Distance Measures for Image Segmentation Evaluation

Fernando C. Monteiro* and Aurélio C. Campilho†

*Polytechnic Institute of Bragança, Portugal
†INEB - Divisão de Sinal e Imagem
Instituto de Engenharia Biomédica,
Universidade do Porto, Faculdade de Engenharia, Porto, Portugal

Abstract. In this paper we present a study of evaluation measures that enable the quantification of the quality of an image segmentation result. Despite significant advances in image segmentation techniques, evaluation of these techniques thus far has been largely subjective. Typically, the effectiveness of a new algorithm is demonstrated only by the presentation of a few segmented images and is otherwise left to subjective evaluation by the reader. Such an evaluation criterion can be useful for different applications: the comparison of segmentation results, the automatic choice of the best fitted parameters of a segmentation method for a given image, or the definition of new segmentation methods by optimization. We first present the state of art of distance evaluation measures, and then, we compare several evaluation criteria.

Keywords: Benchmark, Distance measures, Ground truth, Segmentation evaluation.
PACS: 87.57.nm

INTRODUCTION

Despite the fact that image segmentation algorithms have been, and are still being, widely studied, quantitative evaluation of image segmentation quality is a much harder problem. The history of measures for evaluating segmentation algorithms is as old as the history of segmentation algorithms themselves.

Generally the evaluation methods of image segmentation can be classified into three categories: analytical methods, empirical goodness methods and empirical discrepancy methods. The empirical discrepancy methods have been the most commonly used methods for segmentation evaluation. These methods evaluate segmentation methods by comparing the segmented image against a manually segmented reference image, which is often referred as the ground truth, and computing error measures. One approach is to ask human subjects to segment the images by hand. If a reasonable consensus emerges, the hand segmentations can be treated as ground truth, and compared to the outputs of segmentation schemes. Martin *et al.* [1] presented a database containing hand segmented images from the Corel database, which we used in this study.

A potential problem for a measure of consistency between segmentations is that there is no unique segmentation of an image, since each human perceives the scene differently. For example, the numbers of segments may be different. Humans produce segmentations at different granularities and with different levels of detail, even when they perceive the image as having the same hierarchical tree structure [1].

In addition to a ground truth database, evaluating grouping algorithms requires an error measure. We want a measure exact enough to penalize systematic discrepancies with respect to the ground truth, yet tolerant to inter-subject variations. Under-segmentation is considered to be a serious problem as it is easier to recover true segments through a merging process after over-segmentation rather than trying to split an heterogeneous region. We think that one desirable property of a good evaluation measure is to accommodate refinement only in regions that human segmenters could find ambiguous and to penalize differences in refinements elsewhere. In addition to being tolerant to refinement, any evaluation measure should also be tolerant to different number of segments in each partition.

EVALUATION MEASURES

Reviewing work in the literature, one can find two kinds of empirical discrepancy methods: (1) region-based evaluation [1, 2, 3], which evaluates segmentation consensus in terms of the number of regions, and the locations, sizes and other statistics of the segmented regions, and (2) boundary-based evaluation [1, 2, 4], which evaluates segmentation in terms of both the location and shape accuracies of the extracted region boundaries.

Numerical Analysis and Applied Mathematics ICNAAM 2012
AIP Conf. Proc. 1479, 794-797 (2012); doi: 10.1063/1.4756257

Hamming Distance: Huang and Dom [2] introduced the concept of directional Hamming distance between two segmentations, denoted by $D_H(S \Rightarrow R)$. Let S and R be two segmentations. They began by establishing the correspondence between each region of S with a region of R such that $s_i \cap r_j$ is maximized. The directional Hamming distance from S to R is defined as:

$$D_H(S \Rightarrow R) = \sum_{r_i \in R} \sum_{s_k \neq s_j, s_k \cap r_i \neq 0} |r_i \cap s_k| \ , \tag{1}$$

where $|\cdot|$ denote the size of a set. Therefore, $D_H(S \Rightarrow R)$ is the total area under the intersections between all $r_i \in R$ and their non-maximal intersected regions from S. A region-based evaluation measure based on normalized Hamming distance is defined as

$$p = 1 - \frac{D_H(S \Rightarrow R) + D_H(R \Rightarrow S)}{2 \times |S|} \ , \tag{2}$$

where $|S|$ is the image size and $p \in [0,1]$. The smaller the degree of mismatch, the closer the p is to one.

Local Consistency Error: To compensate for the difference in granularity while comparing segmentations, many measures allow label refinement uniformly through the image. Martin *et al.* [1] proposed an error measure to quantify the consistency between image segmentations of differing granularities - *Local Consistency Error* (LCE) that allows labeling refinement between segmentation and ground truth.

$$LCE(S,R,p_i) = \frac{1}{N} \sum_i \min\{E(S,R,p_i), E(R,S,p_i)\} \ , \tag{3}$$

where $E(S,R,p)$ measures the degree to which two segmentations agree at pixel p, and N is the size of region where pixel p belongs.

Note that the LCE is an error measure, with a score 0 meaning no error and a score 1 meaning maximum error. Since LCE is tolerant to refinement, it is only meaningful if the two segmentations have similar number of segments.

Bidirectional Consistency Error: To overcome the problem of degenerate segmentations, LCE was adapted to a measure that penalizes dissimilarity between segmentations proportional to the degree of region overlap. The *Bidirectional Consistency Error* (BCE) is defined as:

$$BCE(S,R,p_i) = \frac{1}{N} \sum_i \max\{E(S,R,p_i), E(R,S,p_i)\} \ . \tag{4}$$

Partition Distance Measure: Cardoso and Corte-Real [3] proposed a discrepancy measure (d_{sym}) defined as: "given two partitions S_1 and S_2 of S, the partition distance is the minimum number of elements that must be deleted from S, so that the two induced partitions (S_1 and S_2 restricted to the remaining elements) are identical". $d_{sym}(S_1,S_2) = 0$ means that no points need to be removed from S to make the partitions equal, i.e., when $S_1 = S_2$.

Rand Index: Rand index (RI) [5] is the function that converts the problem of comparing two partitions with possibly differing number of classes into a problem of computing pair wise label relationships.

Consider two valid label assignments S_1 and S_2 that assign labels l_i and l'_i, respectively, to each point x_i. The RI can be computed as the ratio of the number of pairs of points having the same label relationship in S_1 and S_2, i.e.,

$$RI(S_1,S_2) = \frac{1}{\binom{N}{2}} \sum_{\substack{i,j \\ i \neq j}} \left[II\left(l_i = l_j \wedge l'_i = l'_j\right) + II\left(l_i \neq l_j \wedge l'_i \neq l'_j\right) \right] \ , \tag{5}$$

where II is the identity function and the denominator is the number of possible unique pairs among N data points.

Distance Distribution Signatures: Let B_S represent the boundary point set derived from the segmentation and B_R the boundary ground truth. A distance distribution signature [2] between the boundary points of two segmentations, denoted $D_B(S_1,S_2)$, is a discrete function whose distribution characterizes the discrepancy between segmentations. The distance from x in set S_1 to S_2 is defined as the minimum distance to S_2, $d(x,S_2) = \min\{d_E(x,y)\}, \forall y \in S_2$, where d_E denotes the Euclidean distance.

In order to normalize the result between 0 and 1, we proposed using $d(x,B_R) = \min\{d_E(x,y), c\}$, where the c value sets an upper limit for the error, allowing the two boundary distances to be combined in a framework similar to the one presented in Eq. (2):

$$b = 1 - \frac{D_B(B_S,B_R) + D_B(B_R,B_S)}{c \times (|R| + |S|)} \ , \tag{6}$$

where $|R|$ and $|S|$ are the number of boundary points in reference mask and segmented mask, respectively.

Precision-Recall Measures: Martin [1], proposed the use of *precision* and *recall* values to characterize the agreement between the oriented boundary edge elements of region boundaries of two segmentations. The two statistics may be distilled into a single figure of merit:

$$F = \frac{PR}{\alpha R + (1-\alpha)P} \ , \tag{7}$$

where α determines the relative importance of each term.

Since these measures are not tolerant to refinement, it is possible for two segmentations that are perfect mutual refinements of each other to have low precision and recall scores.

Earth Mover's Distance: The concept of using Earth Mover's Distance (EMD) to measure perceptual similarity between segmentations was first explored by Monteiro and Campilho[4]. EMD is defined as the the minimal sum of costs incurred to move all the individual points between the signatures. Let $S_1 = \{p_1, ..., p_m\}$ be the first signature with m pixels; the second signature with n pixels is represented by $S_2 = \{q_1, ..., q_n\}$. Let $D = [d_{ij}]$ be the distance matrix where d_{ij} is the distance between p_i and q_j. The flow f_{ij} is the amount of weight moved from p_i to q_j. Then the EMD is defined as the work normalized by the total number of pixels moved f_{ij}, that minimizes the overall cost:

$$EMD(S_1, S_2) = \sum_i \sum_j f_{ij} d_{ij} \Big/ \sum_i \sum_j f_{ij} \ , \tag{8}$$

In order to embed two sets of contour features with different total weights, [4] suggested adding "fake" pixels to the smaller set. The distance between any fake point and any true point is penalized with the maximum possible distance.

Perceptual Discrepancy Measure: This measure applies different weights to false negative and false positive pixels and is supported by research showing that the visual importance of these pixels is not the same and they should be treated differently [4]. As we move away from the border of an object, missing parts are more important than added background, e.g., in medical imaging, it may be enough that the segmented region overlaps with the true region, so the tumor can be located. But if there are missing parts of the tumor the segmentation results will be poor. Therefore, the following weights have been suggested by [4]:

$$w_p = \frac{\alpha_p \log(1 + d_p)}{D} \qquad w_n = \frac{\alpha_n d_n}{D} \ . \tag{9}$$

where d_p be the distance of a false positive pixel from the boundary of the reference region, d_n be the distance of a false negative pixel, and D be the image diagonal distance. Perceptual discrepancy measure is given by $s_w = 1 - \varepsilon_w$, where ε_w is the sum of the weighted functions.

COMPARATIVE STUDY

To achieve comparative results about different evaluation methods, two strategies can be followed: the first one consists in applying the evaluation methods to segmented images obtained from different segmentation approaches. The second one consists in simulating results of segmentation processes. To exempt the influence of segmentation algorithms, the latter has been adopted and a set of images obtained from manual segmentation available in [1] was used. Figure 1 presents an image, from the Berkeley Dataset, with six manual segmentations and three faked segmentations.

Results of boundary-based evaluation on the same set of images of Fig. 1 are reported in Fig. 2. On comparing the results of the boundary-based measures, it is made evident that they are well correlated. EMD tolerates well some amount of deformations that normally happens in the manual segmentation process. However, when the number of pixels in ground truth differs a lot from the number of pixels in the segmented image, EMD gives poor results. Despite its success, the EMD method still needs to be refined to address the limitation in the complexity of algorithm that require to be further reduced. The b-measure gives results similar with F-measure, but is even more intolerant to refinement.

Results of region-based evaluation show that LCE, BCE, d_{sym}, Rand Index and p, are just proportional to the total amount of false detections - different position of those pixels do not affect the similarity. This makes those methods unreliable for applications where the results will be presented to humans. Note that s_w produces results that agree with the visual relevance of errors.

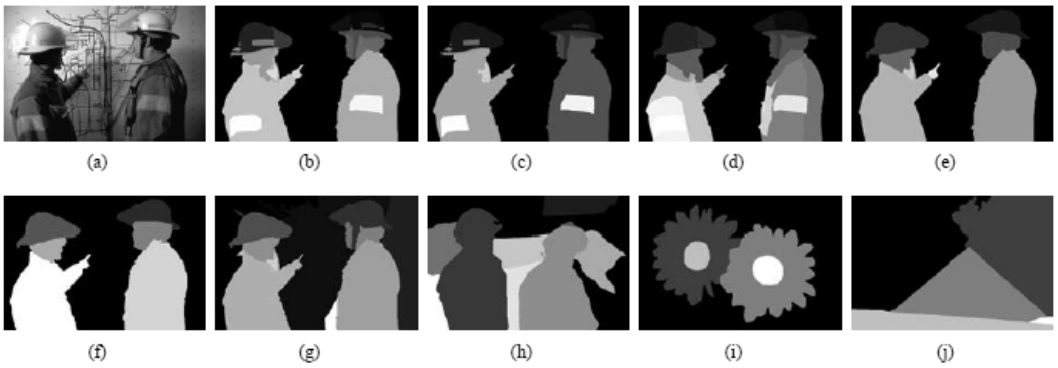
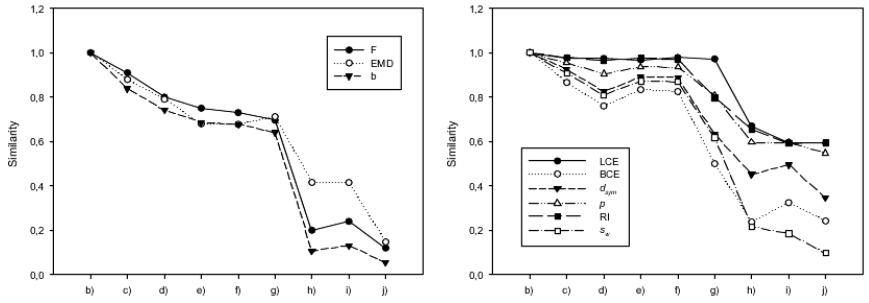


FIGURE 1. The image and its ground truth are shown in (a) and (b), respectively. From (c) to (g) we have different segmentations of image (a). Images (h) to (l) are wrong segmentations.



(a) Region-based (b) Boundary-based

FIGURE 2. Region-based and boundary-based evaluation of images from Fig. 1.

CONCLUSION

Segmentation evaluation is indispensable for improving the performance of existing segmentation algorithms and for developing new powerful segmentation algorithms. In spite of the number of evaluation segmentation algorithms presents in the literature, very few comparative results on evaluation of segmentation algorithms have been proposed. Typically, researchers show their results on a few images and point out why the results are good. We never know from such studies whether the results are best examples or typical examples, whether the technique will work only on images that have no texture, and so on. Moreover the measures used in the evaluations have weaknesses. An ideal measure should incorporate the similarity criteria used by human subjects and have 100% agreement with human subjects when deciding on the most similar shape to a reference.

REFERENCES

1. 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. of IEEE International Conference on Computer Vision*, 2001, vol. II, pp. 416–423, online at: *http://www.cs.berkeley.edu/projects/vision/grouping/segbench/*.
2. Q. Huang, and B. Dom, "Quantitative methods of evaluating image segmentation," in *Proc. of IEEE International Conference on Image Processing*, 1995, vol. III, pp. 53–56.
3. J. Cardoso, and L. Corte-Real, *IEEE Transactions on Image Processing* **14**, 1773–1782 (2005).
4. F. Monteiro, and A. Campilho, "Performance evaluation of image segmentation," in *Proc. of International Conference on Image Analysis and Recognition*, Póvoa de Varzim, Portugal, 2006, vol. 4141 of *LNCS*, pp. 248–259.
5. M. H. R. Unnikrishnan, C. Pantofaru, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **29**, 929–944 (2007).