Performance Evaluation of Image Segmentation

Fernando C. Monteiro[1,2] and Aurlio C. Campilho[1,3]

[1] INEB - Instituto de Engenharia Biomdica
[2] Escola Superior de Tecnologia e de Gesto de Bragana
Campus de Santa Apolnia, Apartado 134, 5301-857 Bragana, Portugal
monteiro@ipb.pt
[3] FEUP - Faculdade de Engenharia, Universidade do Porto
Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
campilho@fe.up.pt

Abstract. In spite of significant advances in image segmentation techniques, evaluation of these methods thus far has been largely subjective. Typically, the effectiveness of a new algorithm is demonstrated only by the presentation of a few segmented images that are evaluated by some method, or it is otherwise left to subjective evaluation by the reader. We propose a new approach for evaluation of segmentation that takes into account not only the accuracy of the boundary localization of the created segments but also the under-segmentation and over-segmentation effects, regardless to the number of regions in each partition. In addition, it takes into account the way humans perceive visual information. This new metric can be applied both to automatically provide a ranking among different segmentation algorithms and to find an optimal set of input parameters of a given algorithm.

1 Introduction

In a conventional sense, image segmentation is the partitioning of an image into regions, where parts within a region are similar according to some uniformity predicate, and dissimilar between neighbouring regions. Due to its importance, many segmentation algorithms have been proposed, and a number of evaluation criteria have also been proposed. In spite of this, very few comparative studies on the methods used for evaluation have been published [14].

Typically, researchers show their segmentation results on a few images and point out why the results 'look good'. We never know from such studies if the results are good or typical examples. Since none of the proposed segmentation algorithms are generally applicable to all images, and different algorithms are not equally suitable for a particular application, there needs to be a way of comparing them, so that the better ones can be selected. The majority of studies proposing and comparing segmentation methods evaluate the results only with one evaluation method. However, results vary significantly between different evaluators, because each evaluator may have distinct standards for measuring the quality of the segmentation.

Only a few evaluation methods actually explore the segments obtained from the segmentation process. Most measures are best suited to evaluate edge detection [12], working directly on the binary image of the regions' boundaries [3]. Although we can always treat segmentation as a boundary map, the problem is in the simplified use of the edge map, as simply counting the misclassified pixels, on an edge/non-edge basis. Pixels on different sides of an edge are different in the sense that they belong to different regions - that is why it may be more reasonable to use the segmentation partition itself.

Evaluation of image segmentation differs considerably from the binary foreground/background segmentation evaluation problem examined in [3,13], in that the correctness of the two class boundary localization is not the only quantity to be measured. This derives from the presence of an arbitrary number of regions in both the reference segmentation and the segmentation to be evaluated.

An evaluation metric is desired to take into account the following effects:

- **Over-segmentation.** A region of the reference is represented by two or more regions in the examined segmentation.
- **Under-segmentation.** Two or more regions of the reference are represented by a single region in the examined segmentation.
- **Inaccurate boundary localization.** Ground truth is usually produced by humans that segment at different granularities.
- **Different number of segments.** We need to be able to compare two segmentations when they have different numbers of segments.

Under-segmentation is considered to be as a much more serious problem as it is easier to recover true segments through a merging process after over-segmentation rather than trying to split an heterogeneous region. One desirable property of a good evaluation measure is to accommodate refinement only in regions that human segmenters could find ambiguous and to penalize differences in refinements elsewhere. In addition to being tolerant to refinement, any evaluation measure should also be robust to noise along region boundaries and tolerant to different number of segments in each partition.

This work will focus on discrepancy evaluation methods, that consist in comparing the results of a segmentation algorithm with a reference and measuring the differences or discrepancy between them. We introduce a new approach for segmentation evaluation that takes into account, using a single metric, not only the accuracy of the boundary localization of the created segments but also the under-segmentation and over-segmentation effects, regardless to the number of regions in each partition. In addition, it takes into account the way humans perceive visual information, given different weights to false positive and false negative pixels. In order to test the accuracy of the proposed measure we compared it with a set of key methods used for the evaluation of image segmentation using real and synthetic images.

The remainder of this paper is organized as follows: in Section 2, previous segmentation evaluation methods are presented. In Sections 3 and 4, we present region-based and boundary-based evaluation methods currently in literature.

The proposed metric for evaluation is presented in Section 5. In Section 6, experimental evaluation is analysed and discussed, and, finally, conclusions are drawn in Section 7.

2 Previous Work

A review on evaluation of image segmentation is presented by Zhang in [14], who classifies the methods into three categories: *analytical*, where performance is judged not on the output of the segmentation method but on the basis of their properties, principles, complexity, requirements and so forth, without reference to a concrete implementation of the algorithm or test data. While in domains such as edge detection this may be useful, in general the lack of a general theory of image segmentation limits these methods; *empirical goodness methods*, which compute some manner of 'goodness' metric such as uniformity within regions [3], contrast between regions [4], shape of segmented regions [12]; and finally, *empirical discrepancy methods*, which evaluate segmentation algorithms by comparing the segmented image against a manually-segmented reference image, which is often referred to as ground truth, and computes error measures.

As stated by Zhang [14], the major difficulty in applying analytical methods is the lack of general theory for image segmentation. The analytical methods may only be useful for simple algorithms or straightforward segmentation problems, where the researchers have to be confident in the models on which these algorithms are based.

Empirical goodness methods have the advantage that they do not require manually segmented images to be supplied as ground truth data. The great disadvantage is that the goodness metrics are at best heuristics, and may exhibit strong bias towards a particular algorithm. For example the intra-region grey-level uniformity metric will assume that a well-segmented image region should have low variance of grey-level. This will cause that any segmentation algorithm which forms regions of uniform texture to be evaluated poorly. Although these evaluation methods can be very useful in some applications, their results do not necessarily coincide with the human perception of the goodness of segmentation. For this reason, when a reference image is available or can be generated, empirical discrepancy methods are preferred.

Empirical discrepancy methods, which compare segmentation output with ground truth segmentation of the test data and quantify the levels of agreement and/or disagreement, have the benefit that the direct comparison between a segmented image and a reference image is believed to provide a finer resolution of evaluation, and as such, they are the most commonly used methods of segmentation evaluation.

Zhang [14] has proposed a discrepancy evaluation method based on misclassified pixels. Yasnoff *et al.* [13], in one of the earliest attempts, have shown that measuring the discrepancy based only on the number of misclassified pixels does not consider the pixel position error. Their solution is based on the number of misclassified pixels and their distance to the nearest correctly segmented pixels.

They only applied it to foreground/background segmentation. Other discrepancy measures calculate the distances between wrong segmented pixels and the nearest correctly segmented pixels [8], thus introducing a spatial component to the measure, or are based on differences between feature values measured from regions of the correctly segmented and output images. Huang and Dom [3] introduced the concept of distance distribution signatures.

3 Region-Based Evaluation

The region-based scheme evaluates the segmentation accuracy in the number of regions, the locations and the sizes. A region-based evaluation between two segmented images can be defined as the total amount of differences between corresponding regions.

3.1 Hamming Distance

Huang and Dom [3] introduced the concept of directional Hamming distance between two segmentations, denoted by $D_H(S_1 \Rightarrow S_2)$. Let S and R be two segmentations. They began by establishing the correpondence between each region of S with a region of R such that $s_i \cap r_j$ is maximized. The directional Hamming distance from S to R is defined as:

$$D_H(S \Rightarrow R) = \sum_{r_i \in R} \sum_{s_k \neq s_j, s_k \cap r_i \neq 0} |r_i \cap s_k| \ , \tag{1}$$

where $|\cdot|$ denote the size of a set. Therefore, $D_H(S \Rightarrow R)$ is the total area under the intersections between all $r_i \in R$ and their non-maximal intersected regions from S. A region-based evaluation measure based on normalized Hamming distance is defined as

$$p = 1 - \frac{D_H(S \Rightarrow R) + D_H(R \Rightarrow S)}{2 \times |S|} \ , \tag{2}$$

where $|S|$ is the image size and $p \in [0, 1]$. The smaller the degree of mismatch, the closer the p is to one.

3.2 Local Consistency Error

To compensate for the difference in granularity while comparing segmentations, many measures allow label refinement uniformly through the image. D. Martin's thesis [6] proposed an error measure to quantify the consistency between image segmentations of differing granularities - *Local Consistency Error* (LCE) that allows labelling refinement between segmentation and ground truth.

$$LCE(S, R, p_i) = \frac{1}{N} \sum_i \min \{E(S, R, p_i), E(R, S, p_i)\} \ , \tag{3}$$

where $E(S, R, p)$ measures the degree to which two segmentations agree at pixel p, and N is the size of region where pixel p belongs.

Note that the LCE is an error measure, with a score 0 meaning no error and a score 1 meaning maximum error. Since LCE is tolerant to refinement, it is only meaningful if the two segmentations have similar number of segments. As observed by Martin in [6], there are two segmentations that give zero error for LCE - one pixel per segment, and one segment for the whole image.

3.3 Bidirectional Consistency Error

To overcome the problem of degenerate segmentations, Martin adapted the LCE formula and proposed a measure that penalizes dissimilarity between segmentations proportional to the degree of region overlap. If we replace the pixelwise minimum with a maximum, we get a measure that does not tolerate refinement at all. The *Bidirectional Consistency Error* (BCE) is defined as:

$$BCE\left(S,R,p_i\right)=\frac{1}{N}\sum_i \max\left\{E\left(S,R,p_i\right),E\left(R,S,p_i\right)\right\}\ . \tag{4}$$

3.4 Partition Distance Measure

Cardoso and Corte-Real [1] proposed a new discrepancy measure - *partition distance* (d_{sym}) defined as: "given two partitions P and Q of S, the partition distance is the minimum number of elements that must be deleted from S, so that the two induced partitions (P and Q restricted to the remaining elements) are identical". $d_{sym}\left(Q,P\right)=0$ means that no points need to be removed from S to make the partitions equal, i.e., when $Q=P$.

4 Boundary-Based Evaluation

Boundary-based approach evaluates segmentation in terms of both localization and shape accuracy of extracted regions boundaries.

4.1 Distance Distribution Signatures

Huang and Dom in [3] presented a boundary performance evaluation scheme based on the distance between distribution signatures that represent boundary points of two segmentation masks.

Let B_S represent the boundary point set derived from the segmentation and B_R the boundary ground truth. A distance distribution signature from the set B_S to the set B_R of boundary points, denoted $D_B\left(B_s,B_R\right)$, is a discrete function whose distribution characterizes the discrepancy, measure in distance, from B_S to B_R. Define the distance from x in set B_S to B_R as the minimum absolute distance from all the points in B_R, $d\left(x,B_R\right)=\min\left\{d_E\left(x,y\right)\right\},\forall y\in B_R$, where d_E denotes the Euclidean distance between points x and y.

The discrepancy between B_S and B_R is described by the shape of the signature, which is commonly measured by its mean and standard deviation. As a rule, $D_B\left(B_s,B_R\right)$ with a near-zero mean and a small standard deviation indicates

high between segmentation masks. Since these measures are not normalized, we cannot determine which segmentation is the most desirable.

We introduce a modification to the distance distribution signature of Huang and Dom, in order to normalize the result between 0 and 1. Doing $d(x, B_R) = \min\{d_E(x, y), c\}$, where the c value sets an upper limit for the error, the two boundary distances could be combined in a framework similar to the one presented in Eq. (2):

$$b = 1 - \frac{D_B(B_S, B_R) + D_B(B_R, B_S)}{c \times (|R| + |S|)} \quad , \tag{5}$$

where $|R|$ and $|S|$ are the number of boundary points in reference mask and segmented mask, respectively.

4.2 Precision-Recall Measures

Martin in his thesis [6], propose the use of *precision* and *recall* values to characterize the agreement between the oriented boundary edge elements (termed *edgels*) of region boundaries of two segmentations. Given two segmentations, S and R, where S is the result of segmentation and R is the ground truth, precision is proportional to the fraction of edgels from S that matches with the ground truth R, and recall is proportional to the fraction of edgels from R for which a suitable match was found in S. Precision measure is defined as follows:

$$Precision = \frac{Matched(S, R)}{|S|} \qquad Recall = \frac{Matched(R, S)}{|R|} \quad , \tag{6}$$

where $|S|$ and $|R|$ are the total amount of boundary pixels. In probabilistic terms, precision is the probability that the result is valid, and recall is the probability that the ground truth data was detected.

A low recall value is typically the result of under-segmentation and indicates failure to capture salient image structure. Precision is low when there is significant over-segmentation, or when a large number of boundary pixels have greater localization errors than some threshold $\delta_{\max}$.

Precision and recall measures have been used in the information retrieval systems for a long time [10]. However, the interpretation of the precision and recall for evaluation of segmentation are a little different from the evaluation of retrieval systems. In retrieval, the aim is to get a high precision for all values of recall. However in image segmentation, the aim is to get both high precision and high recall. The two statistics may be distilled into a single figure of merit:

$$F = \frac{PR}{\alpha R + (1 - \alpha) P} \quad , \tag{7}$$

where α determines the relative importance of each term. Following [6], α is selected as 0.5, expressing no preference for either.

The main advantage of using precision and recall for the evaluation of segmentation results is that we can compare not only the segmentations produced

by different algorithms, but also the results produced by the same algorithm using different input parameters. However, since these measures are not tolerant to refinement, it is possible for two segmentations that are perfect mutual refinements of each other to have very low precision and recall scores.

4.3 Earth Mover's Distance

The concept of using the Earth Mover's Distance (EMD) to measure perceptual similarity between images was first explored by Peleg *et al.* in [9] for the purpose of measuring distance between two grey-scale images. More recently EMD has been used for image retrieval [11].

EMD evaluates dissimilarity between two distributions or *signatures* in some feature space where a distance measure between single features is given. The EMD between two distributions is given by the minimal sum of costs incurred to move all the individual points between the signatures.

Let $P = \{(p_1, w_{p_1}), ..., (p_m, w_{p_m})\}$ be the first signature with m pixels, where p_i is the pixel representative and w_{p_i} is the weight of the pixel; the second signature with n pixels is represented by $Q = \{(q_1, w_{q_1}), ..., (q_n, w_{q_n})\}$; and $D = [d_{ij}]$ the distance matrix where d_{ij} is the distance between two contour points' image coordinates p_i and q_j. The flow f_{ij} is the amount of weight moved from p_i to q_j. The EMD is defined as the work normalized by the total flow f_{ij}, that minimizes the overall cost:

$$EMD\,(P,Q) = \frac{\sum_i \sum_j f_{ij} d_{ij}}{\sum_i \sum_j f_{ij}} \quad , \tag{8}$$

As pointed by Rubner *et al* [11], if two weighted point sets have unequal total weights, EMD is not a true metric. It is desirable for robust matching to allow point sets with varying total weights and cardinalities. In order to embed two sets of contour features with different total weights, we simulate equal weights by adding the appropriate number of points, to the lower weight set, with a penalty of maximal distance. Since normalizing signatures, with the same total weight do not affect their EMD, we made $\sum_{i,j} f_{ij} = 1$. Equation (8) becomes,

$$EMD\,(P,Q) = \sum_i \sum_j f_{ij} d_{ij} \quad , \tag{9}$$

subject to the following constraints: $f_{ij} \geq 0$, $\sum_j f_{ij} = w_{p_i}$ and $\sum_i f_{ij} = w_{q_j}$.

As a measure of distance for the EMD ground distance we use

$$d_{ij} = 1 - e^{-\frac{\|p_i - q_j\|}{\alpha}} \quad , \tag{10}$$

where $\|p_i - q_j\|$ is the Euclidean distance between p_i and q_j and α is used in order to accept some deformation resulted from manual segmentation of ground truth. The exponential map limits the effect of large distances, which otherwise dominate the result.

5 New Discrepancy Measure

In the context of image segmentation, the reference mask is generally produced by humans. There is an agreement that interpretations of images by human subjects differ in granularity of label assignments, but they are consistent if refinements of segments are admissible [6]. One desirable property of a good evaluation measure is to accommodate refinement only in regions that human segmenters could find ambiguous and to penalize differences in refinements elsewhere. In addition to being tolerant to refinement, any evaluation measure should also be robust to noise along region boundaries and tolerant to different number of segments in each partition. The section introduces a new evaluation measure that addresses the above concerns.

For the purpose of evaluating image segmentation results, a correspondence between the examined segmentation mask, S, and the reference mask, R, has initially be established, indicating which region of S better represents each reference region. This is performed by associating each region r_i of mask R with a different region s_j of mask S on the basis of region overlapping, i.e. s_j is chosen so that $r_i \cap s_j$ is maximized. The set of pixels assigned to s_j but not belonging to r_i are false positives, F_p, that can be expressed as $F_p = s_j \cap \overline{r}_i$, where $\overline{r}_i$ denotes the complement of r_i. The pixels belonging to r_i but not assigned to s_j are false negatives, F_n, and can be expressed as $F_n = \overline{s}_j \cap r_i$.

The minimum required overlap between r_i and s_j is 50% of the reference region. Pixels belonging to regions where this ratio is not achieved are considered as false pixels. These measure quantify the errors due to under and over segmentation. Clearly, more visually significant regions that were missed in the segmented mask are assigned a significantly higher error.

The normalized sum of false detections is an objective discrepancy measure that quantifies the deviation of the results of segmentation from the ground truth and can be expressed as:

$$\varepsilon_F = \frac{F_p + F_n}{2N} \ , \tag{11}$$

where N is the set of all pixels in the image. The value of ε_F is proportional to the total amount of errors and indicates the accuracy of region boundaries localization. The quality of the segmentation is inversely proportional to the amount of deviation between the two masks.

In applications where the final evaluator of quality is the human being, it is fundamental to consider human perception to deal with the fact that different kind of errors are not visually significant to the same degree. To accommodate human perception, the different error contributions are weighted according to their visual relevance. Gelasca *el al.* [2], present a psychophysical experiment to assess the different perceptual importance of errors. They conclude that a false positive pixel contributes differently to the quality than a false negative. False negatives are more significant, and the larger the distance, the larger the error.

We use the weighted functions w_p and w_n to deal with that fact. They are normalized by the diagonal distance, D. Let d_p be the distance of a false positive pixel from the boundary of the reference region, and d_n be the distance of a false

negative pixel. The weight function for each false pixel is defined by Eq. (12) and represented in Fig. 1.

$$w_p = \frac{\alpha_p \log(1 + d_p)}{D} \qquad w_n = \frac{\alpha_n d_p}{D} \quad . \tag{12}$$

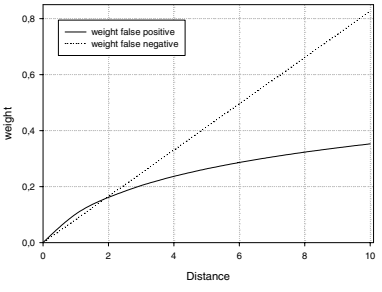

Fig. 1. Weight functions for false negative and false positive pixels

The weights for false negative pixels increase linearly and are larger than those for false positive pixels at the same distance from border. As we move away from the border of an object, missing parts are more important than added background, e.g., in medical imaging, it may be enough that the segmented region overlaps with the true region, so the tumour can be located. But if there are missing parts of the tumour the segmentation results will be poor.

To obtain a measure between $[0, 1]$, we normalize the total amount of weight by the image size. The discrepancy measure of weighted distance, ε_w, becomes:

$$\varepsilon_w = \frac{1}{N} \left(\sum_{f_n} w_n + \sum_{f_p} w_p \right) \quad , \tag{13}$$

where f_n and f_p represent the false pixels. We define a measure of similarity as $s_w = 1 - \varepsilon_w$. The value of $s_w = 1$ indicates a perfect match between the segmentation and the reference mask.

6 Analysis on Evaluation Methods

To achieve comparative results about different evaluation methods, two strategies can be followed: the first one consists in applying the evaluation methods to segmented images obtained from different segmentation approaches. The second one consists in simulating results of segmentation processes. To exempt the influence of segmentation algorithms, the latter has been adopted and a set of images obtained from manual segmentation available in [5] was used (Fig. 2).

A good evaluation measure has to give large similarity values for images (b) to (g) and has to strongly penalize other images. Figure 3.a) shows the comparative

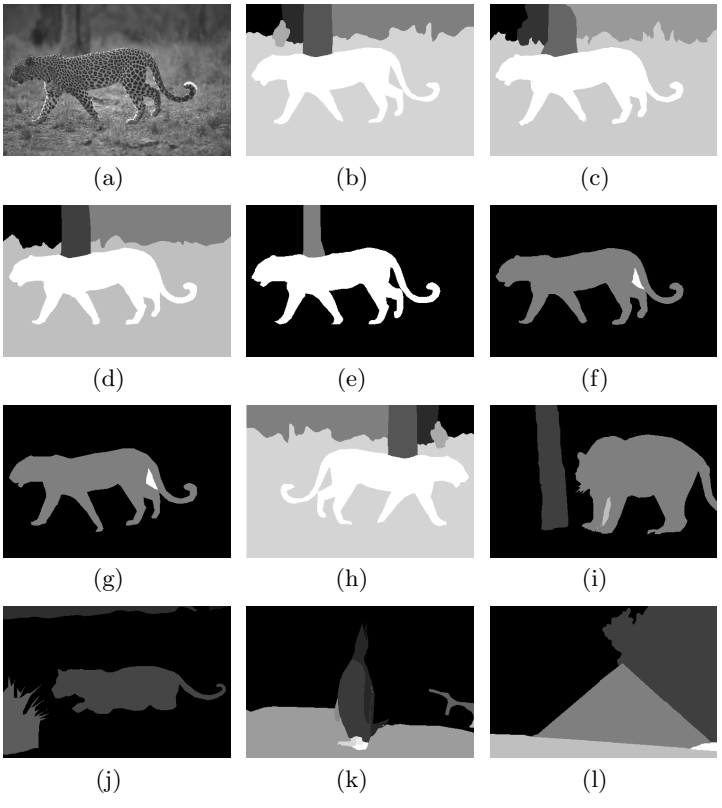

Fig. 2. The image and its ground truth are shown in (a) and (b), respectively. From (c) to (g) we have different segmentations of image (a). (h) is the reflected image of ground truth. Images (i) to (l) are segmentations of other images.

study between the proposed method and the methods presented in Section 3, expressed in terms of region-based evaluation.

Due to its tolerance to refinement, LCE gives low error (high similarity) scores, even when the segmentation is very different from the ground truth. Measure p has a similar behaviour. BCE and d_{sym} give good results for images ((h)-(l)), however, since they are not tolerant to refinement, the results are poor for other images. Note that the proposed measure is tolerant to refinement and at the same time strongly penalizes images ((h)-(l)).

Results of boundary-based evaluation on the same set of images of Fig. 2 are reported in Fig. 3.b). On comparing the results of the boundary-based measures, it is made evident that they are well correlated. EMD tolerates well some amount of deformations that normally happens in the manual segmentation process. However, when the number of pixels in ground truth differs a lot from the number of pixels in the segmented image, EMD gives poor results. Despite its success, the EMD method still needs to be refined to address the limitation in the complexity

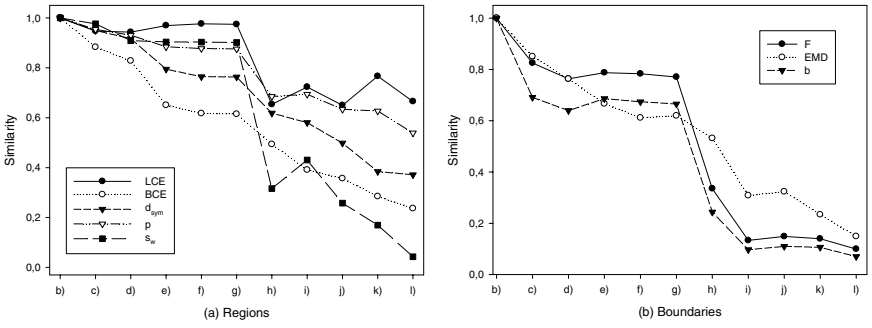

Fig. 3. Evaluation of segmentation, in terms of similarity, from Fig. 2

Fig. 4. Synthetically generated set of segmentations, where (a) is the reference

Table 1. Numerical evaluation of segmentations from Fig. 4

images	LCE	BCE	d_{sym}	p	s_w
(b)	0.99380	0.98088	0.99349	0.99349	0.99741
(c)	0.99380	0.98088	0.99349	0.99349	0.99612
(d)	0.99380	0.98088	0.99349	0.99349	0.99159

of algorithm that require to be further reduced. The b-measure gives results similar with F-measure, but is even more intolerant to refinement.

Table 1 presents the evaluation results obtained from a set of trivial synthetically generated segmentations presented in Fig. 4, where we make constant the number of false detections in each segmentation.

Since LCE, BCE, d_{sym} and p, are just proportional to the total amount of false detections, different position of those pixels do not affect the similarity. This makes those methods unreliable for applications where the results will be presented to humans. Note that s_w produces results that agree with the visual relevance of errors.

7 Conclusion

In this paper, we introduce a new approach for segmentation evaluation that takes into account, using a single metric, not only the accuracy of the bound-

ary localization of the created segments but also the under-segmentation and over-segmentation effects according to the ambiguity of the regions, regardless to the number of segments in each partition. We experimentally demonstrated the efficiency of this new measure against well known methods. This new metric can be applied both to automatically provide a ranking among different segmentation algorithms and to find an optimal set of input parameters of a given algorithm according with their results of segmentation evaluation. Moreover, this paper introduces a modification to the distance distribution signature of Huang and Dom, *b*-measure; it applies the concept of Earth Mover's Distance to the evaluation of image segmentation.

References

1. Cardoso, J.S., Corte-Real, L., Toward a generic evaluation of image segmentation, *IEEE Transactions on Image Processing*, 14(11):1773-1782, (2005).
2. Gelasca, E.D., Ebrahimi, T., Farias, M.C.Q., Carli, M., Mitra, S.K., Towards perceptually driven segmentation evaluation metrics, in *Proc. IEEE Computer Vision and Pattern Recognition Workshop*, Vol. 4, page 52, (2004).
3. Huang, Q., Dom, Byron, Quantitative methods of evaluating image segmentation, in *Proc. IEEE International Conference on Image Processing*, Vol. III:53-56, (1995).
4. Levine, M.D.,Nazif, A.M., Dynamic measurement of computer generated image segmentations, *Trans. Pattern Analysis and Machine Intelligence* 7(2):155-164, (1985).
5. Martin, D., Fowlkes, C., The Berkeley segmentation database and benchmark, online at *http://www.cs.berkeley.edu/projects/vision/grouping/segbench/*.
6. Martin, D., *An empirical approach to grouping and segmentation*, PhD dissertation, University of California, Berkeley (2002).
7. Mezaris, V., Kompatsiaris, I., Strintzis, M.G., Still image objective segmentation evaluation using ground truth, in *Proc. of 5th COST 276 Workshop*, pp: 9-14, (2003).
8. Odet, C., Belaroussi, B., Cattin, H.B., Scalable discrepancy measures for segmentation evaluation, in *Proc. Intern. Conf. on Image Processing*, Vol. I:785-788, (2002).
9. Peleg, S., Werman, M., Rom, H., A unified approach to the change of resolution: Space and gray-level, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 11:739-742, (1989).
10. Raghavan, V., Bollmann, P., Jung, G., A critical investigation of recall and precision as measures of retrieval system performance, *ACM Transactions on Information Systems*, 7(3):205-229, (1989).
11. Rubner, Y., Tomasi, C., Guibas, L.J., The Earth Mover's Distance as a metric for image retrieval, *International Journal of Computer Vision*, 40(2):99-121, (2000).
12. Sahoo, P.K., Soltani, S., Wang, A.K.C., A survey of thresholding techniques, *Computer Vision, Graphics and Image Processing*, 41(2):233-260, (1988).
13. Yasnoff, W.A., Mui, J.K., Bacus, J.W., Error measures in scene segmentation, *Pattern Recognition*, 9(4):217-231, (1977).
14. Zhang, Y.J., A survey on evaluation methods for image segmentation, *Pattern Recognition*, 29(8):1335-1346, (1996).