Toward Objective Evaluation of Image Segmentation Algorithms

Ranjith Unnikrishnan, *Student Member*, *IEEE*, Caroline Pantofaru, *Student Member*, *IEEE*, and Martial Hebert, *Member*, *IEEE*

Abstract—Unsupervised image segmentation is an important component in many image understanding algorithms and practical vision systems. However, evaluation of segmentation algorithms thus far has been largely subjective, leaving a system designer to judge the effectiveness of a technique based only on intuition and results in the form of a few example segmented images. This is largely due to image segmentation being an ill-defined problem—there is no unique ground-truth segmentation of an image against which the output of an algorithm may be compared. This paper demonstrates how a recently proposed measure of similarity, the Normalized Probabilistic Rand (NPR) index, can be used to perform a quantitative comparison between image segmentation algorithms using a hand-labeled set of ground-truth segmentations. We show that the measure allows principled comparisons between segmentations created by different algorithms, as well as segmentations on different images. We outline a procedure for algorithm evaluation through an example evaluation of some familiar algorithms—the mean-shift-based algorithm, an efficient graph-based segmentation algorithm, a hybrid algorithm that combines the strengths of both methods, and expectation maximization. Results are presented on the 300 images in the publicly available Berkeley Segmentation Data Set.

Index Terms—Computer vision, image segmentation, performance evaluation of algorithms.

1 Introduction

Image segmentation is the problem of partitioning an image into its constituent components. In wisely choosing a partition that highlights the role and salient properties of each component, we obtain a compact representation of an image in terms of its useful parts. Depending on the end application, the problem of segmentation can be subjective or objective. For example, the problem of processing an MRI image to separate pixels lying on the ventricle from everything else has a unique solution and is well-defined. This paper focuses on the more general problem of dividing an image into salient regions or "distinguished things" [1], a task which is far more subjective. Since there are as many valid solutions as interpretations of the image, it is an ill-defined problem.

The ill-defined nature of the segmentation problem makes the evaluation of a candidate algorithm difficult. It is tempting to treat segmentation as part of a proposed solution to a larger vision problem (e.g., tracking, recognition, image reconstruction, etc.), and evaluate the segmentation algorithm based on the performance of the larger system. However, this strategy for comparison can quickly become unfair and, more seriously, inconsistent when evaluating algorithms that are tailored to different applications. Furthermore, there are several properties intrinsic to an algorithm that are independent of an end-application. One example of a particularly important such property is an algorithm's *stability* with respect to input image data as well as across its operational parameters. Such properties need to be measured separately to be meaningful.

In the search for an independent ground-truth required by any reliable measure of performance, an attractive strategy is to associate the segmentation problem with perceptual grouping. Much work has gone into amassing hand-labeled segmentations of natural images [1] to compare the results of current segmentation algorithms to human perceptual grouping, as well as understand the cognitive processes that govern grouping of visual elements in images. Yet, there are still multiple acceptable solutions corresponding to the many human interpretations of an image. Hence, in the absence of a *unique* ground-truth segmentation, the comparison must be made against the set of all possible perceptually consistent interpretations of an image, of which only a minuscule fraction is usually available. In this paper, we propose to perform this comparison using a measure that quantifies the agreement of an automatic segmentation with the variation in a set of available manual segmentations.

We 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 output of each algorithm is a label assigned to each pixel of the images. We assume the labels to be nonsemantic and permutable, and make no assumptions about the underlying assignment procedure. The algorithms are to be evaluated by objective comparison of their segmentation results with several manual segmentations.

We caution the reader that our choice of human-provided segmentations to form a ground-truth set is not to be confused with an attempt to model human perceptual grouping. Rather the focus is to correctly account for the variation in a set of acceptable solutions, when measuring their agreement with a candidate result, *regardless* of the cause of the variability. In the described scenario, the variability happens to be generally caused by differences in the attention and level

- *The authors are with the Robotics Institute, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213. E-mail: {ranjith, crp, hebert}@cs.cmu.edu.*

Manuscript received 1 Nov. 2005; revised 3 May 2006; accepted 15 Aug. 2006; published online 18 Jan. 2007.
Recommended for acceptance by S.-C. Zhu.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0592-1105.
Digital Object Identifier no. 10.1109/TPAMI.2007.1046.

of detail at which an image is perceived. Hence, future references to "human subjects" are to be interpreted only as observed instances of this variability.

In the context of the above task, a reasonable set of requirements for a measure of segmentation correctness is:

1. **Nondegeneracy**: The measure does not have degenerate cases where input instances that are not well represented by the ground-truth segmentations give abnormally high values of similarity.
2. **No assumptions about data generation**: The measure does not assume equal cardinality of the labels or region sizes in the segmentations.
3. **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 in the human segmentations. A meaningful measure of similarity should accommodate label refinement *only* in regions that humans find ambiguous and penalize differences in refinement elsewhere.
4. **Comparable scores**: The measure gives scores that permit meaningful comparison between segmentations of different images and between different segmentations of the same image.

In Section 2, we review several previously proposed measures and discuss their merits and drawbacks as performance metrics in light of the above requirements. Section 3 outlines the Probabilistic Rand (PR) index [2], a generalization of a classical nonparametric test called the Rand index [3] and illustrates its properties. Section 4 then describes a scaled version of the measure, termed the Normalized Probabilistic Rand (NPR) index [4], that is adjusted with respect to a baseline common to all of the images in the test set—a step crucial for allowing comparison of segmentation results between images and algorithms. In contrast to previous work, this paper outlines the procedure for quantitative comparison through an extensive example evaluation in Section 5 of some popular unsupervised segmentation algorithms. The results in this paper use the Berkeley Segmentation Data Set [1] which consists of 300 natural images and multiple associated hand-labeled segmentations for each image.

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 Section 1.

We can broadly categorize previously proposed measures as follows:

1. **Region Differencing**: Several measures operate by computing the degree of overlap of 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 [5]. 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, many measures allow label refinement uniformly through the image.

   Martin et al. [1], [6] proposed several 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. The following describes two of the measures more formally.

   Let S and S' be two segmentations of an image $X = \{x_1, \ldots, x_N\}$ consisting of N pixels. For a given 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 [1], the local refinement error (LRE) is then defined at point x_i as:

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

   where $\setminus$ denotes the set differencing operator.

   This error measure is not symmetric and encodes a measure of refinement in one direction only. 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:

$$\mathrm{GCE}(S, S') = \frac{1}{N}\min\left\{\sum_i \mathrm{LRE}(S, S', x_i), \sum_i \mathrm{LRE}(S', S, x_i)\right\}.$$

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

$$\mathrm{LCE}(S, S') = \frac{1}{N}\sum_i \min\{\mathrm{LRE}(S, S', x_i), \mathrm{LRE}(S', S, x_i)\}.$$

   For both the LCE and GCE, a value of 0 indicates no error and a value of 1 indicates maximum deviation between the two segmentations being compared. As LCE $\leq$ GCE, it is clear that GCE is a tougher measure than LCE.

   To ease comparison with measures introduced later in the paper that quantify *similarity* between segmentations rather than error, we define the quantities $\mathrm{LCI} = 1 - \mathrm{LCE}$ and $\mathrm{GCI} = 1 - \mathrm{GCE}$. The "I" in the abbreviations stands for "Index," complying with the popular usage of the term in statistics when quantifying similarity. By implication, both LCI and GCI lie in the range $[0, 1]$ with a value of 0 indicating no similarity and a value of 1 indicating a perfect match.

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

   a. **Degeneracy**: As observed by the authors of [1], [6], 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.

      Work in [6] proposed an alternative measure termed the Bidirectional Consistency Error

(BCE) that replaced the pixelwise minimum operation in the LCE with a maximum. This results in a measure that penalizes dissimilarity between segmentations in proportion to the degree of overap and, hence, does not suffer from degeneracy. But, as also noted by the Martin [6], it does not tolerate refinement at all.

An extension of the BCE to the leave-one-out regime, termed BCE^*, attempted to compensate for this when using a set of manual segmentations. Consider a set of available ground-truth segmentations $\{S_1, S_2, \ldots, S_K\}$ of an image. The BCE^* measure matches the segment for each pixel in a test segmentation S_{test} to the minimally overlapping segment containing that pixel in any of the ground-truth segmentations.

$$\text{BCE}^*(S_{\text{test}}, \{S_k\}) = \frac{1}{N}\sum_{i=1}^{N} \min_k \Big\{ \max \big\{ \text{LRE}(S_{\text{test}}, S_k, x_i), \text{LRE}(S_k, S_{\text{test}}, x_i) \big\} \Big\}.$$

However, by using a hard "minimum" operation to compute the measure, the BCE^* ignores the frequency with which pixel labeling refinements in the test image are reflected in the manual segmentations. As before, to ease comparison of BCE^* with measures that quantify similarity, we will define and refer to the equivalent index $\text{BCI}^* = 1 - \text{BCE}^*$ taking values in [0, 1] with a value of 1 indicating a perfect match.

b. **Uniform penalty**: Region-based measures that the authors are aware of in the literature, with the exception of BCE^*, 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 [7], [8]. Work in [6] 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, termed the *F-measure* 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. Furthermore, for a given matching of edge elements between two images, it is possible to change the locations of the *unmatched* edges almost arbitrarily and retain the same precision and recall score.

3. **Information-based**: Work in [6], [9] proposes to formulate the problem as that of evaluating an affinity function that gives the probability of two pixels belonging to the same segment. They compute the mutual information score between the classifier output on a test image and the ground-truth data, and use the score as the measure of segmentation quality. Its application in [6], [9] is however restricted to considering pixel pairs only if they are in complete agreement in all the training images.

   Work in [10] computes a measure of information content in each of the segmentations and how much information one segmentation gives about the other. The proposed measure, termed the *variation of information* (VI), is a metric and is related to the conditional entropies between the class label distribution of the segmentations. The measure has several promising properties [11] but its potential for evaluating results on natural images where there is more than one ground-truth clustering is unclear.

   Several measures work by recasting the problem as the evaluation of a binary classifier [6], [12] through false-positive and false-negative rates or precision and recall, similarly assuming the existence of only one ground-truth segmentation. Due to the loss of spatial knowledge when computing such aggregates, 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. **Nonparametric tests**: Popular nonparametric measures in statistics literature include Cohen's Kappa [13], Jaccard's index, Fowlkes and Mallow's index, [14] among others. The latter two are variants of the Rand index [3] and work by counting pairs of pixels that have compatible label relationships in the two segmentations to be compared.

   More formally, consider two valid label assignments S and S' of N points $X = \{x_i\}$ with $i = 1 \ldots n$ 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 the same 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' is 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. In Section 3, we describe modifications to the basic Rand index that address these concerns.

3 Probabilistic Rand (PR) Index

We first outline a generalization to the Rand Index, termed the Probabilistic Rand (PR) index, which we previously introduced in [2]. The PR index allows comparison of a test segmentation with multiple ground-truth images through soft nonuniform weighting of pixel pairs as a function of the variability in the ground-truth set [2]. In Section 3.1, we will discuss its properties in more detail.

Consider a set of manual segmentations (ground-truth) $\{S_1, S_2, \ldots, S_K\}$ of an image $X = \{x_1, \ldots, x_N\}$ consisting 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 as p_{ij}. The set $\{p_{ij}\}$ for all unordered pairs (i, j) defines our generative model [4] of correct segmentations for the image X.

The Probabilistic Rand (PR) index [2] is then defined as:

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

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

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

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) and 1 means all segmentations are identical.

Since $c_{ij} \in \{0, 1\}$, (2) can be equivalently written as

$$\mathrm{PR}(S_{\text{test}}, \{S_k\}) = \frac{1}{\binom{N}{2}} \sum_{\substack{i,j \\ i<j}} \left[p_{ij}^{c_{ij}} (1 - p_{ij})^{1 - c_{ij}} \right]. \tag{3}$$

Note that the quantity in square brackets in (3) 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 $\{p_{ij}\}$.

Although the summation in (2) is over all possible pairs of N pixels, we show in the Appendix that the computational complexity of the PR index is $O(KN + \sum_k L_k)$, which is only linear in N, when p_{ij} is estimated with the sample mean estimator. For other choices of estimator (see Section 4.1), we have observed in practice that a simple Monte Carlo estimator using random samples of pixel pairs gives very accurate estimates.

3.1 Properties of the PR Index

We analyze the properties of the PR index in the subsections that follow.

3.1.1 Data Set Dependent Upper Bound

We illustrate the dependence of the upper bound of the PR index on the data set $S_{\{1 \ldots K\}}$ with a toy example. Consider an image X consisting of N pixels. Let two manually labeled segmentations S_1 and S_2 (as shown in Fig. 1) be made available to us. Let S_1 consist of the entire image having one label. Let S_2 consist of the image segmented into left and right 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 can be straightforwardly obtained by inspection as:

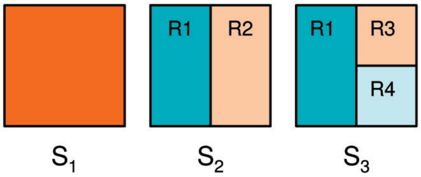


Fig. 1. A toy example of the PR index computed over a manually labeled set of segmentations. See text for details.

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

The above relation encodes that given no information other than the ground-truth set $\{S_1, S_2\}$, it is equally ambiguous as to whether the image is a single segment or two equally sized segments. It can be shown that this defines an upper bound on $\mathrm{PR}(S, S_{1,2})$ over all possible test segmentations S_{test}, and that this bound is attained[1] when the test segmentation S_{test} is identical to S_1 or S_2. The value of the bound is obtained by substituting the above values for p_{ij} into (2), and is given by:

$$\begin{aligned} \max_S \mathrm{PR}(S, S_{1,2}) &= \frac{1}{\binom{N}{2}} \Bigg[\underbrace{\frac{N}{2}\left(\frac{N}{2} - 1\right)}_{\text{pairs with same label in } S_1 \text{ and } S_2} \\ &\quad + \underbrace{\frac{N}{2} \times \frac{N}{2}}_{\text{pairs with different labels}} \times \underbrace{0.5}_{\text{empirical probability}} \Bigg] \\ &= \frac{1}{\binom{N}{2}} \left[\frac{3N^2}{8} - \frac{N}{2} \right]. \end{aligned}$$

Taking limits on the size of the image:

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

Note that this limit value is less than the maximum possible value of the PR index (equal to 1) under all possible test inputs S_{test} and ground-truth sets $\{S_k\}$.

Consider a different S_{test} (not shown) consisting of the image split into two regions, the left region occupying $\frac{1}{4}$ of the image size and the other occupying the remaining $\frac{3}{4}$. It can be shown that the modified measure takes the value:

$$\mathrm{PR}(S_{\text{test}}, S_{1,2}) = \frac{1}{\binom{N}{2}} \left[\frac{3N^2}{16} - \frac{N}{2} \right]$$

with limit $\frac{3}{8}$ as $N \to \infty$.

It may seem unusual that the Probabilistic Rand 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 inherently ambiguous image. Conversely, if the PR index is aggregated over several sets $\{S_{1 \ldots K}\}$, one for each image, the choice of one algorithm over another should be less influenced by an image that human segmenters find ambiguous.

[1] The proof proceeds by first showing that $d(S, S') = 1 - \mathrm{PR}(S, S')$ is a metric, and by then showing that if the PR score of a segmentation S exceeds $\mathrm{PR}(S_1, S_{1,2})$, it will violate the triangle inequality $d(S, S_1) + d(S, S_2) \geq d(S_1, S_2)$.

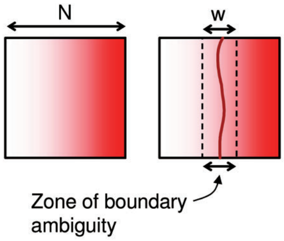


Fig. 2. A toy example of the PR index adapting to pixel-level labeling errors near segment boundaries. The region in the image between the two vertical dashed lines indicates the zone of ambiguity. See text for details.

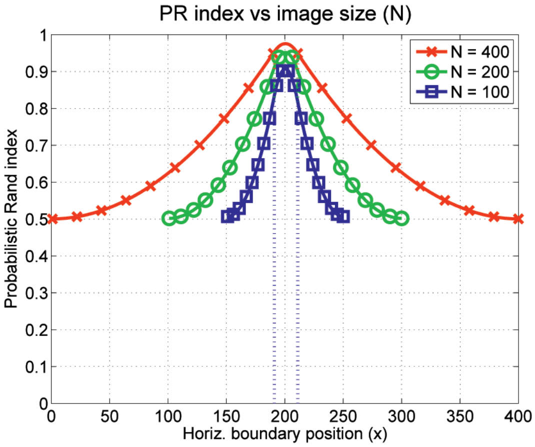


Fig. 3. Plot of PR index computed using (4) for the scenario of Fig. 2 with fixed $w = 20$ and varying image size N. Note that the function profile is maintained while the maximum attainable PR index increases with N.

3.1.2 Region-Sensitive Refinement Accommodation

Another desirable property of a meaningful measure is that it only penalizes fragmentation in regions that are unsupported by the ground-truth images, and allows refinement without penalty if it is consistently reflected in the ground-truth set. Consider now a set of two manually labeled segmentations consisting of S_2 and S_3 (Fig. 1). As seen in Fig. 1, 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 procedure in Section 3.1.1, it can be shown that $\mathrm{PR}(S, sS_{2,3}) \rightarrow \frac{15}{16}$ in upper bound as $N \rightarrow \infty$ 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 $\frac{(1-\alpha)N}{2}$ for $\alpha \in [0, 1]$, it is straightforward to show that the PR index decreases in proportion to $\alpha(1-\alpha)$ as desired.

3.1.3 Accommodating Boundary Ambiguity

It is widely agreed that human segmenters differ in the level of detail at which they perceive images. However, differences exist even among segmentations of an image having equal number of segments [1]. In many images, pixel label assignments are ambiguous near segment boundaries. Hence, one desirable property of a good comparison measure is robustness to small shifts in the location of the boundaries between segments, if those shifts are represented in the manually labeled training set, even when the "true" locations of those boundaries are unknown.

To illustrate this property in the PR index, we will construct an example scenario exhibiting this near-boundary ambiguity and observe the quantitative behavior of the PR index as a function of the variables of interest. Consider an example of the segmentation shown in Fig. 2, where all the human segmenters agree on splitting a $N \times N$ pixel image into two regions (red and white) but differ on the precise location of the boundary. For mathematical clarity, let us adopt a simplified model of the shape of the boundary separating the two segments. We assume the boundary to be a straight vertical line whose horizontal position in the set of available manual segmentations is uniformly distributed in a region of width w pixels.

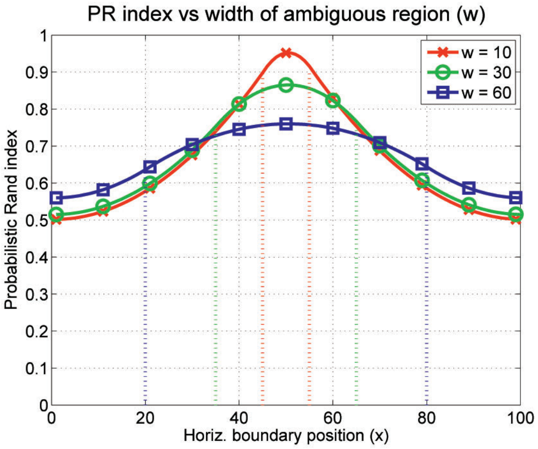


Fig. 4. Plot of PR index computed using (4) for the scenario of Fig. 2 with fixed image size $(N = 100)$ and varying w. Note that the function is everywhere continuous, concave in the zone of ambiguity, and convex elsewhere.

Let the candidate segmentation consist of a vertical split at distance x pixels from the left edge of the image. For a given boundary position x, we can analytically compute, for each pixel pair, the probability p_{ij} of their label relationship existing in the manually labeled images under the previously described boundary model. This essentially involves a slightly tedious counting procedure that we will not elaborate here to preserve clarity. The key result of this procedure for our example scenario in Fig. 2 is an analytical expression of the PR index as a function of x given by:

$$\mathrm{PR}(S(x), \{S'\}) = \begin{cases} A_1 x^2 + C_1 & \text{if } x \in [1, \frac{N-w}{2}] \\ -A_2 x^2 + B_2 x + C_2 & \text{if } x \in [\frac{N-w}{2}, \frac{N+w}{2}] \\ A_1 (N-x)^2 + C_1 & \text{if } x \in [\frac{N+w}{2}, N], \end{cases} \tag{4}$$

where the coefficients A_i, B_2, and $C_i (i = 1, 2)$ are positive valued functions of N and w.

Figs. 3 and 4 plot the expression in (4) for varying values of N and w, respectively. It can be seen that the function is symmetric and concave in the region of boundary ambiguity, and convex elsewhere. Thus, the PR index for the example of Fig. 2 essentially has the profile of a piecewise quadratic inverted M-estimator, making it robust to small local changes in the boundary locations when they are reflected in the manual segmentation set.

Figs. 5 and 6 show (from left to right) images from the Berkeley segmentation database [1], segmentations of those images, and the ground-truth hand segmentations of those

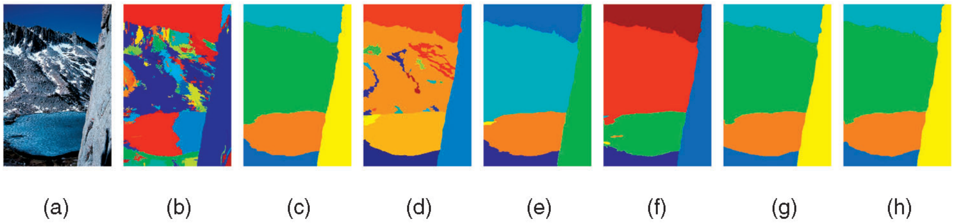


Fig. 5. Example of oversegmentation: (a) Image from the Berkeley segmentation database [1], (b) its mean shift [15] segmentation (using $h_s = 15$ (spatial bandwidth), $h_r = 10$ (color bandwidth)), and (c), (d), (e), (f), (g), and (h) its ground-truth hand segmentations. Average $\mathrm{LCI} = 0.9370$, $\mathrm{BCI}^* = 0.7461$, $\mathrm{PR} = 0.3731$, and $\mathrm{NPR} = -0.7349$.

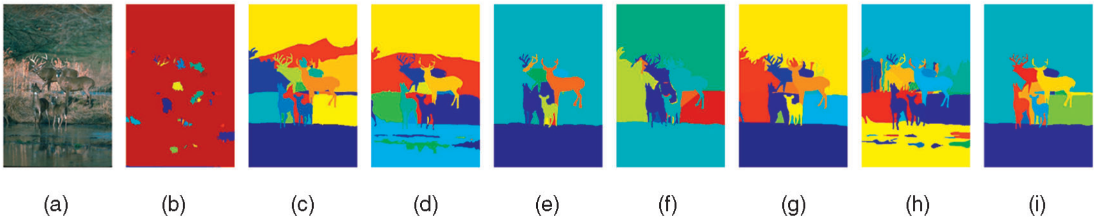


Fig. 6. Example of undersegmentation: (a) Image from the Berkeley segmentation database 1], (b) its mean shift [15] segmentation (using $h_s = 15$, $h_r = 10$), and (c), (d), (e), (f), (g), (h), and (i) its ground-truth hand segmentations. Average $\mathrm{LCI} = 0.9497$, $\mathrm{BCI}^* = 0.7233$, $\mathrm{PR} = 0.4420$, and $\mathrm{NPR} = -0.5932$.

images. The segmentation method we use is mean shift segmentation [15], described briefly in Section 5.1.1. Notice that Fig. 5 is an oversegmentation and Fig. 6 is an undersegmentation. We compare the PR scores to the LCI scores [6] described in Section 2. The LCI measure is tolerant to refinement regardless of the ground truth and, hence, gives high similarity scores of 0.9370 and 0.9497, respectively. On the other hand, the PR does not allow refinement or coarsening that is not inspired by one of the human segmentations. This is correctly reflected in the low PR index (low similarity) scores of 0.3731 and 0.4420, respectively.

At this point, we have successfully addressed Requirements 1 (nondegeneracy), 2 (no assumptions about data generation), and 3 (adaptive accommodation of refinement) for a useful measure, as stated in Section 1.

We have observed in practice, however, that the PR index suffers from lack of variation in its value over images. This is likely due to the smaller effective range of the PR index combined with the variation in maximum value of the PR index across images. Furthermore, it is unclear how to interpret the value of the index across images or algorithms and what a low or high number is. To remedy this, Section 4 will present the Normalized Probabilistic Rand (NPR) index [4], and describe its crucial improvements over the PR and other segmentation measures. It will expand on Requirement 2 and address Requirement 4 (permitting score comparison between images and segmentations).

4 Normalized Probabilistic Rand (NPR) Index

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 [14], [16] is to normalize the index with respect to its baseline as

$$\text{Normalized index} = \frac{\text{Index} - \text{Expected index}}{\text{Maximum index} - \text{Expected index}}. \quad (5)$$

This causes the expected value of the normalized index to be zero and the modified index to have a larger range and hence be more sensitive. There is little agreement in the statistics community [17] regarding whether the value of "Maximum Index" should be estimated from the data or set constant. We choose to set the value to be 1, the maximum possible value of the PR index and avoid the practical difficulty of estimating this quantity for complex data sets.

Hubert and Arabie [16] normalize the Rand index using a baseline that assumes that the segmentations are generated from a hypergeometric distribution. This implies that 1) the segmentations are independent and 2) the number of pixels having a particular label (i.e., the class label probabilities) is kept constant. The same model is adopted for the measure proposed in [14] with an additional, although unnecessary, assumption of equal cardinality of labels. However, as also observed in [10], [17], 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.

To normalize the PR index in (2) as per (5), we need to compute the expected value of the index:

$$\begin{aligned}\mathbb{E}\Big[\mathrm{PR}(S_{\text{test}},\{S_k\})\Big] &= \frac{1}{\binom{N}{2}}\sum_{\substack{i,j\\ i<j}}\Big\{\mathbb{E}\Big[\mathbb{I}\big(l_i^{S_{\text{test}}} = l_j^{S_{\text{test}}}\big)\Big]p_{ij}\\ &\quad + \mathbb{E}\Big[\mathbb{I}\big(l_i^{S_{\text{test}}} \neq l_j^{S_{\text{test}}}\big)\Big](1-p_{ij})\Big\}\\ &= \frac{1}{\binom{N}{2}}\sum_{\substack{i,j\\ i<j}}\Big[p'_{ij}p_{ij} + (1-p'_{ij})(1-p_{ij})\Big].\end{aligned} \quad (6)$$

The question now is: What is a meaningful way to compute $p'_{i,j} = \mathbb{E}[\mathbb{I}(l_i^{S_{\text{test}}} = l_j^{S_{\text{test}}})]$? We propose that for a baseline in image segmentation to be useful, it must be

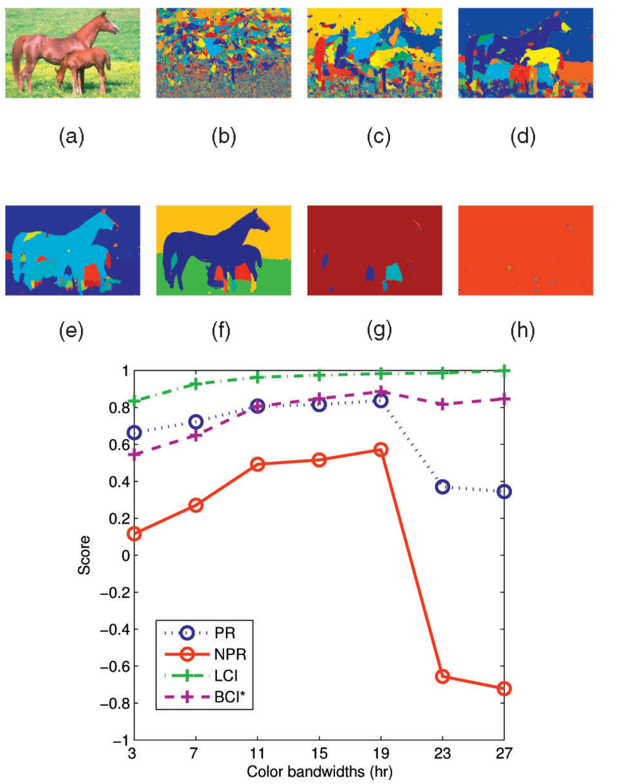


Fig. 7. Example of changing scores for different segmentation granularities: (a) Original image, (b), (c), (d), (e), (f), (g), and (h) mean shift segmentations [15] using scale bandwidth (h_s) 7 and color bandwidths (h_r) 3, 7, 11, 15, 19, 23, and 27, respectively. The plot shows the LCI, BCI^*, PR, and the NPR similarity scores 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.

representative of perceptually consistent groupings of random but *realistic* 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 data set and K_ϕ the number of ground-truth segmentations of image ϕ. Then, p'_{ij} can be expressed as:

$$p'_{ij} = \frac{1}{\Phi}\sum_{\phi}\frac{1}{K_\phi}\sum_{k=1}^{K_\phi}\mathbb{I}\left(l_i^{S_k^\phi} = l_j^{S_k^\phi}\right). \tag{7}$$

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 segmentation S_k^ϕ is expensive, so, in practice, we uniformly sample 5×10^6 pixel pairs for an image size of $321\times481(N=1.5\times10^5)$ instead of computing it exhaustively over all pixel pairs. Experiments performed using a subset of the images indicated that the loss in precision in comparison with exhaustive evaluation was not significant for the above number of samples.

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 [3]. In the Adjusted Rand Index [16], 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.

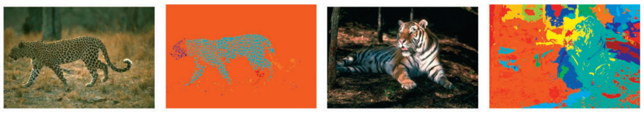


Fig. 8. Examples of segmentations with NPR indices near 0.

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. Fig. 7 demonstrates this behavior. The top two rows show an image from the segmentation database [1] and segmentations of different granularity. Note that the LCI similarity is high for all of the images since it is not sensitive to refinement; hence, it cannot determine which segmentation is the most desirable. The BCI^* measure sensibly reports lower scores for the oversegmented images, but is unable to appreciably penalize the similarity score for the undersegmented images in comparison with the more favorable segmentations. 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 make a judgment as to 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. As intuition, Fig. 8 shows two segmentations with NPR indices close to zero.

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. Fig. 9 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. 10 and Fig. 11 demonstrate the consistency of the NPR. In Fig. 10b, both mean shift [15] segmentations are perceptually equally "good" (given the ground-truth segmentations), and correspondingly their NPR indices are high and similar. The segmentations in Fig. 11b 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. 6b and Fig. 11b are comparable, although the former is an undersegmentation and the latter are oversegmentations.

The normalization step has addressed Requirement 4, facilitating meaningful comparison of scores between different images and segmentations. Note also that the NPR still does not make assumptions about data generation (Requirement 2). Hence, we have met all of the requirements set out at the beginning of the paper.

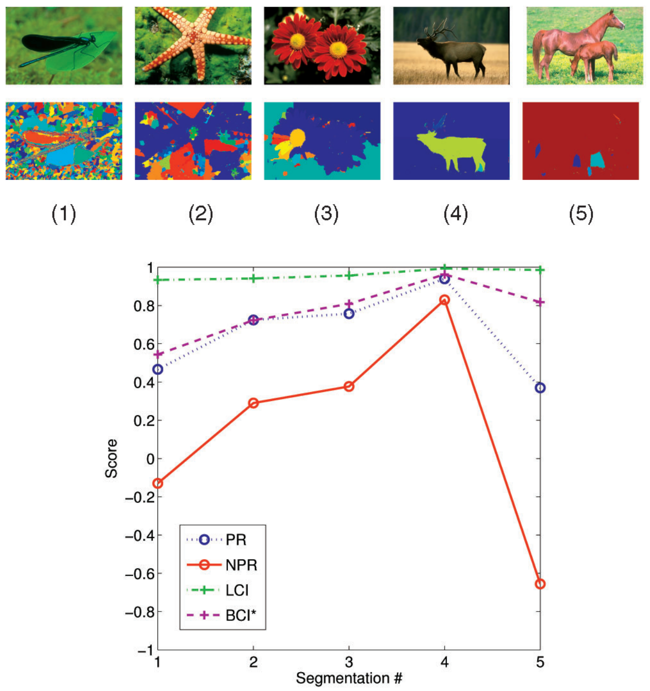


Fig. 9. Example of comparing segmentations of different images: (1), (2), (3), (4), and (5) Top row: Original images, Second row: corresponding segmentations. The plot shows the LCI, BCI*, PR, and the NPR similarity scores for each segmentation as numbered. Note that only the NPR index reflects the intuitive accuracy of each segmentation across images.

In moving from the first-order problem of comparing pixel labels to the second-order problem of comparing compatibilities of pairs of labels, the Rand index introduces a bias by penalizing the fragmentation of large segments more than that of small segments, in proportion to the segment size. To our knowledge, this bias has not deterred the broad adoption of the Rand index in its adjusted form by the statistics community. We have also not observed any practical impact of this in our extensive experimental comparison of algorithms in Section 5.

One way of explicitly tolerating the bias, if required, is to use a spatial prior so as to discount the contribution of pairs of distant pixels in unusually large segments. Another method is to simply give more weight to pixels in small regions that are considered salient for the chosen task. We describe these and other modifications in what follows.

4.1 Extensions

There are several natural extensions that can be made to the NPR index to take advantage of additional information or priors when they are available:

1. **Weighted data points**: Some applications may require the measure of algorithm performance to depend more on certain parts of the image than others. For example, one may wish to penalize unsupported fragmentation of specific regions of interest in the test image more heavily than of other regions. It is straightforward to weight the contribution of points nonuniformly and maintain exact computation when the sample mean estimator is used for p_{ij}.

   For example, let the image pixels $X = \{x_1, \dots, x_N\}$ be assigned weights $W = \{w_1, \dots, w_N\}$, respectively, such that $0 \leq w_i \leq 1$ for all i and $\sum_i w_i = N$. The Appendix describes a procedure for the unweighted case that first constructs a contingency table for the label assignments and then computes the NPR index exactly with linear complexity in N using the values

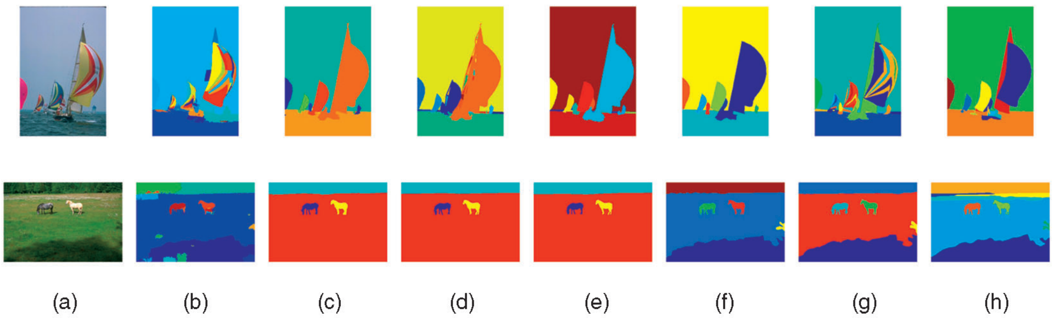


Fig. 10. Examples of "good" segmentations: (a) Images from the Berkeley segmentation database [1], (b) mean shift segmentations [15] (using $h_s = 15$, and $h_r = 10$), and (c), (d), (e), (f), (g), and (h) their ground-truth hand segmentations. Top image: $\mathrm{NPR} = 0.8938$ and bottom image: $\mathrm{NPR} = 0.8495$.

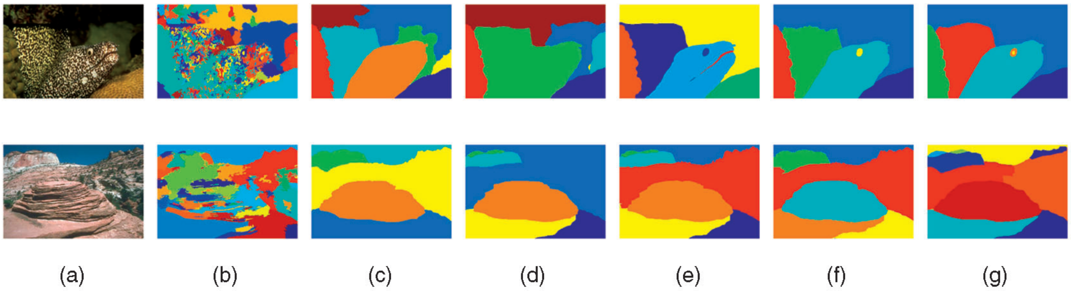


Fig. 11. Examples of "bad" segmentations: (a) Images from the Berkeley segmentation database [1], (b) mean shift segmentations [15] (using $h_s = 15$, and $h_r = 10$), and (c), (d), (e), (f), and (g) their ground-truth hand segmentations. Top image: $\mathrm{NPR} = -0.7333$ and bottom image: $\mathrm{NPR} = -0.6207$.

in the table. For the weighted case, the contingency table can be simply modified by replacing unit counts of pixels in the table by their weights. The remainder of the computation proceeds just as for the unmodified PR index in $O(KN + \sum_k L_k)$ total time, where L_k is the number of labels in the $k\mathrm{th}$ image.

2. **Soft segmentation**: In applications where one wishes to avoid committing 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 kth 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 [1] 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 a 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 parameterized in terms of distance between pixels would be a sensible choice.

5 Experiments

The purpose of creating the NPR index was to facilitate objective evaluations of segmentation algorithms, with the hope that the results of such evaluations can aid system designers in choosing an appropriate algorithm. As an exercise in using the NPR index, we present a possible evaluation framework and give one such comparison. We consider four segmentation techniques: mean shift segmentation [15], the efficient graph-based segmentation algorithm presented in [18], a hybrid variant that combines these algorithms, and expectation maximization [19] as a baseline. For each algorithm, we examine three characteristics which we believe are crucial for an image segmentation algorithm to possess:

1. **Correctness**: The ability to produce segmentations which agree with ground truth. That is, segmentations which correctly identify structures in the image at neither too fine nor too coarse a level of detail. This is measured by the value of the NPR index.
2. **Stability with respect to parameter choice**: The ability to produce segmentations of consistent correctness for a range of parameter choices.
3. **Stability with respect to image choice**: The ability to produce segmentations of consistent correctness using the same parameter choice on different images.

If a segmentation scheme satisfies these three requirements, then it will give useful and predictable results which can be reliably incorporated into a larger system without excessive parameter tuning. Note that every characteristic of the NPR index is required to perform such a comparison. It has been argued that the correctness of a segmentation algorithm is only relevant when measured in the context of the larger system into which it will be incorporated. However, there is value in weeding out algorithms which give nonsensical results, as well as limiting the list of possibilities to well-behaved algorithms even if the components of the rest of the system are unknown.

Our data set for this evaluation is the Berkeley Segmentation Data Set [1]. To ensure a valid comparison between algorithms, we compute the same features (pixel location and color) for every image and every segmentation algorithm. We begin this section by presenting each of the segmentation algorithms and the hybrid variant we considered, and then present our results.

5.1 The Segmentation Algorithms

As mentioned, we will compare four different segmentation techniques, the mean shift-based segmentation algorithm [15], an efficient graph-based segmentation algorithm [18], a hybrid of the previous two, and expectation maximization [19]. We have chosen to look at mean shift-based segmentation as it is generally effective and has become widely-used in the vision community. The efficient graph-based segmentation algorithm was chosen as an interesting comparison to the mean shift in that its general approach is similar, however, it excludes the mean shift filtering step itself, thus partially addressing the question of whether the filtering step is useful. The hybrid of the two algorithms is shown as an attempt at improved performance and stability. Finally, the EM algorithm is presented as a baseline. The following describes each algorithm.

5.1.1 Mean Shift Segmentation

The mean shift-based segmentation technique was introduced in [15] and is one of many techniques under the heading of "feature space analysis." The technique is comprised of two basic steps: a mean shift filtering of the original image data (in feature space), and a subsequent clustering of the filtered data points.

Filtering. The filtering step of the mean shift segmentation algorithm consists of analyzing the probability density function underlying the image data in feature space. In our case, the feature space consists of the (x, y) image location of each pixel, plus the pixel color in L*u*v* space (L^*, u^*, v^*). The modes of the pdf underlying the data in this space will correspond to the locations with highest data density, and data points close to these modes can be clustered together to form a segmentation. The mean shift filtering step consists of finding these modes through the iterative use of kernel density estimation of the gradient of the pdf and associating with them any points in their basin of attraction. Details may be found in [15].

We use a uniform kernel for gradient estimation with radius vector $h = [h_s, h_s, h_r, h_r, h_r]$, with h_s the radius of the spatial dimensions and h_r the radius of the color dimensions. For every data point (pixel in the original image), the gradient estimate is computed and the center of the kernel, $\mathbf{x}$, is moved in that direction, iterating until the gradient is below a threshold. This change in position is the mean shift vector. The resulting points have gradient approximately equal to zero and, hence, are the modes of the density estimate. Each datapoint is then replaced by its corresponding mode estimate.

Finding the mode associated with each data point helps to smooth the image while preserving discontinuities. Let

$S_{\mathbf{x}_j,h_s,h_r}$ be the sphere in feature space, centered at point $\mathbf{x}$ and with spatial radius h_s and color radius h_r. The uniform kernel has nonzero values only on this sphere. Intuitively, if two points $\mathbf{x}_i$ and $\mathbf{x}_j$ are far from each other in feature space, then $\mathbf{x}_i \notin S_{\mathbf{x}_j,h_s,h_r}$ and, hence, $\mathbf{x}_j$ does not contribute to the mean shift vector and the trajectory of $\mathbf{x}_i$ will move it away from $\mathbf{x}_j$. Hence, pixels on either side of a strong discontinuity will not attract each other. However, filtering alone does not provide a segmentation as the modes found are noisy. This "noise" stems from two sources. First, the mode estimation is an iterative process, hence it only converges to within the threshold provided (and with some numerical error). Second, consider an area in feature space larger than $S_{\mathbf{x},h_s,h_r}$ and where the color features are uniform or have a gradient of one in each dimension. Since the pixel coordinates are uniform by design, the mean shift vector will be a 0-vector in this region, and the data points in this region will not move and, hence, not converge to a single mode. Intuitively, however, we would like all of these data points to belong to the same cluster in the final segmentation. For these reasons, mean shift filtering is only a preprocessing step and a second step is required in the segmentation process: clustering of the filtered data points $\{\mathbf{x}'\}$.

Clustering. After mean shift filtering, each data point in the feature space has been replaced by its corresponding mode. As described above, some points may have collapsed to the same mode, but many have not despite the fact that they may be less than one kernel radius apart. In the original mean shift segmentation paper [15], clustering is described as a simple postprocessing step in which any modes that are less than one kernel radius apart are grouped together and their basins of attraction are merged. This suggests using single linkage clustering to convert the filtered points into a segmentation.

The only other paper using mean shift segmentation that speaks directly to the clustering is [20]. In this approach, a region adjacency graph (RAG) is created to hierarchically cluster the modes. Also, edge information from an edge detector is combined with the color information to better guide the clustering. This is the method used in the publicly available EDISON system, also described in [20]. The EDISON system is the implementation we use here as our mean shift segmentation system.

Discussion. Mean shift filtering using either single linkage clustering or edge-directed clustering produces segmentations that correspond well to human perception. However, as we discuss in the following sections, this algorithm is quite sensitive to its parameters. Slight variations in the color bandwidth h_r can cause large changes in the granularity of the segmentation, as shown in Fig. 7. By adjusting the color bandwidth, we can produce oversegmentations as in Fig. 7b, to reasonably intuitive segmentations as in Fig. 7f, to undersegmentations as in Fig. 7g. This instability is a major stumbling block with respect to using mean shift segmentation as a reliable preprocessing step for other algorithms, such as object recognition. In an attempt to improve stability and ease the burden of parameter tuning, we consider a second algorithm.

5.2 Efficient Graph-Based Segmentation

Efficient graph-based image segmentation, introduced in [18], is another method of performing clustering in feature space. This method works directly on the data points in feature space, without first performing a filtering step, and uses a variation on single linkage clustering. The key to the success of this method is adaptive thresholding. To perform traditional single linkage clustering, a minimum spanning tree of the data points is first generated (using Kruskal's algorithm), from which any edges with length greater than a given hard threshold are removed. The connected components become the clusters in the segmentation. The method in [18] eliminates the need for a hard threshold, instead replacing it with a data-dependent term.

More specifically, let $G = (V, E)$ be a (fully connected) graph, with m edges $\{e_i\}$ and n vertices. Each vertex is a pixel, $\mathbf{x}$, represented in the feature space. The final segmentation will be $S = (C_1, \dots, C_r)$, where C_i is a cluster of data points. The algorithm is:

1. Sort $E = (e_1, \dots, e_m)$ such that $|e_t| \le |e_{t'}| \forall t < t'$.
2. Let $S^0 = (\{\mathbf{x}_1\}, \dots, \{\mathbf{x}_n\})$, in other words each initial cluster contains exactly one vertex.
3. For $t = 1, \dots, m$
   a. Let $\mathbf{x}_i$ and $\mathbf{x}_j$ be the vertices connected by e_t.
   b. Let $C^{t-1}_{\mathbf{x}_i}$ be the connected component containing point $\mathbf{x}_i$ on iteration $t-1$ and $l_i = \max_{\text{mst}} C^{t-1}_{\mathbf{x}_i}$ be the longest edge in the minimum spanning tree of $C^{t-1}_{\mathbf{x}_i}$. Likewise for l_j.
   c. Merge $C^{t-1}_{\mathbf{x}_i}$ and $C^{t-1}_{\mathbf{x}_j}$ if
   $$|e_t| < \min\left\{ l_i + \frac{k}{\left|C^{t-1}_{\mathbf{x}_i}\right|}, l_j + \frac{k}{\left|C^{t-1}_{\mathbf{x}_j}\right|} \right\}, \qquad (8)$$
   where k is a constant.
4. $S = S^m$.

In contrast to single linkage clustering which uses a constant K to set the threshold on edge length for merging two components, efficient graph-based segmentation uses the variable threshold in (8). This threshold effectively allows two components to be merged if the minimum edge connecting them does not have length greater than the maximum edge in either of the components' minimum spanning trees, plus a term $\tau = \frac{k}{|C^{t-1}_{\mathbf{x}_i}|}$ As defined here, τ is dependent on a constant k and the size of the component. Note that on the first iteration, $l_i = 0$ and $l_j = 0$, and $\left|C^0_{\mathbf{x}_i}\right| = 1$ and $\left|C^0_{\mathbf{x}_j}\right| = 1$, so k represents the longest edge which will be added to any cluster at any time, $k = l_{max}$. As the number of points in a component increases, the tolerance on added edge length for new edges becomes tighter and fewer mergers are performed, thus indirectly controlling region size. However, it is possible to use any nonnegative function for τ which reflects the goals of the segmentation system. Intuitively, in the function used here, k controls the final cluster sizes.

The merging criterion in (8) allows efficient graph-based clustering to be sensitive to edges in areas of low variability, and less sensitive to them in areas of high variability. However, the results it gives do not have the same degree of correctness with respect to the ground truth as mean shift-based segmentation, as demonstrated in Fig. 12. This algorithm also suffers somewhat from sensitivity to its parameter, k.

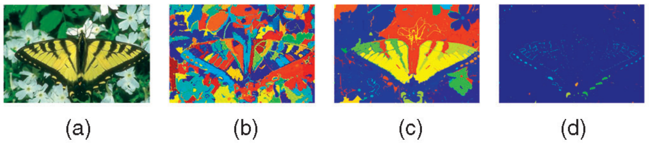


Fig. 12. Example of changing scores for different parameters using efficient graph-based segmentation: (a) Original image, (b), (c), and (d) efficient graph-based segmentations using spatial normalizing factor $h_s = 7$, color normalizing factor $h_r = 7$, and k values 5, 25, and 125, respectively.

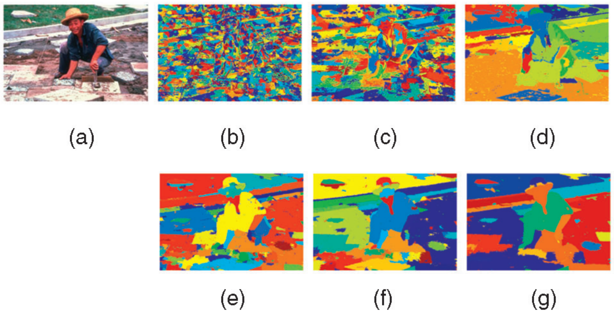


Fig. 13. Example of changing scores for different parameters using a hybrid segmentation algorithm which first performs mean shift filtering and then efficient graph-based segmentation: (a) Original image, (b), (c), (d), (e), (f), and (g) segmentations using spatial bandwidth $h_s = 7$, and color bandwidth (h_r) and k value combinations (3, 5), (3, 25), (3, 125), (15, 5), (15, 25), and (15, 125), respectively.

5.3 Hybrid Segmentation Algorithm

An obvious question emerges when describing the mean shift-based segmentation method [15] and the efficient graph-based clustering method [18]: Can we combine the two methods to give better results than either method alone? More specifically, can we combine the two methods to create more stable segmentations that are less sensitive to parameter changes and for which the same parameters give reasonable segmentations across multiple images? In an attempt to answer these questions, the third algorithm we consider is a combination of the previous two algorithms: First, we apply mean shift filtering and then we use efficient graph-based clustering to give the final segmentation. The result of applying this algorithm with different parameters can be seen in Fig. 13. Notice that for $h_r = 15$, the quality of the segmentation is high. Also, notice that the rate of granularity change is slower than either of the previous two algorithms, even though the parameters cover a wide range.

5.4 EM Segmentation Algorithm

Our final algorithm is the classic Expectation Maximization (EM) algorithm [19], with the Bayesian Information Criterion (BIC) used to select the number of Gaussians in the model. By minimizing the BIC, we attempt to minimize model complexity while maintaining low error. The BIC is formulated as follows:

$$\mathrm{BIC} = n \ln\left(\frac{RSS}{n}\right) + g \ln(n),$$

where n is the sample size, g is the number of parameters, and RSS is the residual sum of squares. We present graphical results for the EM algorithm as a baseline for each relevant experiment, however, we omit it in the detailed performance discussion.

Fig. 14. Examples of images from the Berkeley image segmentation database [1].

5.5 Experiments

Each of the issues raised in the introduction to this section: correctness, stability with respect to parameters, and stability of parameters with respect to different images, is explored in the following experiments and resulting plots. Note that the axes for each plot type are kept constant so plots can be easily compared. In each experiment, the label "EDISON" refers to the publicly available EDISON system for mean shift segmentation [20], the label "FH" refers to the efficient graph-based segmentation method by Felzenszwalb and Huttenlocher [18], the label "MS+FH" refers to our hybrid algorithm of mean shift filtering followed by efficient graph-based segmentation, and the label "EM" refers to the EM algorithm [19]. All of the experiments were performed on the publicly available Berkeley image segmentation database [1], which contains 300 images of natural scenes with approximately five to seven ground-truth hand segmentations of each image. Examples of the images are shown in Fig. 14.

In all of the following plots, we have fixed the spatial bandwidth $h_s = 7$ since it seems to be the least sensitive parameter and removing it makes the comparison more approachable. Also, although the FH algorithm as defined previously only had one parameter, k, we need to add two more. In order to properly compute distance in our feature space $\{x, y, L^*, u^*, v^*\}$, we rescale the data by dividing each dimension by the corresponding $\{h_s, h_r\}$. The same procedure is applied to the EM algorithm. So, each algorithm was run with a parameter combination from the sets: $h_s = 7$, $h_r = \{3, 7, 11, 15, 19, 23\}$, and $k = \{5, 25, 50, 75, 100, 125\}$. We mildly abuse notation by using h_r and h_s to denote parameters for both mean-shift and FH/EM algorithms to avoid introducing extra terms.

5.5.1 Maximum Performance

The first set of experiments examines the correctness of the segmentations produced by the three algorithms with a reasonable set of parameters. Fig. 15a shows the maximum NPR index on each image for each algorithm. The indices are plotted in increasing order for each algorithm, hence image 190 refers to the images with the 190th lowest index for each algorithm, and may not represent the same image across algorithms. Fig. 15b is a histogram of the same information, showing the number of images per maximum NPR index bin.

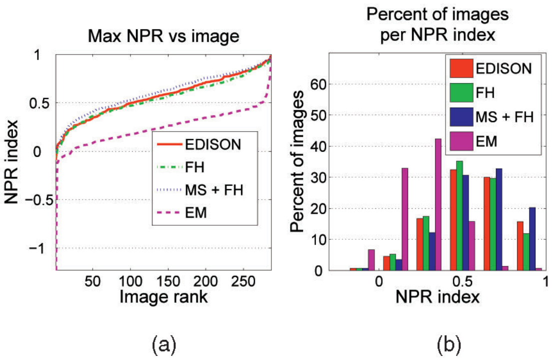


Fig. 15. Maximum NPR indices achieved on individual images with the set of parameters used for each algorithm. Plot (a) shows the indices achieved on each image individually, ordered by increasing index. Plot (b) shows the same information in the form of a histogram. Recall that the NPR index has an expected value of 0 and a maximum of 1.

All of the algorithms, except EM, produce similar maximum NPR indices, demonstrating that they have roughly equal ability to produce correct segmentations with the parameter set chosen. Note that there are very few images which have below-zero maximum NPR index, hence all of the algorithms almost always have the potential to produce useful results. These graphs also demonstrate that our parameter choices for each algorithm are reasonable.

5.5.2 Average Performance per Image

The next set of plots in Figs. 16, 17, and 18 examine correctness through the mean index achieved on each image. The first plot in each row shows the mean NPR index on each image achieved over the set of parameters used (in increasing order of the mean), along with one standard deviation. The second plot in each row is a histogram of the mean information, showing the number of images per mean NPR index bin. An algorithm which creates good segmentations will have a histogram skewed to the right. The third plot in each row is a histogram of the standard deviations.

These plots partially addresses the issue of stability with respect to parameters. A standard deviation histogram that is skewed to the left indicates that the algorithm in question is less sensitive to changes in its parameters. Using the means as a measure certainly makes us more dependent on our choice of parameters for each algorithm. Although we cannot guarantee that we have found the best or worst parameters for any individual algorithm, we can compare the performance of the algorithms with identical parameters.

Average performance over different values of the color bandwidth h_r. We compare the NPR indices averaged over values of h_r, with k held constant. The plots showing this data for the EDISON method are in Fig. 16. Fig. 17 gives the plots for the efficient graph-based segmentation system (FH) and the hybrid algorithm (MS + FH) for $k = \{5, 25, 125\}$. We only show three out of the six values of k in order to keep the amount of data presented reasonable. The most interesting comparison here is between the EDISON system and the hybrid system, which reflects the impact the addition of the efficient graph-based clustering has had on the segmentations produced.

Notice that for $k = 5$, the performance of the hybrid (MS + FH) system is slightly better and certainly more stable than that of the mean shift-based (EDISON) system. For $k = 25$, the performance is more comparable, but the standard deviation is still somewhat lower. Finally, for $k = 125$, the hybrid system performs comparably to the mean-shift based system. Thus, the change to using the efficient graph-based clustering after the mean shift filtering has maintained the correctness of the mean shift-based system while improving its stability.

Looking at the graphs for the efficient graph-based segmentation system alone in Fig. 17, we can see that although for $k = 5$ the mean performance and standard deviation are promising, they quickly degrade for larger values of k. This decline is much more gradual in the hybrid algorithm.

Average performance over different values of k. The mean NPR indices as k is varied through $k = \{5, 25, 50, 75, 100, 125\}$ and h_r is held constant are displayed in figure Fig. 18. Once again, we only look at a representative three out of the six possible h_r values, $h_r = \{3, 7, 23\}$. Since the mean shift-based system does not use k, this comparison is between the efficient graph-based segmentation system and the hybrid system.

The results show that the mean indices of the hybrid system are both higher and more stable (with respect to changing values of k) than those of the efficient graph-based segmentation system. Hence, adding a mean shift filtering preprocessing step to the efficient graph-based segmentation system is an improvement.

5.5.3 Average Performance per Parameter Choice

The final set of experiments look at the stability of a particular parameter combination across images. In each experiment, results are shown with respect to a particular parameter with averages and standard deviations taken over segmentations of each image in the entire image database.

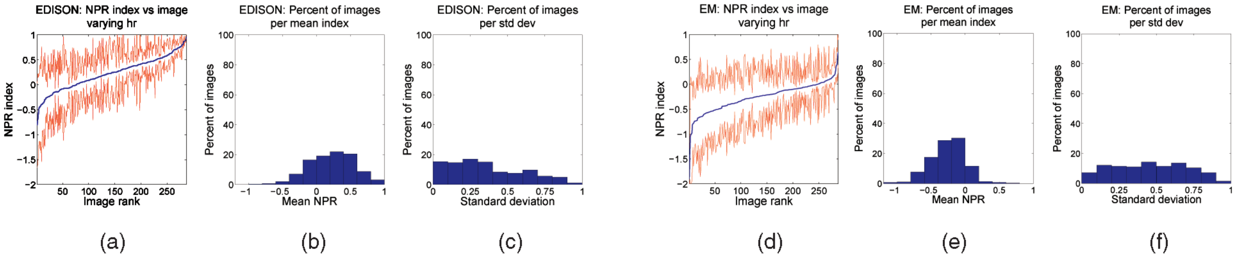


Fig. 16. Mean NPR indices achieved on individual images over the parameter set of all combinations of $h_r = \{3, 7, 11, 15, 19, 23\}$. Results for the mean shift-based system (EDISON) are given in plots (a), (b), and (c), and results for EM are given in (d), (e), and (f). Plots (a) and (d) show the mean indices achieved on each image individually, ordered by increasing index, along with one standard deviation. Plots (b) and (e) show histograms of the means. Plots (c) and (f) show histograms of the standard deviations.

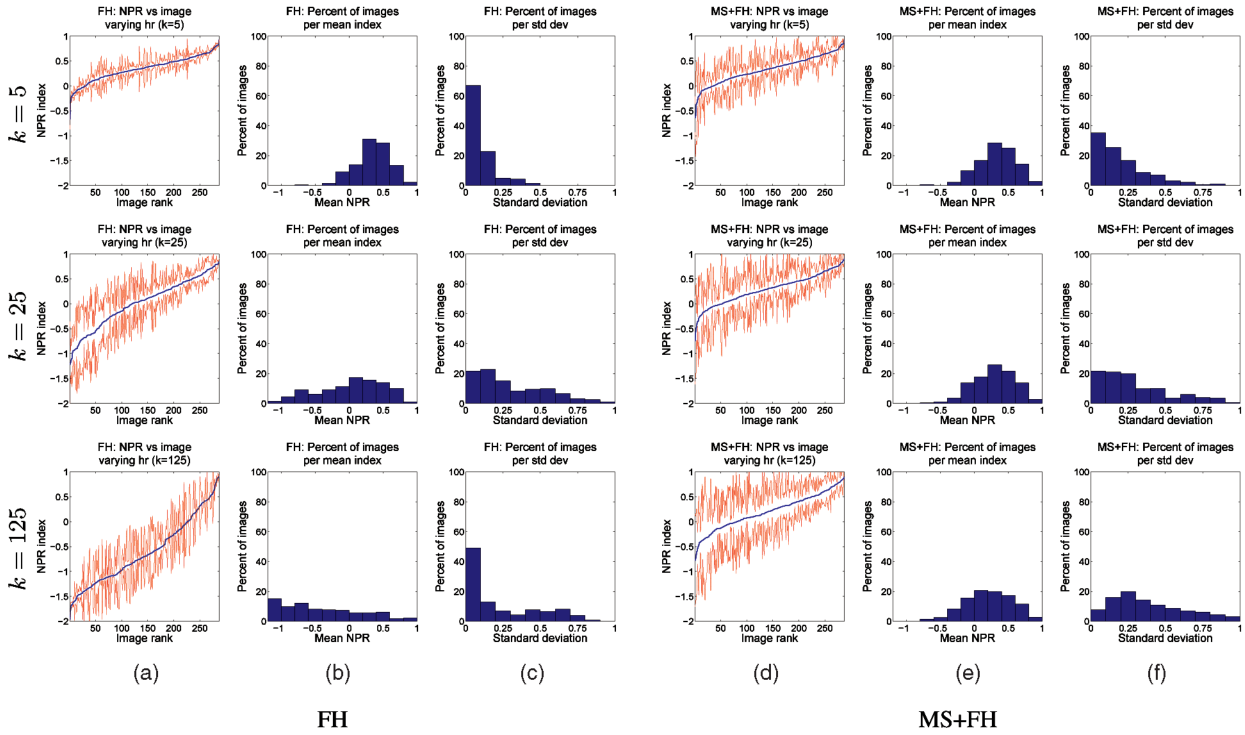


Fig. 17. Mean NPR indices achieved on individual images over the parameter set $h_r = \{3, 7, 11, 15, 19, 23\}$ with a constant k. Results for the efficient graph-based segmentation system (FH) are shown in columns (a), (b), and (c), and results for the hybrid segmentation system (MS+FH) are shown in columns (d), (e), and (f). Columns (a) and (d) show the mean indices achieved on each image individually, ordered by increasing index, along with one standard deviation. Columns (b) and (e) show histograms of the means. Columns (c) and (f) show histograms of the standard deviations.

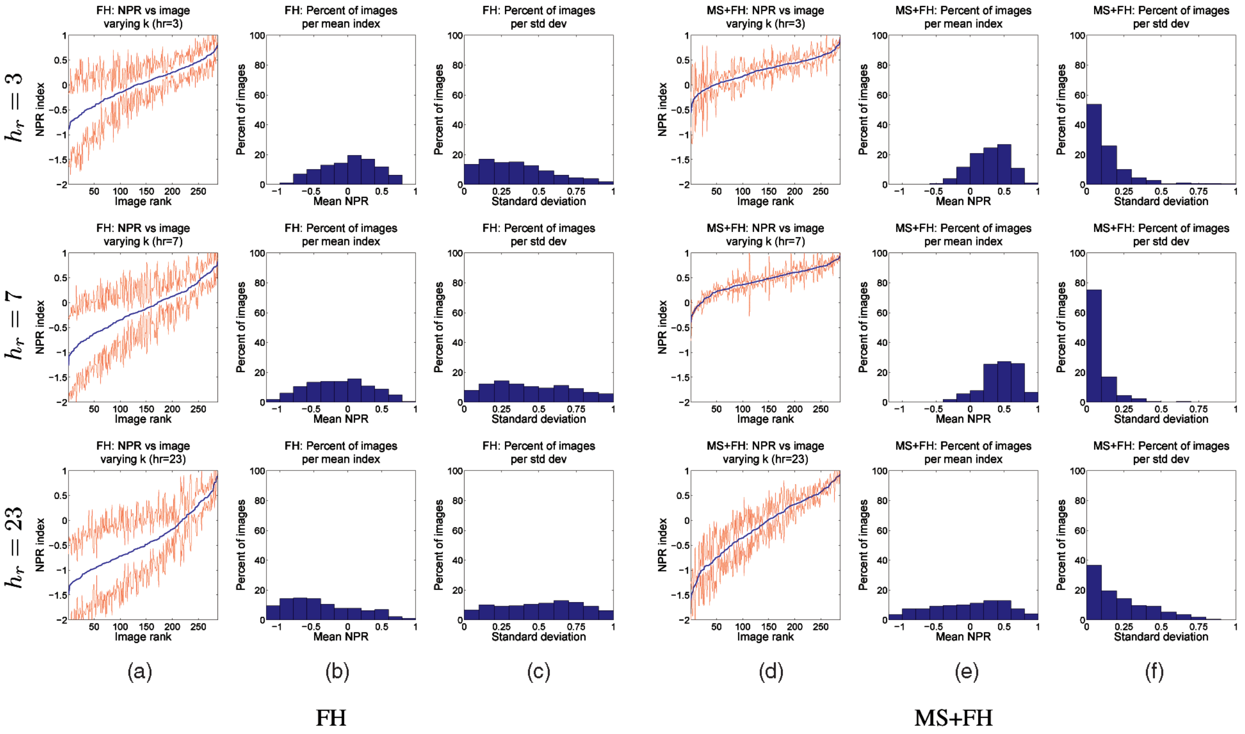


Fig. 18. Mean NPR indices achieved on individual images over the parameter set $k = \{5, 25, 50, 75, 100, 125\}$ with a constant h_r. Results for the efficient graph-based segmentation system (FH) are shown in columns (a), (b), and (c), and results for the hybrid segmentation system (MS+FH) are shown in columns (d), (e), and (f). Columns (a) and (d) show the mean indices achieved on each image individually, ordered by increasing index, along with one standard deviation. Columns (b) and (e) show histograms of the means. Columns (c) and (f) show histograms of the standard deviations.

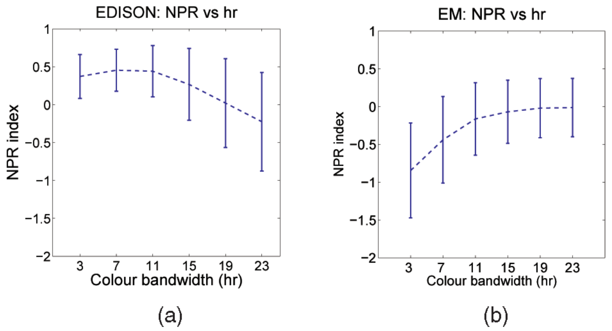


Fig. 19. Mean NPR indices achieved on each color bandwidth (h_r) over the set of images, with one standard deviation. (a) Shows results for the EDISON segmentation system and (b) shows results for EM.

Average performance over all images for different values of h_r. The first three sets of graphs show the results of keeping k constant and choosing from the set $h_r = \{3, 7, 11, 15, 19, 23\}$. Fig. 19 shows the results of running the EDISON system with these parameters, averaged over the image set and with one standard deviation. Fig. 20 shows the same information for the efficient graph-based segmentation (FH) and the hybrid (MS + FH) system on a representative three of the six possible values of k. For completeness, the graphs for the remaining values of k can be found in [21].

As before, we can see that the hybrid algorithm gives slight improvements in stability over the mean shift-based system, but only for smaller values of k. We can also see that, except for $k = 5$, both the mean shift-based system and the hybrid system are more stable across images than the efficient graph-based segmentation system.

Average performance over all images for different values of k. The last two sets of graphs, in Fig. 21, examine the stability of k over a set of images. Each graph shows the average algorithm performance taken over the set of images with a particular h_r and each point shows a particular k. The graphs show a representative subset of the choices for h_r and the remaining graphs can be found in [21]. Once again, we see that combining the two algorithms has improved performance and stability. The hybrid algorithm has higher means and lower standard deviations than the efficient graph-based segmentation over the image set for each k, and especially for lower values of h_r.

5.6 Experiment Conclusions

In this section, we have proposed a framework for comparing image segmentation algorithms using the NPR index, and performed one such comparison. Our framework consists of comparing the performance of segmentation algorithms based on three important characteristics: correctness, stability with respect to parameter choice, and stability with respect to image choice. We chose to compare four segmentation algorithms: mean shift-based segmentation [15], [20], a graph-based segmentation scheme [18], a proposed hybrid algorithm, and expectation maximization [19] as a baseline.

The first three algorithms had the potential to perform equally well on the data set given the correct parameter choice. However, examining the results from the experiments which averaged over parameter sets, the hybrid algorithm performed slightly better than the mean shift algorithm, and both performed significantly better than the graph-based segmentation. We can conclude that the mean shift filtering step is indeed useful, and that the most promising algorithms are the mean shift segmentation and the hybrid algorithm. As expected, EM performed worse than any of the other algorithms both in terms of maximum and average performance.

In terms of stability with respect to parameters, the hybrid algorithm showed less variability when its parameters were changed than the mean shift segmentation algorithm. Although the amount of improvement did decline with increasing values of k, the rate of decline was very slow and any choice of k within our parameter set gave

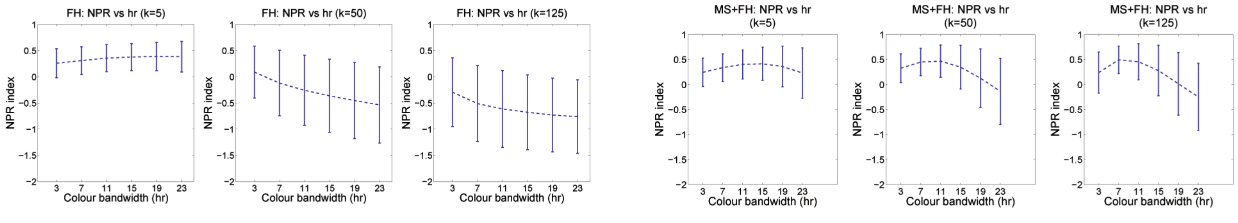


Fig. 20. Mean NPR indices on each color bandwidth $h_r = \{3, 7, 11, 15, 19, 23\}$ over the set of images. One plot is shown for each value of k. Experiments were run with $k = \{5, 25, 50, 75, 100, 125\}$, and we show a representative subsample of $k = \{5, 50, 125\}$. The plots on the left show results achieved using the efficient graph-based segmentation (FH) system and the plots on the right show results fachieved using the hybrid segmentation (MS+FH) system.

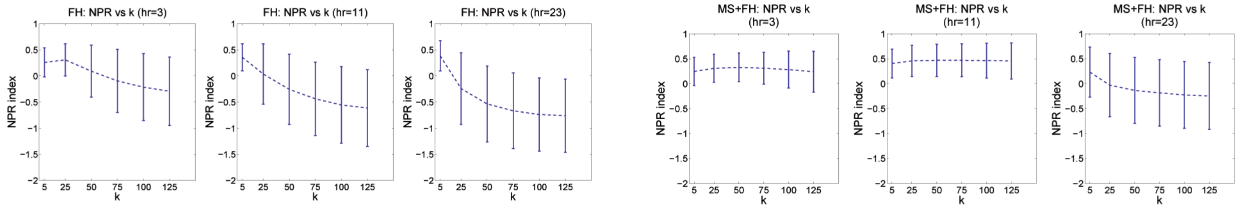


Fig. 21. Mean NPR indices on $k = \{5, 25, 50, 75, 100, 125\}$ over the set of images. One plot is shown for each value of h_r. Experiments were run with $h_r = \{3, 7, 11, 15, 19, 23\}$, and we show a representative subsample of $h_r = \{3, 11, 23\}$. The plots on the left show results achieved using the efficient graph-based segmentation (FH) system and the plots on the right show results achieved using the hybrid segmentation (MS+FH) system.

reasonable results. Although the graph-based segmentation did show very low variability with $k = 5$, changing the value of k decreased its stability drastically.

Finally, in terms of stability of a particular parameter choice over the set of images, we see that the graph-based algorithm has low variability when $k = 5$, however, its performance and stability decrease rapidly with changing values of k. The difference between the mean shift segmentation and the hybrid method is negligible.

We conclude that both the mean shift segmentation and hybrid segmentation algorithms can create realistic segmentations with a wide variety of parameters, however, the hybrid algorithm has slightly improved stability.

6 Conclusions

In this paper, we have presented a measure for comparing the quality of image segmentation algorithms and presented a framework in which to use it. Additionally, we have provided an example of such a comparison.

The proposed measure, the Normalized Probabilistic Rand (NPR) index, is appropriate for segmentation algorithm comparison because it possesses four necessary characteristics: it does not degenerate with respect to special segmentation cases, it does not make any assumptions about the data, it allows adaptive accommodation of refinement, and it is normalized to give scores which are comparable between algorithms and images. We have also demonstrated that the NPR index can be computed in an efficient manner, making it applicable to large experiments.

To demonstrate the utility of the NPR index, we performed a detailed comparison between three segmentation algorithms: mean shift-based segmentation [15], an efficient graph-based clustering method [18], and a hybrid of the other two. We also compared them with a baseline segmentation algorithm based on EM. The algorithms were compared with respect to correctness as measured by the value of the NPR index. Also, two variations of stability were considered: stability with respect to parameters, and stability with respect to different images for a given parameter set. We argue that an algorithm which possesses these three characteristics will be practical and useful as part of a larger vision system. Of course, there is generally a trade-off between these characteristics; however, it is still possible to measure which algorithm gives the best performance. In our experiments, we found that the hybrid algorithm performed slightly better than the mean shift-based algorithm [15] alone, with the efficient graph-based clustering method [18] falling behind the other two.

For future work, it would be interesting to compare other widely used segmentation algorithms such as normalized cuts [22] with the ones presented here. However, many segmentation algorithms have a parameter that explicitly encodes the number of clusters and, yet, do not have well accepted schemes for its selection. Thus, such a comparison would have to be carefully constructed so as not to unfairly bias algorithms either with or without such a parameter.

Appendix

Reduction Using Sample Mean Estimator

We show how to reduce the PR index to be computationally tractable. 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

$$\bar{p}_{ij} = \frac{1}{K}\sum_{k=1}^{K} \mathbb{I}\left(l_i^{S_k} = l_j^{S_k}\right). \tag{9}$$

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:

$$\mathrm{PR}(S_{\text{test}}, \{S_k\}) = \frac{1}{\binom{N}{2}} \sum_{\substack{i,j\\ i<j}} \left[c_{ij}\bar{p}_{ij} + (1-c_{ij})(1-\bar{p}_{ij})\right]. \tag{10}$$

Substituting (9) in (10) and moving the summation over k outward yields

$$\begin{aligned}\mathrm{PR}(S_{\text{test}}, \{S_k\}) = \frac{1}{K}\sum_{k=1}^{K}\Bigg[\frac{1}{\binom{N}{2}}\sum_{\substack{i,j\\ i<j}}\Big[c_{ij}\mathbb{I}\left(l_i^{S_k} = l_j^{S_k}\right)\\ + (1-c_{ij})\mathbb{I}\left(l_i^{S_k} \neq l_j^{S_k}\right)\Big]\Bigg],\end{aligned} \tag{11}$$

which is simply the mean of the Rand index [3] 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 (11) is the number of pairs having the same label in S_{test} and S_k, and is given by

$$\sum_{\substack{i,j\\ i<j}} c_{ij}\mathbb{I}\left(l_i^{S_k} = l_j^{S_k}\right) = \sum_{l,l'} \binom{n^{S_k}(l,l')}{2}, \tag{12}$$

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 (11) 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 of 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

$$\binom{N}{2} + \sum_{l,l'} \binom{n^{S_k}(l,l')}{2} - \sum_{l} \binom{n(l,\cdot)}{2} - \sum_{l'} \binom{n(\cdot,l')}{2}, \quad (13)$$

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

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," *Proc. Int'l Conf. Computer Vision,* 2001.

[2] R. Unnikrishnan and M. Hebert, "Measures of Similarity," *Proc. IEEE Workshop Computer Vision Applications,* 2005.

[3] W.M. Rand, "Objective Criteria for the Evaluation of Clustering Methods," *J. Am. Statistical Assoc.,* vol. 66, no. 336, pp. 846-850, 1971.

[4] R. Unnikrishnan, C. Pantofaru, and M. Hebert, "A Measure for Objective Evaluation of Image Segmentation Algorithms," *Proc. CVPR Workshop Empirical Evaluation Methods in Computer Vision,* 2005.

[5] *Empirical Evaluation Methods in Computer Vision,* H.I. Christensen and P.J. Phillips, eds. World Scientific Publishing, July 2002.

[6] D. Martin, "An Empirical Approach to Grouping and Segmentation," PhD dissertation, Univ. of California, Berkeley, 2002.

[7] J. Freixenet, X. Munoz, D. Raba, J. Marti, and X. Cuff, "Yet Another Survey on Image Segmentation: Region and Boundary Information Integration," *Proc. European Conf. Computer Vision,* pp. 408-422, 2002.

[8] Q. Huang and B. Dom, "Quantitative Methods of Evaluating Image Segmentation," *Proc. IEEE Int'l Conf. Image Processing,* pp. 53-56, 1995.

[9] C. Fowlkes, D. Martin, and J. Malik, "Learning Affinity Functions for Image Segmentation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 2, pp. 54-61, 2003.

[10] M. Meila, "Comparing Clusterings by the Variation of Information," *Proc. Conf. Learning Theory,* 2003.

[11] M. Meilă, "Comparing Clusterings: An Axiomatic View," *Proc. 22nd Int'l Conf. Machine Learning,* pp. 577-584, 2005.

[12] M.R. Everingham, H. Muller, and B. Thomas, "Evaluating Image Segmentation Algorithms Using the Pareto Front," *Proc. European Conf. Computer Vision,* vol. 4, pp. 34-48, 2002.

[13] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement,* pp. 37-46, 1960.

[14] E.B. Fowlkes and C.L. Mallows, "A Method for Comparing Two Hierarchical Clusterings," *J. Am. Statistical Assoc.,* vol. 78, no. 383, pp. 553-569, 1983.

[15] D. Comaniciu and P. Meer, "Mean Shift: A Robust Approach toward Feature Space Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, pp. 603-619, 2002.

[16] L. Hubert and P. Arabie, "Comparing Partitions," *J. Classification,* pp. 193-218, 1985.

[17] D.L. Wallace, "Comments on 'A Method for Comparing Two Hierarchical Clusterings'," *J. Am. Statistical Assoc.,* vol. 78, no. 383, pp. 569-576, 1983.

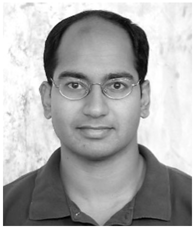
[18] P. Felzenszwalb and D. Huttenlocher, "Efficient Graph-Based Image Segmentation," *Int'l J. Computer Vision,* vol. 59, no. 2, pp. 167-181, 2004.

[19] A. Dempster, N. Laird, and D. Rubin, "Maximum Likelihood From Incomplete Data via the EM Algorithm," *J. Royal Statistical Soc., Series B,* pp. 1-38, 1977.

[20] C. Christoudias, B. Georgescu, and P. Meer, "Synergism in Low Level Vision," *Proc. Int'l Conf. Pattern Recognition,* vol. 4, pp. 150-156, 2002.

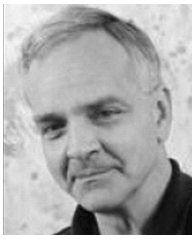
[21] C. Pantofaru and M. Hebert, "A Comparison of Image Segmentation Algorithms," technical report, Robotics Inst., Carnegie Mellon Univ., Sept. 2005.

[22] J. Shi and J. Malik, "Normalized Cuts and Image Segmentation," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, pp. 888-905, 2000.



Ranjith Unnikrishnan graduated from the Indian Institute of Technology, Kharagpur, with the BTech degree (Hons.) in electronics and electrical commnunication engineering in 2000. He received the MS degree from the Robotics Institute at Carnegie Mellon University in 2002 for his work on automated large-scale visual mosaicking for mobile robot navigation. He is currently pursuing the PhD degree at the Robotics Institute working on extending scale theory to problem domains such as nonuniformly sampled 3D data and vector-valued 2D images, and developing new low-level vision algorithms for object recognition and representation in those domains. His research interests also include the development of performance metrics for vision algorithms and new techniques for fast laser-camera calibration. He is a student member of the IEEE.

Caroline Pantofaru received the Honours BSc degree from the University of Toronto in 2001, specializing in mathematics and computer science. She went on to complete the MS degree at the Robotics Institute, Carnegie Mellon University in 2004. Currently, she is pursuing the PhD degree at the Robotics Institute. Her current research interests in computer vision include the incorporation of unsupervised image segmentation into object recognition techniques for robust recognition and precise localization of object classes. She is a student member of the IEEE.



Martial Hebert is a professor at the Robotics Institute, Carnegie Mellon University. His current research interests include object recognition in images, video, and range data, scene understanding using context representations, and model construction from images and 3D data. His group has explored applications in the areas of autonomous mobile robots, both in indoor and in unstructured, outdoor environments, automatic model building for 3D content generation, and video monitoring. He has served on the program committees of the major conferences in computer vision and robotics. He is a member of the IEEE.