A Measure for Objective Evaluation of Image Segmentation Algorithms

R. Unnikrishnan C. Pantofaru M. Hebert
The Robotics Institute
Carnegie Mellon University
Pittsburgh, PA, 15213

Abstract

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. Little effort has been spent on the design of perceptually correct measures to compare an automatic segmentation of an image to a set of hand-segmented examples of the same image. This paper demonstrates how a modification of the Rand index, the Normalized Probabilistic Rand (NPR) index, meets the requirements of large-scale performance evaluation of image segmentation. We show that the measure has a clear probabilistic interpretation as the maximum likelihood estimator of an underlying Gibbs model, can be correctly normalized to account for the inherent similarity in a set of ground truth images, and can be computed efficiently for large datasets. Results are presented on images from the publicly available Berkeley Segmentation dataset.

1. Introduction

Segmentation is a frequent pre-processing step in many image understanding algorithms and practical vision systems. In an effort to compare the performance of current segmentation algorithms to human perceptual grouping as well as understand the cognitive processes that govern grouping of visual elements in images, much work has gone into amassing hand-labeled segmentations of natural images [10].

Quantifying the performance of a segmentation algorithm, however, remains a challenging task. This is largely due to image segmentation being an ill-defined problem – there is no *single* ground truth segmentation against which the output of an algorithm may be compared. Rather the comparison is to be made against the set of all possible perceptually consistent interpretations of the image, of which only a minuscule fraction is usually available. This paper proposes a measure that makes this comparison by quantifying the agreement of an output segmentation with the inherent variation in a set of available manual segmentations.

It is certainly unreasonable to expect a single measure to be valid for every problem instance. For example, figure-ground segmentation for target tracking may value the proximity of the estimated segment to the true target location more than the accuracy of the actual shape of the detected boundary. Measures of similarity that quantify the extent to which two segmentations agree may also depend on the type and cardinality of the labels. For example, supervised segmentations into semantic categories (eg. 'sky', 'road', 'grass', etc.) must be treated differently from unsupervised clustering of pixels into groups with unordered and permutable labels [2]. This work assumes the labels to be non-semantic and permutable, and makes no assumptions about the underlying assignment procedure.

Consider the task where one must choose from among a set of segmentation algorithms based on their performance on a database of natural images. The algorithms are to be evaluated by objective comparison of their segmentation results with manual segmentations, several of which are available for each image. In the context of this task, a reasonable set of requirements for a measure of segmentation correctness are:

I **Non-degeneracy**: It does not have degenerate cases where unrealistic input instances give abnormally high values of similarity.

II **No assumptions about data generation**: It does not assume equal cardinality of the labels or region sizes in the segmentations.

III **Adaptive accommodation of refinement**: We use the term *label refinement* to denote differences in the pixel-level granularity of label assignments in the segmentation of a given image. Of particular interest are the differences in granularity that are correlated with differences in the level of detail at which the image is perceived. While human segmentations of an image differ with interpretation, perceptual grouping is arguably consistent over several large regions. Intuitively, this demands that a perceptually meaningful measure of similarity accommodate label refinement *only* in regions that humans find ambiguous and penalize differences in refinement elsewhere.

IV **Comparable scores**: The measure gives scores that permit meaningful comparison between segmentations

Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)
1063-6919/05 $20.00 © 2005 **IEEE**

of different images and between different segmentations of the same image.

In this paper we introduce a new measure for evaluating segmentations, the Normalized Probabilistic Rand (NPR) index, which is an extension to the Probabilistic Rand (PR) index introduced in [14]. We first show how the PR index meets the first, second, and third requirements listed above. However, the PR index as given in [14] cannot be directly applied to the task of evaluating segmentation algorithms. In order to permit meaningful comparison of scores between images and segmentations (the fourth requirement above), the index must be adjusted with respect to a baseline common to all of the images in the test set. Also, it is necessary to scale the index to reflect the amount of variance inherent in the test set. Hence we extend the PR index [14] to the Normalized Probabilistic Rand (NPR) index and show how it meets all four of the stated requirements for a useful measure.

2. Related work

In this section, we review measures that have been proposed in the literature to address variants of the segmentation evaluation task, while paying attention to the requirements described in the introduction.

We can broadly categorize previously proposed measures as follows:

1. **Region differencing** : Several measures operate by computing the degree of overlap between clusters or the cluster associated with each pixel in one segmentation and its "closest" approximation in the other segmentation. Some of them are deliberately intolerant of label refinement [12]. It is widely agreed, however, that humans differ in the level of detail at which they perceive images. To compensate for the difference in granularity while comparing segmentations, many measures allow label refinement uniformly through the image. D. Martin's thesis [9] proposed two measures – Global Consistency Error (GCE) and Local Consistency Error (LCE) that allowed labeling refinement in either or both directions, respectively.

Measures based on region differencing suffer from one or both of the following drawbacks:

(a) Degeneracy: As observed by the authors of [9, 10], there are two segmentations that give zero error for GCE and LCE – one pixel per segment, and one segment for the whole image. This adversely limits the use of the error functions to comparing segmentations that have similar cardinality of labels.

(b) Uniform penalty: Region-based measures that the authors are aware of in the literature compare one test segmentation to only one manually labeled image and penalize refinement uniformly over the image.

2. **Boundary matching**: Several measures work by matching boundaries between the segmentations, and computing some summary statistic of match quality [6, 7]. Work in [9] proposed solving an approximation to a bipartite graph matching problem for matching segmentation boundaries, computing the percentage of matched edge elements and using the harmonic mean of precision and recall as the statistic. 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.

3. **Information theory**: Work in [11] computes a measure of information content in each of the segmentations and how much information one segmentation gives about the other. The proposed metric measure is termed the *variation of information* (VI) and is related to the conditional entropies between the class label distribution of the segmentations. The measure has several promising properties but its potential for extension to evaluation on real images where there is more than one ground truth clustering is unclear.

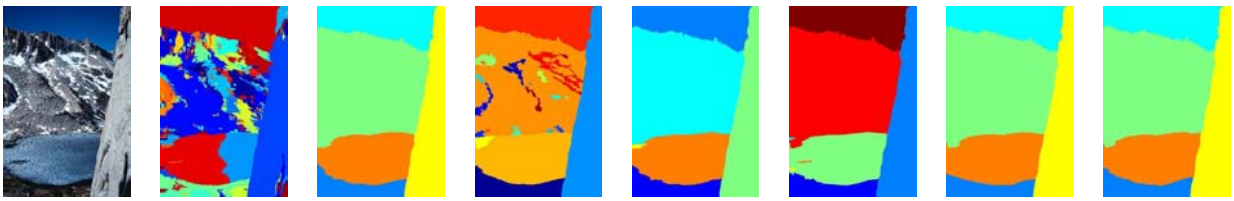
Several measures work by counting the number of false-positives and false-negatives [4] and similarly assume existence of only one ground truth segmentation. Due to the lack of spatial knowledge in the measure, the label assignments to pixels may be permuted in a combinatorial number of ways to maintain the same proportion of labels and keep the score unchanged.

4. **Non-parametric tests**: Popular non-parametric measures in statistics literature include Cohen's Kappa [2], Jaccard's index, Fowlkes and Mallow's index [5] among others. The latter two are variants of the Rand index [13] and work by counting pairs of pixels that have compatible label relationships between the two segmentations to be compared. More formally, consider two valid label assignments S and S' of N points $X = \{x_1, x_2, \ldots x_i, \ldots, x_N\}$ that assign labels $\{l_i\}$ and $\{l'_i\}$ respectively to point x_i . The Rand index R can be computed as the ratio of the number of pairs of points having a compatible label relationship in S and S'. i.e.

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

where $\mathbb{I}$is the identity function, and the denominator is the number of possible unique pairs among N data points. Note that the number of unique labels in S and S' are not restricted to be equal.

Nearly all the relevant measures known to the authors deal with the case of comparing two segmentations, one of which is treated as the singular ground truth. Hence they are not directly applicable for evaluating image segmentations in our framework. Section 3 outlines a modification



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 1: Example of oversegmentation: (a) Image from the Berkeley segmentation database [10],(b) its mean shift [3] segmentation (using hs=15 (spatial bandwidth), hr=10 (color bandwidth)), and (c-h) its ground truth hand segmentations. Average LCE = 0.0630, PR = 0.3731, NPR = -0.7349

(a) (b) (c) (d) (e) (f) (g) (h) (i)

Figure 2: Example of undersegmentation: (a) Image from the Berkeley segmentation database [10],(b) its mean shift [3] segmentation (using hs=15, hr=10), and (c-i) its ground truth hand segmentations. Average LCE = 0.0503, PR = 0.4420, NPR = -0.5932

to the basic Rand index that addresses this concern by soft non-uniform weighting of pixel pairs as a function of the variability in the ground truth set.

3. Normalized Probabilistic Rand (NPR) Index

In this section, we outline the Normalized Probabilistic Rand (NPR) index, an extension to the Probabilistic Rand (PR) index proposed in [14]. Section 3.1 describes the PR index and further discusses its desirable properties. Section 3.2 explains a simplification required for further analysis. Finally, Section 3.3 presents the NPR, describing its crucial improvements over the PR and other segmentation measures.

3.1. Probabilistic Rand Index

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_{test} be the segmentation that is to be compared with the manually labeled set. We denote the label of point x_i by $l_i^{S_{\text{test}}}$ in segmentation S_{test} and by $l_i^{S_k}$ in the manually segmented image S_k. It is assumed that each label $l_i^{S_k}$ can take values in a discrete set of size L_k, and correspondingly $l_i^{S_{\text{test}}}$ takes one of L_{test} values.

We chose to model label relationships for each pixel pair by an unknown underlying distribution. One may visualize this as a scenario where each human segmenter provides information about the segmentation S_k of the image in the form of binary numbers $\mathbb{I}(l_i^{S_k} = l_j^{S_k})$ for each pair of pixels (x_i, x_j). The set of all perceptually correct segmentations defines a Bernoulli distribution over this number, giving a random variable with expected value denoted p_{ij}. Hence the set $\{p_{ij}\}$ for all unordered pairs (i, j) defines a generative model of correct segmentations for the image X.

Consider the Probabilistic Rand (PR) index [14]:

$$\text{PR}(S_{\text{test}}, \{S_k\}) = \frac{1}{\binom{N}{2}} \sum_{\substack{i,j \\ i \prec j}} \Big[\mathbb{I}\Big(l_i^{S_{\text{test}}} = l_j^{S_{\text{test}}}\Big) p_{ij} + \mathbb{I}\Big(l_i^{S_{\text{test}}} \neq l_j^{S_{\text{test}}}\Big) (1 - p_{ij}) \Big] \tag{2}$$

Let c_{ij} denote the event of a pair of pixels i and j having the same label in the test image S_{test}:

$$c_{ij} = \mathbb{I}\Big(l_i^{S_{\text{test}}} = l_j^{S_{\text{test}}}\Big)$$

Then the PR index can be written as:

$$\text{PR}(S_{\text{test}}, \{S_k\}) = \frac{1}{\binom{N}{2}} \sum_{\substack{i,j \\ i \prec j}} [c_{ij} p_{ij} + (1 - c_{ij})(1 - p_{ij})] \tag{3}$$

This measure takes values in $[0, 1]$, where 0 means S_{test} 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.

Since $c_{ij} \in \{0, 1\}$, Eqn (3) can be equivalently written as

$$\text{PR}(S_{\text{test}}, \{S_k\}) = \frac{1}{\binom{N}{2}} \sum_{\substack{i,j \\ i \prec j}} \left[p_{ij}^{c_{ij}} (1 - p_{ij})^{1 - c_{ij}} \right] \quad (4)$$

Note that the quantity in square brackets in Eqn. (4) is the likelihood that labels of pixels x_i and x_j take values $l_i^{S_{\text{test}}}$ and $l_j^{S_{\text{test}}}$ respectively under the pairwise distribution defined by $\{p_{ij}\}$.

Recall that in our segmentation algorithm evaluation environment, a necessary feature of a good measure is a lack of degenerate cases. Figures 1 and 2 show (from left to right) images from the Berkeley segmentation database [10], segmentations of those images, and the ground truth hand segmentations of those images. The segmentation method we use is mean shift segmentation [3], which is a non-parametric kernel density-based segmentation method. Mean shift segmentation has two parameters that provide granularity control: h_s, the bandwidth of the kernel for the spatial features, and h_r, the bandwidth of the kernel for the other features (in our case, color). Now, notice that Fig. 1 is an oversegmentation and Fig. 2 is an undersegmentation. We compare the PR scores to the LCE scores [9, 10]. Note that the LCE is an error, with a score of 0 meaning no error and a score of 1 meaning maximum error. The LCE measure [9, 10] is tolerant to refinement regardless of the ground truth, and hence gives low error (high similarity) scores of 0.0630 and 0.0503, respectively. On the other hand, the PR is a measure of similarity, with a score of 0 meaning no similarity (maximum error) and a score of 1 meaning maximum similarity (no error). The PR does not allow refinement or coarsening that is not inspired by one of the human segmentations, hence the PR index gives low (low similarity, high error) scores of 0.3731 and 0.4420, respectively.

Tolerance to refinement is desired, however, as long as the refinement is inspired by one of the human segmentations. Consider the example in Fig. 3. The image in Fig. 3(a) is the original image, the two stacked images in Fig. 3(b) are two possible segmentations generated by an automatic segmentation algorithm, and the two images in Fig. 3(c) are the ground truths hand-labeled by people. Clearly, one of the hand-segmenters has chosen to segment according to texture, and the other according to color. The topmost automatic segmentation is finer than either of the two hand segmentations, however each of the edges can be found in one of the hand segmentations. Intuitively, it is still a useful segmentation because it only disagrees with the human segmentations in the same places that they are themselves ambiguous. The Probabilistic Rand index [14] would give the same score to either the top image in Fig. 3(b), or either of the hand segmentations. Hence this a permissible refinement. Now, look at the bottom automatic segmentation in Fig. 3(b). It is a further refinement, however the extra boundaries can not be found in either of the hand segmentations. Since it has divided clusters which the hand segmentations unambiguously stated should not be divided, its PR index is lower.

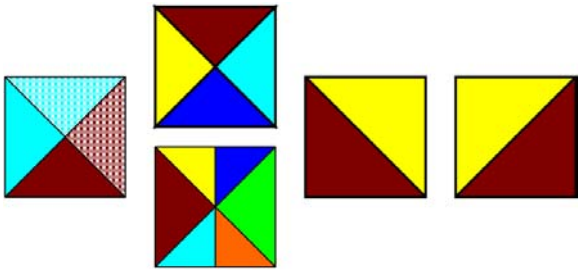


Figure 3: Synthetic example of permissible refinements: (a) Input image, (b) Segmentations for testing, and (c) ground truth set

At this point we have successfully addressed requirements I (non-degeneracy), II (no assumptions about data generation) and III (adaptive accommodation of refinement) for a useful measure, as stated in the introduction. Section 3.3 will expand on requirement II and address requirement IV (permitting score comparison between images and segmentations). Before we can extend the measure, however, we will need to show how to reduce the PR index to be computationally tractable.

3.2. Reduction using sample mean estimator

A straightforward choice of estimator for p_{ij}, the probability of the pixels i and j having the same label, is the sample mean of the corresponding Bernoulli distribution as given by

$$\bar{p}_{ij} = \frac{1}{K} \sum_k \mathbb{I}\left(l_i^{S_k} = l_j^{S_k} \right) \quad (5)$$

For this choice, it can be shown that the resulting PR index assumes a trivial reduction and can be estimated efficiently in time linear in N.

The PR index can be written as:

$$\text{PR}(S_{\text{test}}, \{S_k\}) = \frac{1}{\binom{N}{2}} \sum_{\substack{i,j \\ i \prec j}} \left[c_{ij} \bar{p}_{ij} + (1 - c_{ij})(1 - \bar{p}_{ij}) \right] \quad (6)$$

Substituting Eqn. (5) in Eqn. (6) and moving the summation over k outwards yields

$$\text{PR}(S_{\text{test}}, \{S_k\}) = \frac{1}{K} \sum_k \left[\frac{1}{\binom{N}{2}} \sum_{\substack{i,j \\ i \prec j}} \left[c_{ij} \mathbb{I}\left(l_i^{S_k} = l_j^{S_k} \right) + (1 - c_{ij}) \mathbb{I}\left(l_i^{S_k} \neq l_j^{S_k} \right) \right] \right] \quad (7)$$

which is simply the mean of the Rand index [13] computed between each pair (S_{test}, S_k). We can compute the terms within the square parentheses in $O(N + L_{\text{test}}L_k)$ in the following manner.

Construct a $L_{\text{test}} \times L_k$ contingency table with entries $n^{S_k}(l, l')$ containing the number of pixels that have label l in S_{test} and label l' in S_k. This can be done in $O(N)$ steps for each S_k.

The first term in Eqn. (7) is the number of pairs having the same label in S_{test} and S_k, and is given by

$$\sum_{\substack{i,j \\ i \prec j}} c_{ij} \mathbb{I}\left(l_i^{S_k} = l_j^{S_k}\right) = \sum_{l,l'} \binom{n^{S_k}(l,l')}{2} \tag{8}$$

which is simply the number of possible pairs of points chosen from sets of points belonging to the same class, and is computable in $O(L_{\text{test}}L_k)$ operations.

The second term in Eqn. (7) is the number of pairs having different labels in S_{test} and in S_k. To derive this, let us define two more terms for notational convenience. We denote the number of points having label l in the test segmentation S_{test} as:

$$n(l, \cdot) = \sum_{l'} n^{S_k}(l, l')$$

and similarly, the number of points having label l' in the second partition S_k as:

$$n(\cdot, l') = \sum_{l} n^{S_k}(l, l')$$

The number of pairs of points in the same class in S_{test} but different classes in S_k can be written as

$$\sum_{l} \binom{n(l,\cdot)}{2} - \sum_{l,l'} \binom{n^{S_k}(l,l')}{2}$$

Similarly, the number of pairs of points in the same class in S_k but different classes in S_{test} can be written as

$$\sum_{l'} \binom{n(\cdot,l')}{2} - \sum_{l,l'} \binom{n^{S_k}(l,l')}{2}$$

Since all the possible pixel pairs must sum to $\binom{N}{2}$, the number of pairs having different labels in S_{test} and S_k is given by

$$\binom{N}{2} + \sum_{l,l'} \binom{n^{S_k}(l,l')}{2} - \left(\sum_l \binom{n(l,\cdot)}{2}\right) - \left(\sum_{l'} \binom{n(\cdot,l')}{2}\right) \tag{9}$$

which is computable in $O(N + L_{\text{test}}L_k)$ time. Hence the overall computation for all K images is $O(KN + \sum_k L_k)$.

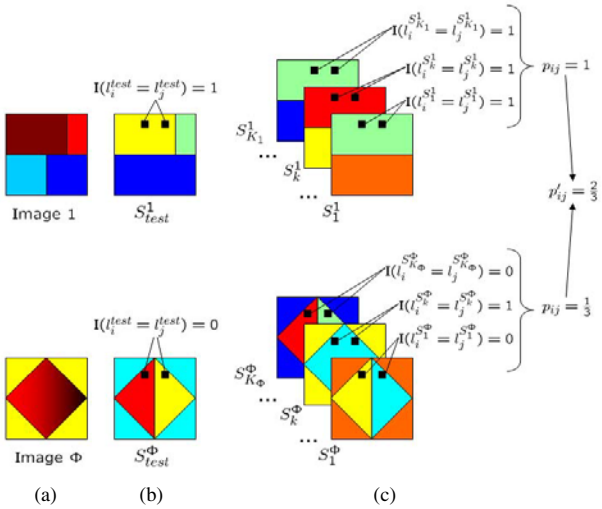


Figure 4: Illustration of the notation used in the Normalized Probabilistic Rand index. Each row ϕ has (a) an associated input Image ϕ, (b) a candidate segmentation S^{ϕ}_{test} and (c) a set of K_ϕ available manual segmentations $\{S^{\phi}_k\}$.

3.3. Normalization

The significance of a measure of similarity has much to do with the baseline with respect to which it is expressed. One may draw an analogy between the baseline and a null hypothesis in significance testing. For image segmentation, the baseline may be interpreted as the expected value of the index under some appropriate model of randomness in the input images. A popular strategy is to use the index normalized with respect to its baseline as

$$\text{Normalized index} = \frac{\text{Index} - \text{Expected index}}{\text{Maximum index} - \text{Expected index}} \tag{10}$$

so that the expected value of the normalized index is zero and it has a larger range and hence is more sensitive.

Hubert and Arabie [8] normalize the Rand index using a baseline that assumes the segmentations are generated from a hypergeometric distribution. This implies that a) the segmentations are independent, and b) the number of pixels having a particular label (the class label probabilities) is kept constant. The same model is adopted for the measure proposed in [5] with an unnecessary additional assumption of equal cardinality of labels. However, as also observed in [11, 15], the equivalent null model does not represent anything plausible in terms of realistic images, and both of the above assumptions are usually violated in practice. We would like to normalize the PR index in a way that avoids these pitfalls.

We will normalize the PR Index in Eqn. (2) using Eqn. (10), so we need to compute the expected value:

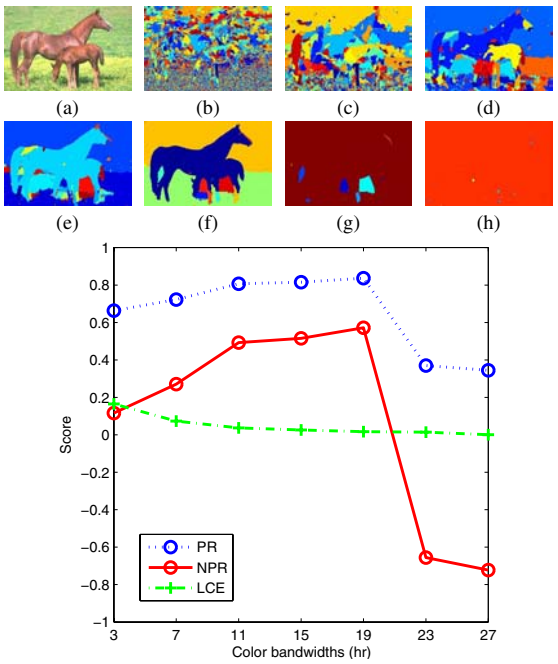


Figure 5: Example of changing scores for different segmentation granularities: (a) Original image, (b)-(h) mean shift segmentations [3] using scale bandwidth (hs) 7 and color bandwidths (hr) 3, 7, 11, 15, 19, 23 and 27 respectively. The plot shows the LCE error, the PR index score and the NPR score for each segmentation. Note that only the NPR index reflects the intuitive accuracy of each segmentation of the image. The NPR index correctly shows that segmentation (f) is the best one, segmentations (d), (e), and (f) are reasonable, and segmentations (g) and (h) are horrible.

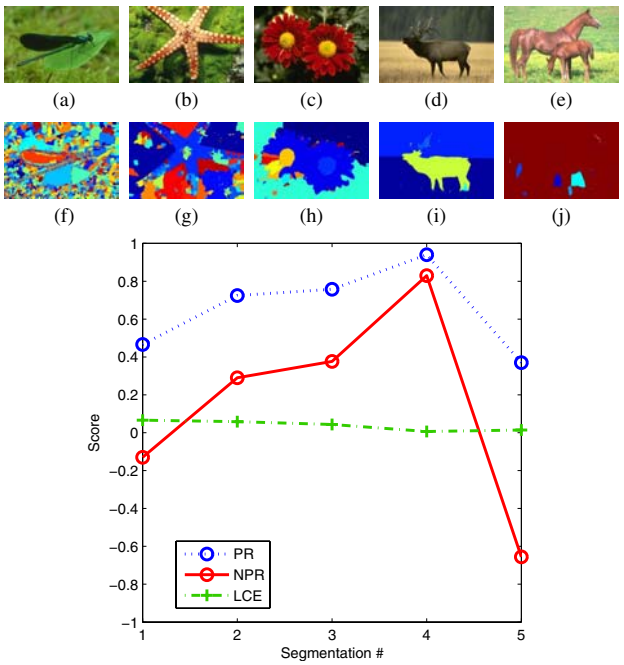


Figure 6: Example of comparing segmentations of different images: (a)-(e) Original images, (f)-(j) segmentations. The plot shows the LCE error, the PR index score and the NPR score for each segmentation. Note that only the NPR index reflects the intuitive accuracy of each segmentation across images

$$\begin{aligned}\mathbb{E}\Big[\mathrm{PR}(S_{\text{test}},\{S_k\})\Big] =& \frac{1}{\binom{N}{2}}\sum_{\substack{i,j\\ i\prec j}}\Big\{\mathbb{E}\Big[\mathbb{I}\Big(l_i^{S_{\text{test}}}=l_j^{S_{\text{test}}}\Big)\Big]p_{ij}\\ &+\mathbb{E}\Big[\mathbb{I}\Big(l_i^{S_{\text{test}}}\neq l_j^{S_{\text{test}}}\Big)\Big](1-p_{ij})\Big\}\\ =&\frac{1}{\binom{N}{2}}\sum_{\substack{i,j\\ i\prec j}}\Big[p'_{ij}p_{ij}+(1-p'_{ij})(1-p_{ij})\Big]\end{aligned}$$

The question is: what is a meaningful way to compute $p'_{i,j}=\mathbb{E}\big[\mathbb{I}\big(l_i^{S_{\text{test}}}=l_j^{S_{\text{test}}}\big)\big]$? We propose that for a baseline in image segmentation to be useful, it must be representative of perceptually consistent grouping of random but *realistic* images. Pair-wise probabilities provide a convenient way to model such segmentations of natural images. This translates to estimating p'_{ij} from segmentations of *all* images for all unordered pairs (i,j). Let Φ be the number of images in a dataset, and K_ϕ the number of ground truth hand segmentations of image ϕ. Then p'_{ij} can be expressed as:

$$p'_{ij}=\frac{1}{\Phi}\sum_{\phi}\frac{1}{K_\phi}\sum_{k=1}^{K_\phi}\mathbb{I}\Big(l_i^{S_k^\phi}=l_j^{S_k^\phi}\Big) \qquad (11)$$

Note that using this formulation for p'_{ij} implies that $\mathbb{E}[\mathrm{PR}(S_{\text{test}},\{S_k\})]$ is just a (weighted) sum of $\mathrm{PR}(S_k^\phi,\{S_k\})$. Although $\mathrm{PR}(S_k^\phi,\{S_k\})$ can be computed efficiently, performing this computation for every hand segmentation S_k^ϕ is expensive, so in practice we uniformly sample 5×10^6 pixel pairs for an image size of 321×481 ($N=1.5\times10^5$) instead of computing it exhaustively over all pixel pairs.

The philosophy that the baseline should depend on the empirical evidence from all of the images in a ground truth training set differs from the philosophy used to normalize the Rand Index [13]. In the Adjusted Rand Index [8], the expected value is computed over all theoretically possible segmentations with constant cluster proportions, regardless of how probable those segmentations are in reality. In comparison, the approach taken by the Normalized Probabilistic Rand index (NPR) has two important benefits:

First, since p'_{ij} and p_{ij} are modeled from the ground truth data, the number and size of the clusters in the images do not need to be held constant. Thus, the error produced by two segmentations with differing cluster sizes can be compared. In terms of evaluating a segmentation algorithm, this allows the comparison of the algorithm's performance with different parameters. Figure 5 demonstrates this behavior. The top two rows show an image from the segmentation

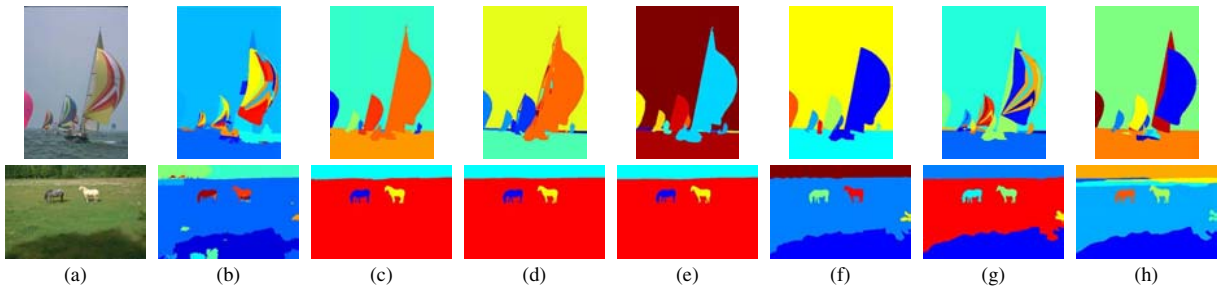


Figure 7: Examples of "good" segmentations: (a) Images from the Berkeley segmentation database [10], (b) mean shift segmentations [3] (using hs=15, hr=10), and (c-h) their ground truth hand segmentations. Top image: NPR = 0.8938, Bottom image: NPR = 0.8495

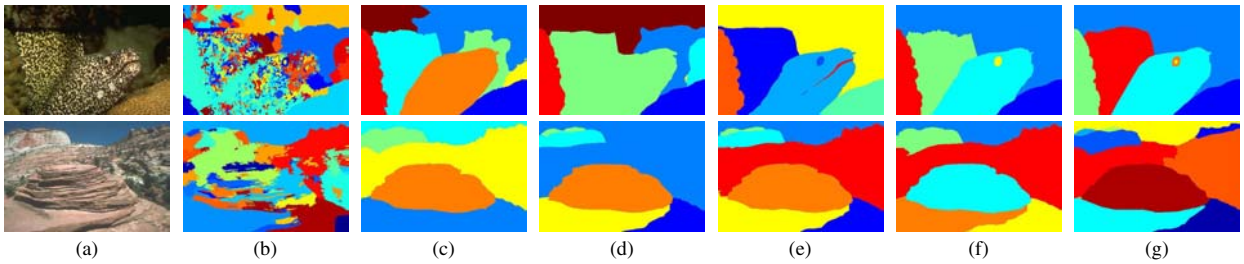


Figure 8: Examples of "bad" segmentations: (a) Images from the Berkeley segmentation database [10], (b) mean shift segmentations [3] (using hs=15, hr=10), and (c-g) their ground truth hand segmentations. Top image: NPR = -0.7333, Bottom image: NPR = -0.6207

database [10] and segmentations of different granularity. Note that the LCE error is low for all of the images since it is not sensitive to refinement, hence it cannot determine which segmentation is the most desirable. The PR index reflects the correct relationship among the segmentations, however its range is small and the expected value is unknown, hence it is difficult to judge what a "good" segmentation is. The NPR index fixes these problems. It reflects the desired relationships among the segmentations with no degenerate cases, and any segmentation which gives a score significantly above 0 is known to be useful.

Second, since p'_{ij} is modeled using all of the ground truth data, not just the data for the particular image in question, it is possible to compare the segmentation errors for different images to their respective ground truths. This facilitates the comparison of an algorithm's performance on different images. Figure 6 shows the scores of segmentations of different images. The first row contains the original images and the second row contains the segmentations. Once again, note that the NPR is the only index which both shows the desired relationship among the segmentations and whose output is easily interpreted.

The images in Fig. 7 and Fig. 8 demonstrate the consistency of the NPR. In Fig. 7(b), both mean shift [3] segmentations are perceptually equally "good" (given the ground truth segmentations), and correspondingly their NPR indices are high and similar. The segmentations in Fig. 8(b) are both perceptually "bad" (oversegmented), and correspondingly both of their NPR indices are very low. Note that the NPR indices of the segmentations in Fig. 2(b) and Fig. 8(b) are comparable, although the former is an undersegmentation and the latter are oversegmentations.

The normalization step has addressed requirement IV, facilitating meaningful comparison of scores between different images and segmentations. Note also that the NPR still does not make assumptions about data generation (requirement II). Hence we have met all of the requirements set out at the beginning of the paper.

3.4. Interpretation as a random field

Consider the labels of image X modeled as a Gibbs distribution with the equivalent random field defined on a *complete* graph with a node for each pixel x_i. The joint likelihood of a segmentation assigning label l_i to each pixel x_i may then be expressed as:

$$P(\{l_i\}) = \frac{1}{Z} \exp\left(\sum_{c \in \mathcal{C}} I_c(\{l_c\})\right) \qquad (12)$$

where $\mathcal{C}$ is the set of cliques in the graph, $-I_c(\{l_c\})$ is the interaction potential as a function of labels at pixels $x_i \in c$ only, and Z is the (constant) partition function.

We assume only pairwise potentials to be non-zero, employing a common restriction placed on model complexity for tractability on k-connected meshes. Taking the loga-

rithm of Eqn. (12) then gives

$$\log P(\{l_i\}) \propto \left(\sum_{\substack{i,j \\ i \prec j}} I_{ij}(l_i, l_j) \right) \tag{13}$$

where $-I_{ij}(l_i, l_j)$ is now a pairwise potential on pair (i, j).

Comparing the RHS of Eqn. (13) to that of the PR index

$$\mathrm{PR}(S_{\text{test}}, \{S_k\}) = \frac{1}{\binom{N}{2}} \sum_{\substack{i,j \\ i \prec j}} \left[p_{ij}^{c_{ij}} (1 - p_{ij})^{1 - c_{ij}} \right] \tag{14}$$

reveals the interaction potential $I_{ij}(l_i, l_j)$ to be proportional to the likelihood of pixels i and j having labels l_i and l_j given the parameters p_{ij} from the manual segmentations.

4. Extensions

There are several natural extensions that can be made to the NPR index to take advantage of side-information or priors:

1. **Weighted data points**: If there are specific regions of interest in the image being tested, it is straightforward to weight the contribution of points non-uniformly and maintain exact computation, assuming the use of a sample mean estimator for p_{ij}.

For example, let the points $X = \{x_1, x_2, \ldots, x_N\}$ be assigned weights $W = \{w_1, w_2, \ldots, w_N\}$ such that $0 < w_i < 1$ for all i and $\sum_i w_i = N$. Then the contingency table in Sec. 3.2 may be modified by replacing unit counts of pixels in the table by their weights. The remainder of the computation proceeds as before in $O(KN + \sum_k L_k)$ complexity.

2. **Soft segmentation**: In a situation where one cannot commit to a hard segmentation, each pixel x_i may be associated with a probability $p_i^{S_k}(l)$ of having label l in the k-th segmentation, such that $\sum_l p_i^{S_k}(l) = 1$. The contingency table can be modified in a similar manner as for weighted data points by spreading the contribution of a point across a row and column of the table. For example, the contribution of point x_i to the entry $n(l, l')$ for segmentation pairs S_{test} and S_k is $p_i^{S_{\text{test}}}(l) p_i^{S_k}(l')$.

3. **Priors from ecological statistics**: Experiments in [10] showed that the probability of two pixels belonging to the same perceptual group in natural imagery seems to follow an exponential distribution as a function of distance between the pixels. In presenting the use of the sample mean estimator for p_{ij}, this work assumed the existence of large enough number of hand-segmented images to sufficiently represent the set of valid segmentations of the image. If this is not feasible, a MAP estimator of the probability parametrized in terms of distance between pixels would be a sensible choice. What influence the choice of prior would have on the measure, particularly with regard to accommodation of label refinement, is the subject of future work.

5. Summary and Conclusions

This paper presented the Normalized Probabilistic Rand (NPR) index, a new measure for performance evaluation of image segmentation algorithms. It exhibits several desirable properties not exhibited together in previous measures. Numbers generated by the NPR index for a variety of natural images correspond to human intuition of perceptual grouping. Also, its flexibility gives it potential applications in related problems where extra domain knowledge is available. Future work includes application to large-scale performance evaluation as well as investigation of its utility as an objective function for training segmentation algorithms.

References

[1] P. Bamford, "Automating Cell Segmentation Evaluation with Annotated Examples", *In Workshop on Digital Image Computing (WDIC)*, 2003, Brisbane, Australia.

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

[3] D. Comaniciu, P. Meer, "Mean shift: A robust approach toward feature space analysis", *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2002, 24, pp. 603-619

[4] M. R. Everingham, H. Muller, and B. Thomas, "Evaluating image segmentation algorithms using the Pareto front", *ECCV*, May 2002, pp. IV:34-48.

[5] E. B. Fowlkes and C. L. Mallows, "A Method for Comparing Two Hierarchical Clusterings", *Journal of the American Statistical Association*, 78 (383), pp. 553–569, 1983.

[6] J. Freixenet, X. Munoz, D. Raba, J. Marti, and X. Cuff, "Yet Another Survey on Image Segmentation" , *ECCV* 2002, pp. 408–422.

[7] Q. Huang, B. Dom, "Quantitative methods of evaluating image segmentation", *IEEE Intl. Conf. on Image Processing*, 1995, pp. 53–56.

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

[9] D. Martin, "An Empirical Approach to Grouping and Segmentation", *Ph.D. dissertation*, 2002, U. C. Berkeley.

[10] 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.

[11] M. Meila, "Comparing clusterings by the variation of information", *Conference on Learning Theory*, 2003.

[12] H. I. Christensen, P. J. Phillips (editors), "Empirical evaluation methods in computer vision", World Scientific Publishing Company, July 2002.

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

[14] R. Unnikrishnan, M. Hebert, "Measures of Similarity", *IEEE Workshop on Applications of Computer Vision*, 2005, pp. 394–400.

[15] D. L. Wallace, "A Method for Comparing Two Hierarchical Clusterings: Comment", *Journal of the American Statistical Association*, 78 (383), pp. 569–576, 1983.