Measures of Similarity

Ranjith Unnikrishnan Martial Hebert
The Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
{*ranjith,hebert*}*@ri.cmu.edu*

Abstract

Quantitative evaluation and comparison of image segmentation algorithms is now feasible owing to the recent availability of collections of hand-labeled images. However, little attention has been paid to the design of measures to compare one segmentation result to one or more manual segmentations of the same image. Existing measures in statistics and computer vision literature suffer either from intolerance to labeling refinement, making them unsuitable for image segmentation, or from the existence of degenerate cases, making the process of training algorithms using the measures to be prone to failure. This paper surveys previous work on measures of similarity and illustrates scenarios where they are applicable for performance evaluation in computer vision. For the image segmentation problem, we propose a measure that addresses the above concerns and has desirable properties such as accommodation of labeling errors at segment boundaries, region sensitive refinement, and compensation for differences in segment ambiguity between images.

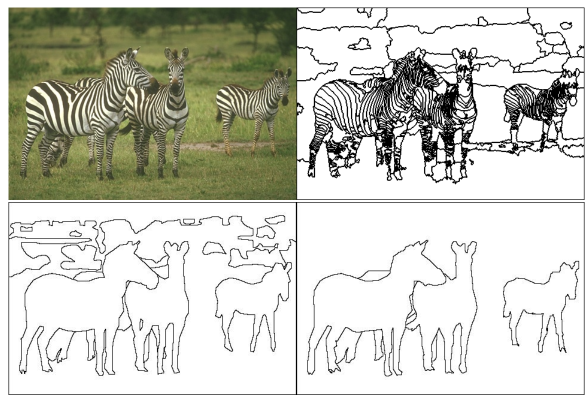

Figure 1: Sample picture (top left) with a boundary map generated from mean-shift clustering in LUV space (top right) with $h_s = 10$, $h_{\text{color}} = 7$. Examples of manual segmentations from the Berkeley dataset are shown in the bottom row.

1. Introduction

Segmentation is an important component of image understanding and data mining systems for discovering groups and identifying interesting distributions and patterns in input data. Recent efforts in amassing hand-labeled segmentations from a variety of natural images [5] have highlighted the need for principled ways to correctly quantify the performance of existing segmentation algorithms. A similar need arises when trying to assess the level of agreement between the results of different clustering algorithms [2] or when computing the stability of a particular label assignment to perturbations in input data (eg. for model selection [3]). In the image analysis domain, the comparison of two segmentations is difficult as image segmentation is inherently ill-defined – there is no *single* ground truth label assignment that can be used for comparison. To compensate for this, proposed measures [4] are often designed to allow pure refinements of label assignments without penalty. This approach inevitably allows degenerate cases to have very high similarity scores, thus making the scores meaningful only when the two segmentations being compared have similar cardinality of classes.

In addition, scores of performance obtained with refinement-invariant measures can be misleading when aggregated over multiple datasets. As an artifact of the scoring system, an algorithm may produce segmentations with high scores of similarity to ground truth for images that humans find ambiguous to segment, as well as "easy" images with unambiguous segment boundaries. Consequently, when interpreting the average score, one algorithm may be chosen over another for its mistakenly high performance on images with inherently ambiguous segments rather than ones with clear segment boundaries. Hence, one desirable property of a good measure is to accommodate refinement *only* in regions that human segmenters find ambiguous and to penalize differences in refinement elsewhere.

In recent work [3], authors have proposed stability [1] as a criterion for assessing the quality of a cluster using a given algorithm. Little variation in label assignment under perturbation of input data indicates the compactness and isolation of the cluster without resorting to an explicit generative model. However, a direct comparison between a cluster and a partition is not straightforward. The approach in [3] is to construct an approximate equivalent of the cluster to be tested by merging clusters in the ground truth partition to obtain the "best" approximating partition. The cluster and its constructed approximation are then compared using a normalized mutual information criterion. This paper explores methods that compare partitions directly irrespective of the difference in cardinality of the labels in each partition.

In Section 2, we survey measures of similarity popular in the statistics and vision literature, and discuss their applicability and associated drawbacks as performance metrics for image segmentation. Section 3 proposes a modified measure with improved properties and presents supporting experimental results.

2. Existing Measures

2.1. Notation

Let $X = \{x_1, x_2, \dots x_i, \dots, x_N\}$ denote data to be labeled (eg. N pixels of an image X). A valid label assignment S maps each pixel x_i to a single label l_i. Let $S_1, S_2 \dots S_K$ denote a set of provided label assignments (eg. manually labeled images), representing (say) the output of K algorithms. In each S_k, pixel labels can take one of L_k values. The subscript in L_k is used to emphasize the possible difference in cardinality of labels in each S_k. The underlying assignment procedure is not of immediate concern, and the sets $\{l_i\}$ are assumed to result from some unknown segmentation, labeling or rating algorithm.

We will refer to a valid assignment as a *segmentation* or a *partition* and look for measures of the form $d(S_1, S_2, \dots, S_K)$. Popular measures typically deal with the case of $K = 2$, to compare two segmentations. In the case of $K > 2$, the measure can be interpreted as quantifying the extent to which the set of segmentations do not conflict. Not surprisingly, these measures are known in the statistics literature under the terms *measures of similarity* or *measures of agreement*.

In our discussion of measures popular in the literature, we pay attention to the following factors that characterize and influence a choice of measure:

- Supported label type: Two types of labels are of particular interest – (a) semantic, non-permutable labels eg. 'grass', 'water', 'horse' etc. , and (b) nominal, permutable labels that are unordered and without semantic meaning, as from unsupervised clustering.
- Requirements in label cardinality: i.e. whether the number of classes, L_k , are equal for all k. While this may seem too strict a requirement for use of the measure, it is useful in scenarios where the number of classes is fixed by design. For example, computing a consistency measure for the output of several k-means algorithms where k is fixed.
- Tolerance to label refinement.
- Whether the comparison can be extended to a set of segmentations with $K > 2$.

2.2. Cohen's Kappa

Cohen's κ statistic [9] is a popular measure for measuring degree of similarity (or agreement) between two raters. Cohen assumed that there were two raters, who rate n subjects into one of m mutually exclusive and exhaustive semantic categories.

Let p_{ij} be the proportion of subjects that were placed in the (i, j)-th cell, i.e. assigned to the ith category by the first rater and the jth category by the second ($i, j = 1 \dots m$). Let $p_{i\bullet} = \sum_{j=1}^{m} p_{ij}$ denote the proportion of subjects placed in the ith row (i.e. ith category by the first rater. Similarly let $p_{\bullet j} = \sum_{i=1}^{m} p_{ij}$. Then, the kappa coefficient proposed by Cohen is

$$\hat{\kappa} = \frac{p_0 - p_c}{1 - p_c} \tag{1}$$

where $p_0 = \sum_{i=1}^{m} p_{ii}$ is the observed proportion of agreement, and $p_c = \sum_{i=1}^{m} p_{i\bullet} p_{\bullet i}$ is the proportion of agreement expected by chance.

Values for $\hat{\kappa}$ range in $[0, 1]$ with $\hat{\kappa} > 0.75$ interpreted as excellent agreement beyond chance. An example application of Cohen's κ is in comparing model-driven segmentations eg. segmentation of an urban image into categories $\{$*sky*, *vegetation*, *road*, *building*, *vehicle*, *other*$\}$ or a natural image into $\{$*man-made*, *natural*$\}$ categories.

Extensions to Cohen's measure have been proposed [10] for more than two raters. Although accepted as a reliable test of rater independence, there seems to be much disagreement about the usefulness of the κ-measure for quantifying levels of agreement. Tests for marginal homogeneity [11] have been proposed along similar lines for 2-way (McNemar test) and K-way (Stuart-Maxwell test) contingency tables. But like the κ statistic, their extension to the labeling problem requires evaluating all permutations of label assignments and possible merges of labels, and is computationally impractical.

2.3. Rand index

William Rand [8] proposed a similarity function that converted the problem of comparing two partitions with possibly differing number of classes into a problem of computing pairwise label relationships.

Consider two valid label assignments S and S' with corresponding labels $\{l_i\}$ and $\{l'_i\}$ of N points $X = \{x_1, x_2, \ldots x_i, \ldots, x_N\}$. The Rand index R can be computed as the ratio of the number of pairs of points having the compatible label relationship in S and S'. i.e

$$R(S, S') = \frac{1}{\binom{N}{2}} \sum_{\substack{i,j \\ i \neq j}} \Big[\mathbf{I}\big(l_i = l_j \wedge l'_i = l'_j\big) + \mathbf{I}\big(l_i \neq l_j \wedge l'_i \neq l'_j\big) \Big] \tag{2}$$

where $\mathbf{I}$ is the identity function, and the denominator is the number of possible unique pairs among N data points. Note that there is no restriction of the number of unique labels in S and S' being the same.

Another way of expressing this quantity is as follows. Let n_{uv} be the number of points having label u in S and label v in S'. We denote the number of points having label u in the first partition S as $n_{u\bullet}$ and the number of points having label v in the second partition S' as $n_{\bullet v}$. Then:

$$n_{u\bullet} = \sum_v n_{uv} \qquad n_{\bullet v} = \sum_u n_{uv}$$

Clearly $\sum_u n_{u\bullet} = \sum_v n_{\bullet v} = N$, the total number of data points.

It can be shown that the Rand index can be written in the form:

$$\mathrm{R}(S, S') = 1 - \frac{\left[\frac{1}{2}\Big(\sum_u n_{u\bullet}^2 + \sum_v n_{\bullet v}^2 \Big) - \sum_{u,v} n_{uv}^2\right]}{N(N-1)/2} \tag{3}$$

which is computationally inexpensive when the number of unique labels in S and S' are much smaller than the number of data points N.

This gives a measure of similarity with value ranging from 0 when the two segmentations have no similarities (i.e. when one consists of a single cluster and the other consists only of clusters containing single points) to 1 when the segmentations are identical.

2.4. Adjusted Rand index

One complaint about Eqn. 3 is that the expected value of the Rand index does not take a constant value. The adjusted Rand index proposed by [7] assumes the generalized hypergeometric distribution as the model of randomness, i.e. the S and S' partitions are picked at random such that the number of points having any particular label is fixed for both segmentations.

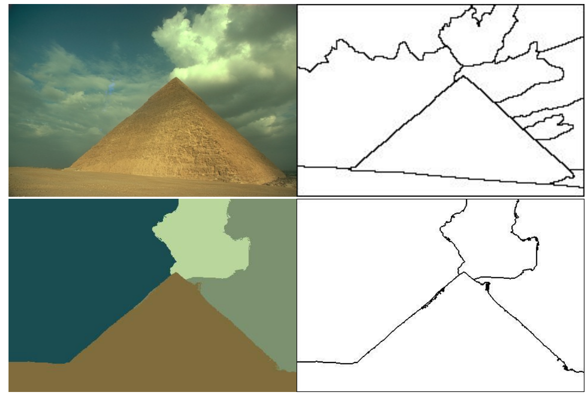

Figure 2: Sample picture (top left) with example of boundary map generated from manual segmentation (top right). Bottom row shows oversegmentation using mean-shift clustering in LUV space $h_s = 30$, $h_{\text{color}} = 20$.

Under the generalized hypergeometric model, it is shown [7] that:

$$\mathbb{E}\left[\sum_{u,v} \binom{n_{uv}}{2}\right] = \left[\sum_u n_{u\bullet} \sum_v n_{\bullet v}\right] / \binom{N}{2}$$

and with some algebra, the Adjusted Rand index can then be expressed as:

$$\mathrm{AR}(S, S') = \frac{\sum_{u,v} \binom{n_{uv}}{2} - \mathbb{E}\left[\sum_{u,v} \binom{n_{uv}}{2}\right]}{\frac{1}{2}\left[\sum_u n_{u\bullet} + \sum_v n_{\bullet v}\right] - \mathbb{E}\left[\sum_{u,v} \binom{n_{uv}}{2}\right]} \tag{4}$$

The adjusted index has the property of having expected value equal to 0 and maximum value of 1. Since the unadjusted Rand index was in the range $[0, 1]$, the adjusted index can take on a wider range of values, increasing the sensitivity of the measure. Because of these properties, the adjusted Rand index is a popular choice in the bioinformatics community (eg. [6]).

In the context of image segmentation, it is generally agreed upon that interpretations of images by human subjects differ in pixel-level granularity of label assignments, but are consistent if refinements of classes are admissible. However, this admissibility is not accommodated in either adjusted or unadjusted Rand measures.

2.5. Boundary-based Segmentation Consistency

D. Martin in his thesis [4] proposed a battery of segmentation comparison measures. One of them was geared to-

Table 1: Properties of various measures of similarity

	Cohen's κ	Rand index	GCE/LCE	Boundary-based	Prob. Rand
Label type	non-permutable	permutable	permutable	permutable	permutable
Required cardinality of labels	equal	any	similar	similar	any
Allows label refinement	✗	✗	✓	✗	✓
Comparison of $K > 2$ segments	✓	✗	✗	✗	✓

ward comparing boundary maps and worked by computing a minimum cost assignment between two sets of oriented boundary edge elements. A bipartite graph was constructed with each set consisting of the boundary edge elements (termed *edgels*) each having the triplet of parameters (x, y, θ) encoding position and orientation, and edge weights proportional to the euclidean distance between the edgels. The minimum cost perfect matching of the graph provides a correspondence between edgels of two candidate segmentations. Outlier nodes were added on both sides of the graph to account for differing cardinality of edge elements and make the overall problem sparse.

Let the two segmentations be S_i and S_j. Consider one of the segmentations, say S_j, to be ground truth. The fraction of matched S_i edgels represented *precision* and the fraction of matched S_j edgels represented *recall*. However, since these measures are not tolerant of refinement, it is possible for two segmentations that are perfect mutual refinements of each other to have very low precision and recall scores.

2.6. Region-based Segmentation Consistency

Martin et al. [5][4] proposed two error measures to quantify the consistency between image segmentations of differing granularities, and used them to compare the results of normalized-cut algorithms to a database of manually segmented images.

Let S and S' be two segmentations as before. For a given point (pixel) x_i, consider the classes (segments) that contain x_i in S and S'. We denote these sets of pixels by $C(S, x_i)$ and $C(S', x_i)$ respectively. Following [5], the local refinement error (LRE) is then defined at point x_i as:

$$\mathrm{LRE}(S, S', x_i) = \frac{|C(S, x_i) \setminus C(S', x_i)|}{|C(S, x_i)|} \tag{5}$$

This error measure is not symmetric and encodes a measure of refinement in one direction only. With this there are two natural ways to combine the LRE at each point into a measure for the entire image. Global Consistency Error (GCE) forces all local refinements to be in the same direction and is defined as:

$$\begin{aligned}\mathrm{GCE}(S, S') = \frac{1}{N} \min\Big\{ &\sum_i \mathrm{LRE}(S, S', x_i), \\ &\sum_i \mathrm{LRE}(S', S, x_i) \Big\}\end{aligned} \tag{6}$$

Local Consistency Error (LCE) allows for different directions of refinement in different parts of the image:

$$\begin{aligned}\mathrm{LCE}(S, S') = \frac{1}{N} \sum_i \min\big\{ &\mathrm{LRE}(S, S', x_i), \\ &\mathrm{LRE}(S', S, x_i) \big\}\end{aligned} \tag{7}$$

As LCE $\leq$ GCE, it is clear that GCE is a tougher measure than LCE.

Both measures have the advantage of being tolerant of refinement. However they are only meaningful if the two segmentations being compared have similar number of segments. As the authors point out [5], there are two segmentations that give zero error – one pixel per segment, and one segment for the whole image. Although they are two degenerate cases, it adversely limits the use of the errors functions. The existence of degenerate solutions that minimize both scores tends to bias algorithms that are trained to explicitly minimize these measures. This can lead to poor segmentation solutions and suboptimal algorithm selection. Furthermore, because the measures are not affected by the intrinsic ambiguity of the image, an aggregate score obtained from comparing several result-ground truth image pairs need not represent the strength of the algorithm. The score proposed in the next section attempts to address these deficiencies.

3. A Probabilistic Rand index

The section introduces a measure that combines the desirable statistical properties of the Rand index with the ability to accommodate refinements appropriately. Since the latter property is relevant primarily when quantifying consistency of image segmentation results, we will focus on that application while describing the measure.

Consider a set of manually segmented (ground truth) images $\{S_1, S_2, \ldots, S_K\}$ corresponding to an image $X =$

$\{x_1, x_2, \ldots x_i, \ldots, x_N\}$, where a subscript indexes one of N pixels. Let S be the segmentation that is to be compared with the manually labeled set. We denote the label of point x_i by l_i in segmentation S and by $l_i^{(k)}$ in the manually segmented image S_k. For convenience of notation, we assume the existence of a set of “true labels”, which we denote by $\hat{l}_i$ for the pixel x_i. Although there is arguably not *one* but many correct sets of labels, the proposed measure only considers the distribution of pairwise relationships between pixels and not the values defined by one dataset. Our goal is to compare a *candidate* segmentation S to this set and obtain a suitable measure of consistency $d(S, S_{1\ldots K})$.

Given the manually labeled images, we can compute the empirical probability of the label relationship of a pixel pair x_i and x_j simply as:

$$\hat{P}(\hat{l}_i = \hat{l}_j) = \frac{1}{K}\sum_{k=1}^{K} \mathbf{I}\left(l_i^{(k)} = l_j^{(k)}\right) \quad (8)$$

and

$$\hat{P}(\hat{l}_i \neq \hat{l}_j) = \frac{1}{K}\sum_{k=1}^{K} \mathbf{I}\left(l_i^{(k)} \neq l_j^{(k)}\right) = 1 - \hat{P}(\hat{l}_i = \hat{l}_j) \quad (9)$$

Consider the probabilistic Rand (PR) index:

$$\mathrm{PR}(S, S_{\{1\ldots K\}}) = \frac{1}{\binom{N}{2}} \sum_{\substack{i,j \\ i \neq j}} \left[\, \mathbf{I}(l_i = l_j)\, P(\hat{l}_i = \hat{l}_j) + \mathbf{I}(l_i \neq l_j)\, P(\hat{l}_i \neq \hat{l}_j) \,\right] \quad (10)$$

This measure takes values in $[0, 1]$ – 0 when S and $\{S_1, S_2, \ldots, S_K\}$ have no similarities (i.e. when S consists of a single cluster and each segmentation in $\{S_1, S_2, \ldots, S_K\}$ consists only of clusters containing single points, or vice versa) to 1 when all segmentations are identical. We analyze the properties of this measure in the subsections that follow.

3.1. Dataset dependent upper bound

We illustrate the dependence of the upper bound on the dataset $S_{\{1\ldots K\}}$ with a toy example. Consider an image X consisting of N pixels. Let two manually labeled segmentations S_1, S_2 (as shown in Fig. 3) be made available to us. Let S_1 consist of the entire image having one label. Let S_2 consist of the image segmented vertically into two equal halves, each half with a different label. Let the left half be denoted region $R1$ and the right half as region $R2$.

The pairwise empirical probabilities for each pixel pair is straightforward and can be summarized as.

$$\hat{P}(\hat{l}_i = \hat{l}_j) = \begin{cases} 1 & \text{if } (x_i, x_j) \in R1 \vee (x_i, x_j) \in R2 \\ 0.5 & \text{if } (x_i \in R1 \wedge x_j \in R2) \\ 0.5 & \text{if } (x_i \in R2 \wedge x_j \in R1) \end{cases}$$

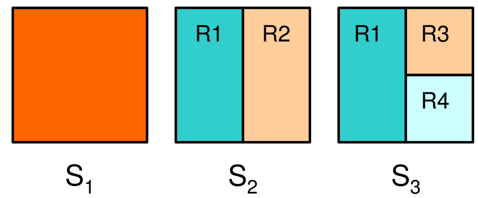

Figure 3: A toy example of the PR index computed for a manually labeled set of segmentations. See text for details.

The above relation encodes that it is equally ambiguous to human subjects as to whether the image is a single segment or two equally sized segments. This sets an upper bound on $\mathrm{PR}(S, S_{\{1,2\}})$ for any candidate S as a function of N which in the limit gives:

$$\lim_{N\to\infty} \mathrm{PR}_{\max}(S, S_{\{1,2\}}) = \frac{3}{4}$$

It can be shown that if the test segmentation S is identical to S_1 or S_2, the measure remains the same in value and limit. Note that this limit value is less than the maximum possible value of the probabilistic Rand index under all possible $\{S, S_1, S_2\}$

Consider a different S consisting of the image split vertically into two regions, the left region occupying $1/4$ th the image size and the other occupying the remaining $3/4$ th. It can be shown that in this case the probabilistic measure takes the limit $\frac{3}{8}$ as $N \to \infty$.

It may seem unusual that the PR index takes a maximum value of 1 only under stringent cases. However we claim that it is a more conservative measure as it is nonsensical for an algorithm to be given the maximum score possible when computed on an intrinsically ambiguous image. Conversely, if the PR index is aggregated over several sets $\{S_{\{1\ldots K\}}\}$, one set for each image, the choice of one algorithm over another should be less influenced by an image that human segmenters find ambiguous.

3.2. Region-sensitive refinement accommodation

Consider an image X consisting of N pixels. Let two manually labeled segmentations S_2, S_3 (Fig. 3) be made available to us. As seen in Fig. 3, the two human segmenters are in “agreement” on region $R1$, but region $R2$ in S_2 is split into two equal halves $R3$ and $R4$.

Following the counting procedure in Sec. 3.1 it can be shown that:

$$\mathrm{PR}(S, S_{\{2,3\}}) \to \frac{15}{16}$$

in upper bound as $N \to \infty$, and is obtained for *both* $S = S_2$ and $S = S_3$. However if a candidate S contained region $R1$ fragmented into (say) two regions of size $\frac{\alpha N}{2}$ and $\frac{(1-\alpha)N}{2}$

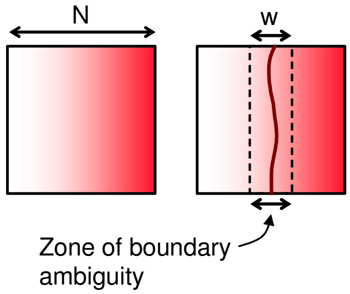

Figure 4: A toy example of the PR index adapting to pixel-level labeling errors near segment boundaries. The region in the image between the two vertical dashed lines indicates the zone of ambiguity. See text for details.

for $\alpha \in [0,1]$, it is straightforward to show that the PR index decreases in proportion to $\alpha(1-\alpha)$ as desired.

3.3. Accommodating boundary ambiguity

It widely agreed that human segmenters differ in the level of detail at which they perceive images, but are consistent [5] if refinements of segmentations are tolerated. Nevertheless, for a given number of class labels it is arguably the case that many images are ambiguous at the level of pixel label assignments near the cluster boundaries. One desirable property of a good comparison measure is accounting for these near-boundary ambiguities, even though the "true" boundary pixels are unknown.

Consider the example of the segmentation shown in Fig. 4 where all the human segmenters agree on splitting the image into two regions (red and white) but differ on the actual location of the boundary. To simplify the analysis, assume that the image is one-dimensional with N pixels. Let the (point) location of the boundary as determined by the set of manual segmentations $\{S'\}$ be given by a uniform distribution in a region of width w pixels. Let the candidate segmentation consist of a split at location x. By counting compatible pairs, it can be shown that the PR index as a function of boundary location x takes the form:

$$\mathrm{PR}\left(S(x),\{S'\}\right) = \begin{cases} A_1 x^2 + C_1 & \text{if } x \in [1, \frac{N-w}{2}] \\ \frac{\left[A_2 x^2 + B_2 x + C_2\right]}{3N(2N-1)(2w+1)} & \text{if } x \in [\frac{N \pm w}{2}] \\ A_1 (N-x)^2 + C_1 & \text{if } x \in [\frac{N+w}{2}, N] \end{cases}$$

where A_i,B_2 and C_i are functions of N and w. Figs. 5 and 6 plot the PR index computed numerically for varying values of N and w. It can be seen that the function is symmetric and concave in the region of boundary ambiguity, and convex elsewhere.

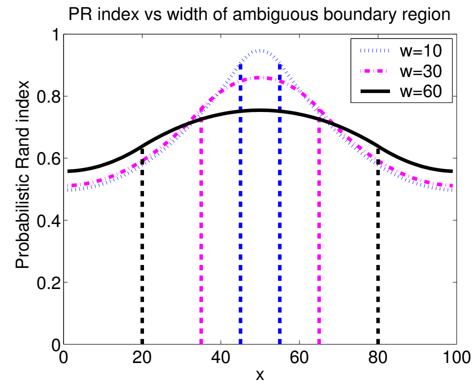

Figure 5: Plot of PR index computed for fixed image size ($N = 100$) and varying w. Function is continuous, concave in zone of ambiguity and convex elsewhere

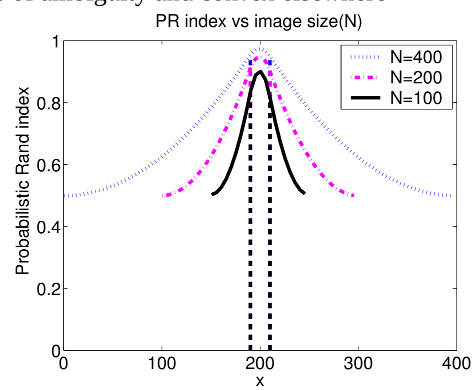

Figure 6: Plot of PR index computed for fixed $w = 20$ and varying image size N. Function profile is maintained while the maximum attainable PR index increases with N

Figure 1 shows an example of an image from the Berkeley database. The manually segmented images illustrate how the human segmenters perceive each zebra in isolation and segment the background at varying resolution. However, poor choice of the space in which clustering is done makes the mean-shift method *oversegment* the image and delineate the individual stripes of the zebra. However it is still a refinement of any one of the manual segmentations. Hence the computed LCE distance value is of the order of 0.063 (or a similarity of $1 - 0.063 = 0.937$) and correspondingly GCE distance is 0.064 (similarity 0.936). However the PR index gives a much lower similarity of 0.577 over all the manual segmentations of the zebra image.

A similar behavior is seen for the images of Figure 2, where all the human segmenters agree on labeling the pyramid and ground differently, but the test image is *undersegmented*. The LCE distance (similarity) is of the order of 0.018 (0.981) and that of GCE is 0.156 (0.844). The PR index gives a more reasonable lower similarity of 0.751 .

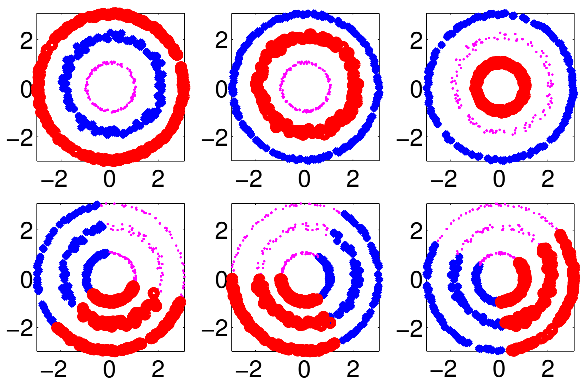

Figure 7: Plot of sample label assignments on subsampled 3-circle dataset using single linkage clustering (top row) and k-means (bottom row).

3.4. Estimating partition consistency

One way of assessing the stability of a clustering algorithm is to compare resulting segmentations obtained with perturbed versions of the input data. A high degree of similarity is indicative of a stable partition and potential of a good model fit. One easy way to perturb the data while maintaining the input data distribution is to sample with replacement and then propagate the labels back to the original unsampled dataset using a nearest-neighbor criterion.

Let the partitions $\{S_1, S_2, \ldots, S_K\}$ correspond to propagated segmentation results from K resampled versions of a dataset X. Consider the leave-one-out estimator of the form:

$$LPR(S_{\{1\ldots K\}}) = \frac{1}{K}\sum_{i=1}^{K} \mathrm{PR}(S_i, S_{\{1,\ldots i-1, i+1,\ldots K\}}) \tag{11}$$

where $\mathrm{PR}(\cdot,\cdot)$ is given in Eqn. 10. By definition, this estimate will be high for sets of partitions that are in complete agreement, and will take values in $[0, 1]$ under the same conditions as the PR index.

Figure 7 shows example results from a 3-circle dataset using k-means and single linkage clustering. Naturally, k-means can be expected to give poor segmentations that are inconsistent with each other, and one would hope that a measure of consistency reflects this. Observe however that due to the symmetry of the dataset, any run of k-means will assign the same label to points lying in a thin sector of the largest circle. This makes all the k-means segmentations, albeit different at the scale of the dataset, similar to a limited extent at a finer scale. The PR index returns a score of 0.985 over 50 segmentation runs for the single-linkage clustering, and 0.504 for the k-means results. The score is not 1 for the single-linkage due to occasional imprecision in propagation of labels from the subsampled to original dataset. For the k-means case, the extent of limited similarity is reflected in the non-zero but significantly lower score.

4. Conclusion

This paper reviewed measures of similarity popular in the statistics and computer vision literature and discussed their shortcomings as performance metrics in computer vision applications. We proposed a new similarity function that accommodates the inherent ambiguity in image segmentation and has additional desirable properties like region sensitivity and compensation for labeling error near class boundaries. We also described its utility for assessing partition consistency and providing sensible aggregate scores over several images for rating different algorithms.

5. Acknowledgments

The authors are grateful to the maintainers of the Berkeley Segmentation Dataset and Benchmark for public availability of the dataset, and to the authors of the EDISON code base at Rutgers. This work was supported in part by the DARPA MARS2020 program under Grant NBCH1020014.

References

[1] T. Lange, M. L. Braun, V. Roth, J. M. Buhmann, "Stability-based model selection", *NIPS*, no. 15, 2003.

[2] M. Halkidi, Y. Batistakis, M. Vazirgiannis., "Cluster Validity Methods: Part I", *ACM SIGMOD Record* June 2002

[3] M. Law, A. P. Topchy, A. K. Jain, "Multiobjective Data Clustering", *CVPR* June 2004.

[4] D. Martin, "An Empirical Approach to Grouping and Segmentation", *Ph.D. dissertation*, 2002, University of California, Berkeley

[5] D. Martin, C. Fowlkes, D. Tal, J. Malik, "A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics", *ICCV*, July 2001.

[6] K. Y. Yeung, W. L. Ruzzo, "Principal Component Analysis for clustering gene expression data", *Bioinformatics*, 17 (9)

[7] L. Hubert, P. Arabie, "Comparing partitions", *Journal of Classification*, 1985, pp. 193–218

[8] W. M. Rand, "Objective criteria for the evaluation of clustering methods", *Journal of the American Statistical Association*, 1971, 66 (336), pp. 846–850.

[9] J. Cohen, "A coefficient of agreement for nominal scales", *Educational and Psychological Measurement*, 1960, pp. 37–46

[10] N. Bendermacher, P. Souren, "Beyond Kappa: Estimating inter-rater agreement with nominal classifications", *Online tech report*

[11] M. Banerjee, M. Capozzoli, L. McSweeney, D. Sinha, "Beyond Kappa: A Review of Interrater Agreement Measures" *Canadian Journal of Statistics*, 1999 pp.2–23