A METHODOLOGY AND METRIC FOR QUANTITATIVE ANALYSIS AND PARAMETER OPTIMIZATION OF UNSUPERVISED, MULTI-REGION IMAGE SEGMENTATION

William B. Kerr
Department of Computer Science
Trinity University
San Antonio, TX
wkerr@trinity.edu

Lucia Dettori
Department of CTI
DePaul University
Chicago, IL
ldettori@cti.depaul.edu

Lindsay Semler
Department of CTI
DePaul University
Chicago, IL
lsemler@cti.depaul.edu

ABSTRACT
While image segmentation makes up a vital step in the process of such tasks in the medical domain as tissue classification, content-based image retrieval, and computer-aided diagnosis, it remains an area of much debate regarding how one interprets the results of machine segmented regions. Many segmentation methods are still evaluated using a subjective human opinion of quality with a lack of quantitative analysis. Ideally, segmentation would be performed on an image with as little aid from a human user as possible, so solid quantitative analysis of results and optimization of user-defined parameters are a must. This paper proposes the use of a methodology based on eight individual performance measures. It then introduces a metric based on a statistical analysis of the overlap between machine segmented and corresponding ground truth images to evaluate and optimize algorithm parameters, and compare inter-algorithm performance for unsupervised segmentation algorithms.

KEY WORDS
Image Segmentation, Performance Evaluation

1 Introduction

Image segmentation is the process of dividing an image into unlabeled regions that have some sort of characteristic that separates them from other regions. Without segmentation, tasks like tissue classification in medical images, computer aided diagnosis in radiology, and content-based image retrieval would not be possible. Segmentation makes up the first step in much image processing work.

Quantitatively stating whether or not the algorithm performs well is difficult and research in this area does not present a way in which algorithms can be evaluated in the general sense. The majority of studies in the area of performance evaluation apply only to segmentations of a single region from its background [1, 5, 7, 13]. For many applications, like content-based image retrieval and computed tomography, it would be more practical to separate a image into multiple regions. First, there may be multiple regions of interest for which individual segmentations would decrease efficiency and increase dependency on human intervention, and second, isolating the background into one homogeneous region may prove daunting. Originally, this research was designed to aid segmentation of computed tomography in medical imaging.

There are two major methods for evaluating segmentation performance [12]. One compares machine segmented results to a ground truth segmentation of the same image [1, 5, 8, 9], while the other uses similarity measures to analyze the machine segmented regions [3, 4, 10, 12, 11]. In general, the latter is less reliable because whatever heuristic one devises to evaluate a segmentation after it is done can be implemented into the algorithm itself to be used during runtime. In other words, why use an evaluation criterion after segmentation has finished, when you could implement it as part of the segmentation itself? In the case of the ground truth, the evaluating criterion is unknown to the algorithm, leaving less potential for bias.

The following methodology builds on region categories that Hoover et al. proposed in their paper on range image segmentation [6] with some slight modifications. Hoover only evaluated pixels on the surfaces of objects that were imaged in true space, whereas we are concerned with the whole image. Chang et al. also used a variation of this method in [2] to evaluate the performance of several segmentation algorithms, however they stopped at categorization, whereas this paper proposes several performance measures and an overall metric based on those categories.

2 Methodology

Our performance evaluation is divided into 3 steps. First, the segmented regions are divided into 5 mutually exclusive categories based on those proposed by Hoover et al. Second, eight performance measures are calculated based on the relation of region categories to the image as a whole, as well as to what degree they fit their classification. Finally, the measures are combined into a metric to produce an overall "Goodness" rating.

2.1 Region Categories

The first step in evaluating performance is to calculate an overlap matrix between the regions from the machine segmentation (MS) and the regions from the ground truth (GT). Here, a region is merely the set of pixels assigned to that region. The matrix consists of GT regions along one axis and MS regions along the other. Each element of the matrix represents the number of pixels in the overlap between the row representing one region and the column representing the other. Naturally, GT regions cannot overlap with GT regions and MS regions cannot overlap with MS regions. Regions are then labeled as one of 5 types: Correctly Detected, Over Segmented, Under Segmented, Missed, or Noise. Each and every region in both GT and MS are categorized. Again, these region catergories originate from [6].

For each category, let T be a threshold representing the quality standard of evaluation where $0.5 \leq T \leq 1.0$. This range, along with the stipulation that correctly detected regions cannot be recategorized, guarantees that a region will never belong to more than one category at a time.

Correct Detection (CD)

Let R_{GT} be a set of pixels in a region of GT, and R_{MS} be a set of pixels in a region of MS.

R_{GT} and R_{MS} are said to be correctly detected if they satisfy two conditions:

1.
$$\frac{|R_{GT} \cap R_{MS}|}{|R_{MS}|} \geq T \tag{1}$$

2.
$$\frac{|R_{GT} \cap R_{MS}|}{|R_{GT}|} \geq T \tag{2}$$

This means that R_{GT} makes up at least $T\%$ of R_{MS} , and R_{MS} makes up at least $T\%$ of R_{GT}. Correct Detection must be categorized first to avoid a special case where a region could belong to two categories at once.

Over Segmentation (OS)

Let R_{GT} be a set of pixels in a region of GT, and $\{R_{MS1}, ..., R_{MSn}\}$ be a set of regions in MS, where $n > 1$.

R_{GT} and each element of $\{R_{MS1}, ..., R_{MSn}\}$ are said to be over segmented if they satisfy two conditions:

1.
$$\frac{|R_{GT} \cap R_{MSi}|}{|R_{MSi}|} \geq T \tag{3}$$

for each i=1,...,n.

2.
$$\frac{|R_{GT} \cap (\bigcup_{i=1}^{n} R_{MSi})|}{|R_{GT}|} \geq T \tag{4}$$

This means that R_{GT} makes up at least $T\%$ of each individual element in $\{R_{MS1}, ..., R_{MSn}\}$, and the union of all elements in $\{R_{MS1}, ..., R_{MSn}\}$ makes up at least $T\%$ of R_{GT}. Any region in GT and group of regions in MS that satisfy those conditions are considered to be over segmented *unless* one of the regions has already been categorized as correctly detected. A region cannot be both an instance of correct detection and over segmentation.

Under Segmentation (US)

Let R_{MS} be a set of pixels in a region of MS, and $\{R_{GT1}, ..., R_{GTn}\}$ be a set of regions in GT, where $n > 1$.

R_{MS} and each element of $\{R_{GT1}, ..., R_{GTn}\}$ are said to be over segmented if they satisfy two conditions:

1.
$$\frac{|R_{MS} \cap R_{GTi}|}{|R_{GTi}|} \geq T \tag{5}$$

for each i=1,...,n.

2.
$$\frac{|R_{MS} \cap (\bigcup_{i=1}^{n} R_{GTi})|}{|R_{MS}|} \geq T \tag{6}$$

This means that R_{MS} makes up at least $T\%$ of each individual element in $\{R_{GT1}, ..., R_{GTn}\}$, and the union of all elements in $\{R_{GT1}, ..., R_{GTn}\}$ makes up at least $T\%$ of R_{MS}. Any region in MS and group of regions in GT that satisfy those conditions are considered to be under segmented *unless* one of the regions has already been categorized as correctly detected. A region cannot be both an instance of correct detection and under segmentation.

Missed

A missed region is a region in GT that does not participate in any instance of correct detection, over segmentation, or under segmentation.

Noise

A noise region is a region in MS that does not participate in any instance of correct detection, over segmentation, or under segmentation.

2.2 Performance Measures

We have derived eight distinct performance measures from the region categories above, seven of which are combined to make up the overall "Goodness" metric. The eight measures are correct detection index, correct detection precision, over segmentation index, over segmentation fragmentation, under segmentation index, under segmentation inclusion, garbage index, and garbage quality.

Correct Detection Index

The correct detection index (CDindex) measures the percentage of correctly detected pixels in the image overall. This includes all pixels in the overlap between correctly detected regions. Let $(R_{GT1}, R_{MS1}), ..., (R_{GTn}, R_{MSn})$

be all the pairs of correctly detected regions and IMG be the set of all pixels in the image.

$$CDindex = \frac{|\bigcup_{i=1}^{n}(R_{GTi} \cap R_{MSi})|}{|IMG|} \quad (7)$$

Correct Detection Precision

Correct detection precision (CDprecision) measures the percentage of overlap between correctly detected regions overall. In other words, the more closely aligned the two regions, the better the precision rating. This measure does not get calculated into the final Goodness Metric, because it measures pixels that are categorized as correctly detected, but excluded from the correct detection index. The penalty for a low correct detection precision is the same penalty the exclusion represents. Let $(R_{GT1}, R_{MS1}), ..., (R_{GTn}, R_{MSn})$ be all the pairs of correctly detected regions and IMG be the set of all pixels in the image.

$$CDprecision = \frac{|\bigcup_{i=1}^{n}(R_{GTi} \cap R_{MSi})|}{|\bigcup_{i=1}^{n}(R_{GTi} \cup R_{MSi})|} \quad (8)$$

Over Segmentation Index

The over segmentation index (OSindex) measures the percentage of over segmented pixels in the image. Let $(R_{GT1}, \{R_{MS1,1}, ..., R_{MS1,m_1}\}), ..., (R_{GTn}, \{R_{MSn,1}, ..., R_{MSn,m_n}\})$ be the groups of over segmented regions where $\{R_{MS1,1}, ..., R_{MS1,m_1}\}$ are the MS fragments that overlap R_{GT1}. Let IMG be the set of all pixels in the image.

$$OSindex = \frac{\left|\bigcup_{i=1}^{n}(R_{GTi} \cap (\bigcup_{j=1}^{m_i} R_{MSi,j}))\right|}{|IMG|} \quad (9)$$

Over Segmentation Fragmentation

Over segmentation fragmentation (OSfrag) measures the percent makeup of all over segmented pixels by a single MS fragment. The fewer the MS regions that overlap a GT region, the better and higher OSfrag will be. Note that the highest this value can be is 0.5, as there can be only as low as two MS regions in a GT region categorized as over segmented. Let $(R_{GT1}, \{R_{MS1,1}, ..., R_{MS1,m_1}\}), ..., (R_{GTn}, \{R_{MSn,1}, ..., R_{MSn,m_n}\})$ be the groups of over segmented regions where $\{R_{MS1,1}, ..., R_{MS1,m_1}\}$ are the MS fragments that overlap R_{GT1}.

$$OSfrag = \sum_{i=1}^{n} \frac{\left|R_{GTi} \cap (\bigcup_{j=1}^{m_i} R_{MSi,j})\right|}{m_i * \left|\bigcup_{k=1}^{n}(R_{GTk} \cap (\bigcup_{j=1}^{m_k} R_{MSi,j}))\right|} \quad (10)$$

Under Segmentation Index

The under segmentation index (USindex) measures the percentage of under segmented pixels in the image. Let $(R_{MS1}, \{R_{GT1,1}, ..., R_{GT1,m_1}\}), ..., (R_{MSn}, \{R_{GTn,1}, ..., R_{GTn,m_n}\})$ be the groups of over segmented regions where $R_{GT1,1}, ..., R_{GT1,m_1}$ are the GT fragments that overlap R_{MS1}. Let IMG be the set of all pixels in the image.

$$USindex = \frac{\left|\bigcup_{i=1}^{n}(R_{MSi} \cap (\bigcup_{j=1}^{m_i} R_{GTi,j}))\right|}{|IMG|} \quad (11)$$

Under Segmentation Inclusion

Under segmentation inclusion (USinclusion) measures the percent makeup of all under segmented pixels by a single GT fragment. The fewer the GT regions that an MS region overlaps, the better and higher USinclusion will be. Note that the highest this value can be is 0.5, as there can be only as low as two GT regions overlapped by an MS region categorized as under segmented. Let $(R_{MS1}, \{R_{GT1,1}, ..., R_{GT1,m_1}\}), ..., (R_{MSn}, \{R_{GTn,1}, ..., R_{GTn,m_n}\})$ be the groups of over segmented regions where $\{R_{GT1,1}, ..., R_{GT1,m_1}\}$ are the GT fragments that overlap R_{MS1}.

$$USinclusion = \sum_{i=1}^{n} \frac{\left|R_{MSi} \cap (\bigcup_{j=1}^{m_i} R_{GTi,j})\right|}{m_i * \left|\bigcup_{k=1}^{n}(R_{MSk} \cap (\bigcup_{j=1}^{m_k} R_{GTi,j}))\right|} \quad (12)$$

Garbage Index

Here, "garbage" refers to both noise and missed regions. The garbage index (Gindex) measures the percentage of overlapping missed and noise pixels in the image. Let $\{R_{MS1}, ..., R_{MSn}\}$ be all noise regions in MS and $\{R_{GT1}, ..., R_{GTn}\}$ be all missed regions in GT. Let IMG be the set of all pixels in the image.

$$Gindex = \frac{\left|(\bigcup_{i=1}^{n} R_{MSi}) \cap (\bigcup_{j=1}^{m} R_{GTj})\right|}{|IMG|} \quad (13)$$

Garbage Quality

Garbage quality (Gquality) is a weight for the garbage index that depends on how "messy" the overlap is between missed and noise regions. Let $\{R_{MS1}, ..., R_{MSn}\}$ be all noise regions in MS and $\{R_{GT1}, ..., R_{GTn}\}$ be all missed regions in GT. Let K be the average number of noise regions overlapping a single missed region and C be the average number of missed regions overlapping a single noise region.

$$Gquality = \frac{\left|(\bigcup_{i=1}^{n} R_{MSi}) \cap (\bigcup_{j=1}^{m} R_{GTj})\right|}{(C+K) * \left|(\bigcup_{i=1}^{n} R_{MSi}) \cup (\bigcup_{j=1}^{m} R_{GTj})\right|} \quad (14)$$

2.3 Overall Goodness Metric

Seven of the eight performance measures are combined to form a metric that produces a single rating. Depending on the application, it may not be suitable to evaluate at such a high level of abstraction, in which case the performance measures can be analyzed individually.

The evaluation metric begins with the percentage of pixels that were correctly detected. All other pixels in the image were essentially segmented incorrectly, however these pixels vary in terms of the degree by which they differ from the ground truth. For example, areas of over segmentation or under segmentation would be preferable to areas of noise and missed, and even more preferable to pixels that do not reside in the overlap between groups (e.g., pixels that are in a correctly detected region, but do not reside in the overlap between the two regions in the correctly detected pair). These differences in quality are captured by weighting the contribution of pixels not correctly detected. The metric consists of three parts: "Good", "Bad", and a weight, υ.

$$Good = CDindex \tag{15}$$

$$Bad = 1 - CDindex \tag{16}$$

$$\upsilon = 1 - (OS_\upsilon + US_\upsilon + G_\upsilon) \tag{17}$$

where

$$OS_\upsilon = (2 * OSfrag) * \frac{OSindex}{(1 - CDindex)} \tag{18}$$

$$US_\upsilon = (2 * USinclusion) * \frac{USindex}{(1 - CDindex)} \tag{19}$$

$$G_\upsilon = Gquality * \frac{Gindex}{(1 - CDindex)} \tag{20}$$

The overall "Goodness" metric is defined as follows:

$$Goodness = Good - \upsilon * Bad \tag{21}$$

where $0 \leq \upsilon \leq 1$

Goodness falls into a range between -1.0 and 1.0 with a local ceiling at CDindex and a floor at 2*CDindex-1 (see fig. 1).

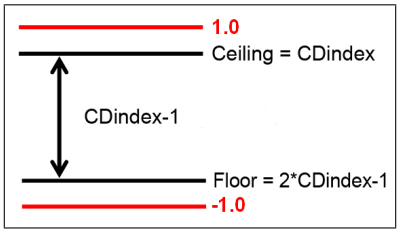

Figure 1: Metric behavior

3 Examples

To illustrate the effectiveness and reliability of the metric, several hand-drawn segmentations were compared to ground truth. The simplicity of hand-drawn examples allows for basic understanding of how the metric would evaluate real machine segmentations. Four example ground truth segmentations were created. Each example has four corresponding segmentations (S-A, S-B, S-C, S-D) representing machine segmentations. For both the ground truth and the example segmentations, each gray level represents a different region. The value for the gray level is irrelevant; only the partitioned space has meaning. For example, a region in the ground truth may be represented by gray level 3 and correspond with a region in the segmentation represented by gray level 1.

Ideally, the metric produces values for the segmentations that correspond with human interpretation of quality. In other words, the quantitative ranking of the metric should correspond to a human qualitative ranking.

The metric is computed for each of the segmentations against their corresponding ground truth image with a Goodness threshold, T, of 0.7 and 0.8. Again, T is the threshold by which regions are categorized.

Example 1

The first example consists of four equally sized regions, each occupying a quarter of the image space (fig. 2). S-A ranks highest followed by S-C at both T=0.7 and T=0.8. S-C ranks second because only 25% of the image has error with a very clean over segmentation of two MS regions in the upper-left. At a first glance, it may appear that S-B should rank higher than S-D, but S-B combined the top-left and bottom-right GT regions into a single MS region resulting in 50% under segmentation. At T=0.7, S-D still contains many regions that classify as correctly detected. If T increases to a value of 0.8, representing a higher quality standard, the "leaking tendrils" of S-D have a more adverse effect on its rating (tables 1 and 2).

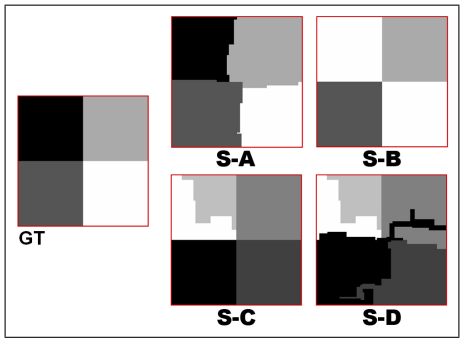

Figure 2: Example 1 ground truth with machine segmentations

Example 2

The second example consists of two equally sized regions on the left and four equally sized regions on the right, making a total of six regions (fig. 3). S-A represents a perfect segmentation and ranks the highest of the four. S-B ranks next, followed by S-C, both examples of under segmentation. S-C ranks lower than S-B because a greater percentage of the image is under segmented. As expected, S-D ranks last because of its noise.

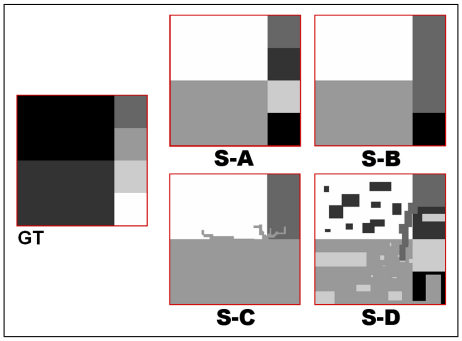

Figure 3: Example 2 ground truth with machine segmentations

Example 3

The third example resembles example 1, except now the upper-left and bottom-right corners are combined into a single ground truth region (fig. 4). S-C ranks first since it almost exactly matches the ground truth segmentation. S-D serves as another example of how the evaluation can change depending on what standard you hold the quality of segmentation to. At T=0.7 S-D ranks second because the MS regions consist of 70% correctly detected pixels. However, at T=0.8 the rating of S-D drops drastically, reducing it to the worst of the four MS segmentations. S-A follows S-D at T=0.7, as 50% of the image has been over segmented by separating the top-left and bottom-right corners into two MS regions. S-B ranks the lowest because it over segments one diagonal and under segments the other. At T=0.8, S-B ranks higher than S-D because the error in S-D consists mostly of noise and missed regions, whereas the error in S-B is primarily over segmentation and under segmentation.

Example 4

The fourth example considers a case with eight equally sized GT regions. S-C ranks first as a perfect match to the ground truth (fig. 5). S-B ranks second despite its poor segmentation on the left, because the four MS regions on the right are all correctly detected. Notice the change in its goodness between T=0.7 and T=0.8. Predominantly under segmented, S-A ranks third with some over segmentation in the second GT region from the left. Lastly, S-D ranks lowest with an abundance of noise and oddly placed MS regions.

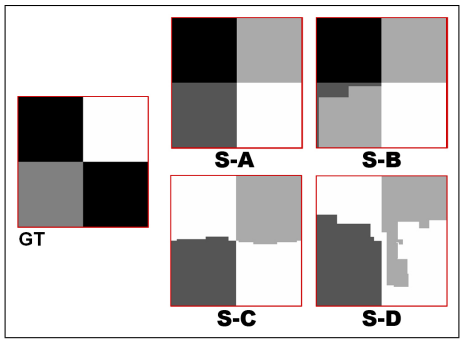

Figure 4: Example 3 ground truth with machine segmentations

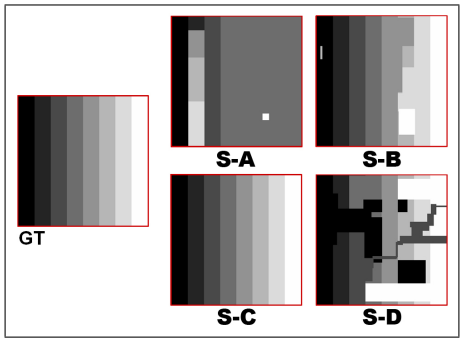

Figure 5: Example 4 ground truth with machine segmentations

T=0.7	S-A	S-B	S-C	S-D
Example 1	0.8771	0.5000	0.7500	0.5611
Example 2	1.0000	0.7500	0.5971	0.2138
Example 3	0.5000	-0.0409	0.9675	0.6605
Example 4	-0.1885	0.7469	1.0000	-0.4922

Table 1: Goodness values at T=0.7 for each of the sixteen segmentations

4 Conclusions and Future Work

These preliminary studies show the Goodness metric to analyze the performance of machine segmentation in a coherent and rational way. Future work into this aspect of its effectiveness might call for a study of human rankings versus goodness ratings. At this stage of the research, however, the metric seems to match human perception (considering T values) fairly well.

The metric provides only a general evaluation of performance, whereas the eight individual performance mea-

T=0.8	S-A	S-B	S-C	S-D
Example 1	0.8771	0.5000	0.7500	0.1828
Example 2	1.0000	0.7500	0.1402	-0.7692
Example 3	0.5000	-0.0409	0.9675	-0.7500
Example 4	-0.1885	0.3445	1.0000	-0.6486

Table 2: Goodness values at T=0.8 for each of the sixteen segmentations

sures describe various aspects of the behavior of the segmentation in more detail. These measures can be combined to provide other results such as the quality with regards to region shape, rather than correct detection priority. For example, this may call for OS and US pixels to be weighted as heavily as CD pixels.

Research is currently underway in terms of using this metric as a means to optimize parameters within a segmentation algorithm. We are currently researching wavelet-based texture segmentation algorithms for medical computed tomography images. A method for creating training images with ground truth based on pure human tissue textures is currently underway. The goodness metric will allow for a quantitative justification of parameter values such as the type of discrete wavelet transformation and the distance threshold in feature space used to assign pixels to regions.

Having a threshold for the Goodness Metric itself may seem to solve a thresholding problem in the algorithm only by creating one for evaluation, however the T value for Goodness differs from an algorithm threshold in that it merely represents the standard of quality that one holds the algorithm to. Parameters and thresholds in an algorithm range anywhere from distances in feature space to the type of transform one uses to extract pixel data. Ideally, T would be set at 1.0, but it is unlikely that any algorithm would perform well at this level of expectation. An algorithm yielding Goodness 0.35 at a higher Goodness T value would be better than an algorithm yielding Goodness 0.35 at a lower Goodness T value. Future work could include quantitatively ascertaining the quality change across Goodness T levels yielding the same Goodness value, standardizing the metric and removing the need for a threshold.

References

[1] Vikram Chalana and Yongmin Kim. A methodology for evaluation of boundary detection algorithms on medical images. *IEEE Transactions on Medical Imaging*, 16(5), October 1997.

[2] Kyong I. Chang, Kevin W. Bowyer, and Munish Sivagurunath. Evaluation of texture segmentation algorithms. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, page 1294, 1999.

[3] C. Erdem, B. Sankur, and A. Tekalp. Performance measures for video object segmentation and tracking. *IEEE Transactions on Image Processing*, 13(7), July 2004.

[4] Mark Everingham, Henk Muller, and Barry Thomas. Evaluating image segmentation algorithms using the pareto front. In *Proceedings of the 7th European Conference on Computer Vision (ECCV2002)*, June 2004.

[5] N. Fatemi-Ghomi and et al. Performance evaluation of texture segmentation algorithms based on wavelets. In *Workshop on Performance Characterization of Vision Algorithms*, April 1996.

[6] Adam Hoover and et al. An experimental comparison of range image segmentation algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(7), July 1996.

[7] Oian Huang and Byron Dom. Quantitative methods of evaluating image segmentation. In *Proceedings of the IEEE International Conference on Image Processing*, 1995.

[8] David R. Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms. Technical report, Berkeley, CA, USA, 2002.

[9] V. Mezaris, I. Kompatsiaris, and M.P. Strintzis. Still image objective segmentation evaluation using ground truth. In *Proceedings of the Fifth COST 276 Workshop on Information and Knowledge Management for Integrated Media Communication*, pages 9–14, October 2003.

[10] M. Sezgin and B. Sankur. Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic Imaging*, 13(1):146–165, January 2004.

[11] Hui Zhang, Jason E. Fritts, and Sally A. Goldman. An entropy-based objective evaluation method for image segmentation. In *Proceedings of the SPIE*, volume 5307, pages 38–49, 2003.

[12] Hui Zhang, Jason E. Fritts, and Sally A. Goldman. A co-evaluation framework for improving segmentation evaluation. In *SPIE Defense and Security Symposium - Signal Processing, Sensor Fusion, and Target Recognition XIV*, March 2005.

[13] Y.J. Zhang. Evaluation and comparison of different segmentation algorithms. *NH Elsevier Pattern Recognition Letters*, 18:963–974, 1997.