Image and Vision Computing 27 (2009) 1223–1227

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Short communication

An evaluation metric for image segmentation of multiple objects ☆

Mark Polak, Hong Zhang *, Minghong Pi

Centre for Intelligent Mining Systems, University of Alberta, Edmonton, Alta., Canada T6G 2E8

ARTICLE INFO

Article history:
Received 7 August 2006
Received in revised form 14 July 2008
Accepted 18 September 2008

Keywords:
Image segmentation
Evaluation
Error measure

ABSTRACT

It is important to be able to evaluate the performance of image segmentation algorithms objectively. In this paper, we define a new error measure which quantifies the performance of an image segmentation algorithm for identifying multiple objects in an image. This error measure is based on object-by-object comparisons of a segmented image and a ground-truth (reference) image. It takes into account the size, shape, and position of each object. Compared to existing error measures, our proposed error measure works at the object level, and is sensitive to both over-segmentation and under-segmentation. Hence, it can serve as a useful tool for comparing image segmentation algorithms and for tuning the parameters of a segmentation algorithm.

1. Introduction

Image segmentation is one of the basic problems in image analysis. Although extensive efforts have been made to develop image segmentation algorithms, much less attention has been paid to evaluating the performance of image segmentation algorithms. In general, evaluation methods for image segmentation can be classified into analytical and empirical evaluation methods [2]. Empirical methods, in turn, can be classified into empirical goodness methods and empirical discrepancy methods.

The analytical methods of evaluation typically focus on analyzing the properties of a segmentation algorithm, such as its processing strategy, complexity, and efficiency. The evaluation is from a theoretical point of view and does not require the actual implementation of the algorithms. However, they work only for particular models or are concerned with certain desirable properties of the algorithm. The empirical goodness methods of evaluation use the original image and the resulting segmented image. Goodness can be expressed in terms of a statistical measure such as the uniformity within segmented regions [3], inter-region contrast [9], or region shape [10]. However, without a reference image, the goodness may not be objective.

The empirical discrepancy methods, to which our proposed metric belongs, explicitly calculate the error between the segmented image and a reference (ground-truth) image. The reference image is often obtained manually with the help of a human expert, and the segmented image is from a segmentation algorithm. Common error measures are the number of mis-segmented pixels [4,5], the position of mis-segmented pixels [6], the number of objects in the image [7,14], or the geometric features of segmented objects such as area, perimeter, or sphericity [8]. Almost all empirical methods are constructed by considering image segmentation as a process of pixel labeling, except for [14] which focuses on the number of objects exclusively with regard to the sizes. Consequently, they are not appropriate for object-level evaluation.

In contrast, Martin et al. [11] proposed an interesting empirical discrepancy measure for evaluating segmentation. It is an object-by-object error measure, and is very useful to quantify the consistency between segmentations manually performed by different people of the same image who view the image at different granularities or scales. Unfortunately, it is inappropriate for segmentation applications in which the details of the segmentation in terms of the exact object boundaries are important. As an example, when an over-segmented image is simply a refined version of an under-segmented image, the Martin error measure would consider the two to be consistent and therefore correct with respect to each other.

This lack of penalty for over or under-segmentation was recognized in [13], which proposed an error measure based on the concept of partition distance. Partition distance counts the number of pixels, normalized with respect to the image size, that must be removed from the interpretation, i.e., segmentation of an image until the induced segmentation agrees with the reference image. Consequently, similar to [11], partition distance considers all levels of over and under refinements to be equally incorrect. In addition, the calculation of partition distance does not weigh objects according to their sizes. However, both of the above properties are important in many applications.

In this paper, we propose a new empirical discrepancy error measure, called object-level consistency error (OCE), which quan-

☆ This research is supported in part by NSERC, iCORE, Syncrude Canada Ltd., Matrikon, and the University of Alberta.

* Corresponding author. Address: Department of Computer Science, University of Alberta, 221 Athabasca Hall, Alta., Canada T6G 2E8. Tel.: +1 780 492 7188.
E-mail addresses: mpolak@cs.ualberta.ca (M. Polak), zhang@cs.ualberta.ca (H. Zhang), minghong@cs.ualberta.ca (M. Pi).

0262-8856/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2008.09.008

tifies the similarity (or discrepancy) between a segmented image and the ground truth image at the object level. The key novelty of the error measure is that it takes into account the existence, size, position, and shape of each fragment and penalizes both over-segmentation and under-segmentation. At the same time, it retains the properties of being normalized ($0 \leqslant OCE(I_g, I_s) \leqslant 1$ and $OCE(I_g, I_s) = 0$ only if $I_g = I_s$), symmetric ($OCE(I_g, I_s) = OCE(I_s, I_g)$), and scale invariant ($OCE(I_g, I_s) = OCE(I_s^{scaled}, I_g^{scaled})$), where I_g and I_s are the segmented and the ground-truth images, and I_g^{scaled} and I_s^{scaled} are their scaled versions, respectively. We argue that our proposed OCE can effectively serve as an objective evaluation of image segmentation algorithms and for tuning the parameters of a segmentation algorithm.

The rest of the paper is organized as follows. Section 2 describes the error measure proposed by Martin et al. Section 3 describes our proposed performance metric. The experimental results are provided in Section 4, followed by the conclusions in Section 5.

2. Martin error measure

Martin et al. [11] proposed an interesting error measure, which takes two images I_g and I_s as input, and produces a real-valued output in the range of $[0,1]$ where 0 signifies no error and 1 worst segmentation. The measure is shown to be effective for qualitative similarity comparison between segmentations by humans, who often produce results with varying degrees of perceived details, which are all intuitively reasonable and therefore "correct". On the other hand, the Martin error measure is sensitive to qualitatively different segmentations.

Assume $I_g = \{A_1, A_2, \ldots, A_M\}$ is a reference image where A_j is the jth fragment in I_g. Assume further that $I_s = \{B_1, B_2, \ldots, B_N\}$ is the segmented image where B_i is the ith fragment in I_s. Let $|A|$ represent the number of pixels in A. Martin et al. [11] define the error between fragment A_j and B_i (in a different but equivalent form) as

$$P_{ji} = \frac{|A_j \setminus B_i|}{|A_j|} \times |A_j \bigcap B_i| = \left(1 - \frac{|A_j \bigcap B_i|}{|A_j|}\right) \times |A_j \bigcap B_i|, \quad (1)$$

where $\setminus$ denotes the set difference operation and $\bigcap$ denotes the intersection. Similarly, the error between fragment B_i and A_j is defined as

$$Q_{ji} = \frac{|B_i \setminus A_j|}{|B_i|} \times |A_j \bigcap B_i| = \left(1 - \frac{|A_j \bigcap B_i|}{|B_i|}\right) \times |A_j \bigcap B_i|. \quad (2)$$

The total area of intersection between I_g and I_s is calculated by

$$n = \sum_{j=1}^{M} \sum_{i=1}^{N} |A_j \bigcap B_i|. \quad (3)$$

There are two variants of the Martin error measure, global consistency error (GCE) and local consistency error (LCE). Specifically,

$$GCE(I_g, I_s) = \frac{1}{n} \min \left\{ \sum_{j=1}^{M} \sum_{i=1}^{N} P_{ji}, \sum_{j=1}^{M} \sum_{i=1}^{N} Q_{ji} \right\}, \quad (4)$$

$$LCE(I_g, I_s) = \frac{1}{n} \sum_{j=1}^{M} \sum_{i=1}^{N} \min(P_{ji}, Q_{ji}). \quad (5)$$

Although these error metrics are calculated by grouping pixels into objects first, they unfortunately tolerate over-segmentation and under-segmentation, as a consequence of their intended purpose for comparing human segmentations. As an example, take Fig. 1 which shows a ground truth image of a single object (I_0) and three possible hypothetical segmentation results (I_1, I_2, and I_3) with varying degrees of over-segmentation. Comparing A_1 and B_1, $M = N = 1$, and $n = |A_1 \bigcap B_1|$. Since A_1 and B_1 are identical, $P_{11} = Q_{11} = 0$ and $GCE = LCE = 0$, as expected of a reasonable error measure. However, for the segmentation in Fig. 1(c), $I_2 = \{C_1, C_2\}$ where C_1 and C_2 are each one half of A_1 (assuming object boundaries are of zero-pixel width) and $M = 1$ and $N = 2$, so that $P_{1i} = \frac{1}{2}|A_1 \bigcap C_i| > 0$ and $Q_{1i} = 0$ $(i = 1, 2)$ and therefore $GCE = LCE = 0$, incorrectly indicating no error. Similarly, in Fig. 1(d), $I_3 = \{D_1, D_2, D_3\}$ where D_1 is a half of A_1 and D_2 and D_3 are a quarter of A_1, respectively, so that $Q_{1i} = 0$ $(i = 1, 2, 3)$ and $GCE = LCE = 0$. As can be seen, the object in Fig. 1(a) can be indefinitely over-segmented, and yet error measures (4) and (5) are insensitive to the extent of over-segmentation. The

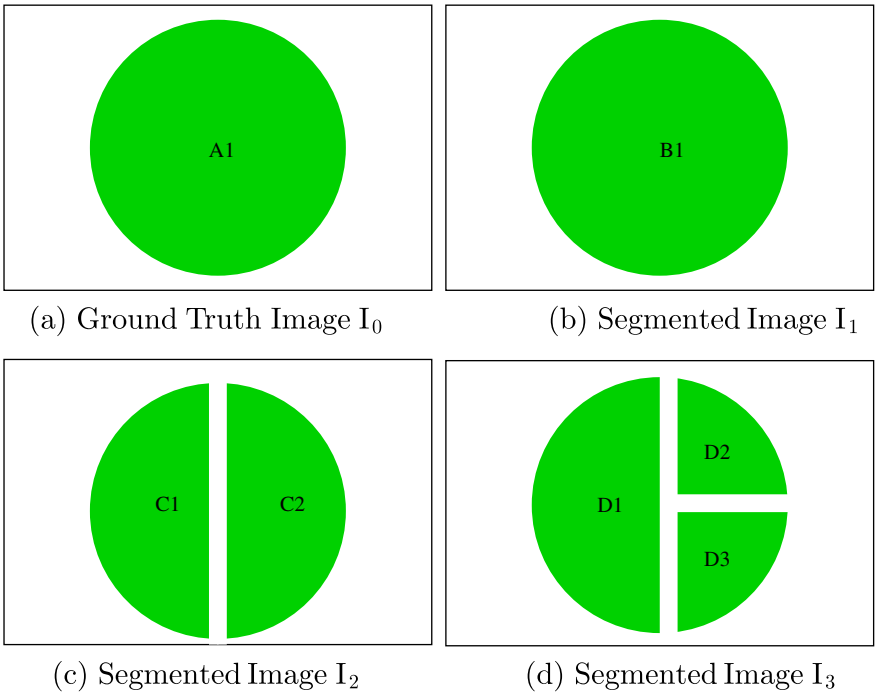

Fig. 1. The ground-truth image and three possible segmentations yield identical error scores according to Eqs. (4) and (5).

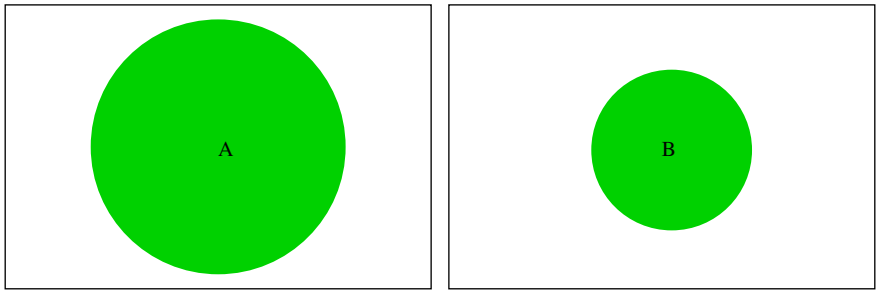

Fig. 2. The ground-truth image (left) and its over-segmented version (right) yielding a perfect score according to Eqs. 4 and 5.

Martin error measure is therefore inappropriate for the purpose of evaluating object-level segmentation where the exact object boundaries are of concern. Partition distance in [13] improves the Martin error and considers Fig. 1(c) and (d) to be incorrect; however, partition distance still considers Fig. 1(c) and (d) to be equally incorrect although Fig. 1(d) should be a worse over-segmentation than Fig. 1(c) in many applications.

In addition, the Martin error measures tolerate under-segmentation. In the case of multiple fragments being incorrectly segmented into one large one when, for example, $I_g = I_2$ and $I_s = I_0$ in Fig. 1, one can easily verify, following the above development, that $GCE(I_g, I_s) = LCE(I_g, I_s) = 0$. As the example of another form of under-segmentation (over-segmentation), refer to Fig. 2 which shows a ground-truth image on the left (right) and its over-segmented (under-segmented) version on the right (left). It is easy to verify that $P_{11} \neq 0$ and $Q_{11} = 0$ so that $GCE = LCE = 0$ in this case.

Note that the tolerance of Martin error measure to over- and under-segmentation – in terms of the number of segmented objects and the object sizes – as illustrated above is by design, as pointed out in [11], as it does not discriminate among segmentations that are simple "refinements" of each other. Since the two images in Fig. 2 are scaled version of each other, GCE and LCE consider them to be *consistent* with respect to each other, and no penalty is incurred. This tolerance however is not appropriate in segmentation applications – such as quantitative cell biology and visual granulometry – in which the exact boundaries or sizes of the fragments are important, giving rise to the need to develop a new error measure. Our error measure, to be introduced in the next section, precisely addresses this problem.

3. Object-level consistency error

To penalize over-segmentation and under-segmentation, we propose a new error measure in this section. To begin, we define a partial error measure as

$$
\begin{aligned}
E_{g,s}(I_g, I_s) &= \sum_{j=1}^{M} \left[1 - \sum_{i=1}^{N} \frac{|A_j \bigcap B_i|}{|A_j \bigcup B_i|} \times W_{ji} \right] W_j, \\
W_{ji} &= \frac{\bar{\delta}(|A_j \bigcap B_i|)|B_i|}{\sum_{k=1}^{N} \bar{\delta}(|A_j \bigcap B_k|)|B_k|}, \\
W_j &= \frac{|A_j|}{\sum_{l=1}^{M} |A_l|},
\end{aligned}
\tag{6}
$$

where I_g, I_s, M, and N are defined as in Section 2, and $\delta(x)$ is the delta function whose value equals 1 if the input is 0 and whose value is 0 otherwise. $\bar{\delta}(x) = 1 - \delta(x)$. We then define the OCE as

$$
OCE(I_g, I_s) = \min(E_{g,s}, E_{s,g}). \tag{7}
$$

The key difference between OCE and GCE (or LCE) is that the denominator of the error formula for a fragment in (6) uses the union of the two fragments that intersect ($|A_j \bigcup B_i|$), rather than one of the two fragments as in (1) or (2). Eq. (6) is based on the so called Jaccard index and allows us to fairly penalize both over- and under-segmentation. In addition, we attach weights to the different terms in (6) that are proportional to the sizes of the fragments in arriving at the final error measure. Specifically, W_{ji} weighs each B_i that intersects with A_j according to the size of B_i relative to all fragments in I_s that intersect with A_j. Similarly, W_j weighs the importance of A_j relative to all fragments in the ground truth image. It is easy to verify that $\sum_i W_{ji} = \sum W_j = 1$.

Note that instead of the Jaccard index, we could also use the Dice's coefficient, by replacing $|A_j \bigcup B_i|$ by $|A_j| + |B_i|$ in Eq. (6). The two similarity measures are similar and either will serve our purpose well. We will compare them empirically in the example section.

To illustrate that the proposed OCE is sensitive to over-segmentation and under-segmentation, again refer to the images in Fig. 1.

Example 1. To calculate, for example, $OCE(I_0, I_2)$, we first need to determine $E_{0,2}$ and $E_{2,0}$.

- For $E_{0,2}$, we have $M = 1$ and $N = 2$, $W_{11} = W_{12} = \frac{1}{2}$, and $W_1 = 1$, so that

$$
E_{0,2} = \left(1 - \sum^{2} \left(\frac{1}{2}\right)\left(\frac{1}{2}\right) \right) \cdot 1 = 0.5.
$$

- For $E_{2,0}$, we have $M = 2$ and $N = 1$, $W_{j1} = 1$ and $W_j = \frac{1}{2}$ $(j = 1, 2)$, so that

$$
E_{2,0} = \sum^{2} \left(\left(1 - \frac{1}{2} \right)\left(\frac{1}{2}\right) \right) = 0.5.
$$

- Overall, $OCE(I_0, I_2) = min(0.5, 0.5) = 0.5$.

Example 2. To show that our OCE can discriminate among various degrees of over-segmentation (or under-segmentation), we can derive $OCE(I_0, I_3)$, where I_3 is more over-segmented than I_2 with respect to the ground truth I_0. In this case

- First, to determine $E_{0,3}$, we note that $M = 1$, $N = 3$, $W_{11} = \frac{1}{2}$, $W_{12} = W_{13} = \frac{1}{4}$, and $W_1 = 1$. With that

$$
E_{0,3} = \left(1 - \left(\frac{1}{2}\frac{1}{2} + \frac{1}{4}\frac{1}{4} + \frac{1}{4}\frac{1}{4} \right) \right) \cdot 1 = 0.625.
$$

- Second, to determine $E_{3,0}$, we note that $M = 3$, $N = 1$, $W_{j1} = 1$ $(j = 1, 2, 3)$, $W_1 = \frac{1}{2}$, and $W_2 = W_3 = \frac{1}{4}$. With that

$$
E_{3,0} = \left(1 - \frac{1}{2} \right)\frac{1}{2} + \left(1 - \frac{1}{4} \right)\frac{1}{4} + \left(1 - \frac{1}{4} \right)\frac{1}{4} = 0.625.
$$

- Overall, $OCE(I_0, I_3) = min(0.625, 0.625) = 0.625 > OCE(I_0, I_2)$.

As expected, our OCE correctly reflects the intuition that I_3 is a worse over-segmentation than I_2, with respect to I_0.

Fig. 3. An aerial photograph of tree canopies and its various segmentations (labeled as white patches): (a) input canopy image, (b) the manual tree-crown segmentation, and (c)–(e) segmentations with machine-learned operator sequences/image processing steps of lengths three, four, and five, respectively [12].

Table 1
Comparison of three segmentation algorithms using Martin's and our OCE error measures in the case of tree images. OCE uses the Jaccard index as in Eq. (6) whereas OCE_D uses the Dice's coefficient

ML IP Algo.	LCE	GCE	OCE	OCE_D
Case (c)	0.00	0.00	0.77	0.67
Case (d)	0.10	0.13	0.48	0.38
Case (e)	0.07	0.07	0.50	0.37

4. Example applications of proposed evaluation error measure

In this section, we will use two examples to illustrate the benefits of our proposed evaluation metric for image segmentation. The first example involves the segmentation of aerial photographs to detect crown canopies of trees and determine both the number and sizes of the canopies, as a way of efficiently assessing a forest and planning its harvest cut. The second example involves segmenting tar sands, an oil-rich mineral, where the distribution of the ore size is to be inferred from the cross-sectional areas of the segmented regions. This information is valuable for optimizing crushing and screening equipments in the mining industry. In both cases, it is important to be able to determine the number of objects as well as their sizes. An objective and sensitive measure of the segmentation quality is critical in order to tune or compare segmentation algorithms.

In the first example, shown in Fig. 3 is an input image (a), together with the manually detected tree crowns (b) (labeled as white patches) – serving as the ground truth – and three possible segmentations (c–e) obtained with three different machine-learned algorithms [12]. Evaluation of these segmentations according to the Martin error measures and our proposed measures is summarized in Table 1. The difference between our OCE and LCE or GCE is clear. First of all, Image (c), obviously a poor segmentation, is incorrectly considered to be error-free according to LCE/GCE as LCE/GCE does not penalize under-segmentation as we have previously pointed out. The difference in segmentation quality, however, is easily uncovered by our proposed score metric, with Fig. 3(c) being singled out as the worst segmentation. Second, since our score metric strikes a fair balance between the accuracy of object sizes and the number of objects, Fig. 3(d) with its better detected object sizes and Fig. 3(e) with a better detected number of objects – with respect to the ground truth – achieve similar scores, as is expected. Overall, our OCE provides better and more intuitive sensitivity than LCE and GCE.

In general, an error measure should be consistent with the subjective evaluation by humans, and this is the case for the proposed OCE. Fig. 3(c) is the worst of all three interpretations with respect to the ground truth, and it has the largest error at 0.77. Interestingly, Fig. 3(c) and (d) produce very similar error scores, where Fig. 3(c) give better boundaries and Fig. 3(d) a better object count. Finally, OCE using Dice's coefficient exhibit similar characteristics as can be observed in the last column of Table 1.

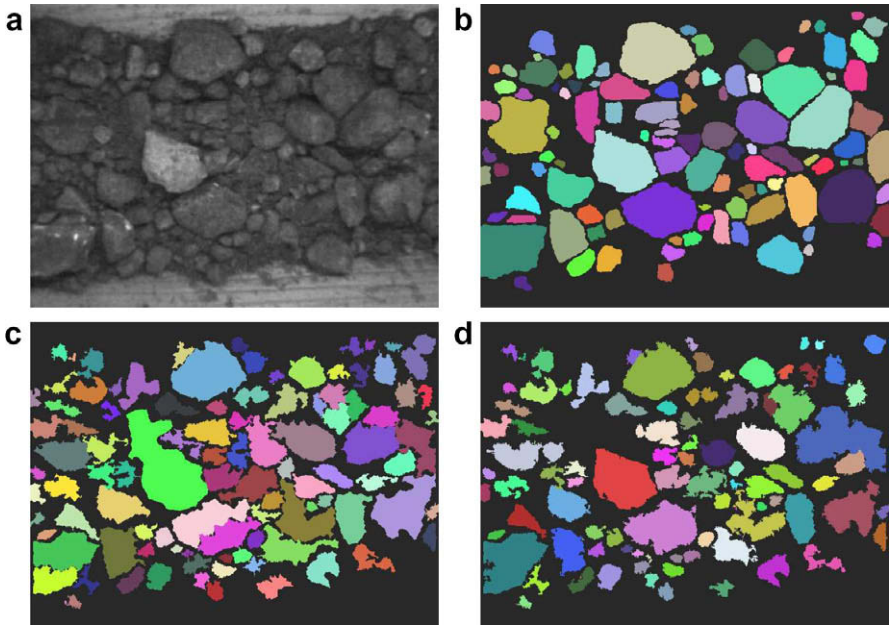

Fig. 4. Sample input image (a), its ground-truth segmentation (b), and segmented images from two segmentation algorithms: watershed transform (c) and level set method (d).

Table 2
Comparison of two segmentation algorithms using Martin's and our error measures across five images

Image	Watershed			Level set		
	LCE	GCE	OCE	LCE	GCE	OCE
A	0.12	0.21	0.53	0.10	0.15	0.49
B	0.06	0.11	0.47	0.07	0.14	0.41
C	0.07	0.17	0.44	0.08	0.13	0.41
D	0.09	0.18	0.53	0.09	0.16	0.50
E	0.07	0.16	0.49	0.07	0.17	0.45
Average	0.08	0.17	0.49	0.08	0.15	0.45

In our second example, we compare OCE with LCE and GCE in their ability to evaluate two segmentation algorithms, a watershed-based method and a level set based method, for measuring ore size in oil sand images. Fig. 4(a) shows an oil–sand image, and Fig. 4(b) a reference segmentation (ground truth) obtained manually. Fig. 4(c) is the segmented image using a marker-driven watershed transform algorithm, and Fig. 4(d) the segmented image using a local level-set method. Table 2 summarizes the comparative result for the three error criteria, GCE, LCE and OCE, over a set of five images (A through E with Fig. 4(a) being a typical one) Two differences between our proposed OCE and GCE (or LCE) are clear. First, our error metric is a more realistic discrepancy measure than GCE and LCE (e.g. 0.5 by OCE versus 0.08 by LCE for the watershed algorithm), which tend to be optimistic (especially LCE) and mistakenly indicating good performance. Secondly, once again, our OCE is more sensitive to changes in performance, with an average change of 0.04 versus 0 and 0.02 in GCE and LCE, when the watershed algorithm and the level set method are compared.

5. Conclusions

In this paper, we have proposed a novel error measure, OCE, for evaluating a segmentation algorithm at the object level. OCE is superior to previous error measures in that it correctly penalizes over- and under-segmentation, and is more sensitive to changes in the segmentation result. These two properties are important in many applications where the purpose of segmentation is to delineate multiple objects of similar geometry but variable sizes such as forest survey, blood cell tracking, and mineral processing. We do not claim, however, that our OCE is a generic error measure that works well in arbitrary application domains - if such an error measure could ever be developed. For example, OCE is not expected to work well with natural scenes where refinement does not cause difficulty in high-level recognition tasks.

We have applied the proposed OCE to the comparative study of different image segmentation algorithms and to the tuning of parameters of an image segmentation algorithm, developed for our visual ore-size measurement system, to demonstrate that OCE can serve the important role of developing an optimal segmentation algorithm for a given application. In our case, the correct delineation of boundaries or sizes of objects in an image are critical, and we have demonstrated that this requirement is satisfactorily met by the proposed OCE.

References

[2] Y.J. Zhang, A survey on evaluation methods for image segmentation, Pattern Recognition 29 (1996) 1335–1346.
[3] J.S. Weszka, A. Rosenfeld, Threshold evaluation techniques, IEEE Trans. Systems, Man and Cybernetics 8 (1978) 622–662.
[4] W.A. Yasnoff, J.K. Mui, J.W. Bacus, Error measures for scene segmentation, Pattern Recognition 9 (1977) 217–231.
[5] S.U. Lee, S.Y. Chung, R.H. Park, A comparative performance study of several global thresholding techniques for segmentation, CVGIP 52 (1990) 171–190.
[6] K. Strasters, J.J. Gerbrands, Three-dimensional image segmentation using a split, merge and group approach, Pattern Recognition Letters 12 (1991) 307–325.
[7] Y. Zhang, Influence of image segmentation over feature measurement, Pattern Recognition Letters 16 (1995) 201–206.
[8] J. Franklin, T. Katsabanis, Measurement of Blast Fragmentation, A.A. Balkema, Rotterdam, 1996.
[9] M.D. Levine, A. Nazif, Dynamic measurement of computer generated image segmentations, IEEE Trans. Pattern Analysis and Machine Intelligence 7 (1985) 155–164.
[10] P.K. Sahoo, S. Soltani, A.K.C. Wong, Y.C. Chen, A survey of thresholding techniques, CVGIP 41 (1988) 233–260.
[11] D. Martin, C. Fowlkes, D. Tal, J. Malik, A database of human segmented natural images and its application to evaluating algorithms and measuring ecological statistics, ICCV (2001) 416–423.
[12] G. Lee, V. Bulitko, GAMM: genetic algorithms with meta-models for vision, Proceedings of the Genetic and Evolutionary Computation Conference (GECCO) (2005) 2029–2036.
[13] Jaime S. Cardoso, Luis Corte-Real, Toward a generic evaluation of image segmentation, IEEE Transactions on Image Processing 14 (11) (2005) 1773–1782.
[14] J.J. Charles, L.I. Kuncheva, B. Wells, I.S. Lim, An evaluation measure of image segmentation based on object centres, in: Proc. of the International Conference on Image Analysis and Recognition, Portugal, September 18–20, 2006.