An Entropy-based Objective Evaluation Method for Image Segmentation

Hui Zhang*, Jason E. Fritts and Sally A. Goldman

Dept. of Computer Science and Engineering, Washington University,
One Brookings Drive, St. Louis, MO USA 63130

ABSTRACT

Accurate image segmentation is important for many image, video and computer vision applications. Over the last few decades, many image segmentation methods have been proposed. However, the results of these segmentation methods are usually evaluated only visually, qualitatively, or indirectly by the effectiveness of the segmentation on the subsequent processing steps. Such methods are either subjective or tied to particular applications. They do not judge the performance of a segmentation method objectively, and cannot be used as a means to compare the performance of different segmentation techniques. A few quantitative evaluation methods have been proposed, but these early methods have been based entirely on empirical analysis and have no theoretical grounding. In this paper, we propose a novel objective segmentation evaluation method based on information theory. The new method uses entropy as the basis for measuring the uniformity of pixel characteristics (luminance is used in this paper) within a segmentation region. The evaluation method provides a relative quality score that can be used to compare different segmentations of the same image. This method can be used to compare both various parameterizations of one particular segmentation method as well as fundamentally different segmentation techniques. The results from this preliminary study indicate that the proposed evaluation method is superior to the prior quantitative segmentation evaluation techniques, and identify areas for future research in objective segmentation evaluation.

Keywords: Image segmentation, objective evaluation, entropy, information theory, minimum description length

1. INTRODUCTION

Image segmentation is a fundamental process in many image, video, and computer vision applications. It is used to partition an image into separate regions, which ideally correspond to different real-world objects. The importance of segmentation has long been recognized, but in previous decades the lack of good segmentation methods was just one of many roadblocks towards making these applications feasible. Today, the problem of generating good segmentations is becoming increasingly critical now that computer processors are able to provide the processing capabilities necessary to make many of these applications feasible. In applications such as content-based image/video retrieval, computer vision for real-world computer interaction, and object- and content-based image/video compression, segmentation has become one of the most important problems that must be solved for successful results.

Innumerable image segmentation methods have been proposed, exploiting a wide variety of image features and characteristics, in the pursuit of more accurate and effective segmentation methods. Unfortunately, segmentation is a complex problem with no exact solution. Research into better segmentation methods invariably encounters two problems: (1) an inability to effectively compare different segmentation methods, or even different parameterizations of any given segmentation method, and (2) an inability to determine whether one segmentation method or parameterization is best for all images or classes of images (e.g. natural images, medical images, etc). Consequently, methods for evaluating different segmentations play a key role in segmentation research.

The majority of studies proposing and comparing segmentation methods evaluate the results only visually or qualitatively. These evaluation methods have many limitations. In particular, visual or qualitative evaluation

*Contact information: E-mail: huizhang@wustl.edu, Telephone: 1 314 935-8561, Fax: 1 314 935-7302

methods are rather subjective. Results vary significantly between different evaluators, because each evaluator may have distinct standards for measuring the quality of the segmentation.

An alternate method popular for system- or application-level studies for systems/applications employing segmentation is 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. However, this evaluation method is indirect. When the steps following segmentation generate superior results, it does not necessarily mean that the segmentation results were superior. Likewise, inferior system results do not necessarily mean that the segmentation results were bad. 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 would create large numbers of segments or irregular regions). Consequently, it is desirable to have a reliable objective evaluation method that evaluates the quality of the segmentation itself and is not tied to any particular system.

A few quantitative evaluation methods have been proposed, but these existing methods were designed using ad-hoc empirical evaluations and have no theoretical foundation. In this paper, we propose a novel objective segmentation evaluation method based on information theory. The new method uses entropy as the basis both for measuring the uniformity of pixel characteristics (luminance is used in this paper) within a segmentation region, and for measuring the complexity of the division of the image into regions. Our new evaluation method provides a quality score that can be used to compare different segmentations of the same image. This method can be used to compare both various parameterizations of one particular segmentation method (including those which differ in terms of the number of regions used in the segmentation) as well as fundamentally different segmentation techniques. Also, the results from this preliminary study indicate that the proposed evaluation method is superior to the prior quantitative segmentation evaluation techniques, and identify areas for future research in objective segmentation evaluation.

The remainder of the paper is organized as follows: In Section 2, we provide an overview of the previous research on objective segmentation evaluation that most closely relate to our proposed objective evaluation method. Then in Section 3, we describe our entropy-based segmentation evaluation method. Experimental results, comparisons, and analysis are presented in Section 4, and Section 5 concludes the paper and discusses future work.

2. RELATED WORK

A variety of techniques have been proposed for quantitatively evaluating segmentation methods. These methods can be classified into three categories. The first category includes analytic methods, in which segmentation algorithms are treated directly by considering some measure (e.g. complexity), which by *a priori* knowledge is assumed to be the appropriate measure. Typically this measure was incorporated into the original segmentation algorithm as well. The second category includes supervised evaluation methods (also known as empirical discrepancy methods).[1,2] In these methods, the results of a segmentation algorithm are compared to a "standard" reference image that is manually segmented beforehand, and the amount of discrepancy becomes the measure of segmentation effectiveness. This is the most commonly used method of objective evaluation. However, manually generating a reference image is a difficult, subjective, and time-consuming job, and for most images, especially natural images, we generally cannot guarantee that one manually-generated segmentation image is better than another. Consequently, comparison using such reference images cannot ensure accurate evaluations. The third category includes unsupervised evaluation methods (also known as empirical goodness methods).[3–7] In these methods, the segmentation results are evaluated by judging the quality of the segmented image directly to evaluate some pre-defined criteria, such as the partitioning of foreground objects from the background. These evaluation measures are typically used with gray-level images and are not designed for general-purpose applications. In particular, such empirical goodness measures are not appropriate for evaluating segmentations of natural images, such as mountain, lake, waterfall, or forestry scenes, since no simple pre-defined criteria exists that separate good segmentations from inferior ones.

In creating an objective quantitative evaluation method, it is desirable that the method be independent of the contents and type of image. Liu and Yang[8] proposed an evaluation function, F, based on empirical studies. Their function requires no user-defined parameters and does not depend on the type of image. However, unless the image has very well defined regions with very little variation in luminance and chrominance, the F evaluation function has a very strong bias towards segmentations with very few regions. Only in images where segmentation will result in very uniform regions, will the F evaluation function prefer segmentations with many regions. Borsotti et al.[9] improved upon Liu and Yang's method, proposing the modified quantitative evaluations, F' and Q. All three of these evaluation functions were generated entirely from the results of empirical analysis, and have no theoretical foundations. More details about F, F', and Q are given in Sections 2.1 and 2.2.

Throughout the remainder of the paper we use the following notation. Let I be the segmented image and let S_I be the area (as measured by the number of pixels) of the full image. 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 $\hat{C}_x(R_j) = \left(\sum_{p \in R_j} C_x(p)\right)/S_j$. The *squared color error* of region j is defined as

$$e_j^2 = \sum_{x \in \{r,g,b\}} \sum_{p \in R_j} (C_x(p) - \hat{C}_x(R_j))^2.$$

We use $N(a)$ to denote the number of regions in the segmented image having an area of exactly a, and $MaxArea$ to denote the area of the largest region in the segmented image.

For all evaluation functions discussed in this paper, segmentations that result in small values are considered more favorable (i.e. better segmentations) than those with high values. This is true of Liu and Yang's F function, Borsotti et al.'s F' and Q functions, and our proposed functions, H_w and E.

2.1. Liu and Yang's evaluation function F

Based on empirical studies, Liu and Yang[8] proposed an objective quantitative evaluation function F for image segmentation:

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

Observe that for any segmentation I in which the color error is zero for all segments (i.e. there is no variance in color within each region), the value of $F(I) = 0$ and hence a segmentation in which each pixel is in its own region will minimize the value of F. Suppose we have a complex image in which all cannot be zero, except for a segmentation in which each pixel is its own region. Still F has two strong biases: segmentations with lots of regions are heavily penalized by $\sqrt{N}$, and segmentations that have regions with large areas are heavily penalized unless the large region is very uniform in color, since the total error (not average error), is used and only divided by the square root of the area of the region (versus being divided by S_j, which would give the average squared error).

2.2. Borsotti, Campadelli, and Schettini evaluation functions, F' and Q

As discussed above, for Liu and Yang's method the presence of many regions in the segmented image is penalized only by the global measure $\sqrt{N}$. When the number of regions in the image is large enough so that the total color error is near zero, the F evaluation measure will be favorable even though the image might be over-segmented. So, Borsotti et al.[9] proposed the following evaluation function, F', to improve upon Liu and Yang's method:

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

Observe that if the segmentation has lots of regions consisting of only one pixel then the multiplicative factor that precedes the summation will be $O(N^2)/S_I$ which is a much larger penalty than the $\sqrt{N}$ that occurs in F.

Thus F' will correctly evaluate segmentations with lots of regions as very poor (unless all color errors are 0) whereas F will incorrectly rank them as good segmentations.

Both F' and F reach their minimum value of zero on an image in which each region is its own pixel. This is not a big problem since one should never consider allowing the number of regions in the segmentation to be as large as the area of the image. However, a more serious problem is both F and F' highly penalize segmentations with a large number of regions and only when the squared error in all regions gets very small (which typically will not happen on natural scenes until the number of regions approaches the image area) will a segmentation with more than a few regions be evaluated as best. Thus, Borsotti et al. further improved upon F and F', and proposed the evaluation function Q*:

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

As in F, Q uses $\sqrt{N}$ to penalize segmentations that have a lot of regions. However, the influence that the $\sqrt{N}$ has is greatly reduced by dividing the squared color error by $1 + \log S_j$ which causes the squared color error to have a much bigger influence in Q as compared to its influence in both F and F'. As an effect of this change, regions with large area that are not uniform in color are penalized even more in Q than in F and F', and so Q has a very strong bias against regions with large area unless there is very little variation in color. Finally, the second term in the summation in the definition of Q adds a small bias against having lots of regions with the same area. However, this term typically has a very small value as compared to the first term in the summation, and so has negligible effect on the evaluation. The only exception to this fact is when $N(S_j)$ gets large for some region j which can only occur when N, the number of regions, is very large. Our experiments, shown in Section 4, demonstrate that such behavior does occur for real images.

While Borsotti et al. concluded that Q produces a better evaluation than F' and F, the equation for Q is still based entirely on empirical analysis, produces inaccurate evaluations in some situations, and is limited by the lack of a theoretical foundation. To overcome these limitations of the existing segmentation evaluation methods, in the next section we propose an information theoretic approach to segmentation evaluation based on entropy. Then we compare our proposed method against the evaluation functions of F, F' and Q in Section 4.

3. OUR NEW ENTROPY-BASED EVALUATION METHOD

A good segmentation evaluation should maximize the uniformity of pixels within each segmented region, and minimize the uniformity across the regions. Consequently, entropy, a measure of the disorder within a region, is a natural characteristic to incorporate into a segmentation evaluation method. Given a segmented image, where region j is a region of the image, we define v as one of the features among those used to describe the pixels in region j, and define $V_j^{(v)}$ as the set of all possible values associated with feature v in region j. Then, for region j of the segmentation and value m of feature v in that region, we use $L_j(m)$ to denote the number of pixels in region j that have a value of m for feature v (e.g. luminance) in the original image. The entropy for region j is defined as:

$$H_v(R_j) = -\sum_{m \in V_j^{(v)}} \frac{L_j(m)}{S_j}\log\frac{L_j(m)}{S_j}.$$

From an information coding theory point of view, $L_j(m)/S_j$ represents the probability that a pixel in region R_j has a luminance (or other feature) value of m. Thus $H_v(R_j)$ is the number of bits† per pixel needed to encode the luminance for region R_j, given that you know region R_j. We shall hereafter use luminance as the feature of choice, and simplify $H_v(R_j)$ to $H(R_j)$ with the default feature v being luminance. Finally, we define the *expected region entropy* of image I as the expected entropy across all regions where each regions has weight (or probability) proportional to its area. That is, the expected region entropy of segmentation I is:

*Throughout this paper we use a base-10 logarithm.

†Since we use a base-10 logarithm, technically we should replace "bits" by "Hartleys". Alternatively, observe that the only difference between a base-2 and base-10 logarithm is a multiplicative constant of $log_2 10$ and this does not have any impact on how images are ranked.

$$H_r(I) = \sum_{j=1}^{N} \left(\frac{S_j}{S_I} \right) H(R_j).$$

The expected region entropy serves in a similar capacity to the term involving the squared color error used in F, F', and Q — it is used as a measure of the uniformity within the regions of I. When each region has very uniform luminance then $H_r(I)$ will be small. And as with the squared color error measure, when all pixels in a region have the same value, then the entropy for the region will be 0. Since an over-segmented image will have a very small expected region entropy, just as done for F, F', and Q, we must combine the expected region entropy with another term or factor that penalizes segmentations having a large numbers of regions since there would otherwise be a strong bias to over-segment an image. One approach would be to use similar ideas to the prior work and multiply the expected region entropy by $\sqrt{N}$ to penalize segmentations with a large numbers of regions. We call this alternative evaluation function the *weighted disorder function*, $H_w(I)$, where:

$$H_w(I) = \sqrt{N} \sum_{j=1}^{N} \left(\frac{S_j}{S_I} \right) \cdot H(R_j) = \sqrt{N} \cdot H_r(I).$$

Just as $\sqrt{N}$ penalizes F too much in segmentations with many regions, we present experiments in Section 4 that show the weighted disorder function, $H_w(I)$, is similarly not a very effective evaluation measure.

While the expected region entropy provides an estimate of the average disorder within a region in a segmented image, fully encoding the information in an image not only entails encoding the feature value of a pixel with a region (i.e. the region entropy), but also encoding a representation for the segmentation. (One can think of $\sqrt{N}$ as one very coarse alternative measure for this representation.) To encode the segmentation representation itself, each pixel must be placed in one of the regions. Whereas the expected region entropy generally decreases with the number of regions, we expect the number of bits for specifying a region for each pixel, a measure we call the *layout entropy*, to increase with the number of regions. Hence the two factors can be used to counteract the effects of over-segmenting or under-segmenting when evaluating the effectiveness of a given segmentation. We define the layout entropy as:

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

Again, when viewed using a coding theory framework, one can view $p_j = S_j/S_I$ as the probability that a each pixel in the image belongs to region j under a probabilistic assumption that each pixel is independently selected to be in region j with probability p_j. Thus the above function indicates the number of bits (or Hartleys when using a base-10 logarithm) per pixel needed to specify a region id of each pixel for a particular segmentation I. Using the layout entropy, we propose an additional entropy-based evaluation function, E, which additively combines both the layout entropy and the expected region entropy in measuring the effectiveness of a segmentation method:

$$E = H_\ell(I) + H_r(I).$$

An alternate view of our evaluation method is obtained by applying the minimum description length (MDL) principle[10] to balance the trade-off between the uniformity of the individual regions with the complexity of the segmentation. When applied to our problem, the MDL principle would select the segmentation that encodes the image in the fewest bits (or Hartleys) where the encoding of the image has two components: the encoding of the segmentation and the encoding of the image given the segmentation. In our formula, $H_\ell(I)$ is the number of Hartleys to encode the segmentation I, and $H_r(I)$ is the number of Hartleys to encode the pixels of the original image given the segmentation I is already known.

Unlike F, F', and Q, our new measure, E, is not minimized when the image is maximally segmented with one pixel per region, since in such an event the layout entropy becomes very large. Furthermore, E will not favor segmentations with very few regions since there the expected region entropy will be high. As desired, E balances these two costs.

Finally, while we have not pursued this direction here, a very natural extension of our work would be to add a user-specified weighting parameter that indicates the relative importance of the expected region entropy measure and the layout entropy. This would enable the a user to tailor the evaluation method to his/her particular subjective preferences. Users favoring fewer, coarser regions would weight the layout entropy more heavily, while users preferring a larger number of regions with less variability (greater uniformity) per region would conversely weight the expected region entropy more heavily. Also, while this paper only explores using the luminance feature for evaluation, the feature choice can be similarly selected by the user, allowing the user to tailor the evaluation method to his/her particular needs.

4. EXPERIMENTAL RESULTS

4.1. Methodology

We empirically studied the objective evaluation methods F, F', Q, H_w, and E, on the segmentation results from two different segmentation algorithms. The first segmentation method, the Edge Detection and Image Segmentation (EDISON) System[11] is a low-level feature extraction tool that integrates confidence-based edge detection and mean shift-based image segmentation. It was developed by the Robust Image Understanding Laboratory at Rutgers University. We used EDISON to generate images that vary in the number of regions in the segmentation to see how the evaluation methods are affected by the number of regions.

The second segmentation algorithm is hierarchical image segmentation method,[12] which builds a segmentation tree, where the leaf nodes of the tree correspond to individual pixels and the root is a segmentation with just one region containing all pixels. At each level of the tree, nodes are grouped together according to the similarities between their feature vectors. The feature vectors are extracted by a fast feature extraction technique[13] applied before tree construction, that uses a threshold to balance the extraction speed and texture extraction coarseness. In the experiments presented here, we vary the threshold paremeter to generate different segmented images, all with the same number of regions, for comparing the evaluation methods when the number of regions is held fixed.

We use these two segmentation methods to do a preliminary study on the effectiveness of these quantitative evaluation methods on different segmentation parameterizations and segmentation techniques. Results from all five evaluation functions are examined and compared. Recall that for all of these measures, a smaller value indicates a better segmentation. We judge the effectiveness of F, F', Q, H_w and E based on their consistency with evaluations provided beforehand by a small group of human evaluators. We generally selected segmentations for which the human evaluators agreed upon which segmentation was best.

4.2. Results

In our first set of experiments, we vary the total number of regions in the segmentation (using EDISON to generate the segmentations) to study the sensitivity of these objective evaluation methods to the number of regions in the segmentation. With an increase in the number of regions, the segmented images clearly look better to the observer, since more details are preserved. However, more regions do not necessarily make a better segmentation, since over-segmentation can occur and the trade-off between the number of regions and the amount of detail preserved can be heavily influenced by the needs of the user. The segmented images with 2, 4, 8, 12, 29, 31, 50 and 415 regions are shown in Figure 1.

In Figure 2 we show the results for the five objective functions over a very wide range of image size (ranging from 1 region to up to 1920 regions). Observe that both F and F' generally increase in value and are almost identical until there about 400 regions are in the segmentation, at which point the differences between F and F' become apparent as the value of F begins to decrease as the squared color error is small enough to counteract the $\sqrt{(N)}$, whereas F' continues to rise due to the penalty added for having lots of small regions. So when the number of regions varies, they give biased evaluation results that favor segmentations with fewer regions. The measure H_w also increases as the number of regions grows which illustrates that introducing H_ℓ is an important component of our evaluation function.

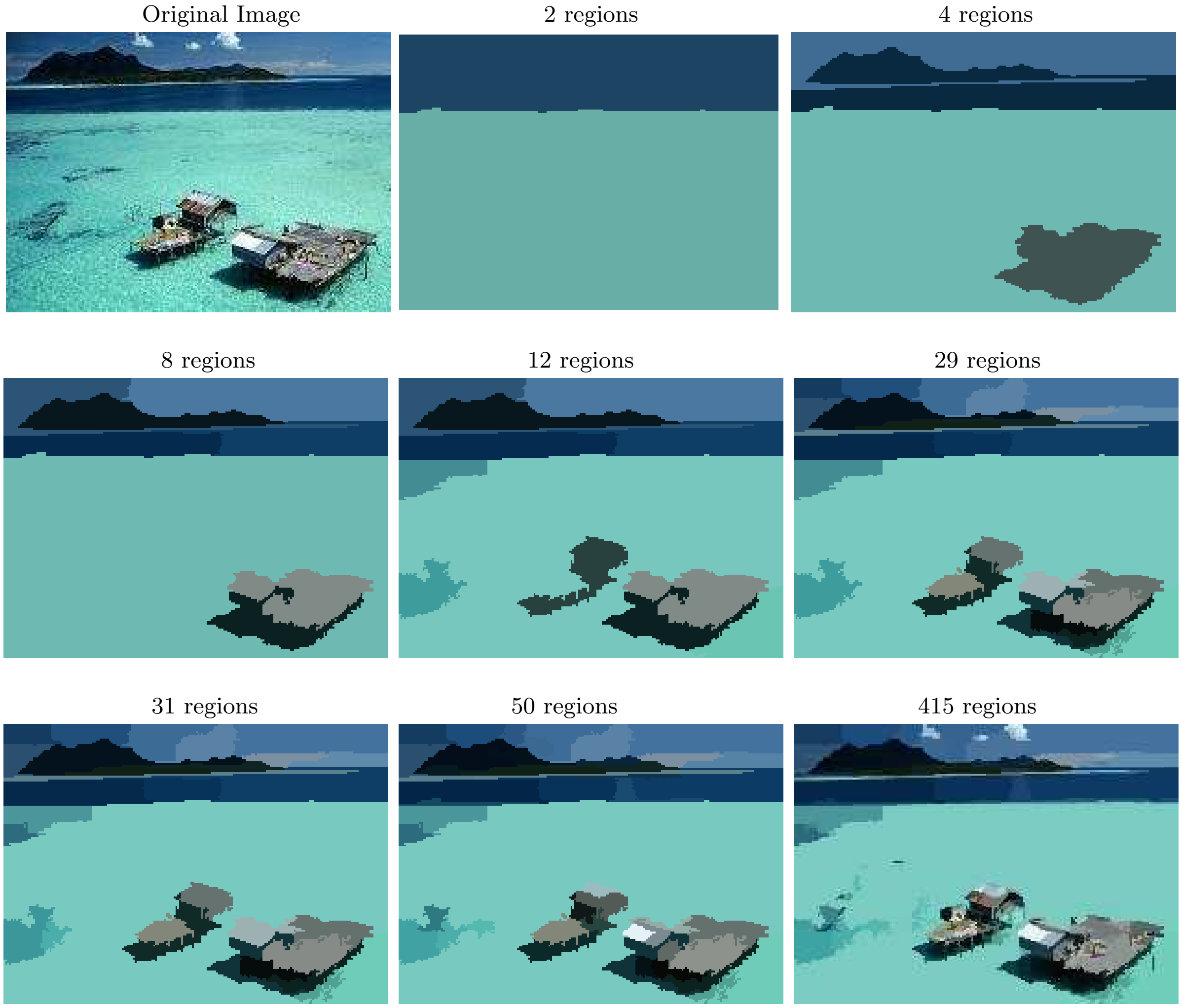

Figure 1. Segmented "sea" images where the number of regions in the segmentation varies.

The measure Q also tends to increase as the number of regions grows, however, unlike the other methods, it has clear local minima which can be viewed as showing a preference for a particular segmentation in a specified range of regions sizes. For example, if an image around 10 regions is desired, then based on Q one would select an image with 12 regions since that corresponds to a local minima in the Q function. Similarly, if a more fine-grained segmentation is desired, then the local minima that occurs at 30 regions could be selected. For this particular image when using EDISON for constructing the segmentations, after 31 regions, the Q measure strictly increases as the number of regions increases.

One important observation to make is that F and F' have their lowest value when the the segmentation has just one region that contains the entire image. While the Q measure ranks the segmentation with 3 regions as slightly better than the segmentation with a single region, in general these measures have a strong bias towards the meaningless segmentation containing a single region. Since the changes in E cannot be seen in Figure 2, we look more closely at E in Figure 3. First, observe that the value of E for images with 4–7 regions is much smaller than its value when there are 3 or less regions. Interestingly, both Q and E have a local minima at 31 regions. However, unlike Q, as the number of regions increases, additional local minima occur in E and so it can be used to pick out the best segmentation over a wider range of desired granularity.

For a greater understanding of the operation of Q and E, we more carefully study the components of these two functions. Q can be broken into two terms $Q_1 = \frac{\sqrt{N}}{1000 S_I} \sum_{j=1}^{N} \left(\frac{e_j^2}{1+\log S_j} \right)$ and $Q_2 = \frac{\sqrt{N}}{1000 S_I} \sum_{j=1}^{N} \left(\frac{N(S_j)}{S_j} \right)^2$.

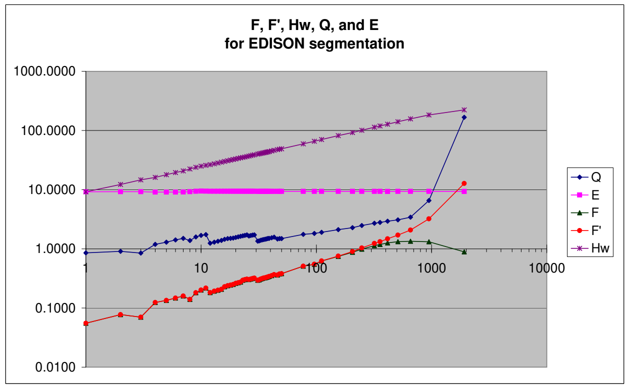

Figure 2. The comparisons of F, F', Q, H_w, and E for segmented "sea' images (when the number of regions varies). The x-axis gives the number of regions. Note that both axes are shown on a log-scale.

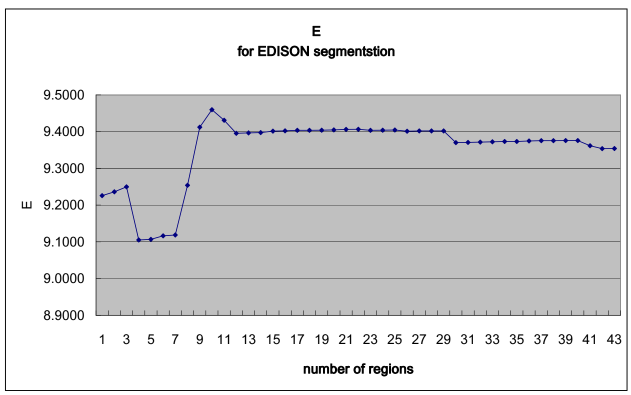

Figure 3. A closer look at E.

In Figure 4, we study E by showing how H_ℓ and H_r change as the number of regions grow, and study Q by showing how Q_1 and Q_2 change as the number of regions grow.

From Figure 4, we see that Q_1 and Q_2 do not complement each other well. Initially, Q_2 has a negligible impact on the value of Q. Only when reaching a very large number of regions does Q_2 have any impact, and at that point Q_2 quickly swamps out Q_1. In contrast, the two components of E complement each other quite nicely and thus together can counteract the effects of over- and under-segmentation. While, an equal weighting of H_ℓ and H_r prefers segmentations with only 4-7 regions, because of the way that these two components interact, we could add a user specified weighting function that provides a trade-off desired between the number of regions and the details of the image preserved. For example, if one preferred a segmentation of the "sea" image that preserved much greater detail, then one could instead put more weight on H_r (by choosing a smaller value of w). This will be studied further in future research.

In our second experiment, we varied the texture feature extraction threshold in the hierarchical segmentation method and generated the set of five images shown in Figure 5. Here, each of the five images has ten regions and thus the differences between the evaluation methods will not be related to the number of regions in the segmentation. All of our human evaluators agreed that image 3 and image 5 are much worse than the others. Image 3 is under-segmented, and the forehead of the lady and part of the background are mistakenly clustered.

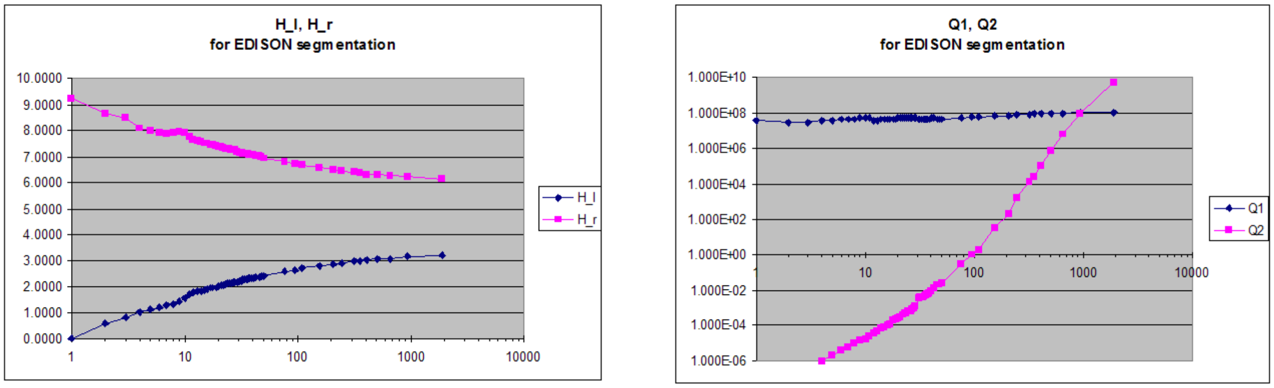

Figure 4. The interactions between H_ℓ and H_r and between Q_1, and Q_2. Recall that $E = H_\ell + H_r$ and $Q = Q_1 + Q_2$.

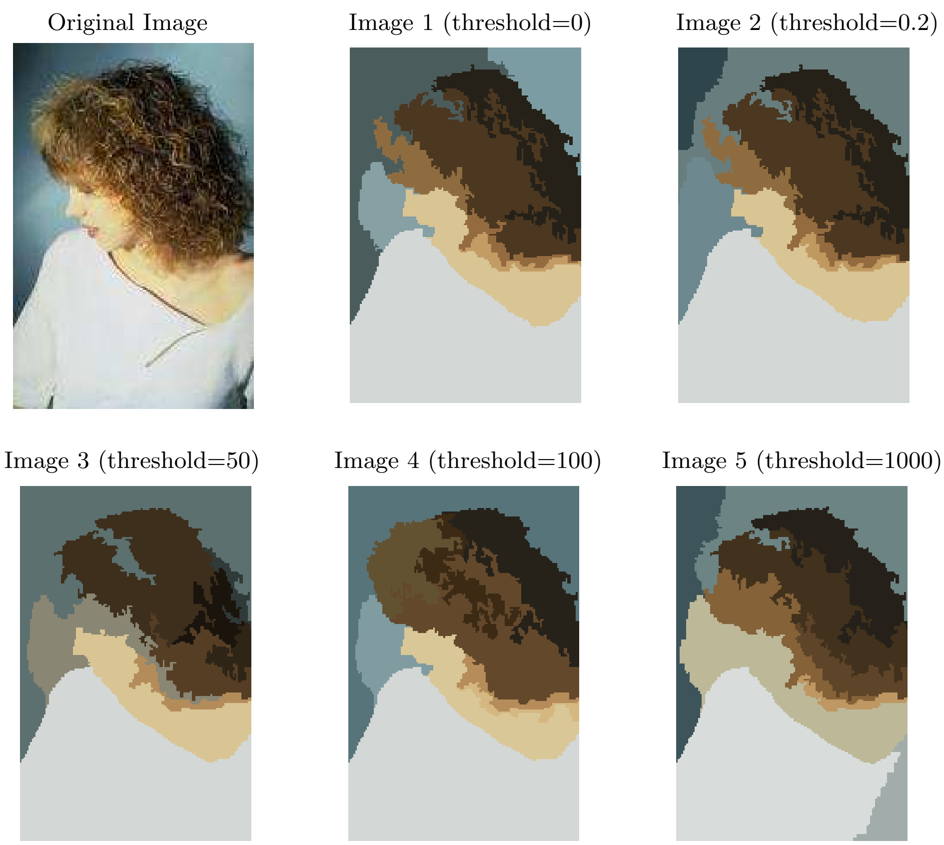

Figure 5. Segmented "lady" images where all images have 10 regions and the variation is created by the threshold parameter of the hierarchical segmentation algorithm.

Table 1. The values of each of the objective evaluaton functions for the five images shown in Figure 5 when all scaled to be in the range [0,1]. They are shown from the highest to lowest ranked image.

evaluation method	ranked 1st	ranked 2nd	ranked 3rd	ranked 4th	ranked 5th
F and F'	1 (0.0)	2 (0.601568)	3 (0.858991)	4 (0.994932)	5 (1.0)
H_w	5 (0.0)	1 (0.334319)	2 (0.419953)	3 (0.54232)	4 (1.0)
Q	1 (0.0)	5 (0.458804)	2 (0.497741)	4 (0.765217)	3 (1.0)
E	1 (0.0)	2 (0.379372)	4 (0.476625)	5 (0.928097)	3 (1.0)

Table 2. The pairwise comparison results of segmented "lady" images given by F, F', Q, H_w, and E.

	pair						
evaluation method	(1,3)	(1,5)	(2,3)	(2,5)	(4,3)	(4,5)	Total correct
F and F'	$\surd$	$\surd$	$\surd$	$\surd$	X	X	4
Q	$\surd$	$\surd$	$\surd$	X	$\surd$	X	4
H_w	$\surd$	X	$\surd$	X	X	X	2
E	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	6

Image 5 is also bad because of its under-segmentation. Image 1, 2 and 4 give relatively good segmentation results. The human evaluators were undecided about whether image 1 and 2 is better, but all agreed that image 4 is best since the lady's hair is properly separated from the background. In Table 1 we show the values for all of evaluation functions, normalized to the range $[0, 1]$ so that they can be more easily compared with each other. Observe that the normalization scaling by a multiplicative constant does not change the rankings given by any of these evaluation functions among a set of segmentations.

As a way to compare how each of F, F', Q, H_w, and E perform we consider the following six image pairs in which there seems to be clear consensus that the first image in the pair is preferable: {Image 1, Image 3}, {Image 1, Image 5}, {Image 2, Image 3}, {Image 2, Image 5}, {Image 4, Image 3}, {Image 4, Image 5}. These results are shown in Table 2 where we use "$\surd$" to indicate that the evaluation function correctly picked the first image of the pair as better, and "X" to denote that the evaluation function incorrectly selected the second image of the pair. The total number of times the evaluation function selected the correct image is shown in the last column. Only E correctly selected the first image in each pair. Notice that Q failed here because it favorably evaluated image 5 thus selecting it over both Image 2 and Image 4, which is obviously incorrect. The reason for this incorrect ranking is that Q uses $(1 + \log S_j)$, which strongly penalizes non-homogeneous regions. When the total number of regions is fixed, it gives a better score to the image whose regions have less variance. Once again, E has superior performance to the other evaluation functions.

One major reason to design an objective quantitative evaluation method is to compare the results from fundamentally different segmentation techniques. In our experiments, we applied the five quantitative evaluation methods to the results from the hierarchical method and from EDISON. Two examples are shown in Figure 6. Sea 1 and Rose 1 are created using the hierarchical segmentation method, and Sea 2 and Rose 2 are created using EDISON. Each of them has ten regions. Sea 1 is clearly better than Sea 2. Rose 2 is better than Rose 1(which is more obvious if printed as color figures), since it preserve the shape of the rose better. In our experiments, only H_w incorrectly evaluated Sea 2 as better, and only F and F' incorrectly evaluated Rose 1 as better. Clearly, more experiments are needed but this preliminary work (and more not shown) indicates that E does not appear to be biased towards segmentations from one segmentation algorithm versus another, and thus can be used effectively

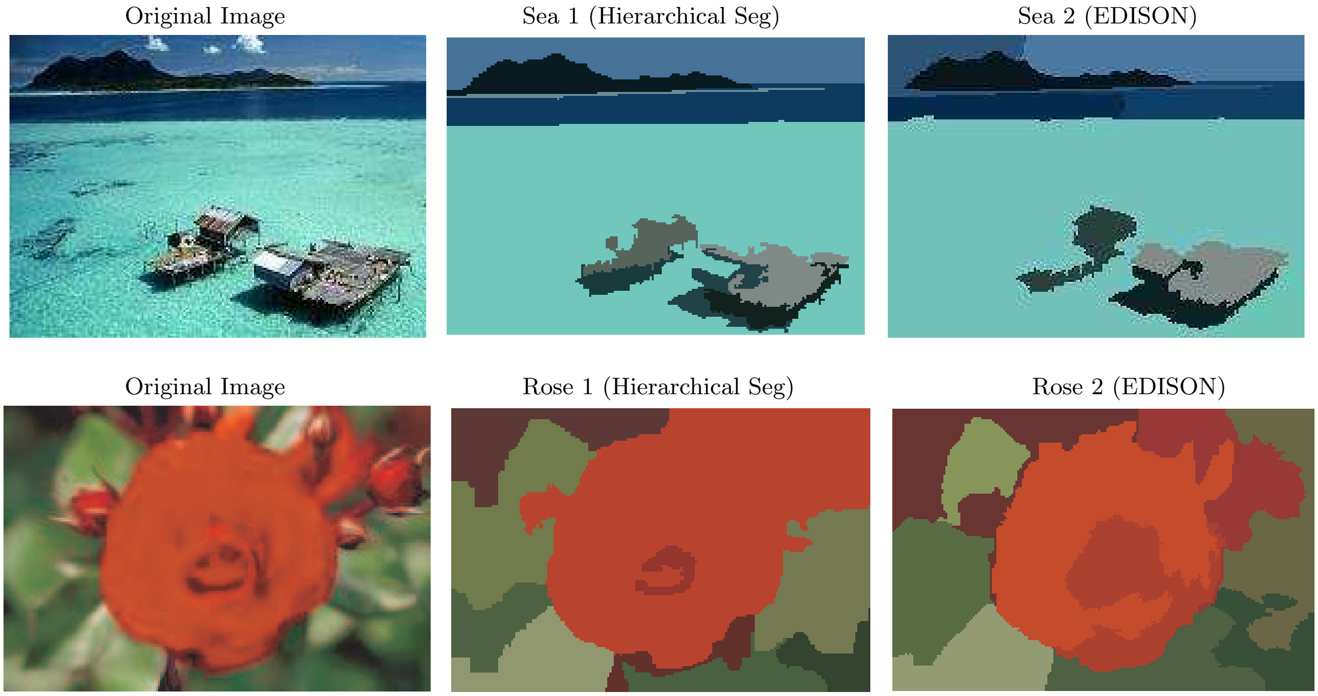

Figure 6. Segmented "sea" and "rose" images where the left one was created by the hierarchical segmentation algorithm and the right one was created by EDISON. All the segmented images have 10 regions.

in cross-algorithm evaluation.

5. CONCLUSIONS AND FUTURE WORK

We have demonstrated via our preliminary experiments that our entropy-based objective segmentation evaluation measure, E, improves upon previously defined evaluation measures in several ways. First, based on the experiments performed so far, E does a better job of selecting images that agreed with our human evaluators than the other methods. In particular, F and F' have a very strong bias towards images with very few regions and thus do not perform well. Q outperforms F and F' but still disagrees with our human evaluators more often than E did. Also, both Q and E need not be viewed as picking a best segmentation — both typically have a set of local minima which can be used to pick a set of preferred segmentations at different segmentation granularities. In our experiments, E was able to indicate local minima over a wider range of values than Q, which did not distinguish local minima when the number of regions was large.

One reason E performs so well is that its two components, one that measures the lack of uniformity in the regions of the segmentation and the other that measures the simplicity of the segmentation itself, counteract each other well. This finding is consistent with what one would expect when you view E according to Rissanen's minimum description length principle(MDL). By comparing E to H_w, we showed that using information theoretic entropy-based definitions for both of these components was important in obtaining our results.

There are many interesting directions for future research. Clearly, more extensive experiments using a wider variety of images and additional segmentation methods are needed. In the experiments we have performed so far, we have weighted two components equally. One future direction of work is to perform experiments using a weighted version of E as mentioned in Section 4. By weighting H_ℓ more strongly, emphasis will be placed on having a low *layout entropy*, creating a bias towards segmentations with fewer regions. Conversely, weighting H_r more strongly will place more emphasis on reducing the *expected region entropy*, causing a bias towards segmentations with more regions.

While the layout entropy provides a much better measure of the segmentation complexity than $\sqrt{N}$, we believe that it needs to be replaced by a better measure. Often E correctly selected the better of two images, but not always. In our experiments with the "lady" image (shown in Figure 5), the value of E favored image 1 and image 2 over image 4. Yet our human evaluators agree that image 4 is best because only in image 4 is the hair correctly separated from the background, while all other aspects of image 4 are similar to images 1 and 2.

One problem with the layout entropy is that it favors very few large regions and many small regions, as opposed to a uniform distribution of equal sized regions. Also since entropy is a global measure, it does not take into account local information or incorporate any measure about the shapes of the regions themselves. For example, a region that is a square of 10 by 10 pixels is treated the same as any region that includes 100 pixels in it(contiguous or non-contiguous). Using the probabilistic view for the definition of the layout entropy, we assume that each pixel is independently selected to be in region j with probability S_j/S_I. Under this model, two neighboring pixels are no more likely to be in the same region than two pixel that are not near each other. Yet a region or subset of a region will most likely be contiguous, and so the assumption that pixels in the image are independent and identically distributed (iid) with respect to the layout entropy is not accurate. To capture the fact that there is correlation between a pixel and its neighbors, a Markov assumption would be more appropriate. Such a solution for this problem would lead to using a conditional probability when defining the entropy.

Finally, in addition to segmentation evaluation, we also plan to utilize our evaluation function to control the segmentation process and dynamically choose a good number of regions, based on local minima in the segmentation evaluation measure.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of NSF grant IIS-0329241. We thank Rouhollah Rahmani and Zara Kahn for many useful discussions. We also thank Rouhollah for his comments on an earlier draft of this paper.

REFERENCES

1. I.E. Abdou, W.K. Pratt, "Quantitative design and evaluation of enhancement/thresholding edge detector," in *Proceeding of IEEE*, **67***(5)*, May 1979.
2. William A. Yansnoff and Jack K. Mui, "Error Measure for scene segmentation," *Pattern Recognition*, **9**, pp. 217–231, 1977.
3. P.K. Sahoo, S. Soltani, A.K.C. Wong, and Y.C. Chen, "Survey: A survey of thresholding techniques," *Computer vision, Graphics and Image Processing*, **41**, pp. 233–260 1988.
4. Martin D. Levine and Ahmed M. Nazif, "Dynamic Measurement of Computer Generated Image Segmentations," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **7***(2)*, pp. 155–164, 1985.
5. Nikhil R. Pal and Sankar K.Pal, "A Review on Image Segmentation Techniques,"*Pattern Recognition*, **26***(9)*, pp. 1277–1294 1993.
6. J.S. Weszka and A. Rosenfeld, "Threshold evaluation techniques," *IEEE Trans, System Man Cybernet*, **8***(8)*, pp. 622–629, 1978.
7. N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, **9***(1)*, pp. 62–66, 1979.
8. Jianqing Liu and Yee-Hong Yang, "Multi-resolution color Image segmentation," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **16***(7)*, pp. 689–700, 1994
9. M. Borsotti, P. Campadelli, and R. Schettini, "Quantitative evaluation of color image segmentation results," *Pattern Recognition Letters*, **19**, pp. 741–747, 1998.
10. J. Rissanen, "A universal prior for integers and estimation by minimum description length," *Annals of Statistics*, **11***(2)*,pp. 416–431.
11. `http://www.caip.rutgers.edu/riul/research/code/EDISON/`
12. Wei Yu, Jason Fritts, and Fangting Sun, "A hierarchical image segmentation algorithm," in *Proceedings of IEEE International Conference on Multimedia and Expo* (ICME'02), Lauzanne, Switzerland, August 2002.
13. Hui Zhang, "A fast texture feature extraction method in hierarchical image segmentation," CBIR-segmentation progress report, 2003.