated by the authors, computes the percentage of cases for which the computer-generated boundaries (CGBs) lie within the interobserver range (IR). Leaving aside details on how these ideas are implemented, it turns out that, under the hypothesis that the CGB is statistically equal to the expert-outlined boundaries (EOBs), some distance measure from the CGB will be identically distributed as these measures from the EOBs. Therefore, a test is built to find whether the measure from the CGB falls within the range of measures from the EOBs in the same proportion as it is theoretically expected.

The authors conclude that the two testing procedures have very different behaviors when tested against the datasets used in their paper, and they make a choice about using WI as opposed to PS on the basis of their results. However, we will hereafter demonstrate that the analysis made by the authors of the PS is not correct, and hence comparative results reached by the authors are not conclusive. This fact, together with the simplicity of the PS with respect to the WI, makes the former an approach, in our opinion, worthtaking.

Manuscript received June 2, 2003; revised January 19, 2004. This work was supported in part by the Spanish Government-CICYT under Research Grant TIC2001-3808-C02. The authors are members of the Network of Excellence. The Associate Editor responsible for coordinating the review of this paper and recommending its publication was N. Ayache. *Asterisk indicates corresponding author.*

*C. Alberola-López is with the Laboratorio de Procesado de Imagen (LPI), ETSI Telecomunicación, University of Valladolid, 47011 Valladolid, Spain (e-mail: caralb@tel.uva.es).

M. Martín-Fernández is with the Laboratorio de Procesado de Imagen (LPI), ETSI Telecomunicación, University of Valladolid, 47011 Valladolid, Spain (e-mail: marcma@tel.uva.es).

J. Ruiz-Alzola is with ETSI Telecomunicación of the University of Las Palmas de Gran Canaria, 35017 Las Palmas, Spain and also with the Surgical Planning Lab, Brigham and Womeńs Hospital, Harvard University, Boston, MA 02115 USA (jruiz@dsc.ulpgc.es).

Digital Object Identifier 10.1109/TMI.2004.826358

II. On the Percent Statistic

A. Definition of Interobserver Range

Denoting by $\mathbf{C}_i$ ($i = \{1, \ldots, n\}$) the EOBs, with n the number of experts, each contour $\mathbf{C}_i$ is a point in the $2m$ dimensional space, with m the number of (two dimensional) points that define the boundary on the image. Therefore, the EOBs create a cloud of points in that space. The authors state that a CGB is within the IR *if it lies within a multidimensional convex polyhedron formed by the observer outlined boundaries*, i.e., if it lies within a polyhedron with vertices being points from the cloud, and that it encloses all the points in the cloud. For implementation purposes the authors use a *quick and easy* approximation to it. Specifically, denoting by $\mathbf{C}_0$ the CGB, this contour is considered to be within the IR if

$$\max_i \{d(\mathbf{C}_0, \mathbf{C}_i)\} \leq \max_{\substack{i,j \\ 1 \leq i,j \leq n}} \{d(\mathbf{C}_i, \mathbf{C}_j)\} \tag{1}$$

with $d(\mathbf{C}_r, \mathbf{C}_s)$ a distance measure between contours $\mathbf{C}_r$ and $\mathbf{C}_s$ (see Section IV for details on the distance measures used by the authors). The authors state that

$$P(\mathbf{C}_0 \in \mathrm{IR}) = \frac{n}{n+1} \tag{2}$$

assuming the CGB to be independent and identically distributed (IID) with respect to the EOBs (which are also considered IID). As we will show this result is not correct.

B. Probabilistic Analysis

Let variable D_{ij} denote the distance between $\mathbf{C}_i$ and $\mathbf{C}_j$, $i \neq j$, and let D_m denote the *maximum distance between any two contours*, i.e.,

$$D_m = \max_{i,j} D_{ij}, \quad i \neq j. \tag{3}$$

If both the CGB and the n EOBs are IID, then the probability that $D_m = D_{ij}$ is equal $\forall (i,j)$ with $i \neq j$, $i, j = \{0, \ldots, n\}$, i.e., any pair of contours is equally likely to be the pair of most separated contours. In addition, we may assume that exactly one pair of contours has distance D_m since the probability that two different continuous random variables D_{ij} and D_{kl} (with either $i \neq k$ or $j \neq l$ and, of course, $i \neq j$ and $k \neq l$) coincide is null [2]. The set of $n+1$ contours may therefore be divided into two sets, namely, the set of the two contours giving rise to D_m and the set of the $n-1$ remaining contours. $\mathbf{C}_0$ is equally likely to be any of these $n+1$ contours; so, the probability of $\mathbf{C}_0$ belonging to the second set, i.e., the probability of $\mathbf{C}_0$ falling within the IR is

$$P(\mathbf{C}_0 \in \mathrm{IR}) = \frac{n-1}{n+1} \tag{4}$$

which is clearly different from the result indicated in (2), originally proposed in [1]. This difference might be negligible for large values of n; however, in a real validation problem, the number of experts giving their opinion will be frequently small. Actually, the authors in [1] use $n = 4$; for this case the authors claim that the expected probability is $4/(4+1) = 0.8$. However, it is clear from our result that

$$P(\mathbf{C}_0 \in \mathrm{IR}) = \frac{4-1}{4+1} = \frac{3}{5} = 0.6 \tag{5}$$

which is quite a remarkable difference.

Finally, in order for a statistical test to be fully specified we need to define the acceptance region of the hypothesis "the CGB lies within the IR" out of the values of a statistic derived from the image data. The authors in [1] compute two-sided confidence intervals (CIs) to check whether they include the expected value. We understand a one-sided

hypothesis test is more appropriate for this type of problem. In addition, the radius of the CIs in [1] depend on the estimated quantity, while the acceptance region, as we will show next, is unique for each problem configuration, a fact that makes the whole process simpler.

III. Defining a Test for Sample Results

Assume we have N images and we apply the PS method to check whether the algorithm performs similarly as the board of experts. To check that this is true we need to calculate the fraction of times that the CGB lies within the IR. The value of this fraction should be equal to that in (4), and deviations from this value are only due to the finiteness of the sample size N. We need to find a region of acceptance of the values of this fraction assuming that the $n+1$ contours are IID. To that end we will consider that the test is not passed if the fraction of times that the CGB lies within the IR is too small with respect to its nominal value. If that value is larger than the expected value there is no reason to justify that the CGB is not correct, so the test should be passed. This is why we understand a unilateral test is more appropriate than its bilateral counterpart.

Let variable X_j denote the event: "contour $\mathbf{C}_0$ lies within the IR for the $j-th$ image" ($j = \{1,\ldots,N\}$). Variable X_j is a Bernoulli variable with parameter $p = (n-1)/(n+1)$. Let $q = 1-p$. Let Z be the arithmetic mean of the variables X_j. The purpose is to find a value ϵ for which the identity $P(p - Z > \epsilon) = \alpha$ holds. α is a prespecified level for the test (typically $\alpha = 0.05$). This is a well-known problem [3], the solution of which is[1] $\epsilon = \sqrt{(pq/N)}z_{1-\alpha}$, with $z_{1-\alpha}$ the value for which $G(z_{1-\alpha}) = 1-\alpha$ and $G(\cdot)$ the distribution function of a normal standard variable. The acceptance region consists of the values of Z that are greater than the critical value $p-\epsilon$. For $\alpha = 0.05$ it turns out that $z_{1-\alpha} = 1.6449$, and for the values used in [1] (see next section for details), the critical values are indicated in Table I.

A common question is how to choose the number of images that should be used in order to get reliable conclusions, i.e., to be sure that the test is passed because the CGB is similar enough to the EOBs, and not because the test itself is excessively loose. This problem is typically solved [3] by defining a power for the test or, complementarily, by limiting the probability of a different value from p, say p_1, being accepted as being p. p_1 would be an expected value of CGBs falling within the IR when they are not identically distributed to the EOBs. For such contours, the test should be passed with a low probability, say (using the customary terminology in hypothesis testing) $1-\beta$. Once these two parameters are set, it is simple to see [3] that the sample size N should be

$$N = \left(\frac{z_{1-\alpha}\sqrt{pq} + z_{\beta}\sqrt{p_1 q_1}}{p - p_1}\right)^2 \tag{6}$$

with $q_1 = 1 - p_1$ and $G(z_\beta) = \beta$. We will elaborate on how to set parameters p_1 and β in Section VI.

IV. Comments on the Experiments in [1]

A. Datasets and Distances Used by the Authors

The authors in [1] apply their methodology for algorithm performance comparison on two different image datasets.

- The first dataset consists of $N = 44$ ultrasound short-axis cardiac images at end diastole, in which $n = 4$ independent experts have manually drawn both the epicardial and the endocardial boundaries. The authors also calculate the areas enclosed by these two boundaries. Both the WI and the PS are given, together with their two-sided CIs at a 5% confidence level both for the contours themselves and for the areas calculated from them. Some other statistics are given as well, but we will not be concerned with them.
- The second dataset consists of $N = 30$ ultrasound fetal images; in this case, $n = 4$ experts have manually drawn the skull and abdomen contours and have measured the biparietal diameter (BPD), the head circumference (HC), and the abdomen circumference (AC). As before, both the WI and the PS are given both for contours and for measurements, together with their two-sided CIs.

TABLE I
Critical Values for the two Datasets Used in [1] (See Section IV for Details on the Datasets). The Test Is Passed if $Z > p - \epsilon$

Image	N	$p-\epsilon$
Heart	44	0.4785
Fetus	30	0.4529

For the two datasets the authors compare the CGBs —given by some algorithm based on active contours— with the EOBs. The measurements described above (areas, BPD, HC, and AC) are calculated automatically from the CGBs.

As for the distance measures, two of them are used in the authors' analysis. The first one is called *average distance*, and it is defined as the average of the distance between every pair of corresponding points in the two contours under comparison. Corresponding points are found by an algorithm proposed by the authors. As for the second distance, it is called *Haussdorf distance*, and it is calculated in two steps; given two contours, say $\mathcal{A}$ and $\mathcal{B}$, the first step is to calculate the distance of every point in contour $\mathcal{A}$ to contour $\mathcal{B}$ and the distance from every point in contour $\mathcal{B}$ to contour $\mathcal{A}$. Denoting by $\mathbf{a}_i$ and $\mathbf{b}_i$ the ith point in contours $\mathcal{A}$ and $\mathcal{B}$, respectively (with the number of points in both not necessarily equal), these distances are defined

$$d(\mathbf{a}_i, \mathcal{B}) = \min_j \|\mathbf{b}_j - \mathbf{a}_i\| \tag{7}$$

$$d(\mathbf{b}_i, \mathcal{A}) = \min_j \|\mathbf{a}_j - \mathbf{b}_i\|. \tag{8}$$

The second step is to define the distance between the contours themselves. This is done by picking the maxima of the distances defined above, i.e.,

$$d(\mathcal{A}, \mathcal{B}) = \max\left(\max_i d(\mathbf{a}_i, \mathcal{B}), \max_j d(\mathbf{b}_j, \mathcal{A})\right). \tag{9}$$

B. Implications of Our Result in the Authors' Experiments

The authors' experiments consists of several numerical calculations; as for the WI and the PS, the authors report nine comparisons. Within these comparisons, in all the cases in which the WI test is not passed the PS test is not passed either. However, there are three cases in which the former is passed while the latter is not. We understand this is why the authors state that the test they build on the PS *is a very stringent test* (page 647, fourth line in first column). However, we cannot agree with such an statement, at least not from the results presented in that paper. We will go over these three cases.

- For the first dataset, contour comparison of the epicardial boundary based on the Haussdorf distance gives a value of $\mathrm{PS} = 0.614$, while the CI is (52.7,70.0) (see [1, Table I, page 649, first row, two rightmost columns]). Since this CI does not include the value expected by the authors (recall from Section II-B that this value is 0.8) the authors take this test as rejected. However, it is clear from (4) and (5) that this value of PS is fairly equal to the true expected value 0.6. The two-sided CI given by the authors

[1]We have used an approximate analysis based on the DeMoivre-Laplace theorem [2], which can be easily shown to hold [3] for the values of the parameters p and N that will be used in Section IV.

does include this value and, in addition, the value $\mathrm{PS} = 0.614$ is within the acceptance region of the one-sided hypothesis test defined above (see the critical value in Table I, first row).

- For the second dataset the PS for BPD is 0.485, with CI (0.339,0.631) (see [1, Table V, page 650, first row, two rightmost columns]). Once again, the upper level of this CI does not reach 0.8. However, the upper value of this CI does include the true expected value 0.6 and this value of PS falls within the acceptance region of our test (see the critical value in the second row of Table I). As for the AC (see [1, Table V, third row, two rightmost columns]), the value of PS is 0.514, with CI (0.373,0.655); the upper level of this CI does not include the value 0.8 but it does include 0.6; as for our test, the acceptance region includes the value obtained by PS in this case as well.

Therefore, it is our understanding that the considerations made in [1] about the comparative behavior of the two tests are not conclusive from the data presented in that paper.

V. An Additional Discussion

Apart from the main point of this correspondence, which is the result indicated in (4), two additional issues will be commented. The first one has to do with the number of images N to be used to build the tests. The second one is a brief discussion on the distances used by the authors.

About the former, the sample size can be determined using (6); however, two parameters must be set (both p_1 and β) beforehand. How to choose these parameters is not a simple problem; one obvious solution is to set these parameters arbitrarily, using uniquely common sense. For instance, stating that a probability p_1 10% lower than p should be accepted with a probability as low as $1 - \beta = 0.1$. If this was the choice, (6) would give a (rounded) value of $N = 580$. The problem with this procedure is that it is unclear how different CGBs are from EOBs when the value of p_1 resulting from them is 10% lower than the expected value when these contours are IID. Another possibility is to build a procedure with the help of experts to estimate these parameters on the basis of their physical meaning; for instance, one of the experts could draw contours with slight deformations intentionally added. Some of the other experts may admit them and some may not. The fraction of contours admitted by the experts would give a hint on the value of $1-\beta$. Then, the number of these deformed contours falling within the IR could give an estimate of p_1. In any case, and just to get an idea of the validity of the authors' analysis, recall the first dataset used in [1] ; since it consists of $N = 44$ images, it gives a result of accepting a value p_1 approximately 36% lower than $p = 0.6$, with a probability $1 - \beta = 0.1$. For the second dataset, the number of images is $N = 30$, which results in admitting a p_1 about 43% lower than $p = 0.6$ for this same value of $1 - \beta$.

A final remark on the measures used by the authors to quantify contour differences may be of interest. As previously indicated, the authors use a scalar value of contour difference that stems from some distance measure. This causes the effect of summarizing the information of two contours, i.e., $4m$ scalars, into a single scalar, which, for instance, impedes to distinguish whether the CGB is globally different from the EOBs, or whether large local deviations have occurred though a significant part of the CGB may be similar to the EOBs; obviously, this has the side effect of not being able to identify the particular regions of the CGB that are similar to the EOBs, and those that are dissimilar. About the distances themselves, the average distance is a global measure, so local deviations could be obscured in the average. The Hausdorf distance, on the other hand, has a more local character, since it takes into account maximum deviations; however, the global contour behavior is ignored. In addition, it should be pointed out that (7) and (8) related to the Haussdorf distance are not calculated between corresponding points, but between *every* two points in the two contours; therefore, there is not guarantee that comparable entities are compared, a fact that could be easily solved by calculating this distance between corresponding points. As a conclusion of this analysis, it is our understanding that extending the methodology proposed by the authors to perform multiple local inferences (not necessarily on distances, but also on any parameter defined out of the contour point positions) may solve some of the above mentioned problems; if distances are used, we maintain that they should be preferably calculated on corresponding points.

VI. Conclusion

This paper demonstrates that the expression on which the authors in [1] base one of their statistical tests is not correct, so conclusions derived from their analysis are, in our opinion, not correct either. As a matter of fact, using PS—with our analysis—on the data presented in [1] gives the same results as the WI in terms of test acceptance/rejection, not to mention the far less computational load required by the former with respect to the latter. Even though additional experiments should be conducted to get more solid conclusions on their comparative behavior, it is our opinion, on the basis of our mathematical analysis, that the PS is preferable in terms of computation and without a demonstrated poorer performance than the WI.

Acknowledgment

The authors thank both Dr. Chalana and Dr. Kim for their valuable contribution to the field of validation of medical image analysis algorithms.

References

[1] V. Chalana and Y. Kim, "A methodology for evaluation of boundary detection algorithms on medical images," *IEEE Trans. Med. Imag.*, vol. 16, pp. 642–652, Oct. 1997.

[2] A. Papoulis, *Probability, Random Variables and Stochastic Processes*. New York: McGraw-Hill, 1990.

[3] B. Rosner, *Fundamentals of Biostatistics*. Pacific Grove, CA: Duxbury Thomson Learning, 2000.

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