QUANTITATIVE METHODS OF EVALUATING IMAGE SEGMENTATION

Qian Huang

Siemens Corporate Research, Inc.
755 College Road East
Princeton, NJ 08540
huang@scr.siemens.com

Byron Dom

IBM Research Division, Almaden Research Center,
650 Harry Road
San Jose, CA 95120-6099
dom@almaden.ibm.com

Abstract

Two sets of measures are proposed in this paper for quantitatively evaluating segmentation results. The first set is designed for the situation where ground truth is available; while the second for the situation where ground truth is not available. Based on a test bank of more than 50 images for which ground truth is available, we computed both sets of evaluation measures and then correlated the two sets. Experimental results show that the first set of measures proposed agree with human (subjective) visual evaluation and the second set of measures correlates well with the first set, an indication of the usefulness of this set of measures in assessing the quality of image segmentation results even when ground truth is not available.

1 Introduction

While the field of computer vision is growing into a discipline of science and engineering, quantitative approaches for evaluating image segmentations are needed. In this paper, we propose several quantitative measures for cases when ground truth is and is not available.

When ground truth is available, *parameter-based*, *boundary-based*, and *region-based* performance evaluation schemes are proposed. The first scheme is suitable only when the regions of interests can be modeled in a parametric form. The segmentation quality is evaluated in terms of the discrepancy between the model and ground truth parameters. Boundary-based approach evaluates segmentation in terms of both localization accuracy and the shape accuracy of extracted regions. Region-based approach assesses the segmentation quality in terms of both the size and location of the segmented regions. The measures of the last two schemes may fluctuate drastically in some extreme cases with the number of points involved in ground truth. Nevertheless, they serve what is intended in most situations.

We also propose quantitative measures for the situation in which ground truth is not available. Such measures are important because of (1) the difficulty in obtaining ground truth in practice and (2) the demand from many applications for a reasonable assessment on segmentation results. We propose performance measures that evaluate a segmentation in terms of the contrast between the heterogeneity among different regions versus the homogeneity within individual regions. This principle of evaluating segmentation results can be applied to both boundary and region based schemes once the definitions of homogeneity and heterogeneity are well understood within the context of applications.

2 Performance Evaluation Against Ground Truth

When ground truth is available, segmentation can be evaluated in terms of *accuracy* and *robustness*. ‘ The accuracy reflects the precision of segmentation with respect to ground truth. The robustness is related to the accuracy degradation with respect to the degradation of the quality of test data.

2.1 Parameter-Based Evaluation

Assume a ground truth image region, either a closed boundary or a connected component, can be represented by a set known parameters, $\mathcal{G}^P = \{\mathcal{G}^{p_1}, ..., \mathcal{G}^{p_k}\}$. $\mathcal{G}^P$ can be as simple as a constant intensity value or more complex such as the coefficients of a polynomial function approximating the intensity surface of the region or Fourier coefficients describing the shape of the closed boundary of the region. Let the corresponding region identified by a segmentation algorithm be represented by a parallel set of parameters $P = \{p_1, ..., p_k\}$. The *discrepancy* between $\mathcal{G}^P$ and P, defined as $\delta(\mathcal{G}^P, P) = \{\delta(\mathcal{G}^{p_i}, p_i) \mid 1 \leq i \leq k\}$, where $\delta(\mathcal{G}^{p_i}, p_i) = \mid \mathcal{G}^{p_i} - p_i \mid$, specifies the quality of the segmentation. A smaller deviation indicates a better performance. The amount of increase in these deviations with respect to the degree of degradation of the input image reflects the sensitivity of the algorithm to noise.

2.2 Boundary-Based Evaluation

This scheme is intended to assess segmentation qual-

0-8186-7310-9/95 $4.00 © 1995 IEEE

ity in terms of the accuracy of the extracted region boundaries. Let B be the boundary point set derived from the segmentation. Assume the boundary ground truth, denoted by $\mathcal{G}^B$, is available. The goal is to describe the discrepancy between $\mathcal{G}^B$ and B. We propose two *distance distribution signatures*, one from ground truth to the estimated, denoted by $\mathcal{D}_{\mathcal{G}}^{B}$, and the other from the estimated to ground truth, denoted by $\mathcal{D}_{B}^{\mathcal{G}}$. Several measures are further designed to characterize the discrepancy.

Distance Distribution Signatures

A distance distribution signature from a set B_1 to a set B_2 of boundary points, denoted by $\mathcal{D}_{B_1}^{B_2}$, is a discrete function whose distribution characterizes the discrepancy, measured in distance, from B_1 to B_2. Define the distance from an arbitrary point $\mathbf{x}$ in set B_1 to B_2 as the minimum absolute distance from $\mathbf{x}$ to all the points in B_2, $d(\mathbf{x}, B_2) = min\{d_E(\mathbf{x}, \mathbf{y})\}$, $\forall \mathbf{y} \in B_2$, where $d_E(\mathbf{x}, \mathbf{y})$ denotes the Euclidean distance between points $\mathbf{x}$ and $\mathbf{y}$. Signature $\mathcal{D}_{B_1}^{B_2}$ can be established from the distance histogram from individual $\mathbf{x} \in B_1$ to B_2, which may be estimated through a distance transformation with respect to B_2.

The shape of signature $\mathcal{D}_{B_1}^{B_2}$ describes the discrepancy or ultimately the degree of match (or mismatch) between B_1 and B_2. A number of statistics can be computed to reflect their shapes. The most common measures are its mean and standard deviation Since means are known to be sensitive to outliers, the median may serve the purpose better. In general, $\mathcal{D}_{B_1}^{B_2}$ is skewed. Therefore, skewness can also be used to characterize $\mathcal{D}_{B_1}^{B_2}$. A perfect match between B_1 and B_2 should yield zero mean and zero standard deviation, signaling that B_1 and B_2 completely coincide with each other. Generally, a $\mathcal{D}_{B_1}^{B_2}$ with a near-zero mean and a small standard deviation indicates high quality of the image segmentation. A large standard deviation may reveal the existence of outliers. In that case, median provides a better indication in terms of the accuracy of the segmentation.

Weighted Boundary Segmentation Error Rates

There are two types of errors in boundary segmentation: *missing boundary rate* e_B^m and *false boundary rate* e_B^f. The former specifies the percentage of the points on $\mathcal{G}^B$ that are mistakenly classified as non-boundary points; while the latter, e_B^f, indicates the percentage of the points in B that are actually false alarms. Therefore, $e_B^m = \frac{|T1|}{|\mathcal{G}^B|}$ and $e_B^f = \frac{|T2|}{|B|}$, where $T1 = \{\mathbf{x} \mid (\mathbf{x} \in \mathcal{G}^B) \wedge (\mathbf{x} \notin B)\}$, and $T2 = \{\mathbf{x} \mid (\mathbf{x} \in B) \wedge (\mathbf{x} \notin \mathcal{G}^B)\}$. Weighted boundary segmentation error rates are defined as weighted e_B^m and e_B^f, denoted by (e_B^m, w_B^m) and (e_B^f, w_B^f), where weights are the average distances between misclassified points to the ground truth boundary. Such weighted errors are useful because merely a percentage indicates only how many points do not coincide but not how far apart they are to the right positions which is now captured by weights w_B^m and w_B^f.

2.3 Region-Based Evaluation

This approach evaluates the segmentation accuracy in the number of regions, the locations, and the sizes. Let the segmentation be S and the corresponding ground truth be $\mathcal{G}^S$. Both S and $\mathcal{G}^S$ are functions on the image plane with labels as their function values. The goal is to quantitatively describe the degree of mismatch between S and $\mathcal{G}^S$.

Overall Performance Measure

We first introduce the concept of *directional Hamming distance* [3] from one segmentation $S_1 = \{R_1^1, R_1^2, ..$ to another segmentation $S_2 = \{R_2^1, R_2^2, .., R_2^{n_2}\}$, denoted by $D_H(S_1 \Longrightarrow S_2)$. In region-based evaluation, we first establish the correspondence between the labels of two given segmentations using the following scheme: associate each region R_2^i from S_2 with a region R_1^j from S_1 such that $R_2^i \cap R_1^j$ is maximal. Directional Hamming distance from S_1 to S_2 is defined as:

$$D_H(S_1 \Longrightarrow S_2) = \textstyle\sum_{R_2^i \in S_2} \sum_{R_1^k \neq R_1^j, R_1^k \cap R_2^i \neq \emptyset} \mid R_2^i \cap R_1^k \mid,$$

where $|.|$ denote the size of a set. Therefore, $D_H(S_1 \Longrightarrow S_2)$ is the total area under the intersections between all $R_2^i \in S_2$ and their non-maximal intersected regions R_1^k's from S_1. The reversed distance $D_H(S_2 \Longrightarrow S_1)$ can be similarly computed. Define a region-based performance measure based on *normalized Hamming distance* as $p = 1 - \frac{D_H(S \Longrightarrow \mathcal{G}^S) + D_H(\mathcal{G}^S \Longrightarrow S)}{2 \times |S|}$, where $\mid S \mid$ is the image size and $p \in [0, 1]$. The smaller the degree of mismatch, the closer the p is to one.

Region Segmentation Error Rates

We define two types of errors in region segmentation: *missing rate* e_R^m and *false alarm rate* e_R^f. The former indicates the percentage of the points in $\mathcal{G}^S$ being mistakenly segmented into the regions in S that are non-maximal with respect to the corresponding region in $\mathcal{G}^S$; while the latter describes the percentage of the points in S falling into the regions of $\mathcal{G}^S$ that are non-maximal intersected with the region under consideration. With previously defined directional Hamming distance, we therefore have

$$e_R^m = \frac{D_H(S \Longrightarrow \mathcal{G}^S)}{|S|}, \text{ and } e_R^f = \frac{D_H(\mathcal{G}^S \Longrightarrow S)}{|S|}.$$

3 Performance Evaluation Without Ground Truth

When ground truth is not available, evaluating segmentation results is similar to evaluating cluster validity. A segmentation algorithm partitions an image plane F, yielding M non-overlap regions, $\{R_1, .., R_M\}$, such that $\cup_{i=1}^{i=M} R_i = F$, $R_i \cap R_j = \emptyset, \forall i \neq j$. A generic criterion of creating such a partition is that the homogeneity within segmented regions and the heterogeneity among different regions are simultaneously maximized.

Let H_w denote the total within-region heterogeneity and H_b denote the total between-region heterogeneity. The first generic measure we propose for evaluating a segmentation without ground truth is $p = (H_b + 1)/(H_b + H_w + 1)$. This ratio reaches one as the internals of all regions are completely homogeneous ($H_w = 0$) and close to zero when there is simultaneously no discriminability between regions ($H_b = 0$) and no internal homogeneity (large H_w). Another generic measure we propose is a score from a statistical test that indicates how far the distribution of within-region heterogeneity aparts from the distribution of between-region heterogeneity. A good segmentation corresponds to two significantly far apart distributions. Various statistical test methods can be used. We currently use the Kolmogorov-Smirnov Test which is effective due to the large sample set in the context of image processing. While these generic measures are flexible because conceptually they can be applied to all three evaluation schemes discussed in section 2, they are computable only when the definition of heterogeneity is made explicit in application.

4 Experimental Results of Evaluating Image Segmentations

Due to the space limit, we show only some applications of these measures to illustrate their usefulness. In More applications are described in[1], Figure 1 gives the segmentations of two images derived at different stages of an incremental segmentation scheme[2] and their ground truth. Table 1 lists the boundary-based evaluation using their ground truth, where the left group of measures (columns 2-4) are from the initial segmentations (column 3 in Figure 1) and the second group (column 5-7) from the final segmentations (column 4 of Figure 1). As we can see, these measures agree with human visual evaluation because while the algorithm incrementally improves the segmentation (from initial to final), the weighted error rates drop and the confidence measure C (by combining different measures, see[1]) increases. The plots in Figure 2 correspond to boundary and region evaluation on 30 segmentations using Kolmogorov-Smirnov test scores. In each plot, the solid curve is the evaluation using ground truth and the dashed curve using segmentations only. In the case of boundary-based evaluation (Figure 2(a)), heterogeneity is defined as edgeness measured in gradient magnitude and the correlation between the two curves is 0.953723. In the case of region-based evaluation (Figure 2(b)) where the heterogeneity is defined as the deviation of intensity values in color bands from a reconstructed background, the correlation is 0.790447. The significant correlations between two distributions in both plots indicates that the measures proposed in Section 3 are informative in assessing segmentation results when ground truth is not available.

5 Concluding Remarks

In this paper, we proposed two groups of measures for evaluating image segmentation results, one for the situation where ground truth is available and the other for the situation where ground truth is not available. The first group of measures are useful because we can use them to assess the performance and the robustness of the algorithm from which the segmentation is produced. The second group is useful because with the difficulty or expensive cost of obtaining ground truth, many applications need a reasonable estimate, without using ground truth, about the quality of the segmentation before it can be utilized. We applied the proposed measures to some segmented images and present some of the results in this paper to show that (1) the first group of measures agrees with human subjective visual evaluation and (2) the second group of measures correlates well with the first group, indicating that they can provide a reasonable assessment, without using ground truth, about the quality of the segmentation to be evaluated.

References

[1] Qian Huang. *Hierarchical Token Grouping in Extracting Tubular Objects.* PhD thesis, Michigan State University, 1994.

[2] Qian Huang, Byron Dom, David Steele, Jon Ashley, and Wayne Niblack. Foreground/background segmentation of color images by integration of multiple cues. In *Proceedings of 1995 IEEE Conference on Image Processing*, Washington D.C., U.S.A., 1995.

[3] Tapas Kanungo, Byron Dom, Wayne Niblack, and David Steele. A fast algorithm for mdl-based multiband image segmentation. Technical Report RJ 9754 (84640), IBM Research Division, Almaden Research Center, San Jose, CA, 1994.

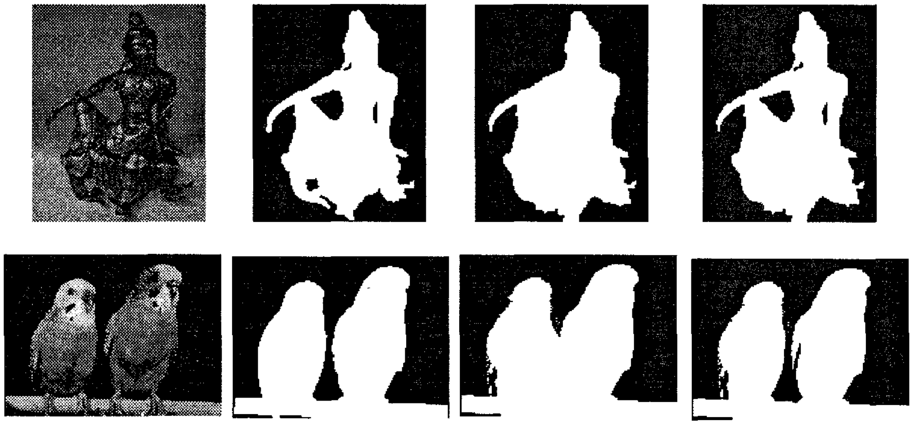
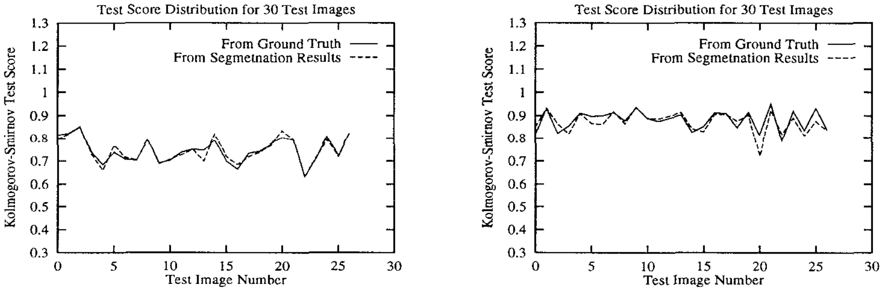


Figure 1: Two sets of segmentation results. Column 1: input images; column 2: ground truth; column 3: initial segmentations; column 4: final segmentations.

Table 1: Boundary-based evaluation on two sets of input images using ground truth.

Image	$(e_B^m, w_B^m)_{init}$	$(e_B^f, w_B^f)_{init}$	C_{init}	$(e_B^m, w_B^m)_{fnl}$	$(e_B^f, w_B^f)_{fnl}$	C_{fnl}
Statue	(0.058, 4.268)	(0.107, 15.489)	0.9286	(0.072, 4.390)	(0.008, 3.444)	0.9763
Birds	(0.078, 3.227)	(0.205, 9.864)	0.9202	(0.069, 3.190)	(0.111, 6.153)	0.9563



(a) (b)

Figure 2: With and without ground truth evaluation based on (a) boundary segmentation, (b) region segmentation.