Design and Perceptual Validation of Performance Measures for Salient Object Segmentation

Vida Movahedi James H. Elder
Centre for Vision Research, York University
Toronto, Canada
{movahedi,jelder}@yorku.ca

Abstract

Empirical evaluation of salient object segmentation methods requires i) a dataset of ground truth object segmentations and ii) a performance measure to compare the output of the algorithm with the ground truth. In this paper, we provide such a dataset, and evaluate 5 distinct performance measures that have been used in the literature practically and psychophysically. Our results suggest that a measure based upon minimal contour mappings is most sensitive to shape irregularities and most consistent with human judgements. In fact, the contour mapping measure is as predictive of human judgements as human subjects are of each other. Region-based methods, and contour methods such as Hausdorff distances that do not respect the ordering of points on shape boundaries are significantly less consistent with human judgements. We also show that minimal contour mappings can be used as the correspondence paradigm for Precision-Recall analysis. Our findings can provide guidance in evaluating the results of segmentation algorithms in the future.

1. Introduction

Segmentation is one of the fundamental problems of computer vision and often the major first step in computer vision applications. An important question is how to evaluate performance of segmentation methods. In this paper, we focus on evaluation of *salient object segmentation algorithms*, where the goal is to detect the most salient object or objects in a scene.

The history of evaluating segmentation algorithms is as old as the history of segmentation algorithms themselves. Zhang [26] published a survey on these methods in 1996. He has classified the evaluation methods into three main categories:
(1) *The analytical methods.* These evaluation methods consider the algorithms without considering their output. The major difficulty in evaluation by analytical methods is the lack of a general theory for image segmentation.
(2) *The empirical goodness methods.* These evaluation methods are based on the outputs of the segmentation algorithms. For example the outputs can be compared based on the intra-region uniformity of the segments, or the inter-region contrast between the segments. These evaluation methods are not suitable for evaluation of salient object segmentation approaches, since there is no reported goodness measure that can generalize to segmentation of all images and all objects. For example, in Figure 1(a), the intra-region uniformity inside the penguin and the inter-region contrast between penguin and background are both low.
(3) *The empirical discrepancy methods.* These methods compare the outputs of segmentation algorithms with ground truth. In order to use the empirical discrepancy methods to evaluate performance of object segmentation algorithms, there is the need for i) a dataset of ground truth boundaries of salient objects in diverse imaging conditions, and ii) a suitable error measure. This is the focus of the present paper.

Existing ground truth segmentation datasets are insufficient for the evaluation of salient object segmentation for two reasons. First, there is generally not a 1:1 mapping between segments and objects: often a single object is broken into multiple segments. Second, these datasets provide no information about the salience of each segment. Among a few existing salient object datasets, none provides subjective variations among segmentations. In order to satisfy the first requirement for performance evaluation, we have constructed a new segmentation dataset called the Salient Object Dataset (SOD). The SOD is built upon the Berkeley Segmentation Dataset (BSD) [17], and provides ground truth for the boundaries of salient objects perceived by humans in natural images.

Our second requirement is a suitable error measure. In this paper we review a number of measures in use in the literature and identify potential weaknesses. Based on this analysis, we propose a contour mapping measure which can

978-1-4244-7030-3/10/$26.00 ©2010 IEEE

be seen as a particular form of elastic matching [3]. This simple measure finds the optimal mapping between the two point cycles forming the boundaries of the 2D shapes being compared, where optimality is defined in terms of the sum of Euclidean distances between corresponding points. While the correspondence must be monotonic, respecting the ordering of the points, it need not be 1:1, thus allowing for differences in level of detail. We demonstrate how the method addresses issues arising with previous measures. Finally, employing a test database of human and algorithm segmentations, we psychophysically evaluate and compare the proposed measure against 4 other measures from the literature.

We emphasize that our goal is not to assess theories of human shape perception, which is a very rich topic in its own right (*e.g*. [2, 4]). Rather, our practical aim is to assess the degree to which measures of salient object segmentation align with human judgement, with the hope of aiding automatic evaluation of object segmentation algorithms.

The structure of the paper is as follows. In Section 2, we discuss the need for a ground truth dataset and introduce the Salient Object Dataset (SOD). In Section 3, we review and analyze error measures in the literature and introduce the contour mapping measure. We report results of our psychophysical experiments in Section 4. We refer to Precision-Recall analysis in Section 5. Some suggestions for future work are given in Section 6 and conclusions are drawn in Section 7.

2. The Salient Object Dataset (SOD)

For the specific purpose of salient object segmentation, existing ground truth datasets have limitations. For example, the Berkeley Segmentation Dataset (BSD) [17] provides a suitable set of 300 images and segmentations by up to 30 subjects. Segments generally correspond to areas in the image with homogenous color or texture but do not necessarily define regions corresponding to objects. Moreover there is no distinction between object and background segments (see *e.g*., Figure 1(a)). Dataset reported in [15] provides rectangles around the salient objects, not the detailed boundary. Also [7] provides local figure-ground information on contour segments, but does not indicate the location of salient objects. The PETS dataset [25], Goldmann's dataset [12], and the ground truth dataset of lake boundaries used in [5] are limited to their respective domains of application.

The most appropriate datasets we are aware of are due to Ge *et al*. [10] and Alpert *et al*. [1], which introduced ground truth datasets containing salient object segmentations. However, in these datasets the subjects are either working together or the most likely foreground is reported based on subjects' votes. Therefore these datasets do not provide enough data to capture subjective variations.

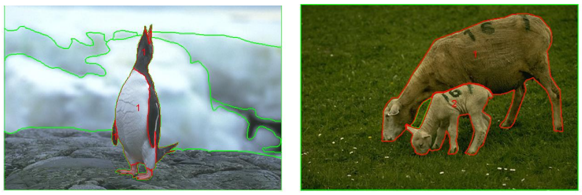

(a) (b)

Figure 1. Example objects from the SOD dataset.

To overcome these limitations and yet make use of the extensive work done in creating the BSD dataset, we constructed our new SOD dataset based on the human segmentations in the BSD. Specifically, we employed a set of human subjects, and presented each with a random subset of the BSD boundaries superimposed on the corresponding images. Each subject could then identify the object(s) they perceived as most salient by clicking on the BSD segment(s) that comprised each object. Variations over both BSD subjects and SOD subjects for the same image provides a reasonable measure of subject variability. We employed 7 human subjects to construct the SOD database.

Figure 1 shows examples of the visual interface used to produce the SOD database. Each subject could combine several regions to form one object (Figure 1(a)), and could identify multiple salient objects in the same image (Figure 1(b)). In the latter case, subjects were required to rank the identified objects according to their salience. All 300 images of the BSD where employed, and about half of the BSD segmentations were randomly selected and shown to each subject, resulting in a total of 12,110 salient object boundaries selected by 7 subjects. This dataset is available at `http://elderlab.yorku.ca/SOD`.

3. Error Measures

In addition to a ground truth database, evaluating object segmentation algorithms requires an error measure. Measures in the literature can be categorized into i) region-based measures, ii) boundary-based measures, iii) mixed measures.

3.1. Region-Based Error Measures

Region-based error measures consider the consistency between the pixels comprising algorithm and ground truth segments (Figure 2). For example, the *regional coincidence accuracy* $P(A;B)$ proposed by Ge *et al*. [10] is defined as:

$$P(A;B) = \frac{|R_A \cap R_B|}{|R_A \cup R_B|} \qquad (1)$$

where R_A and R_B are the pixels within algorithm (A) and ground truth (B) segments. A *Region Intersection (RI)* measure of error is then given as

$$RI(A,B) = 1 - P(A;B) \qquad (2)$$

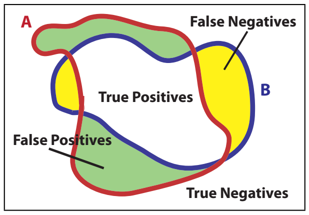
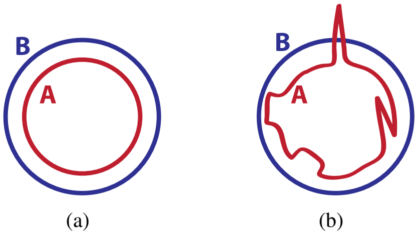


Figure 2. Region-based error between algorithm (A) and ground-truth (B) segments.



(a) (b)

Figure 3. Limitations of region-based measures. Regional measures are not sensitive to spikes, wiggles and some large shape features. Based on regional measures, the boundary in (b) is almost as good as the boundary in (a) when compared with the ground truth circle, although it has spikes, wiggles and shape differences.

Region-based measures are usually symmetric with respect to the two segments and therefore treat false positives and false negatives in the same way. Other examples of region-based measures include the Negative Rate Metric [12], the Hamming Distance [13], the Local Consistency Error [17], the Bidirectional Consistency Error [19], the regional Precision-Recall measures [17].

There are reasons to expect that regional measures will not be optimal for evaluating object segmentation algorithms, as they are not sensitive to spikes, wiggles, and major shape differences [20]. For example, most region-based measures would predict that the algorithm segments (A) in Figure 3(a) and (b) are comparable in their consistency with ground truth (B), whereas to the human eye, the algorithm result is better in (a) than in (b).

3.2. Boundary-Based Error Measures

Boundary-based measures evaluate segmentations based on the accuracy of their boundaries. The algorithm boundary is compared with a ground-truth boundary. The error measure is usually some aggregate measure of distance between points on the two boundaries. In particular, for each point a on the boundary A, a distance to boundary B, denoted as $d_B(a)$, can be defined as the minimum distance of point a to all points on B.

$$d_B(a) = \min_{b \in B} (d(a,b)), \qquad a \in A \tag{3}$$

Consideration of the distance of all points in A from B yields a *distance distribution signature* (SD) [19]: $SD_B(A,B) = \{d_B(a), a \in A\}$. In a similar fashion, $d_A(b), b \in B$ and $SD_A(B,A)$ can be defined. Aggregate values of these distributions can be used as a measure of distance. For example, letting $D_B(A,B)$ represent the mean distance of points on the boundary of A from B, *i.e.* $D_B(A,B) = \overline{SD_B(A,B)}$, leads to a *mean distance (MD)* error measure:

$$MD(A,B) = \frac{1}{2}\left(D_B(A,B) + D_A(B,A)\right) \tag{4}$$

When a measure with less sensitivity to outliers is needed, the median of the distribution can also be used.

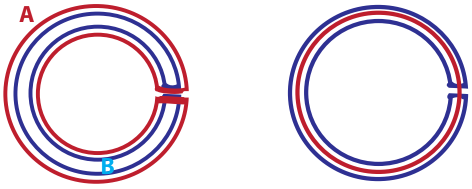
The *Hausdorff distance (HD)* [14], on the other hand, is based upon the maximum value of the distance distribution signatures :

$$h(A,B) = \max\left(SD_B(A,B)\right) = \max_{a \in A} \min_{b \in B} d(a,b) \tag{5}$$

$$HD(A,B) = \max\left(h(A,B), h(B,A)\right) \tag{6}$$

Since the Hausdorff distance only looks at the maximum value in the distance distribution, two contours having the same worst case distance are evaluated as being the same, irrespective of other distances.

While these boundary-based measures do not suffer from the problem depicted in Figure 3, they are subject to a different problem. In Figure 4, these boundary-based measures would assign similar errors to the algorithm boundaries (A), whereas the algorithm boundary in (b) has far greater perceptual error.



(a) (b)

Figure 4. Problems with boundary-based distance measures. Very different shapes can produce very small errors.

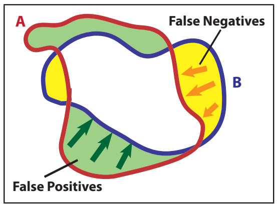


Figure 5. Measures using a mixture of regional and boundary information. The false positive and false negative regions are penalized by their distance from the intersection region.

3.3. Mixed Measures

One can attempt to solve the shortcomings of region and boundary based error measures by combining the two to

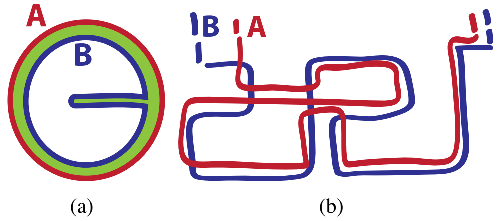


Figure 6. Mapping limitations. The mixture measures cannot penalize all shape differences effectively.

form a *mixed measure (MM)* as follows (see Figure 5):

$$MM(A,B) = \frac{1}{2D_{diag}} \times \left(\frac{1}{N_{fn}} \sum_{j=1}^{N_{fn}} d_A(p_j) + \frac{1}{N_{fp}} \sum_{k=1}^{N_{fp}} d_B(q_k) \right) \quad (7)$$

where N_{fn} is the number of false negative pixels, and $d_A(p_j)$ is the distance of the jth false negative pixel, p_j, from the algorithm boundary A. Similarly N_{fp} is the number of false positive pixels and $d_B(q_k)$ is the distance of the kth false positive pixel, q_k, from the ground truth boundary B. D_{diag} is the diagonal size of the image and can be used to normalize the distance values. Other error measures suggested in this category are the rate of misclassification metric [25] and the weighted quality measure metric [25].

Although the above measures are sensitive to wiggles and spikes, they still are not sensitive to some important shape differences. For example, in Figure 6(a) the green region is penalized by distances to the intersection area. Since the pixels in this region are close to B, the above measures do not effectively penalize the difference in the shapes. A more exaggerated case is shown in Figure 6(b). These examples demonstrate that very different shapes can produce arbitrarily small error measures. The core problem appears to be that these measures do not enforce a direct monotonic mapping of boundary points, allowing shapes to diverge even when error is constrained.

3.4. Contour Mapping Measure

Elastic matching methods[11, 3, 9, 8, 22, 23] directly align two contours by determining a mapping between the points on the contours that minimizes a matching cost. Typically, the matching cost is based on two components: 1) dissimilarity of local properties of matched points, *e.g.*, tangent orientations, and 2) dissimilarity of matched curve segments, *i.e.*, the cost of deforming one curve segment (stretching, bending or compressing) to match the other curve segment. Equipped with a translation-, rotation- and scale-invariant cost function, these measures have proven effective in image database search applications and for clustering of shape databases [21]. Optimization is often based on cyclic string correction [16] methods and its variants [8, 21]. The contour mapping measure we propose is in the spirit of these elastic measures. The cost function is simply the Euclidean distance of matched points.

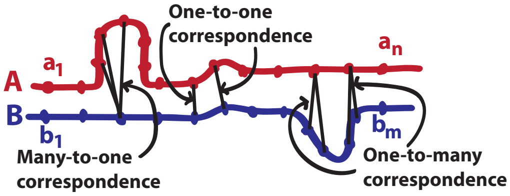

One-to-one correspondence
A
a_1
a_n
B
b_1
b_m
Many-to-one correspondence
One-to-many correspondence

Figure 7. Bimorphism

Following the notation of Maes *et al.* [16], we represent shape boundaries A and B as strings of points, $A = a_1a_2...a_n$, $B = b_1b_2...b_m$. A mapping between point a and point b is denoted by $s : a \leftrightarrow b$ (Figure 7). To avoid the problems illustrated in Figures 4 and 6, the order of the mapping must be monotonic. In other words, if $a_i \leftrightarrow b_m$ and $a_j \leftrightarrow b_n$ then $i < j \Rightarrow m \leq n$ and $m < n \Rightarrow i \leq j$. For closed boundaries, the indices are assigned cyclically.

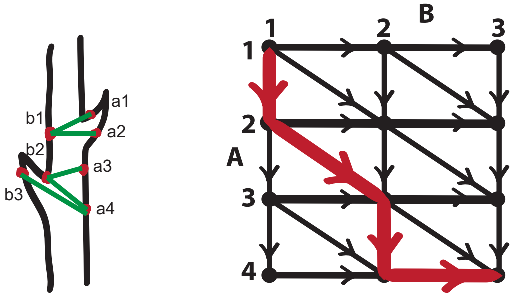
Note that although the mapping is monotonic, it is not necessarily strictly monotonic, and thus need not be 1:1. The boundaries being compared can have different levels of detail and very different total arc lengths. Therefore the mapping can be one-to-one, many-to-one, or one-to-many. Tagare *et al.* [24] call this class of correspondence a bimorphism (Figure 7). We define a mapping sequence $S = s_1s_2...s_k$ as a mapping between A and B in which all points in A are mapped to at least one point in B and vice versa. The cost of this sequence is $\gamma(S) = \sum_{i=1}^{k} \gamma(s_i)$, where $\gamma(s_i)$ is simply the Euclidean distance between the points. The mapping distance, $\delta(A,B)$, is defined as the minimum cost of mapping A and B [16]:

$$\delta(A,B) := \min_S \gamma(S) \quad (8)$$

A trace T from A to B is the set of k ordered pairs of integers $(i,j), i \in 1..n, j \in 1..m$ corresponding to the k mappings in a mapping sequence. Note that for two distinct pairs (i_1,j_1) and (i_2,j_2), $i_1 < i_2 \Rightarrow j_1 \leq j_2$. Since all points on A and B have a match, we have:

$$\forall i \in [1..n], \exists j' \in [1..m] : (i,j') \in T \text{ and}$$
$$\forall j \in [1..m], \exists i' \in [1..n] : (i',j) \in T \quad (9)$$

Potential mappings and associated costs can be represented as a graph (Figure 8). Moving down the graph corresponds to advancing on the boundary A and moving right on the graph is equivalent to advancing on the boundary B. The set of points traversed on a path from the upper left corner of this graph to the lower right corner defines a trace starting from (a_1,b_1) and ending at (a_n,b_m) and represents the mapped point pairs on the two boundaries. Such path also ensures that all points on the two boundaries have a match. The monotonicity condition constrains each edge of the path to have only down and/or rightward components. Since the total matching cost is defined as a sum,



(a) (b)

Figure 8. Mapping graph for calculating CM. (a) Two boundaries $A = a_1a_2a_3a_4$ and $B = b_1b_2b_3$. (b) Associated mapping graph. The red path corresponds to the sequence of mappings $S = (a_1 \leftrightarrow b_1, a_2 \leftrightarrow b_1, a_3 \leftrightarrow b_2, a_4 \leftrightarrow b_2, a_4 \leftrightarrow b_3)$ and the trace $T = (1,1),(2,1),(3,2),(4,2),(4,3)$ of size $|T| = 5$ with mapping distance of $\gamma(S) = d(a_1,b_1)+d(a_2,b_1)+d(a_3,b_2)+d(a_4,b_2)+d(a_4,b_3)$. If this sequence has the lowest mapping distance among all possible mapping sequences between A and B, then $\delta(A,B) = \gamma(S)$.

if edges are weighted by the matching costs the shortest path from (a_1, b_1) to (a_n, b_m) corresponds to the minimum cost matching. Since the mapping costs are symmetric, *i.e.* $\gamma(a \leftrightarrow b) = \gamma(b \leftrightarrow a)$, the mapping distance is also symmetric and we have $\delta(A,B) = \delta(B,A)$.

In the preceding discussion, we assumed that the first (and last) points on the two boundaries were matched. Since the points on the boundaries of the two shapes form cycles, we must consider all possible cyclical shifts of the boundaries. A cyclical shift σ^k of size k of the boundary $A = a_1a_2...a_n$ is defined by $\sigma^k(a_1a_2...a_n) = a_{k+1}...a_na_1...a_k, 1 \le k \le n$, and $\sigma^0(A) = A$. The equivalence class of A defined by k cyclic shifts will be denoted by $[A]$. Therefore:

$$\delta([A],[B]) := \min\left\{\delta\Big(\sigma^k(A),\sigma^l(B)\Big)\Bigg| \; 0 \le k < n, 0 \le l < m\right\} \qquad (10)$$

It can also be shown that $\delta([A],[B]) = \delta(A,[B])$. We define the *contour mapping measure (CM)* as the normalized mapping distance between the boundaries A and B:

$$CM(A,B) = \frac{1}{|T|}\delta([A],[B]) \qquad (11)$$

where T is the trace corresponding to the optimal mapping sequence and $|T|$, the size of the trace, is the number of mapped point pairs.

The distance $\delta(A,B)$ can be obtained by shortest path methods in the mapping graph, as explained above, or can be solved using dynamic programming, since the problem can be broken into sub-problems as follows. We define $A_i = a_1...a_i$ and $B_j = b_1...b_j$. We have $\delta(A_1,B_1) = \gamma(a_1 \leftrightarrow b_1) = d(a_1,b_1)$. For $i \in [2..n]$ and $j \in [2..m]$, we have:

$$\delta(A_i,B_j) = d(a_i,b_j) + \min\begin{cases}\delta(A_{i-1},B_{j-1})\\ \delta(A_{i-1},B_j)\\ \delta(A_i,B_{j-1})\end{cases} \qquad (12)$$

Using Dynamic Programming to find $\delta(A,B)$ has a complexity of $O(mn)$ since the distance calculation between n points on A and m points on B is $O(mn)$ and the dynamic programming table itself is of size mn. Assuming $m \le n$, constructing the same table for the m cyclic shifts of B will result in a complexity of $O(m^2n)$. Using a method similar to the method proposed by Maes [16] for string editing, the complexity can be reduced to $O(nm \log m)$.

By requiring explicit monotonic correspondence between points on the two shapes, the contour mapping measure (CM) avoids the problems experienced by other boundary measures (Figures 4 and 6), in which the error of very different shapes is minimized by implicitly mapping the same or nearby points on one shape to points that are widely separated (in arc-length) on the other curve. The code for the CM measure is available at `http://elderlab.yorku.ca/ContourMapping`.

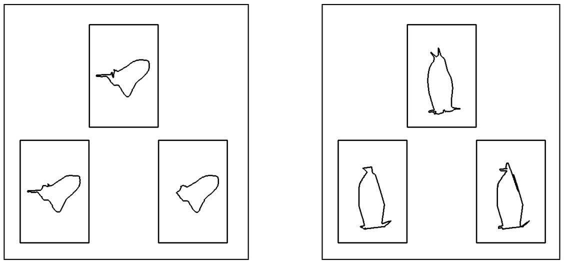
4. Psychophysical Experiments

Mumford [20] raised this question: *"There are many mathematical ways to define a numerical measure of the similarity of 2 shapes: do any of these approximate the human idea of similarity?"* Here we report the results of two psychophysical experiments that address this question. Specifically, we compare human judgements of shape similarity with decisions made based on the region intersection measure (RI) (Eq. 2), mean distance (MD) (Eq. 4), Hausdorff distance (HD) (Eq. 6), mixed measure (MM) (Eq. 7), and the contour mapping measure (CM) (Eq. 11).

4.1. General Methods

Both experiments consisted of a set of trials in which the subject was shown a reference shape A and two test shapes B and C (Figure 9), and asked to indicate which of the test shapes appeared more similar to the reference. The shapes were drawn from 30 of the 300 images in the BSD/SOD database that contain at least one completely unoccluded salient object . The shapes were displayed as outlines. (In preliminary experiments we found that judgements were similar for outlines and silhouettes).

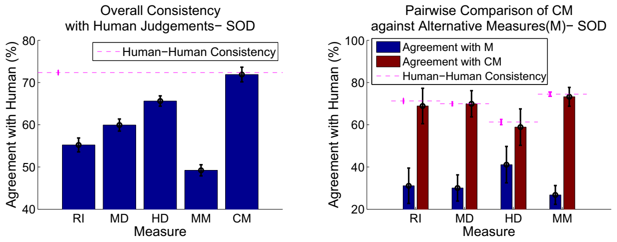
In both experiments the reference shape was an object segmentation from SOD. The two experiments differed only in the nature of the test shapes. In Experiment 1, the test shapes were other segmentations of the same object by other human subjects. In Experiment 2, they were machine-generated approximations of the reference shape. We used these two very different sets of stimuli in order to judge how well the results are likely to generalize.



(a)

(b)

Figure 9. Psychophysical displays. In both experiments, the reference shape A was shown at the top of the display, and the two test shapes B and C were shown at the bottom. (a) Experiment 1: reference and test shapes are segmentations of the same object by different human observers. (b) Experiment 2: reference is a human segmentation, while test shapes are approximations generated by an automatic algorithm. See text for details.



(a)

(b)

Figure 10. Results of Experiment 1. (a) Overall consistency with human subjects. (b) Consistency with human subjects for trials on which CM disagrees with each other measure $M \in \{RI, MD, HD, MM\}$.

Experiment	RI	MD	HD	MM
SOD	1.4e-4	1.1e-5	1.5e-2	2.8e-7
ALG	5.7e-2	1.6e-1	2.8e-3	3.5e-3

Table 1. p-values for pairwise repeated measures t-tests of CM versus the other four error measures.

The 9 subjects who participated in the two experiments were naïve to the exact purpose of the experiments. There was no time limit: subjects could view the shapes for as long as they wanted. In order to assess the consistency of each measure described in Section 3 with human judgements, we ran each measure as an 'observer' for the two experiments. On each trial, each of the measures was used to compute the similarity of the two test shapes to the reference shape, and the test shape with the higher similarity was 'selected' by the measure. Agreement with human subjects was then computed as the percentage agreement in the test shape selected over all trials.

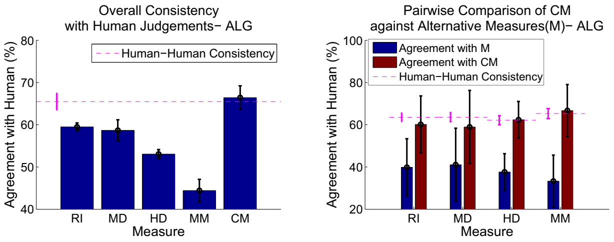
4.2. Experiment 1- SOD hand drawn boundaries

Stimuli: The three shapes within each trial were selected from different human segmentations of the same object in 30 images of the BSD images. The shape differences were thus due to inter-subject variations in the original BSD segmentations and/or the SOD constructions. Of the possible triplets, only those that generated disagreements between at least one pair among the 5 error measures were considered, since inclusion of pairs on which all measures agree would not serve to discriminate the measures. We used an automatic method to select from this subset 170 test pairs which were maximally different, in order to ensure they would be visually discriminable by our human subjects (see supplementary material for details).

Results: Figure 10(a) shows the overall consistency of each measure with the human subjects. The results show that the contour mapping measure (CM) is the most consistent with human judgements among all five measures. Since each of the 9 subjects saw the same stimuli, we could also compute an overall average consistency between our human subjects (pink dashed line). Remarkably, the CM algorithm is as predictive of human judgements as human subjects are of each other. The HD measure is the closest competition in this experiment. Figure 10(b) shows pairwise comparisons between the CM measure and each of the alternative measures on the subset of trials on which they disagreed. These differences are all statistically significant at the $\alpha < .05$ level (Table 1).

4.3. Experiment 2- Algorithm boundaries

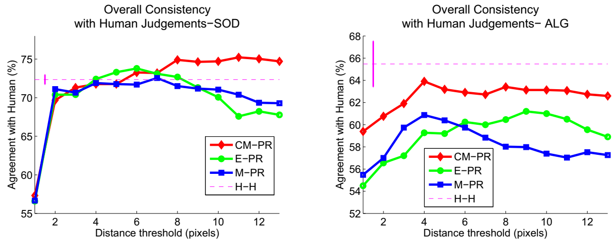
Stimuli: In Experiment 2, the reference shapes were again human segmentations drawn from the SOD database (Figure 9(b)). The two test shapes, however, were algorithm-generated boundaries approximating the reference shapes. The algorithm for shape approximation (explained in detail in supplementary material) takes as input a set of line segments automatically detected in the image, as well as the hand-drawn reference shape. The goal of the iterative algorithm is to find the cycle of line segments that best approximates the reference shape according to a specified error measure. Test shapes were shapes produced during intermediate iterations of the above algorithm, using each of the 5 measures. In particular, the two test shapes B and C were always selected to be intermediate shapes produced using the same measure, at least 2 and at most 10 iterations apart. We used an automatic procedure to select test shapes that were maximally different from each other, to ensure they would be discriminable by our human subjects, but still reasonably similar to the reference shape, so that the judgement was still meaningful (see supplementary material for details). From the resulting set of shape triplets, 600 were randomly selected for each subject. The trials had a minimum of 20 percent overlap between pairs of subjects to allow estimation of inter-subject consistency. 9 subjects participated in the experiment.



(a)

(b)

Figure 11. Results of Experiment 2. (a) Overall consistency with human subjects. (b) Consistency with human subjects for trials on which CM disagrees with each other measure $M \in \{RI, MD, HD, MM\}$.



(a)

(b)

Figure 12. Consistency with human subjects for F-measures based on different implementations of the Precision-Recall measures. The pink H-H line denotes human-human consistency.

Results: Figure 11 shows the results of this experiment. We find again that the CM algorithm is most consistent with human judgements, and again is as predictive of human judgements as humans are of each other. We find that pairwise differences between CM and the other measures are statistically significant at the $p < .05$ level for the HD and MM measures, but not for the RI and MD measures (Table 1).

Interestingly, while the Hausdorff (HD) measure ranked second in Experiment 1, the Region Intersection (RI) and Mean Distance (MD) measures score better in Experiment 2. This shows how the appropriateness of some measures can vary with the stimuli. At the same time, the CM measure performs well in both cases, suggesting that it may generalize well.

5. Precision-Recall Analysis

Precision and recall measures [18] have also been widely used for evaluation of segmentation algorithms. For an algorithm boundary A and a ground-truth boundary B, Precision is the proportion of boundary points on A that are true positives: Precision $= \frac{\text{Matched}(A,B)}{|A|}$, and Recall is the proportion of boundary points on B that are actually detected: Recall $= \frac{\text{Matched}(B,A)}{|B|}$. High precision corresponds to a low false positive rate, whereas high recall corresponds to a low false negative (miss) rate. In order to calculate these measures, a method for matching points on the two boundaries is required.

In their original Precision-Recall approach (M-PR), Martin et al. [18] solved the correspondence problem as a minimum cost bipartite matching, where the cost of matching two points is proportional to the distance between them. A 1:1 matching is possible by adding outlier nodes. Any match to an outlier or beyond some distance threshold is counted as a mismatch.In a recent variation on this approach (E-PR), Estrada et al. [6] include 'no intervening contours' and 'same side' constraints. While these constraints serve to encourage ordering consistency between the two contours being matched, neither approach strictly enforces ordering consistency in a global sense.

We can assess the significance of the ordering constraint within the Precision-Recall framework by using the CM method of section 3.4 to match points, and pruning matches beyond a distance threshold. Among multiple matches incident on one point, only the one with the shortest distance between the matched points is preserved and the rest are pruned. Since the matching step is independent of the distance threshold, changing the distance threshold does not require re-computation of matchings as required by Estrada's method.

To evaluate each of these P-R measures against our human data, we ran each measure through our experiments (Section 4), selecting the test shape that yielded the highest F-measure on each trial, where $F = \frac{PR}{\alpha R + (1-\alpha)P}$ with $\alpha = 0.5$, varying the distance threshold from 1 to 13 pixels. The results in Figure 12 show that best agreement with human judgements is obtained using the CM matching method, suggesting that the global ordering constraint is still important within the Precision-Recall framework. Note that performance is best at high threshold values, indicating the importance of allowing quite distant matches. Note also that the CM matching method appears to be relatively stable in consistency with human judgements once the distance threshold is sufficiently high, whereas the competing measures have a narrower 'sweet spot' at an intermediate threshold.

6. Future Work

Although the CM measure appears to model perceptual error for salient shape segmentation well for our stimuli, there are still challenges that remain to be addressed:

1. Mumford [20] has suggested that shape judgements are asymmetric and can depend upon context. All of the measures we consider here are symmetric, and are functions only of the segmented shapes.
2. Previous research shows the visual importance of

false-negative and false-positive pixels is not necessarily the same [19, 25]. Specifically, missing object parts tend to be more important than added background. The measures we consider treat shapes symmetrically and therefore ignore this perceptual difference.

3. The current form of our CM measure is limited to simply-connected shapes. For example shapes with holes cannot be compared using this measure. However the measure could potentially be generalized to other topologies by mapping each bounding contour separately, while enforcing topological constraints.

7. Conclusions

Empirical performance evaluation of object segmentation algorithms requires a dataset of ground truth object segmentations and an appropriate error measure. In this paper we have constructed a dataset of ground truth object segmentations that can be used for this purpose. We then considered 5 error measures that have appeared in various forms in the literature, and analyzed their potential strengths and weaknesses. Finally, we psychophysically evaluated these measures using two distinct types of stimuli. Our results show that a Contour Mapping measure based upon contour bimorphisms between the boundaries of the object segmentations under comparison were most consistent with human judgements, and, amazingly, were as predictive of human judgements as human subjects were of each other. We also proposed using the same matching paradigm in Precision-Recall analysis.

We believe that the perceptual consistency of the contour mapping measure derives from its sensitivity to prominent shape features that may have small area and its topological strictness, which requires that boundary points be considered as shapes rather than a scatter of points.

Acknowledgements This research was supported by GEOIDE, NSERC and OCE.

References

[1] S. Alpert, M. Galun, R. Basri, and A. Brandt. Image segmentation by probabilistic bottom-up aggregation and cue integration. In *CVPR 2007*.

[2] F. Attneave and M. D. Arnoult. The quantitative study of shape and pattern perception. *Psychological Bulletin*, 53(6):452–471, 1956.

[3] R. Basri, L. Costa, D. Geiger, and D. Jacobs. Determining the similarity of deformable shapes. *Vision Research*, 38:2365–2385, 1998.

[4] H. P. O. D. Beeck, K. Torfs, and J. Wagemans. Perceived shape similarity among unfamiliar objects and the organization of the human object vision pathway. *J of Neurosience*, 28(40):10111–10123, 2008.

[5] J. H. Elder, A. Krupnik, and L. A. Johnston. Contour grouping with prior models. *IEEE TPAMI*.

[6] F. Estrada and A. D. Jepson. Benchmarking image segmentation algorithms. *Int J Comut Vis*, 85:167–181, 2009.

[7] C. Fowlkes, D. Martin, and J. Malik. Local figure–ground cues are valid for natural images. *J of Vision*, 8:1–9, 2007.

[8] M. Frenkel and R. Basri. Curve matching using the fast matching method. *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 35–51, 2003.

[9] Y. Gdalyahu and D. Weinshall. Flexible syntactic matching of curves and its application to automatic hierarchical classification of silhouettes. *IEEE TPAMI*.

[10] F. Ge, S. Wang, and T. Liu. Image segmentation evaluation from the perspective of salient object extraction. *CVPR06*.

[11] D. Geiger, A. Gupta, L. A. Costa, and J. Vlontzos. Dynamic programming for detecting, tracking and matching deformable contours. *IEEE TPAMI*.

[12] L. Goldmann, T. Adamek, and P. Vajda. Towards fully automatic image segmentation evaluation. *Lecture Notes in Computer Science (LNCS)*, 5259:566–577, 2008.

[13] Q. Huang and B. Dom. Quantitative methods of evaluating image segmentation. In *Proc. 1995 Int. Conf. Image Processing*, volume 3, pages 53–56 vol.3, 1995.

[14] D. P. Huttenlocher, G. A. Klanderman, G. A. Kl, and W. J. Rucklidge. Comparing images using the hausdorff distance. *IEEE TPAMI*.

[15] T. Liu, J. Sun, N. Zheng, X. Tang, and H.-Y. Shum. Learning to detect a salient object. In *CVPR 2007*.

[16] M. Maes. On a cyclic string-to-string correction problem. *Information Processing Letters*, 35:73, 1990.

[17] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. 8th IEEE Int. Conf. Computer Vision*, volume 2, pages 416–423, 2001.

[18] D. R. Martin, C. C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE TPAMI*, 26(5):530–549, 2004.

[19] F. C. Monteiro and A. C. Campilho. Performance evaluation of image segmentation. *Lecture Notes in Computer Science*, 4141:248–259, 2006.

[20] D. Mumford. Mathematical theories of shape: do they model perception? In *Proc SPIE Vol 1570 Geometric Methods in Computer Vision*, pages 2–10, 1991.

[21] F. R. Schmidt, D. Farin, and D. Cremers. Fast matching of planar shapes in sub-cubic runtime. In *Proc IEEE 11th Int Conf Computer Vision, ICCV 2007*, pages 1–6, 2007.

[22] C. Scott and R. Nowak. Robust contour matching via the order-preserving assignment problem. In *IEEE Transactions on Image Processing*, volume 15, pages 1831–1837, 2006.

[23] T. B. Sebastian, P. N. Klein, and B. B. Kimia. On aligning curves. *IEEE TPAMI*.

[24] H. D. Tagare, D. O'shea, and D. Groisser. Non-rigid shape comparison of plane curves in images. *J of Mathematical Imaging and Vision*, 16:57, 2002.

[25] D. P. Young and J. M. Ferryman. Pets metrics: On-line performance evaluation service. In *VS-PETS*, 2005.

[26] Y. J. Zhang. A survey on evaluation methods for image segmentation. *Pattern Recognition*, 29(8):1335–1346, 1996.