An Entropy-based Objective Evaluation Method for Image Segmentation

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

Department of Computer Science and Engineering
Washington University in St. Louis

Introduction

- Research into better segmentation encounters two problems:
  1. cannot effectively compare different segmentations
     - different segmentation methods
     - different parameterizations of a method
  2. cannot determine whether one segmentation method is better than another for classes of images (e.g. natural images, medical images, etc.)

- Current segmentation evaluation methods are subjective or system-specific

- Objective segmentation evaluation methods are greatly needed

Current Evaluation Methods

(Subjective / System-Level Evaluation)

- Evaluate segmentation visually/qualitatively
  - Large amount of human involvements
  - Rather subjective

- Evaluate segmentation by its effectiveness on the subsequent processing steps
  - Only good for systems/applications employing segmentation
  - Indirect, not necessarily correlated in a positive way

Current Evaluation Methods

(Objective Evaluation)

- **Analytic methods**
  - judge segmentation method's effectiveness by conceptual elegance, mathematical sophistication or computational complexity, etc.

- **Supervised methods** (a.k.a. Empirical discrepancy methods)
  - comparing results to manually-segmented reference image
  - generating reference image is difficult, subjective, time-consuming, and inaccurate for most images, e.g. for natural images.

- **Unsupervised methods** (a.k.a. Empirical goodness methods)
  - evaluating results by measuring various image features, such as smoothness or continuity of the edge, etc.
  - often used with simple images; not designed for general applications.

Current Evaluation Methods

(Quantitative objective evaluation)

- Good segmentation evaluation methods must:
  - accurately judge the segmentation performance
  - have minimal human involvement
  - be independent of the contents and type of image
  - be independent of the segmentation method being evaluated

- Current quantitative objective evaluation methods:
  - *F*, proposed by Liu and Yang
  - *F'* and *Q*, proposed by Borsotti, Campadelli and Schettini
  - based on empirical analysis; little grounding in theory

Liu and Yang's F Function

$$F(I) = k\sqrt{N}\sum_{i=1}^{N}\frac{e_i^2}{\sqrt{S_i}}$$

- *F(I)* is biased towards small numbers of segments or large numbers of small segments
  - *F(I)* is *0* when the color error is zero for all segments, which occurs when each pixel is its own region
  - large numbers of regions in the segmented image is penalized only by the global measure $\sqrt{N}$.
  - segmentations that have regions with large areas are heavily penalized unless the region is very uniform in color

- Based on empirical analysis

Borsotti et. al 's F' Function

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

- Better than F when the segmentation has lots of regions consisting of small number of pixels
- Problems:
  - Reaches minimum value of zero when segmented such that each region is its own pixel
  - Heavily penalizes segmentations with a very large number of regions

Borsotti et. al 's Q Function

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

- Segmentations with large numbers of regions are not penalized as heavily
- Problems:
  - Very strong bias against regions with large area unless there is very little variation in color
  - Second term in the summation typically has a very small value as compared to the first term, so has negligible effect on evaluation results

Entropy-based Evaluation

- A good segmentation should maximize the uniformity of pixels within each region, and minimize the uniformity across the regions.
- Hence, entropy is a natural characteristic to be incorporated in evaluation function.

Entropy for region *j*:

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

Expected region entropy:

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

Entropy-based Evaluation

- Expected region entropy has a strong bias to over-segment, we must combine the expected region entropy with another term or factor that penalizes segmentations having a large numbers of regions.
- One approach would be to multiply the expected region entropy by $\sqrt{N}$ to penalize segmentations with a large numbers of regions.

Weighted disorder function: $H_w(I) = \sqrt{N} \sum_{j=1}^{N} \left(\frac{S_j}{S_I} \right) H\left(R_j\right) = \sqrt{N} H_r(I)$

Entropy-based Evaluation

- Regard segmentation as a modeling process.
- According to minimum description length (MDL) principle, if we balance the trade-off between the uniformity of the individual regions with the complexity of the segmentation, the minimum description length corresponds to the best segmentation.
- A measure of segmentation complexity:

Layout Entropy: $H_l(I) = -\sum_{i=1}^{N} p_i \log p_i = -\sum_{N} \left(\frac{S_i}{S_I}\right) \log \left(\frac{S_i}{S_I}\right)$

- Our Evaluation function, E, based on MDL:

$$E = H_l(I) + H_r(I) = -\sum_{N} \left(\frac{S_i}{S_I}\right) \log \left(\frac{S_i}{S_I}\right) + \sum_{N} \left(\frac{S_i}{S_I}\right) H(R_i)$$

Experimental Results

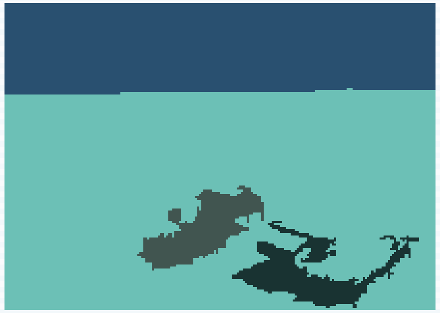
- Evaluation effectiveness when the number of regions in the segmentation varies
- Evaluation effectiveness when the number of regions is fixed
- Evaluation effectiveness when work on theoretically different segmentation methods

When number of regions varies

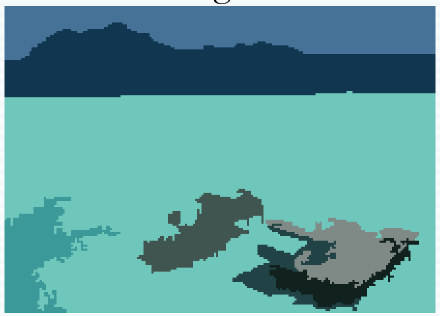
Original image

2 regions

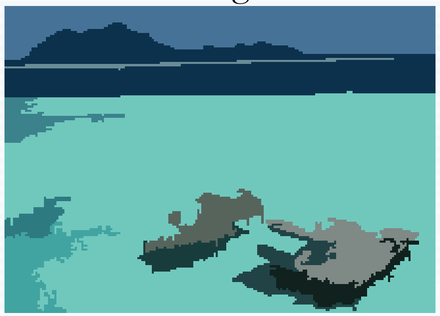
4 regions



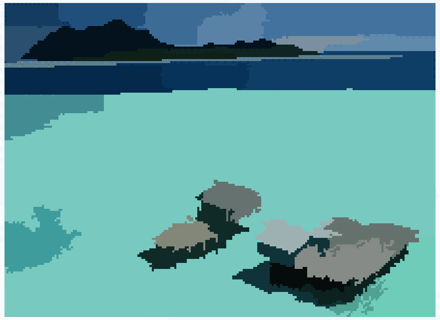
8 regions



12 regions

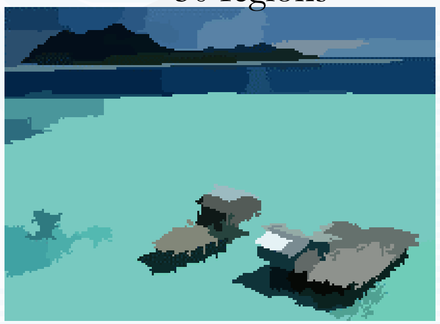


29 regions

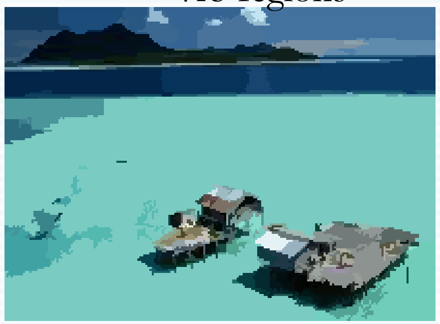


31 regions

50 regions



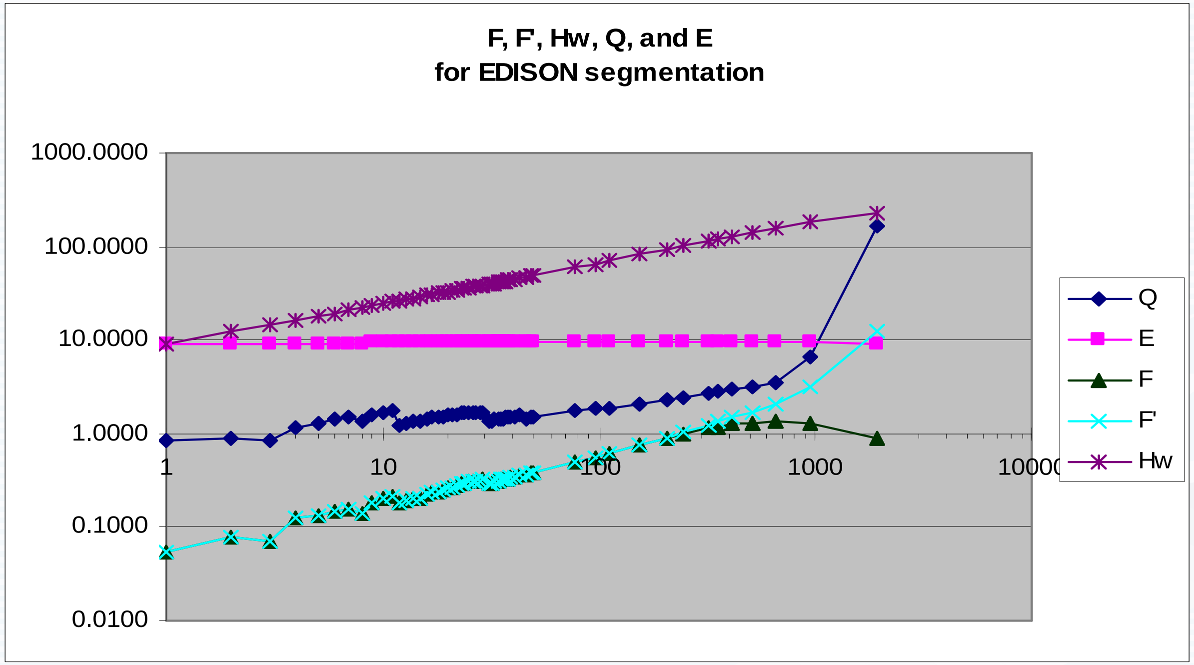
415 regions



Generated with EDISON

When number of regions varies

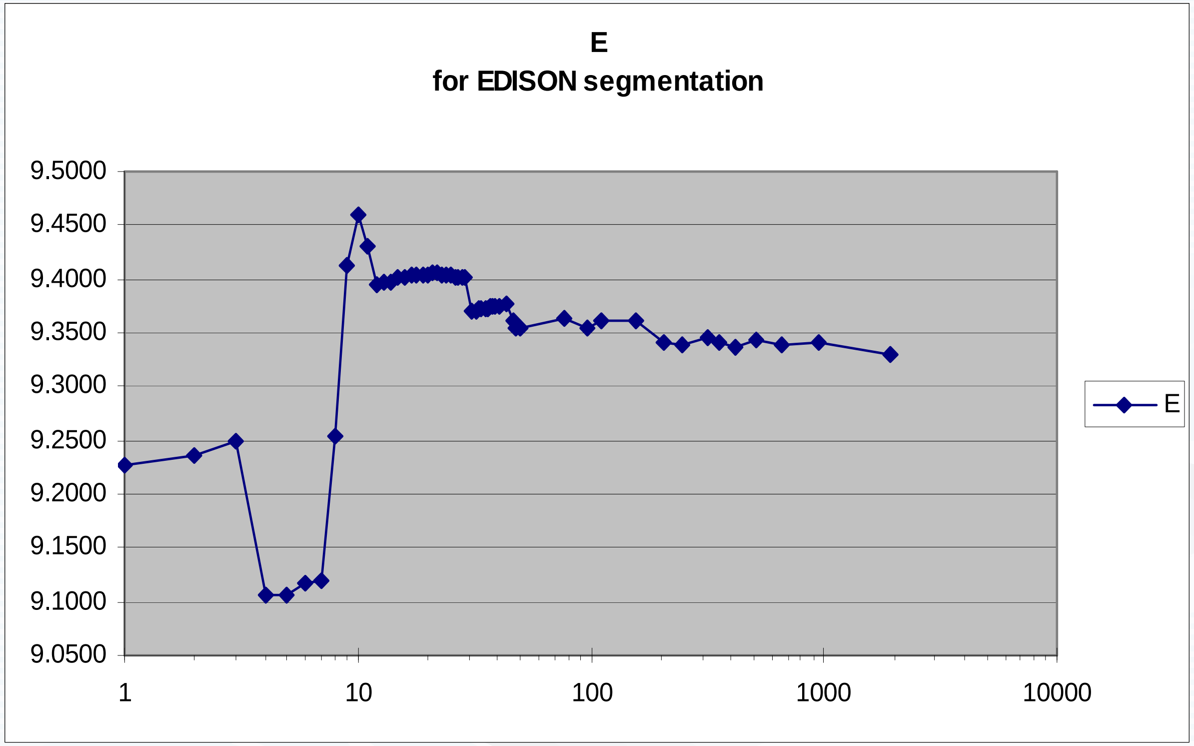
(Comparison of F, F', Hw, Q and E)



- *F* and *F'* generally increase and are almost identical until about 400 regions are in the segmentation
- *Q* also tends to increase as the number of regions grows, but it has clear local minima
- *F*, *F'* and *Q* have a strong bias towards the meaningless segmentation containing a single region

When number of regions varies

(A Closer Look at E)



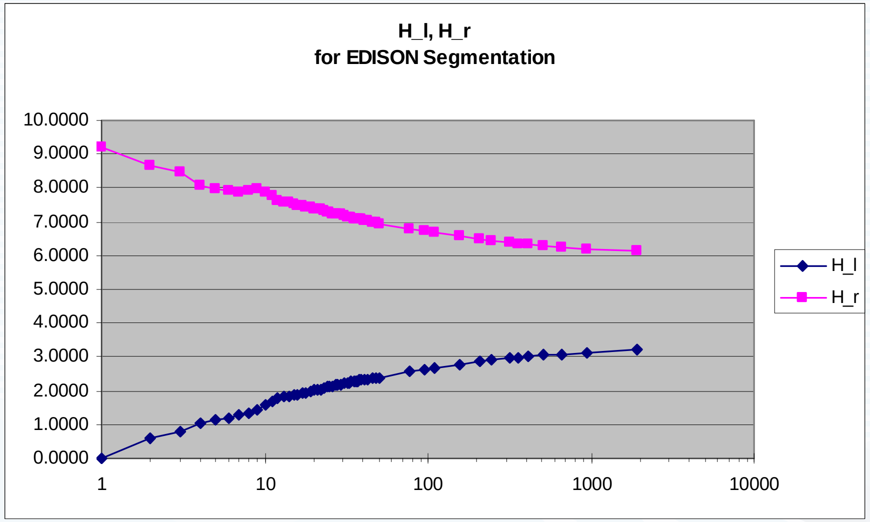
- E can be used to pick out the best segmentation over a wider range of desired granularity
- E does not have a strong bias towards the segmentation containing a single region

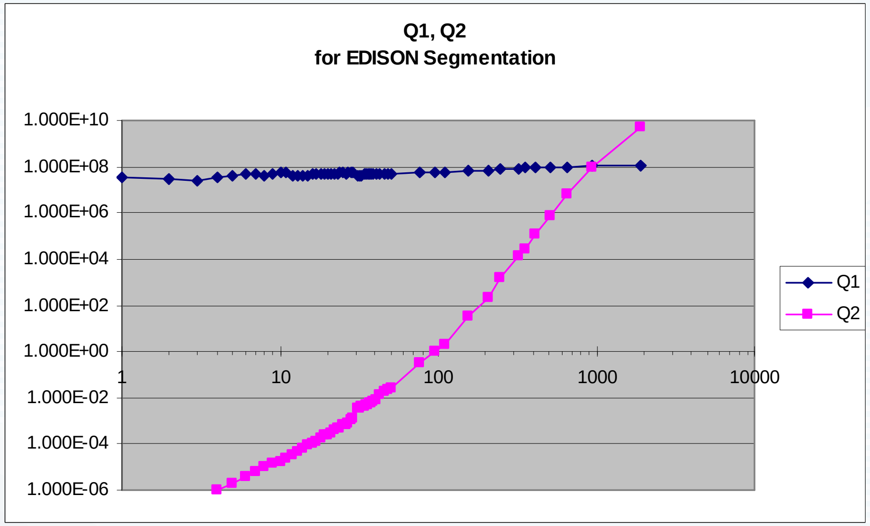
When number of regions varies

(The interaction of components in E and Q)

- *Q* can be broken into two terms: $Q_1 = \sqrt{N}\sum_{j=1}^{N}\left(e_j^2/(1+\log S_j)\right)$ $\quad Q_2 = \sqrt{N}\sum_{j=1}^{N}\left(C(S_j)/S_j\right)^2$

- The interactions between *Hl* and *Hr* and between *Q1* and *Q2* :





- *Q1* and *Q2* do not complement each other well. In contrast, the two components of *E* complement each other quite nicely and thus together can counteract the effects of over- and under-segmentation.

Experimental Results

- Evaluation effectiveness when the number of regions in the segmentation varies
- Evaluation effectiveness when the number of regions is fixed
- Evaluation effectiveness when work on theoretically different segmentation methods

When the number of regions is fixed

Original image

Image 1 (thresh.= 0)

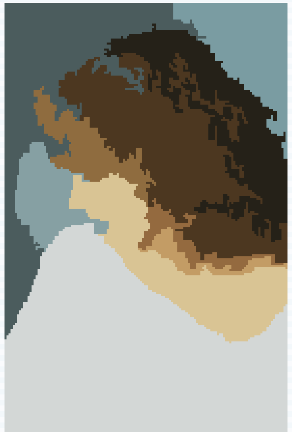


Image 2 (thresh.= 0.2)

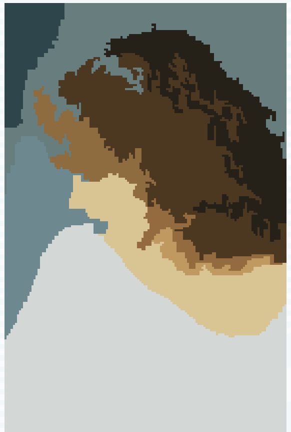


Image 3 (thresh.= 50)

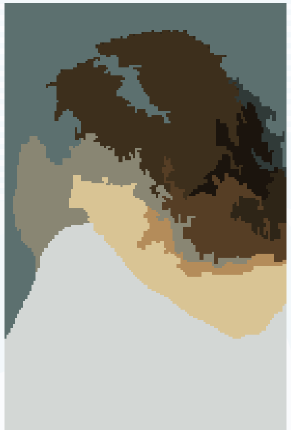


Image 4 (thresh.= 100)

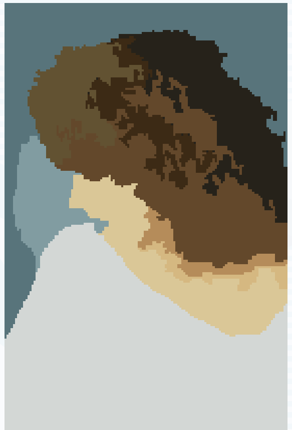
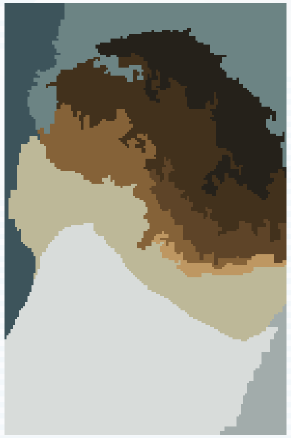


Image 5 (thresh.= 1000)



All five segmented images have 10 regions.

(Generated with hierarchical image segmentation methods with different fast feature extraction threshold)

When the number of regions is fixed

- Images are paired into 6 groups: {Image 1, Image 3}, {Image 1, Image 5}, {Image 2, Image 3}, {Image 2, Image 5}, {Image 4, Image 3}, {Image 4, Image 5}.
- Based on clear consensus of human evaluators, the first image in each pair is preferable.
- The pair-wise comparison results of segmented ``lady'' images given by *F*, *F'*, *Hw*, *Q* and *E*

Evaluation method	(1,3)	(1,5)	(2,3)	(2,5)	(4,3)	(4,5)	Total correct
F and F'	√	√	√	√	×	×	4
Q	√	√	√	×	√	×	4
Hw	√	×	√	×	×	×	2
E	√	√	√	√	√	√	6

Experimental Results

- Evaluation effectiveness when the number of regions in the segmentation varies
- Evaluation effectiveness when the number of regions is fixed
- Evaluation effectiveness when work on theoretically different segmentation methods

- In the above examples, all but *Hw* correctly evaluated Sea 1 is better. All but *F* and *F'* correctly evaluated Rose 2 is better.
- More experiments are needed, but preliminary results showed that *E* is not biased towards some segmentation, thus can be used in cross-segmentation evaluation.

Conclusion

- E does a better job of selecting images that agreed with our human subjective evaluation
  - 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
  - Q and E have a set of local minima which can be used to pick a set of preferred segmentations at different segmentation granularities
  - E was able to indicate local minima over a wider range of parameterizations than Q

Future Research

- More extensive experiments using a wider variety of images and additional segmentation methods are needed.
- Add user-specified weighting parameter to expected region entropy and the layout entropy, thus enable user to tailor the evaluation method to his/her particular subjective preferences.
- Use Markov assumption instead of iid (*independent and identically distributed*) assumption for layout entropy
- Improve layout measure to take into account local information and incorporate measure about the shapes of the regions, and to diminish the effects of region sizes.
- Utilize evaluation function to control the segmentation process and dynamically choose the optimal number of regions.