Available online at www.sciencedirect.com
ScienceDirect
Computer Vision and Image Understanding 110 (2008) 260–280
Computer Vision and Image Understanding
www.elsevier.com/locate/cviu

Image segmentation evaluation: A survey of unsupervised methods

Hui Zhang [a,*], Jason E. Fritts [b], Sally A. Goldman [a]

[a] Department of Computer Science and Engineering, Washington University, St. Louis, MO 63130, USA
[b] Department of Mathematics and Computer Science, Saint Louis University, St. Louis, MO 63103, USA

Received 5 February 2007; accepted 21 August 2007
Available online 20 September 2007

Abstract

Image segmentation is an important processing step in many image, video and computer vision applications. Extensive research has been done in creating many different approaches and algorithms for image segmentation, but it is still difficult to assess whether one algorithm produces more accurate segmentations than another, whether it be for a particular image or set of images, or more generally, for a whole class of images. To date, the most common method for evaluating the effectiveness of a segmentation method is subjective evaluation, in which a human visually compares the image segmentation results for separate segmentation algorithms, which is a tedious process and inherently limits the depth of evaluation to a relatively small number of segmentation comparisons over a predetermined set of images. Another common evaluation alternative is *supervised* evaluation, in which a segmented image is compared against a manually-segmented or pre-processed reference image.

Evaluation methods that require user assistance, such as subjective evaluation and supervised evaluation, are infeasible in many vision applications, so *unsupervised* methods are necessary. Unsupervised evaluation enables the objective comparison of both different segmentation methods and different parameterizations of a single method, without requiring human visual comparisons or comparison with a manually-segmented or pre-processed reference image. Additionally, unsupervised methods generate results for individual images and images whose characteristics may not be known until evaluation time. Unsupervised methods are crucial to real-time segmentation evaluation, and can furthermore enable self-tuning of algorithm parameters based on evaluation results.

In this paper, we examine the unsupervised objective evaluation methods that have been proposed in the literature. An extensive evaluation of these methods are presented. The advantages and shortcomings of the underlying design mechanisms in these methods are discussed and analyzed through analytical evaluation and empirical evaluation. Finally, possible future directions for research in unsupervised evaluation are proposed.

Keywords: Image segmentation; Objective evaluation; Unsupervised evaluation; Empirical goodness measure

1. Introduction

Image segmentation is a fundamental process in many image, video, and computer vision applications. It is often used to partition an image into separate regions, which ideally correspond to different real-world objects. It is a critical step towards content analysis and image understanding.

Many segmentation methods have been developed, but there is still no satisfactory performance measure, which makes it hard to compare different segmentation methods, or even different parameterizations of a single method. However, the ability to compare two segmentations (generally obtained via two different methods/parameterizations) in an application-independent way is important: (1) to autonomously select among two possible segmentations within a segmentation algorithm or a broader application; (2) to place a new or existing segmentation algorithm on a solid experimental and scientific ground [1]; and (3) to monitor segmentation results on the fly, so that segmentation performance can be guaranteed and consistency can be maintained [2].

* Corresponding author.
E-mail address: huizhang@wustl.edu (H. Zhang).

1077-3142/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2007.08.003

Designing a good measure for segmentation quality is a known hard problem—some researchers even feel it is impossible. Each person has his/her distinct standard for a good segmentation and different applications may function better using different segmentations. While the criteria of a good segmentation are often application-dependent and hard to explicitly define, for many applications the difference between a favorable segmentation and an inferior one is noticeable. It is possible (and necessary) to design performance measures to capture such differences.

Although development of image segmentation algorithms has drawn extensive and consistent attention, relatively little research has been done on segmentation evaluation. Most evaluation methods are either subjective, or tied to specific applications. Some objective evaluation methods have been proposed, but the majority of these have been in the area of *supervised objective evaluation*, which are objective methods that require access to a ground truth reference, i.e. a manually-segmented reference image. Conversely, the area of *unsupervised objective evaluation*, in which a quality score is based solely on the segmented image, i.e. it does not require comparison with a manually-segmented reference image, has received little attention.

The key advantage of unsupervised segmentation evaluation is that it does not require segmentations to be compared against a manually-segmented reference image. This advantage is indispensable to general-purpose segmentation applications, such as those embedded in real-time systems, where a large variety of images with unknown content and no ground truth need to be segmented. The ability to evaluate segmentations independently of a manually-segmented reference image not only enables evaluation of any segmented image, but also enables the unique potential for self-tuning.

The class of unsupervised objective evaluation methods is the only class of evaluation methods to offer segmentations algorithms the ability to perform self-tuning. Most segmentation methods are manually tuned; the parameters for the segmentation algorithm are determined during system development, prior to system deployment, based on the set of parameters that generate the best overall segmentation results over a predetermined set of test images. However, these parameters might not be appropriate for the segmentation of later images. It would be preferable to have a self-tunable segmentation method that could dynamically adjust the segmentation algorithm's parameters in order to automatically determine the parameter options that generate better results. Pichel et al. [72] recently proposed one such system, which uses unsupervised evaluation methods to evaluate and merge sub-optimal segmentation results in order to generate the final segmentation. Supervised segmentation evaluation methods only enable this capability on images for which a manually-segmented reference image already exists. Only unsupervised objective evaluation methods, which do not require a reference image for generating a segmentation evaluation metric, offer this ability for any generic image.

This paper provides a survey of the unsupervised evaluation methods proposed in the research literature. It presents a thorough analysis of these methods, categorizing the existing methods based on their similarities, and then discusses their specific differences. A number of empirical evaluations are performed, comparing the relative performance of nine of these unsupervised evaluation methods. Finally, based on the analysis and experimental results, we propose possible future directions for research in unsupervised segmentation evaluation.

The remainder of this paper is organized as follows: In Section 2, we provide an overview of different kinds of segmentation evaluation methods. In Section 3, we give a detailed analysis of the unsupervised evaluation methods that have been proposed in the literature, categorizing the different methods based on the techniques they use to generate their evaluation scores. Section 4 performs a number of experiments that empirically evaluate nine of the existing unsupervised segmentation evaluation methods in a variety of different situations. Further analysis of these methods are presented in Section 5. Section 6 reviews new multi-level unsupervised evaluation methods that combine the results of the existing methods to achieve better overall performance. Finally, conclusions and future directions for research in unsupervised evaluation methods are discussed in Section 7.

2. Segmentation evaluation

2.1. The hierarchy of current evaluation methods

Many image segmentation methods have been proposed over the last several decades. As new segmentation methods have been proposed, a variety of evaluation methods have been used to compare new segmentation methods to prior methods. These methods are fundamentally very different, and can be partitioned based on five distinct methodologies, as shown in Fig. 1.

Depending on whether a human evaluator examines the segmented image visually or not, these evaluation methods can be divided into two major categories: *Subjective Evaluation* and *Objective Evaluation*. In the objective evaluation category, some methods examine the impact of a segmentation method on the larger system/application employing

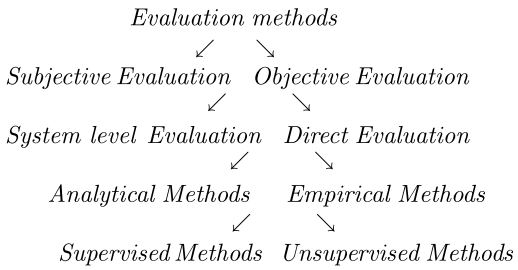

Fig. 1. The hierarchy of segmentation evaluation methods. Our emphasis in this paper is on the unsupervised objective evaluation.

this method, while others study the segmentation method independently. Thus, we divide objective evaluation methods into *System-level Evaluation* and *Direct Evaluation*. The direct objective evaluation can be further divided into *Analytical Methods* and *Empirical Methods*, based on whether the method itself, or the results that the method generated are being examined. Finally, the empirical methods are divided into *Unsupervised Methods* and *Supervised Methods*, based on whether the method requires a ground-truth reference image (as described later) or not.

Notice that these categories are not mutual exclusive. Evaluation methods might use techniques from multiple categories. For example, Shin et al. [68,69] use both supervised evaluation and system-level evaluation. As discussed below, evaluation measures from each category have their own particular limitations. Using evaluation methods that combine techniques from multiple categories is encouraged.

The details of each of these categories is discussed below.

2.2. Subjective evaluation

The most widely used type of evaluation method is subjective evaluation, in which the segmentation results are judged by a human evaluator. The disadvantage of such methods is that visual or qualitative evaluation are inherently subjective (hence their namesake). Subjective evaluation scores may vary significantly from one human evaluator to another, because each evaluator has their own distinct standards for assessing the quality of a segmented image. Furthermore, the results of the evaluation can depend upon the order in which evaluators observe the segmentation results, so obtaining an unbiased understanding of the effectiveness of a segmentation algorithm is a difficult task. It requires a large intensive visual evaluation study. To minimize bias, such a study necessarily involves visual comparative evaluation of an algorithm's segmentation results over a large set of test images by a large group of human subjects. The set of test images must be sufficiently large to be representative of the category of images targeted by the segmentation algorithm. Likewise, the group of human evaluators must be sufficiently large to be representative of the typical human observer. And to reduce favoritism between different algorithms and parameterizations, the testing must be performed under a well-designed set of guidelines [3]. Consequently, subjective evaluation is a very tedious and time-consuming process, and intrinsically, such methods cannot be used in a real-time system to pick between segmentation algorithms or even different parameterizations of a single segmentation algorithm.

2.3. System-level evaluation

Alternate methods popular in systems/applications employing segmentation are to examine the impact of different segmentation methods on the overall system. This approach enables the researchers or system designers to argue that one segmentation method is better than another on the basis of the empirical system results (e.g.[68] compares edge-based methods in an object recognition system).

Unfortunately, this evaluation method is indirect. When the steps following segmentation generate superior results, it does not necessarily mean that the segmentation results were superior, and vice versa. The system-level results from different segmentation methods simply indicate that the characteristics of the results were more favorable for that particular system (e.g. a system might favor fewer regions or rectangular regions, even if more accurate segmentations have larger numbers of segments or irregularly-shaped regions).

2.4. Analytical methods

Analytic methods [4,71] assess segmentation algorithms independently of their output, evaluating them based on certain properties of the segmentation algorithms, such as processing strategy (parallel, sequential, iterative, or mixed), processing complexity, resource efficiency, and segmentation resolution, which are usually not deemed effective for assessing the segmentation quality (e.g. in Liedtke [11]). In other words, analytical methods are only applicable for evaluating algorithmic or implementation properties of segmentation algorithms. These properties are generally independent of the quality of an algorithm's segmentation results, so these properties are not considered effective at characterizing the performance difference between segmentation algorithms.

2.5. Supervised evaluation methods

Supervised evaluation methods [5,6], also known as *relative* evaluation methods [7] or *empirical discrepancy methods* [4], evaluate segmentation algorithms by comparing the resulting segmented image against a manually-segmented reference image, which is often referred to as a *gold standard* [8] or *ground-truth*. The degree of similarity between the human and machine segmented images determines the quality of the segmented image.

One potential benefit of supervised methods over unsupervised methods (discussed below) is that the direct comparison between a segmented image and a reference image is believed to provide a finer resolution of evaluation, and as such, discrepancy methods are commonly used for objective evaluation. However, manually generating a reference image is a difficult, subjective, and time-consuming task.[1] Besides, for most images, especially natural images, we usually cannot guarantee that one manually-generated segmentation image is better than another. In this sense,

[1] While the use of synthesized images as reference images for discrepancy testing offers one potential solution to this problem, Haralick [9] argues that evaluations based on synthetic data can seldomly be generalized.

comparisons based on such reference images are somewhat subjective.

A variety of discrepancy measures have been proposed for segmentation evaluation. Most early discrepancy methods evaluated segmented images based on the number of misclassified pixels versus the reference image, with penalties weighted proportional to the distance to the closest correctly classified pixel for that region [10–14]. Another group of discrepancy methods are based on the differences in the feature values measured from the segmented images and the reference image [15–20]. These methods have been extended to accommodate the problem when the number of objects differs between the segmented and reference images [21–25]. There are also a variety of discrepancy methods for the evaluation of edge-based image segmentation methods [26–28,68,30–35]. Finally, Everingham et al. [36] proposed a method to comprehensively evaluate segmentation algorithms using the Pareto front. Instead of using a single discrepancy metric and evaluating effectiveness in a discrepancy space, it performs evaluation in a multi-dimensional fitness/cost space with multiple discrepancy metrics.

3. Unsupervised evaluation methods

Whereas supervised methods evaluate segmented images against a reference image, unsupervised evaluation methods [45], also known as *stand-alone* evaluation methods [38] or *empirical goodness methods* [4] do not require a reference image, but instead evaluate a segmented image based on how well it matches a broad set of characteristics of segmented images as desired by humans.

Unsupervised evaluation is quantitative and objective. It has distinct advantages, perhaps the most critical of which is that it requires no reference image. A manually-created reference image is intrinsically subjective and creating such a reference image is tedious and time-consuming, and for many applications, it is hard or maybe even impossible. The ability to work without reference images allows unsupervised evaluation to operate over a wide range of conditions (or systems) and with many different types of images. This property also makes unsupervised evaluation uniquely suitable for automatic control of online segmentation in real-time systems, where a wide variety of images, whose contents are not known beforehand, need to be processed.

3.1. Current unsupervised methods

Although supervised methods are the most widely used objective quantitative evaluation methods, some unsupervised methods have been proposed. Many of the early methods in this area focused only on the evaluation of foreground-background segmentation, or only on gray-level images. However, many of these methods contain theory that is beneficial to multi-segment images, and may be adapted to color image segmentation evaluation by revisiting the fundamental theory and re-engineering the methods according to the new constraints.

We first summarize the unsupervised evaluation methods proposed in the literature, and then describe the criteria they use in more depth. These methods are named and listed in Table 1.

- D_{WR} measures the gray-level difference between the original image and the output image after thresholding. It was proposed to evaluate thresholding-based segmentation techniques that separate the foreground object from the background.
- *Busy* is based on the measure of "*busyness*" in the image, with the assumption that the ideal objects and background are not strongly textured and have simple compact shapes.
- η measures both intra- and inter-region variance of the foreground object and the background, allowing the segmentation algorithm to select the threshold that maximizes the inter-region variance.
- PV is a set of segmentation measures that constitute a *performance vector (PV)*. The PV vector stores the factors characterizing the segmentation, including region uniformity, region contrast, line contrast, line connectivity, and texture.
- NU improved upon PV by enhancing the region uniformity measure in PV to use a normalized region uniformity measure.
- SM is a shape measure. It is defined as the sum of the gradients at each pixel whose feature value exceeds both the segmentation threshold and the average value of its neighbors.
- SE is an entropy-based segmentation evaluation measure for intra-region uniformity based on the second-order local entropy.
- F measures the average squared color error of the segments, penalizing over-segmentation by weighting proportional to the square root of the number of

Table 1
The unsupervised evaluation methods

Name	Source	Publication date
D_{WR}	Weszka, Rosenfeld [11]	1978
Busy	Weszka, Rosenfeld [11]	1978
η	Otsu [39]	1979
PV	Levine and Nazif [40]	1985
NU	Sahoo et al. [41]	1988
SM	Sahoo et al. [41]	1988
SE	Pal and Bhandari [16]	1993
F	Liu and Yang [42]	1994
F'	Borsotti et al. [43]	1998
Q	Borsotti et al. [43]	1998
F_{RC}	Rosenberger and Chehdi [44]	2000
V_{CP}	Correia and Pereira [18]	2003
Zeb	Chabrier et al. [45]	2004
E_{CW}	Chen and Wang [46]	2004
E	Zhang et al. [47]	2004
V_{EST}	Erdem et al. [2]	2004

segments. It requires no user-defined parameters and is independent of the contents and type of image.

F' was proposed to improve F, because F was found to have a bias towards over-segmentation, which is the characteristic of producing many more regions than desired within a single real-world object. Since F favors segmentations with a large number of small regions, F' extended F by penalizing segmentations that have many small regions of the same size.

Q improves upon F' by decreasing the bias towards both over-segmentation and under-segmentation (i.e. having too few regions to represent all the real-world objects in the image).

F_{RC} is an evaluation criterion which takes into account both the global intra-region homogeneity and the global inter-region disparity. F_{RC} has two implementations, one designed for non-textured images and one for textured images.

Zeb is an evaluation criterion based on the internal and external contrast of the regions measured in the neighborhood of each pixel.

E_{CW} is a composite evaluation method for color images. It uses *intra-region visual error* to evaluate the degree of under-segmentation, and uses *inter-region region visual error* to evaluate the degree of over-segmentation.

E is an evaluation function based on information theory and the minimum description length principle (MDL). It uses *region entropy* as its measure of intra-region uniformity, which measures the entropy of pixel intensities within each region.[2] It uses *layout entropy*, the entropy indicating which pixels belong to which regions,[3] to penalize over-segmentation when the region entropy becomes small. There is no explicit metric for inter-region disparity, rather the inter-region disparity measure is implicit in the combination of region entropy and layout entropy, which counteract each other to provide a balance between over-segmentation and under-segmentation.

A few evaluation metrics have also been designed to evaluate the segmentation performance of video. These methods use similar metrics to image segmentation evaluation, but typically extend them with metrics to account for inter-frame similarities and differences, such as that attributed to object motion. By modifying these metrics to eliminate the temporal inter-frame metrics, these methods can also be used for image segmentation evaluation. In these methods:

V_{CP} consists of a set of metrics for both intra-object measures (e.g. shape regularity, spatial uniformity, etc.) and inter-object measures (such as contrast). Furthermore, each object in the image is weighted according to its *Relevance*, which is an estimate of how much the human reviewer's attention is attracted to that object.

V_{EST} is a metric measuring the spatial color contrast along the boundary of each object.

3.2. The criteria of unsupervised evaluation

What constitutes a good segmentation? Haralick and Shapiro [48] proposed four criteria:

(i) Regions should be uniform and homogeneous with respect to some characteristic(s)
(ii) Adjacent regions should have significant differences with respect to the characteristic on which they are uniform
(iii) Region interiors should be simple and without holes
(iv) Boundaries should be simple, not ragged, and be spatially accurate

The first two criteria examine characteristics of objects in the image, so we call them *Characteristic Criteria*, whereas the last two criteria are based on how likely each region is regarded as a single object by people, thus we call them *Semantic Criteria*. Many segmentation evaluation methods are based, either explicitly or implicitly, upon the characteristic criteria, perhaps because the semantic criteria are highly application- or object-dependent. For example, criterion (3) may not hold for the segmentation of strongly textured images, and (4) is usually not appropriate for natural images.

These criteria have become the *de facto* standard for unsupervised image segmentation evaluation. Although not all evaluation methods in Table 1 explicitly claim what criteria their metrics are based on, these metrics can be largely divided into three categories: those for measuring *intra-region uniformity* (criterion 1), those for measuring *inter-region disparity* (criterion 2), and those for measuring *semantic cues* of objects, such as shape (criterion 3 and 4). These metrics are then combined in some fashion, such as through the weighted sum of inter- and intra-region metrics or through the division of intra-region metrics by inter-region metrics, to give a composite effectiveness measure.

These metrics are instantiated differently for each method. These metrics, and how they are utilized in each of the existing methods, are summarized in Table 2. The details of these metrics are presented in the following subsections and their mathematical definitions are given in Appendix A (For the complete mathematical definitions of these evaluation methods, please refer to the original papers.).

3.3. Intra-region uniformity metrics

Intra-region uniformity metrics are based on criterion (1). It is an intuitive and effective way to evaluate segmen-

[2] In other words, *region entropy* is the "intensity" entropy of a region.
[3] In other words, *layout entropy* is the "size" entropy of a region.

Table 2
The details of proposed unsupervised evaluation methods

Name	Intra- region		Inter-region		Intra-and inter-region	Semantic metrics
	Metrics	Combination	Metrics	Combination	Combination	
D_{WR} [11]	Color error	Sum	—	—	—	—
Busy [11]	Texture	Sum	—	—	—	—
η [39]	Squared color error	Sum_w (size)	Region color difference	Sum_w (size)	Intra ÷ inter	—
PV [40]	Squared color error	Sum_w (size)	Region color difference	Sum_w (HVS)	Show both	—
	Texture	Sum				
NU [41]	Squared color error	Sum	—	—	—	—
SM [41]	—	—	—	—	—	Shape
SE [16]	Entropy	—	—	—	—	—
F [42]	Squared color error	Penal (sum)	—	—	—	—
F′ [43]	Squared color error	Penal (sum)	—	—	—	—
Q [43]	Squared color error	Penal (sum)	—	—	—	—
F_{RC} [44]	Squared color error	Sum_w (size)	Region color difference	Sum_w (size)	Intra − inter	—
	Texture	Sum_w (size)	Barycenter distance	Sum_w (size)		
V_{CP} [18]	Texture	Sum_w (HVS)	Local color difference	Sum_w (HVS)	Sum_w (weights)	Shape
Zeb [45]	Color error	Sum_w (size)	Local color difference	Sum_w (length)	(Intra÷inter)|inter	—
E_{CW} [46]	Color error	Sum	Region color difference	Sum_w (length)	Sum| (show both)	—
E [47]	Entropy	Sum_w (size)	Entropy	—	Sum	–
V_{EST} [2]	—	—	Local color difference	Sum_w (length)	—	—

"|" is or; sum is unweighted sum; $\text{sum}_w(x)$ denotes summing per-region measures weighted by x; and penal(sum) denotes summary with some additional function applied to improve performance.

tation performance by measuring its intra-region uniformity, so almost all unsupervised methods contain metrics to capture it. While a variety of intra-region uniformity metrics have been proposed, all are based on four quantities: color error, squared color error, texture, and entropy.

3.3.1. Metrics based on color error

Evaluation method E_{CW} computes the intra-region color error, E_{intra} (see Eq. (A4) in Appendix A), as the proportion of misclassified pixels in an image. A misclassified pixel is defined as a pixel whose color error (in L^*a^*b space) between its original color and the average color of its region is higher than a pre-defined threshold.

Evaluation method *Zeb* uses internal contrast, I_i (A5), to measure the uniformity of each region. I_i is defined as the average *MaxContrast* in that region, where *MaxContrast* is the largest luminance difference between a pixel and its neighboring pixels in the same region.

Evaluation method D_{WR} (A3) is designed for evaluating foreground/background segmentation methods based on thresholding. It measures the difference between the gray-level of the original image and the segmented image after thresholding.

3.3.2. Metrics based on squared color error

Evaluation method η was also defined for evaluating foreground/background segmentation methods based on thresholding. Its intra-region uniformity measure, the *within-class variance*, σ_W^2 (A6), is the sum of the squared color error of the foreground object and the background, weighted by their respective sizes.

Evaluation method F_{RC} uses $D(I)$ (A9) as its measure of intra-region uniformity. In the version of F_{RC} designed for non-textured images, $D(I)$ is computed as the average squared color error of each region weighted by its size.

Evaluation method *PV* uses the gray-level uniformity measure, U (A7), to describe intra-region uniformity.[4] For a gray-scale image, U is a measure of the weighted sum of the squared gray-level error of each region.

Evaluation method *NU* was also defined for evaluating foreground/background segmentation methods based on thresholding. It's region uniformity measure, the *normalized uniformity measure*, *NU* (A8), is the normalized sum of the squared color error of the foreground object and the background.

Methods *F* (A10), *F*′ (A11) and *Q* (A12) are based on the average squared color error of each region, although different penalties, either additive and multiplicative, are used to counteract over-segmentation (and under-segmentation, in the case of *Q*). Both *F* and *F*′ use the sum of the squared color error of each region, averaged by the square root of region size, whereas *Q* averages the squared color error by the logarithm of its size (plus 1).

3.3.3. Metrics based on texture

Evaluation method *Busy* measures the "busyness" of an image, assuming that a "smoother" image is preferred. The "busyness" of an image is computed as either the sum of the absolute values of 4- (or 8-) neighbor Laplacians, or is based on the gray-level co-occurrence matrix of the image. Both metrics actually measure the texture or edges

[4] Although all six metrics constituting *PV* are defined on a per-area basis, the ones other than U and texture measure R_α are related to boundaries and pixels on different sides of boundaries, thus inter-region in nature.

across the whole image, so *Busy* effectively only measures global texture uniformity, not individual region uniformity.

Evaluation method *PV* used the texture measure, *R* (A13), to describe intra-region texture uniformity. It computes the texture uniformity based on the average number of regions per section of the segmented image. Like *Busy*, *PV* also only provides a measure of global texture uniformity, not individual region uniformity.

Evaluation method V_{CP} uses *spatial uniformity* to evaluate the intra-region uniformity, which includes two metrics, *SI* (A14) and *textvar* (A15). *SI* measures the standard deviation of the Sobel coefficients of each region, and *text var* is computed as the weighted sum of the variances of the pixels' color (*YUV* components) in each region. So V_{CP} uses the texture within each region as the measure of uniformity.

Evaluation method F_{RC}, as mentioned above, uses $D(I)$ (A9) as its measure of intra-region uniformity. In the version of F_{RC} designed for textured images, the same equation is used for $D(I)$, but instead of computing the squared error from the color components, the squared error is computed from a set of texture attribute vectors computed over a sliding window.

3.3.4. Metrics based on entropy

Evaluation method *SE* uses an entropy-based segmentation evaluation metric, $H^{(2)}$ (A16), as its measure of intra-region uniformity. $H^{(2)}$ is based on the second-order local entropy. It measures intra-region uniformity as the entropy over the co-occurrence matrix containing the probabilities for pixel intensity pairs i and j, for all values of i and j.

Evaluation method *E* uses *region entropy*, H_r (A17), as the measure of intra-region uniformity, which is computed as the entropy for the pixels' luminance values over all pixels within a region.

3.4. Inter-region disparity metrics

Inter-region disparity metrics are based on criterion (2). All inter-region disparity metrics basically use one of four features: average color difference between regions, local color difference along boundaries, barycenter distance, and layout entropy.

3.4.1. Metrics based on average color between regions

Evaluation method η uses *the between-class variance*, σ_B^2 (A18), as the disparity measure, which is the squared difference of the average color between the foreground object and the background.

Evaluation method *PV* uses the region contrast, *C* (A19), to describe the inter-region disparity. *C* is the sum of the per-region contrast measures, weighted by a function approximating the human contrast sensitivity curve. The per-region contrast measure is the weighted sum of the differences between the average color of this region and its adjacent regions divided by the sum of their average colors.

Evaluation method F_{RC} uses $\overline{D}(I)$ (A20) as its measure of global inter-region disparity. In the version of F_{RC} designed for non-textured images, $\overline{D}(I)$ is computed as the average of the weighted sum of $\overline{D}(R_i)$ over all regions, R_i. For each region, $\overline{D}(R_i)$ is computed as the difference in the average gray-level between region R_i and other regions in the image, divided by the number of gray levels in the image.

Evaluation method E_{CW} uses E_{inter} (A21) to measure the inter-region color difference, which is defined as the weighted proportion of pixels whose color difference between its original color and the average region color in the other region is less that a pre-defined threshold. The weights are based on the boundary length between the region and each of the separate regions.

3.4.2. Metrics based on difference of local color along boundaries

Evaluation method V_{CP} uses contrast (A22) as the measure of inter-object disparity. It is defined as the normalized sum of the *local contrast* for the pixels on the boundary of a region, where the local contrast of each pixel is the sum of the largest differences between its *Y*, *U* and *V* components and that of its four neighbors.

Evaluation method *Zeb* uses external contrast E_i (A23) to measure the inter-region disparity. E_i is defined as the average *MaxBorderContrast* for all border pixels in that region, where *MaxBorderContrast* is the largest difference in luminance between a pixel and its neighboring pixels in separate regions.

Evaluation method V_{EST} (A24) measures the spatial color contrast along the boundary of each region. Its key component is the difference between the average colors of the pixel neighborhoods (a pixel and its neighboring pixels in the same region) on opposing sides of a boundary line, averaged by the total number of normal lines [2] drawn on the object boundary.

3.4.3. Metrics based on barycenter distance

Evaluation method F_{RC}, as mentioned above, uses $\overline{D}(I)$ (A25) as its measure of global inter-region disparity. In the version of F_{RC} designed for textured images, an alternate definition of $\overline{D}(I)$ is used, which is computed as the sum of the disparity between two regions. The disparity is computed as the Euclidean distance between the barycenters of the two regions, divided by the magnitude of their barycenters. $\overline{D}(I)$ indirectly measures the complexity of the segmentation.

3.4.4. Metrics based on layout entropy

Evaluation method *E* uses *layout entropy*, H_l (A26), as the measure of inter-region disparity. H_l is defined as the entropy of the pixels in a segmentation layout.[5] H_l does

[5] A segmentation layout is an image used to describe the result of a segmentation. It has the same dimensions as the segmented image, and uses different colors to denote different segments. In a segmentation layout, any two pixels in the same segment have the same color, and any two pixels in different segments have different colors.

not evaluate inter-region disparity directly, but instead works together with the *region entropy*, H_r (A17), to take disparity into account.

3.5. Shape measures

Shape measures are more semantically meaningful than uniformity or disparity measures, but they are highly dependent on the applications and the type of images. For example, while shape information is very beneficial for the segmentation of "sunset" images, it will work poorly when it is added to evaluate the segmentation of "mountain" or "waterfall" images.

Evaluation method *SM* (A27) utilizes a shape measure that is the sum of the gradients at each pixel whose feature value exceeds the segmentation threshold and the average value of its neighbors.

Evaluation method V_{CP} uses a few *shape regularity* metrics to measure geometrical properties of objects, such as *compactness* (A28), *circularity* (A29), and *elongation* (A30).

3.6. Combining into composite metrics

The metrics just defined in Sections 3.3, 3.4 and 3.5 are usually defined on a per-region basis. The evaluation of the whole image requires the combination of the metrics for each individual region. Furthermore, most of the unsupervised evaluation methods consist of both inter-region metrics and intra-region metrics. Some of them also incorporate shape metrics. The way each unsupervised measure combines these metrics is critical for its evaluation performance.

There are two key aspects regarding how the various measures are combined. The first aspect addresses how the individual measures for each region are combined into one composite metric for intra-region uniformity, inter-region disparity, or shape. The second aspect addresses how the separate composite metrics (intra-region uniformity, inter-region disparity, and shape) are combined into a single overall evaluation metric.

For the first aspect, which addresses how the individual measures for each region are combined, there are five different combination methods used. These are discussed below, and also detailed for each specific evaluation measure in Table 2:

3.6.1. Unweighted sum of the individual per-region measures

Evaluation methods D_{WR}, *Busy*, *NU*, E_{CW}, and *PV* use an unweighted sum to combine the individual measures for each region into a single composite metric for intra-region uniformity. D_{WR} sums the gray-level color error for each region of all regions. Similarly, E_{CW} sums the per-region color differences, and *NU* sums the per-region squared color errors over all regions. And finally, *Busy* and *PV* sum their per-region texture metrics across all regions.

3.6.2. Sum of the individual per-region measures, weighted by their size

Most methods based on squared color error weight the regions according to their size in order to sum the individual measures. Such methods include the intra-region measures of η and F_{RC}, and the gray-level uniformity measure (U) of *PV*. *Zeb* and *E*[6] use this method for combining their per-region intra-region uniformity measures. η and F_{RC} also use this method to sum their per-region inter-region disparity measures.

3.6.3. Sum of the individual per-region measures, weighted by the Human Visual System (HVS)

Combining the measures for each region using equal weight, as in (1), or weighting them by their size, as in (2) are straightforward, but oftentimes different objects in an image may attract different degrees of attention from human viewers. Consequently, some measures compute weights based on the HVS, such as the weights that approximate the human contrast sensitivity curve for C in *PV*, and the *relevance* weight in V_{CP}. The relevance reflects the importance of an object in terms of the HVS, and can be computed by the combination of a set of metrics expressing the features that tends to capture the viewer's attention, including texture, compactness, circularity, elongation, size of region, and the average value of Y and V components for every pixel in the region.

3.6.4. Sum of the individual per-region measures, weighted by boundary length

Some of the methods generate their composite inter-region disparity measure by weighting the per-region inter-region disparity values by the length of the adjacent boundaries. This is particularly common in those methods that use the local color difference in the neighborhood of the boundaries, such as in *Zeb*, E_{CW} and V_{EST}.

3.6.5. Sum of the individual per-region measures, with a penalty

Some of the methods do not measure inter-object disparity, but instead use a penalty to make over-segmented images unfavorable. These methods include F, F', and Q.

For the second aspect of combining, which addresses how to combine the intra-region measures with the inter-region measures, there are currently four different ways to combine them. These are discussed below, and also detailed for each specific evaluation measure in Table 2:

(i) one approach is to sum the metrics so that inter- and intra-region metrics can complement each other, either in an unweighted fashion, such as in E and E_{CW}, or using weights, as in V_{CP}.
(ii) another alternative is to simply return the individual metrics separately (i.e. essentially not combining the metrics), either in a table, as in *PV*, or in a graph,

[6] E indirectly weights measures based on size, through entropy

as in E_{CW}.When both inter- and intra-region metrics are measuring color differences, larger inter-region errors and smaller intra-region errors are preferred. Consequently, such methods may combine them in one of the following ways:

(iii) taking the ratio of the intra-region and inter-region measures by dividing them, as in η and Zeb:

$$(\textit{intra-region unif}) \div (\textit{inter-region disp})$$

(iv) taking the difference of the intra-region and inter-region measures by subtracting them, as in F_{RC}:

$$(\textit{intra-region unif}) - (\textit{inter-region disp})$$

3.7. Edge-based segmentation evaluation

Image segmentation methods can be largely divided into three categories: *pixel-based methods*, *region-based methods*, and *boundary-based methods*. *Pixel-based methods* group the pixels with similar features, such as color or texture, without considering the spatial relationship among pixel groups (consequently, regions formed with these segmentation methods can be non-contiguous.) Examples of these methods include clustering [49], adaptive K-means method [50], and histogram thresholding [51], among others. In *region-based methods*, objects are defined as regions of pixels which have homogeneous characteristics. Region-based methods group the pixels according to their similarities and spatial connectedness. Examples of these methods include split-and-merge methods [52], and region-growing methods [53], among others.

The third category of segmentation methods, *boundary-based methods*, are quite distinct from pixel- and region-based methods. In boundary-based methods, objects are defined as pixels surrounded by closed boundaries. In contrast with pixel-based and region-based methods, boundary-based segmentation methods offer the potential advantage that pixels within a closed boundary can have significant variations in their characteristics; i.e. regions may be more heterogeneous in feature values, whereas pixel- and region-based methods are more homogeneous in feature values. Hence, boundary-based methods offer the potential for isolating complex or compound objects into a single region, whereas pixel- and region-based methods are usually unable to do this. Examples of boundary detection methods include edge-flow [54], and color snakes [55], among others.

Unfortunately, the advantage of boundary-based methods for segmentation presents a problem for unsupervised segmentation evaluation methods. This is the problem of discerning whether an edge corresponds to a region boundary, or is simply an intra-region edge. Consequently, to date, edges have not been used with unsupervised evaluation methods. Edges have been used in some supervised evaluation methods [56], which have ground truth reference images that include edges, but in unsupervised evaluation methods, edges are currently used only for measuring the characteristic features of regions, after the regions have already been determined. An example is the PV evaluation method, which uses edges as a measure of texture, specifying line contrast and line connectivity. Again, the problem with edges in unsupervised evaluation is determining whether an edge is a region boundary or simply an edge within a region. Edges (before they are connected as boundaries to form regions) are simply an intermediate product of feature extraction. While this is not necessarily an insurmountable problem, it is one that has as yet not been tackled by the unsupervised segmentation evaluation research community.

4. Experiments

We performed four sets of experiments to examine the performance of these unsupervised segmentation evaluation metrics. These four experiments were designed to demonstrate the effectiveness and bias of these evaluation measures over various types of images and across different segmentation algorithms. The first experiment examines the performance of the evaluation measures on synthetic images. The second and third experiments examine the performance of the metrics on machine segmentations, with the former comparing segmentations produced by the same segmentation algorithm (with varying numbers of segments), while the latter compares segmentations produced by different segmentation algorithms. The final experiment examines the performance of the evaluation metrics in comparing machine segmentations to manually-generated human segmentations.

Since subjective segmentation evaluation (over a sufficient-sized set of human evaluators) is commonly accepted as producing the highest-quality evaluation results, in each of these experiments we used a group of human evaluators (with diverse backgrounds) to subjectively evaluate the segmentation results. The consensus of these evaluators provides a subjective measure of the quality of each segmented image. These subjective evaluations are then used for comparison with the objective quality results from the unsupervised evaluation methods on each segmented image. The better unsupervised evaluation methods are those that demonstrate performance closer to the subjective evaluation results.

These experiments examine the performance of nine evaluation metrics: F, F', Q, E, Vs, Vm, E_{CW}, Zeb, and F_{RC}. The remaining evaluation measures are not used because they are unsuited to general image segmentation. The majority of the other evaluation measures only target evaluation of gray-scale images or foreground-background segmentations. The one exception, V_{EST}, is also not used in these experiments, as it is only a partial evaluation metric, measuring only inter-region disparity; it does not consider intra-region uniformity. So, the metrics F, F', Q, E, V_{CP}, E_{CW}, Zeb, and F_{RC} constitute the full set of metrics that are suitable for use with standard color images and multi-region segmentation algorithms. For V_{CP}, we imple-

mented both variations, *Vs* and *Vm*, which differ in the weights used to combine their intra- and inter-region metrics [57].

4.1. Experiment 1: Synthetic images

First we compare the performance of the unsupervised segmentation evaluation methods on three sets of synthetic images from the Brodatz album [58]. Each of the three sets of images contains 100 images, and each synthetic image has five regions. The images in set 1 contain five regions, each with uniform color. For the images in set 2, two of their five regions are highly textured, and the three remaining regions are of uniform color, with noise. Finally, in set 3, all five regions in each of the images are highly textured. Example images from these sets are shown in Fig. 2. Image sets 2 and 3 are from [59].

Since there are 5 distinct regions in each of the synthetic images, it is evident that the optimal segmentation for all of these images is segmentation layout 5 in Fig. 3, which contains exactly 5 segments, one for each region in the synthetic images. So clearly this segmentation result should be identified by the evaluation measures as the best segmentation for the synthetic images. In addition to the optimal segmentation, we use three under-segmented layouts and three over-segmented layouts for examining the performance of the evaluation metrics. To produce the under-segmented layouts, we merged some of the regions in the optimal segmentation (segmentation layout 5) to generate the three under-segmented layouts, which are shown as layouts 2–4 in Fig. 3, with 2, 3 and 4 segments, respectively. In a similar fashion, we generated the three over-segmented layouts by further dividing some of the regions in the optimal segmentation layout. These are shown as layouts 6–8 in Fig. 3, with 6, 7 and 8 segments, respectively.

These 7 segmentation layouts provide 7 different possible segmentations for the images in the three synthetic image sets. Upon applying the unsupervised segmentation evaluation measures to these 7 segmentation layouts, each

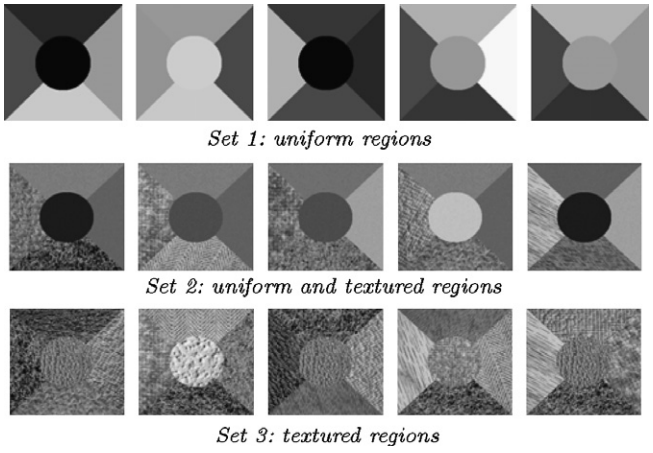

Fig. 2. Example images from the three sets of synthetic images.

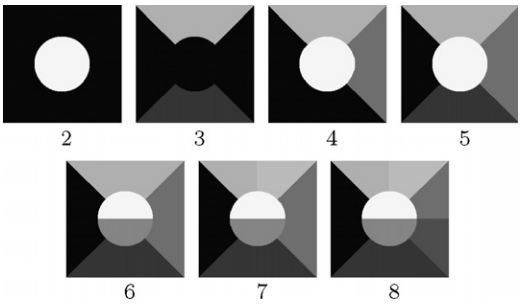

Fig. 3. Segmentation layouts for synthetic images.

evaluation measure gives a quality score to each layout, denoting its "goodness". Since segmentation layout 5 is the optimal segmentation, it should receive the best score among the 7 layouts. However, many of the evaluation measures frequently selected other segmentation layouts. In the event an evaluation measure gives the best score to a layout between 2 and 4, it means that the evaluation measure favors under-segmentation. Furthermore, the lower the layout number is, the more bias the evaluation measure has towards under-segmentation. Conversely, if the best score goes to a layout between 6 and 8, it means that the evaluation measure favors over-segmentation. Again, the higher the layout number is, the more bias the evaluation measure has towards over-segmentation. Consequently, we can use the synthetic image sets and the 7 segmentation layouts to examine the bias of these evaluation methods.

To quantitatively measure the bias, we first define the *bias distance* of an evaluation measure. If N denotes the layout number that an evaluation measure selects as the best segmentation for an image, then we can define *bias distance* $= N - 5$. So, a negative bias distance means a bias towards under-segmentation, and a positive one means a bias towards over-segmentation.

The average bias distances for the nine evaluation measures on the three synthetic image sets are shown in Fig. 4. For set 1, since all the segments are of uniform color, this experiment predominantly examines the bias introduced by

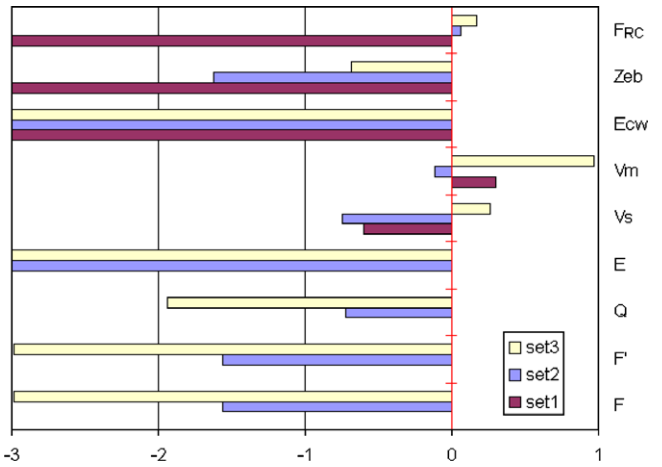

Fig. 4. The average bias distances for 9 evaluation measures on the images in experimental one.

the intra-region disparity metrics. As evident, *Vs* and *Vm* have small biases towards under-segmentation and over-segmentation, respectively. Since *Vs* and *Vm* are simply two variations of V_{CP} that differ only in their weighting mechanisms, we can conclude that for uniform regions the intra- and inter-region metrics in V_{CP} counteract each other well. Conversely, in examining E_{CW}, *Zeb*, and F_{RC}, it is clear that they are all strongly-biased towards under-segmentation for images with uniform regions. However, since few real images contain such uniform regions, the performance results of the evaluation measures on sets 2 and 3 will be more representative for most images.

Also notice that for set 1, no bias distance is reported for F, F' and Q. In regions with perfectly uniform color, their intra-region uniformity measures, which use the squared color error, become zero both for the optimal segmentation layout as well as the over-segmented layouts. So, F, F' and Q achieve their minimum (best) value for all segmentation layouts between 5 and 8. No bias distance is reported for E either, because E is constant when all segments are uniform. In other words, these evaluation methods are not sufficiently discriminative to evaluate images consisting of perfectly uniform regions. From a different perspective, if N is randomly chosen from the set of segmentations that are tied as the best results by an evaluation method, the bias distance for F, F', Q and E would be 1.5. So, they all statistically favor over-segmentation when the regions are uniform.

Fig. 4 also presents the results for the nine evaluation measures on the images in set 2 (textured images) and set 3 (highly-textured images). The results show that F, F', E and E_{CW} all have fairly strong biases towards under-segmentation for textured and highly-textured images. *Zeb* and Q are also moderately biased towards under-segmentation, while *Vs*, *Vm* and F_{RC} are much more balanced, with only small biases towards over-segmentation.

The results from this first experiment demonstrate how the content of the image, and specifically, the amount of texture, affects its performance. E_{CW} is consistently strongly biased towards under-segmentation, regardless of the degree of texture. E, F, F' and Q are biased towards under-segmentation, in a degree proportional to the amount of texture in the image, with E being more biased, and Q being less biased. *Zeb* is also biased towards under-segmentation, but its bias is inversely proportional to the amount of texture. F_{RC} has a negligible bias when the images are textured, but has a large bias towards under-segmentation when the regions are uniform. *Vs* and *Vm* have only minimal bias, slight favoring under-segmentation and over-segmentation in textured and highly-textured images, respectively.

4.2. Experiment 2: Machine vs. Machine segmentation by the same segmentation method

While experiment one was insightful in delineating the biases in the segmentation evaluation measures under varying degrees of texture, the images used are not representative of most real-world images. The images were synthetic and the segmentations ideal. In this and subsequent experiments, we therefore shift over to real-world images for testing the effectiveness of the evaluation measures. In particular, this second experiment examines the performance of the evaluation measures in discerning which segmentations are better among segmentations produced by the same segmentation algorithm.

The test images in this experiment are from the aircraft images in the Military Graphics Collection [60]. For each image we create a series of segmentations where the number of segments varies from 2 to 20, using the Improved Hierarchical Segmentation (IHS) [61] algorithm with fast texture feature extraction [62]. Separately, a subjective evaluation was performed in which each human evaluator selected the best three segmentations and the worst three segmentations in his/her judgment. From the set of best and worst segmentations for each human evaluator, the segmentations from all the best sets for the seven human evaluators were aggregately combined into *best set*, B. Similarly, the segmentations from all the worst sets for the seven human evaluators were aggregately combined into *worst set*, W. For each original image, a segmentation in B is paired with a segmentation in W. We created 250 pairs of segmentations using this approach. Examples are shown in Fig. 5. The images in the leftmost column are the original images, those in the middle column are segmentations from B, and the rightmost column shows segmentations from W.

The nine evaluation methods were applied to these image pairs, and their results were then compared to human evaluation results. For each image pair, an evaluation measure should give the better score to the segmentation from the *best set*, B. An evaluation measure that gives the better score to the segmentation from the *worst set*, W, runs contrary to the subjective evaluation results. The effectiveness of all nine evaluation measures for this experiment are shown in Table 3. The effectiveness is described by *Accuracy*, which is defined as the percentage of the number of times the evaluation measure correctly matches human

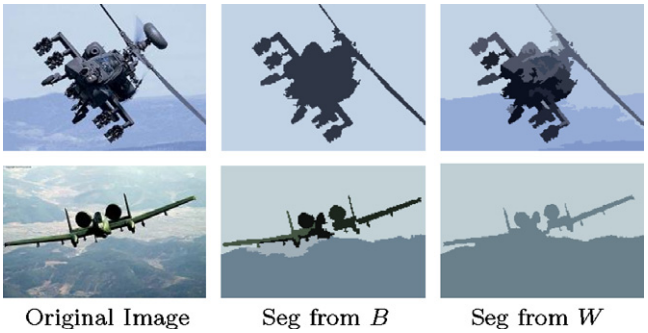

Fig. 5. IHS segmentations of different resolution (shown as a segmentation layouts displaying the average color for each region) (For interpretation of the references to colours in this figure legend, the reader is referred to the web version of this paper.).

Table 3
Accuracy (%) of the evaluation measures in Experiment 2

F	F'	Q	E	Vs	Vm	E_{CW}	Zeb	F_{RC}
47.2	47.2	74.0	33.6	60.8	60.4	33.6	68.4	61.6

evaluation result (i.e. the better score was given to the image in the pair from the *best set*, B), divided by the total number of comparisons in the experiment (i.e. 250 here).

The results, given in Table 3, once again demonstrate the bias of many of the evaluation methods towards under-segmentation. In 66.7% of the 250 segmentation pairs, the better segmentation (the segmentation from the *best set*, B) has a greater number of segments, so those evaluation methods that are biased strongly towards under-segmentation on textured images (as determined in experiment one), namely F, F', E, and E_{CW}, achieve low accuracy in this experiment. On the other hand, those measures that are more balanced or less biased towards under-segmentation, i.e. F_{RC}, Vm, Vs, Q, and Zeb, achieve higher accuracy. Overall, Q performs best here.

4.3. Experiment 3: Machine vs. machine segmentation by different segmentation algorithms

While accurately comparing different segmentations produced by the same segmentation algorithm is sufficiently difficulty, we expect that it is even more difficult to compare segmentations produced by different segmentation algorithms. To examine this, we performed a third experiment using the both the IHS and the Edge Detection and Image Segmentation (EDISON) System [64], which is a low-level feature extraction tool that integrates confidence-based edge detection and mean shift-based image segmentation. For this experiment, we used images from the Berkeley Segmentation Dataset [63] as our test images. 296 images from this database were segmented by both IHS and EDISON. The IHS and EDISON segmentations for each image were paired together, resulting in 296 segmentation pairs for experiment three. Two of the sample images and their segmentations by IHS and EDISON are shown in Fig. 6.

Like the last experiment, this experiment similarly compares the objective results from the nine unsupervised segmentation evaluation measures to subjective evaluation results. The subjective evaluation results were produced by a group of six human evaluators that, for each pair, compared the segmentations from both algorithms and selected the one that they considered better. Only those images where at least four evaluators agreed which segmentation is best were used.

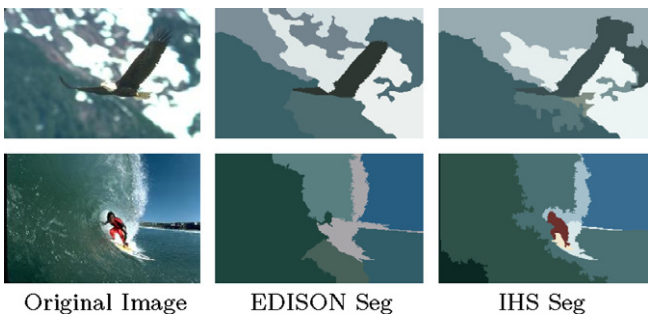

Fig. 6. Image examples segmented by EDISON and IHS.

Table 4 demonstrates the performance of these evaluation measures in experiment three. Results show that the accuracies for Vs, Vm, F_{RC} and Zeb are around 55%, while the accuracies for the other evaluators are even lower.

This experiment indicates once again that many of the evaluation methods are biased towards under-segmentation. Because IHS and Edison commonly produce segmentations with different numbers of regions, it was readily apparent that many of the evaluation methods favored the segmentation in each pair with fewer segments. Upon close examination of the results we found that the percentage of the tests in which the segmentation with fewer regions was judged the better segmentation is: 74.0% for F and F', 78.7% for E_{CW}, 62.16% for E, 61.8% for Q and about 50% for the others, whereas the human evaluators found the segmentations with fewer segments to be better in only 36.8% of the segmentation pairs. In other words, F, F', E_{CW} again demonstrate a strong bias towards under-segmentation, which is why their accuracies are so much lower than the other evaluators in this experiment. Q and E are also moderately biased towards under-segmentation, similarly accounting for their lower accuracies. Consequently, these methods all performed poorly in this experiment since they prefer under-segmented images while the human evaluators preferred more highly-segmented results.

Another reason we anticipate this experiment resulted in the lower accuracies than the prior experiments is the fact that, in many cases the two segmentations in a pair are so similar that it was hard even for human evaluators to determine which one is better. And since the human evaluators had such difficulty, it is unsurprising that the evaluation measures had similar difficulties, resulting in the low accuracies on this experiment.

4.4. Experiment 4: Human vs. machine segmentation

Lastly, we performed a final experiment that contrasts machine segmentations versus the ideal segmentations of those images, as specified by humans. This is a crucial experiment, as it indicates which, if any, of the evaluation measures can distinguish the human ideal for segmentation as the better segmentation, over machine segmentations. Human segmentations differ from machine segmentations

Table 4
The accuracy (%) of 9 evaluation measures in Experiment 3

F	F'	Q	E	Vs	Vm	E_{CW}	Zeb	F_{RC}
38.9	38.9	47.0	41.2	54.4	54.4	30.4	56.1	52.7

in that they are determined based on human perceptual organization, thus their segments are semantically more meaningful.

In this experiment, we again use the images from the Berkeley Segmentation Dataset [63] as our test images. The Berkeley Segmentation Dataset is particularly useful for this experiment, in that it already provides manually-segmented versions (often multiple manually-segmented versions) of the images in the database. For this experiment, we use these human-generated segmentations for comparison with machine segmentations produced by the EDISON segmentation algorithm. For each human-generated segmentation in the database, we generate a machine segmentation with an equal number of segments. Each human segmentation is then paired with the corresponding machine segmentation with the same number of segments. There are 196 pairs of segmentations in our experiments, two of which are shown in Fig. 7.

Again, we apply the nine evaluation measures on these image pairs and compare their results to human evaluation. The results from a group of six human evaluators confirmed that for each image, the human segmentation is better than the machine segmentation.

The performance of these evaluation measures are shown in Table 5. These results demonstrate that most of the evaluation methods disagree with humans, instead selecting the machine segmentation as the better segmentation. There are two major reasons for this result. The first reason behind this disparity is that most of the evaluation measures are based on the same type of feature metrics as the machine-based segmentation algorithms themselves. Such measures as intra-region uniformity and inter-region disparity are based on maximizing or minimizing features such as color error, texture, etc., which are the same features used frequently in machine algorithms for segmenting images. Consequently, it is not surprising that most evaluation metrics prefer machine segmentations. In fact, many of the existing segmentation algorithms could be re-engineered to function as segmentation evaluation methods, as discussed below in Section 7.

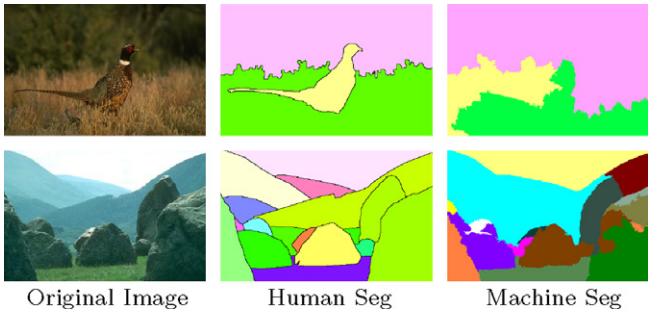

Fig. 7. Human and machine segmentation with the same number of segments (shown as layout where the colors are only used to differentiate segments) (For interpretation of the references to colours in this figure legend, the reader is referred to the web version of this paper.).

Table 5
The accuracy (%) of 9 evaluation measures in Experiment 4

F	F'	Q	E	Vs	Vm	E_{CW}	Zeb	F_{RC}
19.4	15.3	4.1	82.1	1.0	4.1	17.9	9.2	76.5

The second, and foremost, reason behind this disparity is that the manually-segmented images are based on humans' semantic understanding of the real-world objects in the image. Human viewers can and do segment out real-world objects which contain multiple disparate sub-regions. For example, if you consider the bird in Fig. 7, there are four major sub-regions that comprise the bird. The first is the head region, which is predominantly black with little texture. The second is the neck ring, which is white with negligible texture. The chest area is an orange-brown color with some texture. And finally, the back and tail comprise a well-textured coloring of various shades of brown and tan. Human viewers combine these four sub-regions into a single segment based off their recognition of the object as a bird. Conversely, machine algorithms will much more commonly segment these sub-regions separately, or potentially combine them with the background regions, as illustrated in the machine segmentation for the bird in Fig. 7.

A particular reason that most of the segmentation evaluation methods perform poorly when evaluating human-segmented images is evident in an analysis of the mathematical equations defining the evaluation metrics. Notice from Table 2 that for intra-region homogeneity, the majority of the methods use color error or squared color error. Likewise, for the inter-region disparity metric, the majority of the methods use region color difference. The use of these metrics is problematic for human segmentations as that they all implicitly assume a region has a single indicative average value. Further, the color error and squared color error metrics in intra-region homogeneity assume an intensity histogram with a single Gaussian-like distribution. Since human segmentations frequently contain segments with multiple sub-regions, it is clear that these assumptions are invalid for such segments. In these cases, the average values and squared color errors will clearly produce values that poorly represent the sub-regions of such segments.

Given the above understanding, it is not surprising that the majority of the segmentation evaluation measures strongly favored machine segmentations over human segmentations. Of the nine evaluation measures, seven of the measures, F, F', Q, Vs, Vm, E_{CW}, and Zeb, fall into this category. F, F', and Q are strictly based on squared color error for intra-region homogeneity, while E_{CW} uses color error for intra-region homogeneity and region color difference for inter-region disparity, so all four of these measures were prone to problem of an invalid assumption for the intra-segment intensity distribution. Zeb uses local color difference for inter-region disparity, which is a better metric since it only considers pixels on the boundary edges between regions, but it also uses color difference for intra-region homogeneity, and so also fell victim to the

invalid distribution assumption. Like *Zeb*, *Vs* and *Vm* also use local color difference for inter-region disparity, so their disparity metric is safe from the distribution assumption problem. However, while they use a texture measure for intra-region homogeneity, instead of color error or squared color error, their texture measures are largely based on the assumption of a single underlying Gaussian-like distribution, so *Vs* and *Vm* are also prone to the distribution assumption error.

Of the remaining two evaluation measures, E clearly performed the best, achieving dramatically better results than most of the other measures. The reason behind this is two-fold. First, E does not use color error or squared color error for intra-region homogeneity, and similarly does not rely upon region averages for inter-region disparity. Instead, it uses a more flexible distribution assumption, assuming only that good segments will have a set of pixel intensities such that pixels with those intensities will occur frequently in the segment. This assumption is able to accurately model segments containing multiple sub-regions because it does not place any assumptions on the correlation of pixels with different intensities, so it can model multiple distributions, each of which has a subset of pixels with a set of frequently occurring intensities.

The second reason E performs well for the human segmentations is due to E's bias towards unequal-sized segments, which is very complementary with the manner in which humans interpret images. Images frequently contain real-world objects that very widely in size, and because humans generally interpret images by grouping semantically-related regions into larger, more meaningful segments corresponding to real-world objects (e.g. grouping the separate regions of a human body corresponding to limbs, torso, face, hair, etc. into a single complex object representing a human being), regions in human segmentations tend to vary much more dramatically in size than machine-based segmentations, such as those produced by EDISON. E likes the unequal-sized regions in the human segmentation better than the more equal-sized regions in the machine segmentations, because its layout entropy, by definition, favors unequal-sized regions. Mathematically, given a number of segments, X, within an image (or subset of an image) the logarithm of the probability with which a pixel belongs to a region gives a higher entropy value (i.e. a poorer evaluation score) to X regions that are of similar size, and gives lower values (i.e. better evaluation scores) to X segments that vary widely in size. Consequently, in comparing two segmentations, it tends to favor the segmentation in which a few segments dominate, which often matches how humans define objects.

F_{RC} is the one other segmentation evaluation method that performed well in distinguishing that human segmentations are better than machine segmentations. At first glance, it appears surprising that F_{RC} performed well, considering that it may use squared color error for intra-region homogeneity, and region color difference for inter-region disparity. However, recall from Section 3 that F_{RC} has two versions, one for non-textured images and one for textured images. Since many of the textures are well textured and F_{RC} is less prone to the distribution assumption problem in its textured version. Additionally, the combining method used by F_{RC} helps compensate to some degree for the distribution assumption problem (in both the textured and non-textured versions). Finally, in its textured version, the inter-region disparity metric is computed as the distance between the barycenters of regions, which indirectly captures and factors into account the diversity in region sizes. And as we saw with E, this conforms well to the way humans interpret images, which is particularly beneficial to its performance.

Note that the evaluators perform significantly differently in Experiment 4 as they do in Experiment 2 and Experiment 3, although in all experiments their performances are compared against ground-truths from human evaluators. As we analyze in this section (especially in Section 4.5), the supervised evaluation methods experimented in this paper are generally biased. Their intra-region and inter-region measures are not balanced. The changes in inter-region measures easily swamp changes in intra-region measures, so they are generally biased towards under-segmentation. In Experiment 2 and 3, we are comparing two machine segmentations. They usually have different number of segments. The biases towards under-segmentation cause these evaluations methods have higher accuracy, as we discussed in Section 4.2 and 4.3. While in Experiment 4, we compare machine segmentations with human segmentations with the same number of segments. Now the difference between their inter-region measures are very small (smaller or comparable to difference between intra-region measures), so generally these methods work poorly. The exceptions are E and F_{RC}, but they work better here only because of their designs of inter-region measures indirectly prefer machine segmentations.

4.5. Results and discussion

As we saw above, many of the segmentation algorithms exhibited biases or trends under certain circumstances. Here we summarize those findings and discuss the reasons behind them. In doing so, we will discuss those measures first that were more biased towards under-segmentation, and then discuss the remaining measures, which were generally more balanced with respect to over-segmentation versus under-segmentation.

4.5.1. F, F', Q, E and E_{CW}

The first three experiments all demonstrated that five of the evaluation measures, F, F', Q, E and E_{CW}, all have biases (and in some cases strong biases) towards under-segmentation, particularly for textured images.

To examine why these methods are biased towards under-segmentation, recall (as discussed earlier and shown in Table 2) that for their intra-region uniformity measures,

F, F', Q and E_{CW} use the squared color error or color differences for each segment, while E uses region entropy. These metrics for intra-region uniformity generate large values in noisy and highly-textured regions. As a result, as the number of regions decreases the intra-region uniformity values do not increase significantly in textured or noisy images. Conversely, the inter-region disparity measures decrease much more quickly as the number of regions decreases. The degree to which this occurs varies between these five metrics, but they all commonly display this effect where the inter-region values decrease faster than the intra-region values increase as the number of segments decrease. The end result is that the overall goodness measure computed by these metrics is frequently lower (indicating a "better" segmentation) in segmentations with fewer regions in noisy/textured images, which explains the bias towards under-segmentation.

F, F' and Q are further biased towards under-segmentation because they use a weighting factor to penalize against over-segmentation. The weighting factor is a variable value which is a function of the sizes of the segments and the total number of segments. In general however, this weighting factor is much larger than it needs to be (if in fact it is needed at all), which further explains the bias of these three methods towards under-segmentation.

Finally, we also saw from experiment four that E is biased towards segmentations with unequal-sized regions. This was found to be beneficial for the evaluation measure since it corresponds well with human interpretation of images, which frequently contain real-world objects that vary widely in size. As a result, E performed much better in experiment four than most of the other evaluation measures. However, it is also straightforward to envision scenarios containing images with many similar-sized regions, in which case E would not prove as accurate because of its bias towards unequal-sized segments.

4.5.2. Zeb, F_{RC}, Vs and Vm

Zeb, F_{RC}, *Vs* and *Vm* were all more balanced with respect to under-segmentation and over-segmentation, with only slight or negligible biases one way or the other.

Although *Zeb* uses the average color difference as intra-region uniformity measure, it uses the average local color differences along boundaries as the inter-region disparity measure. Moreover, *Zeb* combines metrics by dividing the inter-region metric by the intra-region metric, so it is less biased than F, F', Q, E and E_{CW}.

For textured images, F_{RC} uses texture as the uniformity measure, which is less sensitive to noise than both squared color error and color difference, if an appropriate texture measure is used. It uses Euclidean distance between barycenters of regions as the disparity measure, which indirectly describes disparity. Although F_{RC} performs better than the others in the experiments, its uniformity measure and disparity measure are not guaranteed to counteract each other reliably, so may achieve poorer results under different circumstances.

Vs and *Vm* use local color difference along the boundaries as disparity measure, which is more effective as compared to color difference between neighboring regions. They also use simple shape measures to counteract their disparity measure's bias towards over-segmentation, and accommodate the Human Visual System by varying the weights for different segments based on the estimated importance of each region to human viewers, which is likely why they achieve relatively good evaluation performance in the experiments.

5. Summary analysis

Both the analysis and the experiments demonstrate that the existing unsupervised evaluation methods are far from perfect. As the experimental results demonstrate, the existing approaches are most effective at comparing segmentations produced by different parameterizations of the same segmentation algorithm, they are much less effective at comparing segmentations from different algorithms, and most are particularly poor at distinguishing human segmentations versus machine segmentations. The initial experiments demonstrated one of the major problems with current methods, which is the bias in many of the methods towards under-segmentation. As discussed in Section 4.1, the bias of many of the uniformity measures becomes particularly strong in textured or noisy images, which stems from the fact that the various intra-region uniformity measures are all too sensitive to noise.

The second major problem became readily apparent from the final experiment, which is that most of the methods assume a single underlying distribution, usually Gaussian-like, of the pixels in each segment. Since human segmentations essentially define the ideal target for image segmentation, this is a particularly significant problem, particularly in image segmentation algorithms, such as edge-based methods, that enable compound segments containing multiple sub-regions.

Another important problem, which is not as apparent from the experimental results, is that the existing inter-region disparity measures frequently do not complement the intra-region measures well. While texture as a uniformity measure and local color difference as a disparity measure outperformed the others in the experiments, there is no effective way to combine them to ensure a reliable overall measure. Examining the intra-region homogeneity metrics and inter-region disparity metrics, we found that they commonly have different value ranges, which accounts in part for biases towards over-segmentation and under-segmentation. Most evaluation methods currently combine the two in an additive fashion, in which case the ranges of the two parts should have equivalent ranges in order to counteract each other well. Alternatively, different methods for combining the two parts, such as subtractive or divisive, as used by F_{RC} and *Zeb*, respectively, may prove more effective. As we saw with F_{RC} in the fourth experiment, it was still effective at distinguishing human versus

machine segmentations even though it was still prone to the distribution assumption problem for non-textured images.

The final, and likely most difficult problem, is essentially the same problem that image segmentation itself frequently encounters, which is that the methods currently rely solely on low-level feature extraction, and do not consider the semantic meanings of segments. As a result, the best segmentation result as identified by these evaluation methods may well not be the best segmentation in a human's judgment. It is well known in image segmentation research that purely data-driven segmentation methods based on simple assumptions (such as partitioning an image into different homogeneous regions) is likely to fail in non-trivial situations. Consequently, measuring only the uniformity and/or heterogeneity of simple features, it is unlikely that a segmentation evaluation method can achieve performance comparable to a human evaluator.

Resolving these problems is important to the success of unsupervised segmentation evaluation methods.

6. Evaluation using machine learning

All methods discussed in Section 3 are one-level methods. More specifically, in those unsupervised methods, the constituent metrics are combined in some fixed and predetermined way, as described in Section 3.6, without further analysis of the results, or learning from their previous behavior. Since these measures usually examine different fundamental criteria of the objects, or examine the same criteria in a different fashion, they usually work well in some cases, but poorly in the others.

Zhang et al. [65] propose a *Meta-evaluation* technique (and a *Co-evaluation framework* [57] as a precursor), in which different measures evaluate a segmentation in different ways in the first level, then in the second level a meta-learner generates the final judgment by combining all first-level evaluation results. In the training process, first-level evaluation results, image features, and the corresponding human evaluation are all sent to the meta-learner, enabling the meta-learner to learn how to coalesce the results from the constituent measures under different circumstances, i.e. it learns for what types of images each evaluation measure generates good results and which images generate bad results. This enables it to leverage the appropriate evaluation measures to achieve reliable overall results.

Because of the structure and working mechanism of this method, different evaluation methods can be used together to improve the evaluation performance. The resulting Meta-evaluation is unsupervised if all of its constituent first-level evaluation methods are unsupervised.

Examining the results of the Meta-evaluation for three of the four experiments in the last section, we found the following: For experiment two, using E, F, Q, Vs and Vm as the first level evaluation measures, the Meta-evaluation achieves an accuracy of 85.53% (versus 74.0% from the best result without Meta-evaluation). For experiment three, using F, Q, E, Vs and F_{RC} as the first level evaluation measures, the Meta-evaluation achieves an accuracy of 73.86% (versus 56.1% from the best result without Meta-evaluation). And finally, for experiment four, using F, F', E, E_{CW} and F_{RC} as the first level evaluation measures, the Meta-evaluation achieves an accuracy of 95.87% (versus 82.1% from the best result without Meta-evaluation). The accuracy of the Meta-evaluation in each of these three experiments is significantly better than both the accuracy of any of the constituent first level measures (as shown in Tables 3–5), and the accuracy of any unsupervised one-level evaluation method examined in this paper.

7. Conclusion and future directions

In this paper, we examine the breadth of existing unsupervised methods that objectively evaluate image segmentation. We first present the full range of segmentation evaluation methodologies, and discuss the advantages and shortcomings of each type of evaluation, including subjective, supervised, system-level, and unsupervised evaluation, among others. Subjective and supervised evaluation are currently the two most popular methods, but they have their disadvantages. Subjective evaluation demands time-consuming human studies in which a large body of human subjects evaluates segmentations over a wide variety of images. Supervised methods necessitate comparison against a manually-segmented reference image, which are tedious to produce and can vary widely from one human to another. Unsupervised segmentation evaluation methods offer the unique advantage that they are purely objective and do not require a manually-segmented reference image. This advantage is crucial to general-purpose segmentation applications, such as those embedded in real-time systems, where a large variety of images with unknown content and no ground truth need to be segmented.

We comprehensively analyze the advantages and shortcoming of the underlying design mechanisms of the various unsupervised segmentation evaluation measures through analytical evaluation and experimentation. We classify these methods according to their evaluation criteria, how they define their metrics, and how they combine their individual metrics. These underlying metrics and combination methods help determine the performance of an evaluation measure. We also implemented nine of the evaluation methods suitable for color images, and tested their performance with four different experiments. Finally, we reviewed a promising recent technique employing machine learning to coalesce the results of multiple evaluators to provide much greater overall evaluation accuracy.

The empirical results demonstrate that unsupervised segmentation evaluation performs reasonably-well in evaluating segmentations produced by the same segmentation algorithm, but have much more modest performance in comparing evaluation methods produced by different algorithms and in comparing human versus machine segmentations. We have identified four of the major problems in the current methods, which include:

(i) The existing intra-region uniformity metrics are too sensitive to noise and are biased towards under-segmentation.
(ii) Most existing metrics assume a single underlying distribution, usually Gaussian-like, of pixels in a segment.
(iii) The homogeneity and disparity metrics are frequently not balanced and do not complement each other effectively.
(iv) All the evaluation methods use only low-level features and do not incorporate semantic information.

All four of these are important problems, and need to be addressed in future methods in order to make unsupervised segmentation evaluation a truly viable and robust technology.

7.1. Future directions

In working towards resolving some of these problems, one alternative is to use more sophisticated feature representations. Instead of using a purely data-driven evaluation using basic image features, higher-level information about regions could be used. For instance, the image epitome [66] could be used to measure the similarity between regions. Since an image epitome provides a composite description of shape and appearance, it is possible to achieve a better measure of homogeneity/heterogeneity of the segments. Furthermore, this metric unifies the measures for intra-region homogeneity and inter-region heterogeneity should be unified, enabling them to counteract each other nicely and provide a reliable overall measure.

As a first step towards resolving the semantic information problem, when possible, prior knowledge, especially application-dependent knowledge, should be incorporated into an evaluation method so that the evaluation method knows the preferred characteristics of a segment (as corresponding to a real-world object). Two methods can be applied to include prior knowledge about a preferred segmentation. One method is to design object models. In the example of human face segmentation, using a model defining the configuration of a human face could enable an evaluation method to more effectively human bodies and faces.

The other method to obtain prior knowledge is though machine learning. An evaluation method learns the knowledge about a good segmentation in the training process in which examples and their correct human labels are provided. With such knowledge, an evaluation method differentiates segmentations based not only on the low-level feature characteristics, but also on how human evaluators subjectively rank them. The Meta-evaluation in Section 6 is one example of learning-aided evaluation methods. Since the prior knowledge is gained through the training process, not through manually-segmented reference images, these methods are still unsupervised evaluation methods.

One final possible direction for future research is in re-engineering existing image segmentation methods to perform segmentation evaluation. Since existing segmentation algorithms predominantly perform image segmentation by performing a sequence of decisions in identifying pixel regions based on quantitative image, region, and pixel feature data, many of these algorithms could be re-designed to serve as evaluators. The key difference between segmenting an image and evaluating a segmented image is that in evaluation, the completed segmentation is provided and the quantitative score must be computed based on knowledge of the final result, not on knowledge of the sequence of segmentation steps. Consequently, if the approximate sequence of segmentation steps can be extrapolated from the final segmentation, and the series of segmentation steps can be effectively quantified into a segmentation score, then the image segmentation algorithm can be re-engineered to serve as an unsupervised segmentation evaluation method.

Appendix A. Evaluation metric equations

We use the following notation for the evaluation metrics. Let I be the segmented image with the height I_h and width I_w. Let S_I be the area (as measured by the number of pixels) of the full image (i.e. $S_I = I_h \times I_w$). Observe that S_I is independent of the segmentation itself. We define a segmentation as a division of the image into N arbitrarily-shaped (and possibly non-contiguous) regions. We use R_j to denote the set of pixels in region j, and use $S_j = |R_j|$ to denote the area of region j. For component x (e.g. x might be the red, green, or blue intensity value) and pixel p, we use $C_x(p)$ to denote the value of component x for pixel p. We define the average value of component x in region j by

$$\widehat{C}_x(R_j) = \left(\sum_{p\in R_j} C_x(p)\right)/S_j \tag{A1}$$

The *squared color error* of region j is defined as

$$e_x^2(R_j) = \sum_{p\in R_j} (C_x(p) - \widehat{C}_x(R_j))^2 \tag{A2}$$

We use $N(a)$ to denote the number of regions in the segmented image having an area of exactly a, *MaxArea* to denote the area of the largest region in the segmented image, and Z as a normalization factor. The subscript "*gl*" denotes gray-level, subscript "*o*" means those measures for object, and subscript "*b*" means those for background.

A.1. Intra-region uniformity metrics

A.1.1. D_{WR}

$$\text{Discrepancy} = \sum_{i}^{I_h} \sum_{j}^{I_w} (C_{gl}(i,j) - L(i,j)) \tag{A3}$$

where $C_{gl}(i,j)$ is the gray-level value of pixel $p(i,j)$ on original image and $L(i,j)$ is the gray-level value of $p(i,j)$ on the image after thresholding.

A.1.2. E_{intra} of E_{CW}

$$E_{\text{intra}} = \frac{\sum_{p\in I}\mu\left(\|C_x^o(p) - C_x^s(p)\|_{L*a*b} - \text{TH}\right)}{S_I} \tag{A4}$$

where $C_x^o(p)$ and $C_x^s(p)$ are pixel feature value (color components in *CIEL*a*b* space) for pixel p on original and segmented image, respectively, TH is the threshold to judge significant difference, and $\mu(t) = 1$ when $t > 0$, otherwise $\mu(t) = 0$.

A.1.3. I_j of Zeb

$$I_j = \frac{1}{S_j}\sum_{s\in R_j}\max\{\text{contrast}(s,t), t \in W(s)\bigcap R_j\} \tag{A5}$$

where $W(p)$ is the neighborhood of the p, and contrast$(s,t) = |C_x(s) - C_x(t)|$ is the contrast of pixel s and t.

A.1.4. σ_W^2 of η

$$\sigma_W^2 = \frac{S_b}{S_I}e_{gl}^2(R_b) + \frac{S_o}{S_I}e_{gl}^2(R_o). \tag{A6}$$

A.1.5. Gray-level uniformity measure (U) of PV

$$U = 1 - \sum_{j=1}^{N}\frac{e_{gl}^2(R_j)\times W_j}{Z} \tag{A7}$$

where W_j is a weighting factor.

A.1.6. Normalized uniformity measure (NU)

$$NU = 1 - \frac{e_{gl}^2(R_o) + e_{gl}^2(R_b)}{Z} \tag{A8}$$

A.1.7. D(I) of F_{RC}

$$D(I) = \frac{1}{N}\sum_{j=1}^{N}\frac{S_j}{S_I}e_x^2(R_j) \tag{A9}$$

where $x\in$ {color components} (RGB in our experiments).

A.1.8. F

$$F(I) = \sqrt{N}\sum_{j=1}^{N}\frac{e_j^2}{\sqrt{S_j}}. \tag{A10}$$

A.1.9. F′

$$F\prime(I) = \frac{1}{1000\cdot S_I}\sqrt{\sum_{a=1}^{\text{MaxArea}}[N(a)]^{1+1/a}}\sum_{j=1}^{N}\frac{e_j^2}{\sqrt{S_j}} \tag{A11}$$

A.1.10. Q

$$Q(I) = 0\frac{\sqrt{N}}{1000\cdot S_I}\sum_{j=1}^{N}\left[\frac{e_j^2}{1+\log S_j} + \left(\frac{N(S_j)}{S_j}\right)^2\right] \tag{A12}$$

A.1.11. R_α of PV

$$R_\alpha = \frac{NR_\alpha/S_\alpha}{NR_I/S_I} \tag{A13}$$

where NR_α means the number of regions in area α.

A.1.12. SI and text_var of V_{CP}

$$SI_j = \sqrt{\frac{1}{N_j}\sum_j\sum_k\text{sobel}_j^2 - \left(\frac{1}{N_j}\sum_j\sum_k\text{sobel}_j\right)^2} \tag{A14}$$

$$text\_var(R_j) = \frac{1}{5}\left(3\sigma_{Y(R_j)}^2 + \sigma_{U(R_j)}^2 + \sigma_{V(R_j)}^2\right) \tag{A15}$$

where SI_j is the standard deviation of the Sobel coefficients of region R_j after Sobel operator being applied, and $\sigma_{Y(R_j)}^2$, $\sigma_{U(R_j)}^2$ and $\sigma_{V(R_j)}^2$ are the variances of the Y, U and V components of the pixels in region R_j.

A.1.13. $H^{(2)}$ of SE

$$H^{(2)}(R_k) = -\sum_{i=0}^{T}\sum_{j=0}^{T}p_{ij}\ln p_{ij} \tag{A16}$$

where p_{ij} is the probability from the co-occurrence matrix for pixel intensities i and j, and T is the assumed threshold.

A.1.14. H_r of E

Given a segmented image, define V_j as the set of all possible values associated with the luminance in region j. Then, for region j of the segmentation and value m of the luminance in that region, $L_j(m)$ denotes the number of pixels in region j that have a value of m for luminance in the original image. The entropy for region j is defined as:

$$H_r(R_j) = -\sum_{m\in V_j}\frac{L_j(m)}{S_j}\log\frac{L_j(m)}{S_j}. \tag{A17}$$

A.2. Inter-region disparity metrics

A.2.1. σ_B^2 of η

$$\sigma_B^2 = \frac{S_b}{S_I}\cdot\frac{S_o}{S_I}\cdot(\widehat{C}_{gl}(R_o) - \widehat{C}_{gl}(R_b))^2 \tag{A18}$$

A.2.2. Region contrast (C_α) of PV

For an area α, C_α is computed as:

$$C_\alpha = \left(\sum_{R_j\in\alpha}v_j\sum_{\text{adj}R_i}p_{ij}\frac{|\widehat{C}(R_i) - \widehat{C}(R_j)|}{\widehat{C}(R_i) + \widehat{C}(R_j)}\right)\Big/\sum_{R_j\in\alpha}v_j \tag{A19}$$

where p_{ij} is the adjacency value used for weighting the contrast between regions, and v_j is the weight for region R_j, using a function approximating human contrast sensitive curve.

A.2.3. $\overline{D}(I)$ of F_{RC}

$\overline{D}(I)$ is defined as the average of the weighted sum of $\overline{D}(R_i)$ for each of the region R_i. $\overline{D}(R_i)$ is defined as the disparity between two regions. For two uniform regions R_i and R_j, the disparity is:

$$D(R_i, R_j) = \frac{|\widehat{C}_{gl}(R_i) - \widehat{C}_{gl}(R_j)|}{NG} \tag{A20}$$

where NG is the number of gray levels in the image.

A.2.4. E_{inter} of E_{CW}

$$E_{\text{inter}} = \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} \left[\mu(\text{TH} - ||C_x^o(p) - C_x^s(p)||_{L*a*b}) \cdot w_{ij}/(S_I \cdot Z)\right] \tag{A21}$$

where w_{ij} denotes the jointed length between R_i and R_j, and TH is the threshold to judge significant difference.

Also, as in Eq. (A4), $C_x^o(p)$ and $C_x^s(p)$ are pixel feature value (color components in *CIEL*a*b* space) for pixel p on original and segmented image, respectively. TH is the threshold to judge significant difference, and $\mu(t) = 1$ when $t > 0$, otherwise $\mu(t) = 0$.

A.2.5. Contrast of V_{CP}

$$\text{contrast} = \frac{\sum_{i,j}(2\max(D_{i,j}^Y) + \max(D_{i,j}^U) + \max(D_{i,j}^V))}{4 \cdot 255 \cdot N_b} \tag{A22}$$

where N_b is the number of border pixels for the object, and $D_{i,j}^X$ is the differences between the X component ($X \in \{Y, U, V\}$) of an object's border pixel and its four neighbors.

A.2.6. E_j of Zeb

$$E_j = \frac{1}{N_b(R_j)} \cdot \sum_{s \in n(R_j)} \max\{\text{contrast}(s,t), t \in W(s), t \notin R_j\} \tag{A23}$$

where $n(R_j)$ is the set of pixels on the border of R_j, and $N_b(R_j)$ is the total length of the border of R_j.

A.2.7. V_{EST}

V_{EST} measures the spatial color contrast along the boundary of each object. It is defined as:

$$d_{\text{color}} = 1 - \frac{1}{K_t} \sum_{i=1}^{K_t} \frac{||C_O^i - C_I^i||}{\sqrt{(3 \times 255^2)}} \tag{A24}$$

where C_O^i and C_I^i are the average color calculated in the neighborhood of outside and inside pixel, respectively, and K_t is the total number of normal lines drawn on the object boundary.

A.2.8. $\overline{D}(I)$ of F_{RC}

$\overline{D}(I)$ is the average of $D(R_i,R_j)$ and

$$D(R_i, R_j) = \frac{d(B_i, B_j)}{||B_i|| + ||B_j||} \tag{A25}$$

where B_i is the barycenter of region R_i and $d(.,.)$ is the Euclidean distance.

A.2.9. H_l of E

$$H_\ell(I) = -\sum_{j=1}^{N} \frac{S_j}{S_I} \log \frac{S_j}{S_I}. \tag{A26}$$

A.3. Semantic matrics

A.3.1. SM

$$SM = \frac{1}{C} \sum_{(x,y)} \{\text{Sgn}[C(x,y) - C_{N(x,y)}]\delta(x,y) \text{Sgn}[C(x,y) - T]\} \tag{A27}$$

where $\delta(x,y)$ is the gradient at pixel (x,y), T is segmentation threshold and $C_{N(x,y)}$ is the average value of the neighbors of pixel (x,y).

A.3.2. Shape regularity of V_{CP}

For an object j,

$$\text{compactness} = \frac{p_j^2}{S_j}; \tag{A28}$$

$$\text{circularity} = \frac{4 \cdot \pi \cdot S_j}{p_j^2}; \tag{A29}$$

$$\text{elongation} = \frac{S_j}{(2 \cdot \text{thickness}_j)^2}, \tag{A30}$$

where p_j is the perimeter of object R_j, thickness being the number of morphological erosion steps [67] that can be applied to the object until it disappears.

References

[1] K.W. Bowyer, P.J. Phillips, Empirical evaluation techniques in computer vision, Wiley-IEEE Computer Society Press, 1998.
[2] C.E. Erdem, B. Sanker, A.M. Tekalp, Performance measures for video object segmentation and tracking, IEEE Transactions on Image Processing 13 (2004) 937–951.
[3] E.D. Gelasca, T. Ebrahimi, M. Farias, M. Carli, S. Mitra, Towards perceptually driven segmentation evaluation metrics, in: Proceedings of Conference on Computer Vision and Pattern Recognition Workshop (CVPRW04) vol. 4, 2004.
[4] Y. Zhang, A survey on evaluation methods for image segmentation, Pattern Recognition 29 (8) (1996) 1335–1346.

[5] L. Yang, F. Albregtsen, T. Lonnestad, P. Grottum, A supervised approach to the evaluation of image segmentation methods, Computer Analysis of Images and Patterns (1995) 759–765.
[6] S. Chabrier, H. Laurent, B. Emile, C. Rosenburger, P. Marche, A comparative study of supervised evaluation criteria for image segmentation, in: EUSIPCO, 2004, pp. 1143–1146.
[7] P. Correia, F. Pereira, Objective evaluation of relative segmentation quality, in: ICIP00, 2000, vol I, pp. 308–311.
[8] C. Graaf, A. Koster, K. Vincken, M. Viergever, Validation of the interleaved pyramid for the segmentation of 3d vector images, Pattern Recognition Letters 15 (1994) 467–475.
[9] R.M. Haralick, Validating image analysis algorithms, in: Keynote Address at SPIE Medical Imaging 2000, February 2000, pp. 2–16.
[10] W.A. Yasnoff, J.K. Mui, J.W. Bacus, Error measure for scene segmentation, Pattern Recognition 9 (1977) 217–231.
[11] J. Weszka, A. Rosenfeld, Threshold evaluation techniques, IEEE Transactions on Systems, Man and Cybernetics 8 (8) (1978) 622–629.
[12] S. Lee, S. Chung, R. Park, A comparative performance study of several global thresholding techniques for segmentation, Computer Vision, Graphs, and Image Processing 52 (1990) 171–190.
[13] Y. Lim, S. Lee, On the color image segmentation algorithms based on the thresholding and fuzzy c-means techniques, Pattern Recognition 23 (1990) 935–952.
[14] M. Wollborn, R. Mech, Refined procedure for objective evaluation of video generation algorithms, Doc. ISO/IEC/JTC1/SC29/WG11 M3448, March 1998.
[15] K. Strasters, J. Gerbrands, Three-dimensional image segmentation using split, merge and group approach, Pattern Recognition Letters 12 (1991) 307–325.
[16] N. Pal, D. Bhandari, Image thresholding: some new techniques, Signal Processing 33 (2) (1993) 139–158.
[17] Y. Zhang, J. Gerbrands, Objective and quantitative segmentation evaluation and comparison, Signal Processing 39 (1994) 43–54.
[18] P. Correia, F. Pereira, Objective evaluation of video segmentation quality, IEEE Transactions on Image Processing 12 (2) (2003) 186–200.
[19] M. Van Droogenbroeck, O. Barnich, Design of statistical measures for the assessment of image segmentation schemes, in: Proceedings of International Conference on Computer Analysis of Images and Patterns, 2005(280).
[20] F. Ge, S. Wang, T. Liu, Image-segmentation evaluation from the perspective of salient object extraction, in: Proceedings of IEEE Internatioanl Conference on Computer Vision and Pattern Recognition, 2006, vol. I, pp. 1146–1153.
[21] W. Yasnoff, J. Bacus, Scene segmentation algorithm development using error measures, AOCH 6 (1984) 45–58.
[22] V. Mezaris, I. Kompatsiaris, M. Strintzis, Still image objective segmentation evaluation using ground truth, in: Proceedings of Fifth COST 276 Workshop on Information and Knowledge Management for Integrated Media Communication, October 2003, pp. 9–14.
[23] P. Villegas, X. Marichal, A. Salcedo, Objective evaluation of segmentation masks in video sequences, in: Workshop on Image Analysis for Multimedia Interactive Services, May 2001.
[24] J.S. Cardoso, L. Corte-Real, Toward a generic evaluation of image segmentation, in: IEEE Transanctions on Image Processing (14) No. 11, pp. 1773–1782.
[25] R. Unnikrishnan, C. Pantofaru, M. Hebert, A measure for objective evaluation of image segmentation algorithms, Empirical Evaluation Methods in Computer Vision (2005) 34.
[26] R.M. Haralick, J.S. Lee, Context dependent edge detection and evaluation, Pattern Recognition 23 (1/2) (1990) 1–19.
[27] T. Kanungo, M. Jaisimha, J. Palmer, R.M. Haralick, A methodology for quantitative performance evaluation of detection algorithms, IEEE Transactions on Image Processing 4 (12) (1995) 1667–1673.
[28] K. Bowyer, C. Kranenburg, S. Dougherty, Edge detector evaluation using empirical roc curves, Computer Vision and Image Understanding 84 (1) (2001) 77–103.
[30] Y. Yitzhaky, E. Peli, A method for objective edge detection evaluation and detector parameter selection, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (8) (2003) 1027–1033.
[31] T. Nguyen, D. Ziou, Contextual and non-contextual performance evaluation of edge detectors, Pattern Recognition Letters 21 (9) (2000) 805–816.
[32] J. Fram, E. Deutsch, On the quantitative evaluation of edge detection schemes and their comparison with human performance, IEEE Transactions on Computers C-24 (1975) 616–628.
[33] W. Pratt, Digital Image Processing, John Wiley and Sons, New York, 1978.
[34] C. Odet, B. Belaroussi, H. Benoit-Cattin, Scalable discrepancy measures for segmentation evaluation, in: Proceedings of the IEEE International Conference on Image Processing, vol. 1, September 2002, pp. 785–788.
[35] F.J. Estrada, A.D. Jepson, 'Quantitative evaluation of a novel image segmentation algorithm, in: Proceedings of IEEE Internatioanl Conference on Computer Vision and Pattern Recognition, 2005, vol II, pp. 1132–1139.
[36] M. Everingham, H. Muller, B. Thomas, Evaluating image segmentation algorithms using the pareto front, in: Proceedings of the 7th European Conference on Computer Vision, June 2002, pp. 34–48.
[38] P. Correia, F. Pereira, Stand-alone objective segmentation quality evaluation, JASP 2002 (4) (2002) 389–400.
[39] N. Otsu, A threshold selection method from gray-level histograms, IEEE Transactions on Systems, Man and Cybernetics 9 (1) (1979) 62–66.
[40] M.D. Levine, A.M. Nazif, Dynamic measurement of computer generated image segmentations, IEEE Transactions on Pattern Analysis and Machine Intelligence 7 (2) (1985) 155–164.
[41] P. Sahoo, S. Soltani, A. Wong, Y. Chen, A survey of thresholding techniques, Computer Vision, Graphics, and Image Processing 41 (2) (1988) 233–260.
[42] J. Liu, Y.-H. Yang, Multi-resolution color image segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 16 (7) (1994) 689–700.
[43] M. Borsotti, P. Campadelli, R. Schettini, Quantitative evaluation of color image segmentation results, Pattern Recognition Letters 19 (8) (1998) 741–747.
[44] C. Rosenberger, K. Chehdi, Genetic fusion: application to multi-components image segmentation, in: Proceedings of ICASSP-4 Istanbul, Turkey, 2000.
[45] S. Chabrier, B. Emile, H. Laurent, C. Rosenberger, P. Marche, Unsupervised evaluation of image segmentation appliation to multi-spectral images, in: Proceedings of the 17th international conference on pattern recognition, 2004.
[46] H.-C. Chen, S.-J. Wang, The use of visible color difference in the quantitative evaluation of color image segmentation, in: Proceedings of ICASSP, 2004.
[47] H. Zhang, J. Fritts, S. Goldman, An entropy-based objective evaluation method for image segmentation, in: Proceedings of SPIE-Storage and Retrieval Methods and Applications for Multimedia, 2004.
[48] R. Haralick, L. Shapiro, Survey: image segmentation techniques, Computer Vision, Graphics and Image Processing 29 (1985) 100–132.
[49] V. Meas-Yedid, S. Tilie, J.-C. Olivo-Marin, Color image segmentation based on markov random field clustering for histological image analysis, in: 16th International Conference on Pattern Recognition, 2002.
[50] C.J. Darken, J. Moody, Fast adaptive k-means clustering: some empirical results, in: Proceedings of International Conference on Neural Network, vol. 2, 1990, pp. 233–238.
[51] J.C. Russ, The Image Processing Handbook, CRC Press-IEEE Press, 1998.
[52] X. Wu, Adaptive split-and-merge segmentation based on piecewise least-square approximation, in: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 15, 1993, pp. 808–815.

[53] A. Thakur, R.S. Anand, A local statistics based region growing segmentation method for ultrasound medical images, International Journal of Signal Processing 1 (2) (2004).
[54] W. Ma, B. Manjunath, Edge-flow: a framework of boundary detection and image segmentation, in: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 1997.
[55] G. Sapiro, Color snakes, in: HP Labs Technical Reports, HPL-95-113, 1995.
[56] D. Martin, C. Fowlkes, J. Malik, Learning to detect natural image boundaries using local brightness, color and texture cues, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (1) (2004).
[57] H. Zhang, J. Fritts, S. Goldman, A co-evaluation framework for improving segmentation evaluation, in: Proceedings of SPIE- Defense and Security, Signal Processing, Sensor Fusion, and Target Recognition XIV, 2005.
[58] P. Brodatz, Textures, a photographic album for artistes and designers, Dover, New York, 1966.
[59] Base of 300 synthetic images, http://www.ensibourges.fr/LVR/SIV/interpretation/evaluation/.
[60] Military Graphics Collection, http://www.locked.de/en/index.html.
[61] H. Zhang, J. Fritts, S. Goldman, Improved hierarchical segmentation, Washington University CSE Technical Report, 2005.
[62] H. Zhang, J. Fritts, S. Goldman, A fast texture feature extraction method in hierarchical image segmentation, in: Proceedings of SPIE—Image and Video Communications and Processing, 2005.
[63] 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, in: Proceedings of 8th International Conference Computer Vision, vol. 2, July 2001, pp. 416–423.
[64] Edge Detection and Image Segmentation System, http://www.caip.rutgers.edu/riul/research/code/EDISON.
[65] H. Zhang, S. Cholleti, S. Goldman, J. Fritts, Meta-evaluation of image segmentation using machine learning, in: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2006.
[66] N. Jojic, B.J. Frey, A. Kannan, Epitomic analysis of appearance and shape, in: Proceedings of International Conference on Computer Vision, 2003.
[67] J. Serra, Image Analysis and Mathematical Morphology, Academic, New York, 1993.
[68] M.C. Shin, D.B. Goldgof, K.W. Bowyer, Comparison of edge detector performance through use in an object recognition task, Computer Vision and Image Understanding 84 (1) (2001).
[69] M.C. Shin, D.B. Goldgof, K.W. Bowyer, S. Nikiforou, Comparison of edge detector algorithms using a structure from motion task, in: IEEE Transactions in Systems, Man and Cybernetics, vol. 4, 2001.
[71] J.S. Cardoso, L. Corte-Real, Toward a generic evaluation of image segmentation, IEEE Transactions on Image Processing 14 (11) (2005).
[72] J.C. Pichel, D.E. Singh, F.F. Rivera, Image segmentation based on merging of sub-optimal segmentations, Pattern Recognition Letters 10 (2006).