Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition Letters 27 (2006) 1916–1926
Pattern Recognition Letters
www.elsevier.com/locate/patrec

On the use of the overlapping area matrix for image segmentation evaluation: A survey and new performance measures

Alberto Ortiz *, Gabriel Oliver

University of the Balearic Islands, Department of Mathematics and Computer Science, Palma de Mallorca, Spain

Received 8 November 2005; received in revised form 19 April 2006
Available online 30 June 2006

Communicated by Y.J. Zhang

Abstract

The development of common and reasonable criteria for evaluating and comparing the performance of segmentation algorithms has always been a concern for researchers in the area. As it is discussed in the paper, some of the measures proposed are not adequate for general images (i.e. images of any sort of scene, without any assumption about the features of the scene objects or the illumination distribution) because they assume a certain distribution of pixel gray-level or colour values for the interior of the regions. This paper reviews performance measures not performing such an assumption and proposes a set of new performance measures in the same line, called the *percentage of correctly grouped pixels* (CG), the *percentage of over-segmentation* (OS) and the *percentage of under-segmentation* (US). Apart from accounting for misclassified pixels, the proposed set of new measures are intended to compute the level of fragmentation of reference regions into output regions and vice versa. A comparison involving similar measures is provided at the end of the paper.

Keywords: Performance evaluation; Image segmentation; Overlapping area matrix

1. Introduction

Since the innumerable edge detection and segmentation algorithms published are all based on their own working hypotheses, the development of common and reasonable criteria for evaluating and comparing their performance has always been a concern for researchers in the area. In the often-cited survey (Zhang, 1996), Zhang distinguishes between two general approaches: analytical and empirical assessment. On the one hand, *analytical methods* refer to the analysis of the segmentation algorithm on the basis of its properties, such as: (1) the type and amount of a priori knowledge that has been incorporated into the algorithm; (2) the processing strategy, whether it is parallel, sequential, iterative or mixed; or (3) the processing complexity and efficiency. *Empirical methods*, on the other hand, judge the algorithms analyzing their output. Among them, some are oriented towards measuring the "goodness" of the results (*empirical goodness methods*) while others produce some discrepancy measures between the algorithm output and the expected output, often referred to as ground truth data (*empirical discrepancy methods*).

This paper discuses the behaviour of some traditional measures for checking the performance of a segmentation algorithm when general images (i.e. images of any sort of scene, without any assumption about the features of the scene objects or the illumination distribution) are involved in the comparison and also proposes a set of new performance measures independent of the sort of test images used in the performance analysis. The paper is organized as follows: Section 2 reconsiders the different sorts of performance evaluation methods of Zhang and discusses their

* Corresponding author. Address: Edificio Anselm Turmeda – Campus UIB, Cra. Valldemossa, km 7.5, 07071 Palma de Mallorca, Spain. Tel.: +34 971 172992; fax: +34 971 173003.
E-mail addresses: alberto.ortiz@uib.es (A. Ortiz), goliver@uib.es (G. Oliver).
URL: http://dmi.uib.es/~aortiz (A. Ortiz).

0167-8655/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2006.05.002

use when general images must be utilized for comparing the performance of segmentation algorithms; Section 3 reviews several empirical discrepancy measures based on the use of the so-called confusion matrix; Section 4 proposes new metrics based on a variant of the confusion matrix called the overlapping area matrix (OAM), and reviews some recent measures which can also be expressed using the OAM; Section 5 presents a simple example to illustrate the use of the new measures, and compares them with some of the other measures reviewed; finally, Section 6 concludes the paper.

2. Discussion on general work on segmentation performance evaluation

In the classification of Zhang (1996), analytical methods avoid the implementation of the algorithm and, thus, reduce the influence of the experimental arrangement. However, not all properties of a segmentation algorithm can be captured by analytical studies. In this sense, empirical methods are of larger scope and more general at the expense of having to implement the algorithm. Within this class of methods, empirical goodness methods evaluate some region properties and compare them with some expected values considered good for the output regions. The *intra-region uniformity* and the *inter-region contrast* are region-oriented goodness measures proposed in many studies (Levine and Nazif, 1985; Pal and Bhandari, 1993; Pal and Pal, 1989; Sahoo et al., 1988; Weszka and Rosenfeld, 1978). An empirical discrepancy method can still be more general since the evaluation is based on contrasting the algorithm's output with an already computed reference. The number of pixels incorrectly classified as edge pixels or the number of incorrectly segmented pixels, their position and the number of regions are among the different discrepancy measures proposed in the literature (Baddeley, 1992; Goumeidane et al., 2003; Heyden, 1989; Huang and Dom, 1995; Lewis and Brown, 2001; Lim and Lee, 1990; Odet et al., 2002; Pratt, 1978; Rees et al., 2002; Roman-Roldan et al., 2001; Strasters and Gerbrands, 1991; Weszka and Rosenfeld, 1978; Yasnoff and Bacus, 1984; Yasnoff et al., 1977). Additional discrepancy measures based on region features such as area, eccentricity or perimeter, among others, have also been considered (Zhang, 1995; Zhang and Gerbrands, 1992, 1994).

A single evaluation measure can perhaps not be enough to judge the result of a segmentation algorithm. Haralick and Shapiro (1985) stated this idea in the form of three heuristic criteria for image segmentation evaluation: regions must be uniform and homogeneous, regions shape must be simple, without many holes, and adjacent regions must present significatively different values for the feature defining the partitioning process. Liu and Yang, 1994 proposed then a function based on these criteria:

$$F(I) = \frac{1}{1000(N \cdot M)} \sqrt{R} \sum_{i=1}^{R} \frac{e_i^2}{\sqrt{A_i}}, \quad (1)$$

where $N \times M$ is the size of the image, R is the number of regions, A_i is the area of region i and e_i is the *colour error*, calculated as the sum of the differences between region pixel colour values and the colour attributed to region i. The more recent work (Borsotti et al., 1998) improved this expression after proving empirically that the expression by Liu and Yang did not satisfy, in general, the criteria by Haralick and Shapiro. Some of the studies cited above propose other measures which also combine different aspects of the output of a segmentation algorithm in a single expression (Pal and Bhandari, 1993; Strasters and Gerbrands, 1991; Yasnoff and Bacus, 1984; Zhang, 1993).

Either in isolation or as part of more elaborated performance evaluation expressions, goodness measures such as the *intra-region uniformity*, the *inter-region contrast* or the *color error* e_i of Eq. (1) were clearly conceived for images where the segmentation output was expected to consist of regions of uniform gray-level or uniform colour. In other words, the images to be processed by the assessed segmentation algorithms were hypothesized as consisting of noisy bi-dimensional constant piecewise functions. In case the segmentation algorithm was devised to produce regions assimilable to objects or part of objects of uniform colour, this scenario takes place only when shadows or the effects of objects curvature, in the form of mainly shading and specular reflection, are not noticeable in the image. However, for general images (i.e. images of any sort of scene, without any assumption about the features of the scene objects or the illumination distribution) the aforementioned measures clearly do not make sense from a performance evaluation viewpoint. This is precisely the sort of images which segmentation algorithms based on physics-based models of image formation try to solve. On those cases, the goodness measures mentioned above could be of use only if pixel colour values were first transformed to a new colour space invariant to the scene factors leading to changes in intensity not related with object boundaries (shading due to objects curvature, illumination changes, shadows, specularities, inter-reflections, etc.). Nevertheless, although several photometric invariants have been proposed to tolerate some of those changes (see, for instance, the work by Gevers and colleagues (Gevers, 2002, 2004; Gevers and Stokman, 2003a,b)), none of them is able to deal with all of them simultaneously (see Ortiz, 2005, chapter 3). From this point of view, and in order to be able to use the performance evaluation method for any sort of algorithm and test image, it is obvious that, in general, it is preferable that the evaluation measures do not include any reference to the particular pixel colour values present in the output regions. In this sense, it seems more appropriate to resort to empirical discrepancy measures whenever possible.

3. Image segmentation evaluation and the confusion matrix

A certain amount of empirical discrepancy methods are based on a matrix C which gathers information about the

distribution of pixels between the different regions of the image, either real or predicted. An entry of C, say (i,j), contains the number of pixels belonging to region i in the reference segmentation (*reference region* from now on) such that are labeled as pertaining to region j in the algorithm output (*output region* from now on). A particular case of this matrix, in which, among other features, C is square, is known as *confusion matrix*, *error matrix* or *contingency matrix* according to the terminology employed in the evaluation of classifiers performance (Stehman, 1997).

In a classification context, the percentage of correctly classified pixels or, conversely, the percentage of misclassified pixels are the discrepancy measures that come most readily to mind and that can be easily obtained from C. In a multi-class case, say N classes, C is an $N \times N$ matrix and the percentage of correctly classified pixels is given by $\sum_{i=1}^{N} C_{ii} / \sum_{i,j=1}^{N} C_{ij}$. As for the misclassified pixels, the *multi-class type I error* for class k or percentage of false positives for class k is expressed as (Yasnoff et al., 1977; Zhang, 1996) (see Fig. 1):

$$M_{\mathrm{I}}^{(k)} = \frac{\left(\sum_{i=1}^{N} C_{ki}\right) - C_{kk}}{\sum_{i=1}^{N} C_{ki}} \times 100, \tag{2}$$

while the *multi-class type II error* for class k or percentage of false negatives for class k is given by

$$M_{\mathrm{II}}^{(k)} = \frac{\left(\sum_{i=1}^{N} C_{ik}\right) - C_{kk}}{\left(\sum_{i=1}^{N}\sum_{j=1}^{N} C_{ij}\right) - \sum_{i=1}^{N} C_{ki}} \times 100. \tag{3}$$

In a recent paper, Lewis and Brown introduce what they claim to be a *generalized confusion matrix* for producing accurate estimates of the area covered by each class in an image, typically resulting from remote sensing (Lewis and Brown, 2001). Essentially, they put matrix C as the combination of a *target matrix* T and an *output matrix* Y, where both matrices have as many rows as pixels in the image and each cell of both is the membership of the corresponding pixel to a class: T contains reference memberships, while Y contains the memberships produced by the classification algorithm. Finally, each row of both matrices sum to unity and $C = T^{\mathrm{T}} Y$ is defined. With the introduction of the membership concept, they contemplate real situations where several classes may occur in the field of view of the satellite sensor represented by a single pixel. Besides, the authors introduce an *error matrix* E as the difference between the generalized confusion matrix and a *reference matrix* $R = T^{\mathrm{T}} T$. Finally, they propose several measures that summarize the information contained in the error matrix, such as the average error among all the pixels, or the error for a given class.

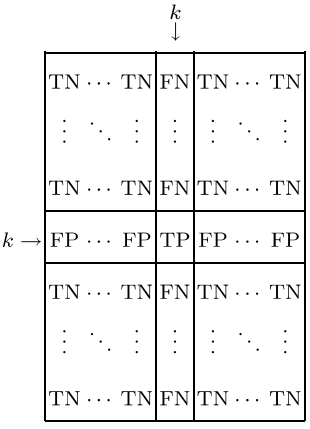


Fig. 1. Confusion matrix for a multi-class classification problem. In the matrix, TP (true positives), TN (true negatives), FP (false positives) and FN (false negatives) correspond to class k.

To finish with this brief review on the use of the confusion matrix, it is necessary to cite a new discrepancy evaluation strategy rooted on the same principles of the confusion matrix and known as *receiver operating characteristic* (ROC) curve analysis. ROC analysis comes from psychophysics and signal detection theory and has received an important amount of attention within the vision community (Alexander and Buxton, 1997; Bowyer et al., 2001; Edwards et al., 1999; Forbes and Draper, 2000; Ramesh and Haralick, 1992; Rees et al., 2002; Shin et al., 2001; Southall et al., 2000; Yitzhaky and Peli, 2003). A ROC curve plots true positive rates against false positive rates as a decision parameter of the classifier is varied (i.e. the classification threshold). It is therefore mainly suited for two-class classification problems, such as edge detection, although (Rees et al., 2002) addressed multi-class classification evaluation by means of ROC analysis. As can be observed, ROC analysis does not provide new performance measures, it is rather a change of philosophy, as Bowyer et al. remarked regarding the evaluation of edge detectors (Bowyer et al., 2001, p. 80).

4. A new discrepancy-based strategy for performance evaluation

4.1. The overlapping area matrix

Although it is true that sometimes a classification task follows a segmentation task in the pipeline of a vision application, and that, therefore, the evaluation can be applied directly to the output of the classifier, on many occasions it can be important to evaluate the segmentation output per se. An important difference which arises then is that, unlike the case of a typical classification task, the final number of output regions/classes does not need to agree with the number of reference regions/classes. This fact can be the consequence of two misbehaviours of the segmentation algorithm, which can happen alone or in combination: (1) a reference region can appear split in several output regions, which is known as *over-segmentation*; and (2) several reference regions can result merged in an only output region, which is known as *under-segmentation*. A non-null degree of over- and/or under-segmentation

derives in the following consequences from a practical point of view:

(1) the confusion matrix is no longer square, and, in fact, the term *overlapping area matrix* (OAM) has been introduced to refer to this new structure (Beauchemin and Thomson, 1997);
(2) the region numbering of the reference segmentation rarely coincides with the region numbering of the algorithm's output, and
(3) due to (1) and (2), the correspondence between reference and output regions is not trivial (i.e. the kth row and the kth column of the confusion matrix do not refer to the same class, unlike the matrix of Fig. 1).

4.2. New empirical discrepancy measures

A set of new performance measures are introduced next in accordance with the special features of the OAM. For notation purposes, a reference region i will be denoted as R_i, while output region j will be written as $\widehat{R}_j$. Every measure is summarized in Table 1 and explained in the following:

- The *percentage of Correctly Grouped pixels*, CG, aims at accounting for those pixels which, belonging to a reference region R_i, are put together in a single output region $\widehat{R}_j$. That is to say, the goal is to determine to which degree the segmentation algorithm does not mix pixels from different reference regions into a single output region. Being strict, only those output regions which completely overlap with a reference region should be considered in *CG*. However, even in the best of the segmentations, a certain level of dispersion can be found, in the sense that small fractions of regions $\widehat{R}_j$ fall outside the main region R_i. Because of this, the definition of CG has been relaxed through the parameter p, by which CG(p) represents the amount of pixels which belong to a region $\widehat{R}_j$ which is mostly included in a region R_i, where the meaning of *mostly* depends on p. Therefore, the amount of dispersion which is tolerated is expressed through parameter p, being $p = 100$ the most restrictive case which leads to the first definition of CG. Observe that this scattering of pixels can be easily detected looking at the columns of the confusion matrix, which will contain more than one non-null entry in case a region $\widehat{R}_j$ groups pixels from several regions R_i.

  As can be easily guessed, CG tries to determine the correspondence between reference and output regions. With the introduction of the parameter p, this is particularly true when $p > 50$, since, on those occasions, $\mathrm{SRa}(\cdot, \widehat{R}_j, p)$ takes value 1 only for one R_i, which would correspond to the region R_i which mostly overlaps with region $\widehat{R}_j$. For $p \leqslant 50$, $\mathrm{SRa}(\cdot, \widehat{R}_j, p)$ can take value 1 for several R_i, which makes CG(p) account for several fractions of $\widehat{R}_j$; since the sense of correspondence has been lost, CG can be said less meaningful for those values of p.

  As a final remark, notice that, however, even for $p > 50$, several output regions $\widehat{R}_j$ can correspond to a single reference region R_i. As was said in the first definition of CG, this measure just tries to determine to which degree the segmentation algorithm does not put together pixels belonging to different reference regions. Therefore, a segmentation with a 100% of correctly grouped pixels can be over-segmented (see later in cases (2) and (3) of the example, Section 5).
- The *percentage of under-segmentation*, US, represents the amount of pixels of the image which have been assigned to regions $\widehat{R}_j$ which cover several reference regions R_i. As before, parameter p allows relaxing the definition of US to tolerate slight segmentation errors, so that US(p) accounts for regions $\widehat{R}_j$ whose overlapping with a reference region R_i is below p%. To this end, function $\mathrm{SRb}(\widehat{R}_j, p)$ works over the columns of the OAM signaling those output regions $\widehat{R}_j$ whose degree of coincidence with at least one reference region is above p%; when this condition is not met for an output region $\widehat{R}_j$, $\mathrm{SRb}(\widehat{R}_j, p) = 0$, what makes US accumulate all the pixels of $\widehat{R}_j$.

  Observe that US comprises all the pixels of output regions under-segmenting reference regions, contrary to CG, which accumulates fractions of output regions. This

Table 1
Discrepancy measures formulae

$n(R_i) = \sum_{k=1}^{N_o} C_{ik}$	Size of region R_i
$n(\widehat{R}_j) = \sum_{k=1}^{N_r} C_{kj}$	Size of region $\widehat{R}_j$
$n(I) = \sum_{k=1}^{N_r} n(R_k) = \sum_{k=1}^{N_o} n(\widehat{R}_k)$	Size of image
$\mathrm{SRa}(R_i, \widehat{R}_j, p) = \begin{cases} 1 & \text{if } \frac{C_{ij}}{n(\widehat{R}_j)} \times 100 \geqslant p \\ 0 & \text{otherwise} \end{cases}$	p% pixels of $\widehat{R}_j$ are concentrated in a single reference region R_i
$\mathrm{SRb}(\widehat{R}_j, p) = \begin{cases} 1 & \text{if } \frac{\max_{k=1,\ldots,N_r}\{C_{kj}\}}{n(\widehat{R}_j)} \times 100 \geqslant p \\ 0 & \text{otherwise} \end{cases}$	At least p% pixels of $\widehat{R}_j$ are concentrated in a single reference region R_i
$\mathrm{SO}(R_i, p) = \begin{cases} 1 & \text{if } \frac{\max_{k=1,\ldots,N_o}\{C_{ik}\}}{n(R_i)} \times 100 \geqslant p \\ 0 & \text{otherwise} \end{cases}$	At least p% pixels of R_i are concentrated in a single output region $\widehat{R}_j$
$\mathrm{CG}(p) = \frac{\sum_{i=1}^{N_r}\sum_{j=1}^{N_o}\mathrm{SRa}(R_i, \widehat{R}_j, p) \times C_{ij}}{n(I)} \times 100$	Percentage of correctly grouped pixels, at level p
$\mathrm{US}(p) = \frac{\sum_{j=1}^{N_o}(1 - \mathrm{SRb}(\widehat{R}_j, p)) \times n(\widehat{R}_j)}{n(I)} \times 100$	Percentage of under-segmentation, at level p
$\mathrm{OS}(p) = \frac{\sum_{i=1}^{N_r}(1 - \mathrm{SO}(R_i, p)) \times n(R_i)}{n(I)} \times 100$	Percentage of over-segmentation, at level p

N_r is the number of reference regions, while N_o is the number of output regions.

is because the concept of under-segmentation is relative to a segmentation: one would say that the output of a segmentation algorithm under-segments an image. CG, however, accounts for groups of pixels which are correct from the point of view of not belonging to several reference regions. In this sense, US has to take into account whole output regions which under-segment the image.

- The *percentage of over-segmentation*, OS, accounts for pixels of output regions $\widehat{R}_j$ which split a reference region R_i. In this case, the key role is played by the rows of the OAM. If a row, say i, of C contains several non-null entries is because the reference region R_i appears divided into the regions $\widehat{R}_j$ corresponding to those non-null entries. Again, the amount of dispersion tolerated is controlled through parameter p now within function $\mathrm{SO}(R_i,p)$. This function takes value 1 whenever the coincidence between a reference region R_i and at least one output region is above $p\%$. In this way, OS just has to accumulate the pixels of reference regions R_i not meeting the above condition, i.e. $\mathrm{SO}(R_i,p)=0$.

4.3. Related work

Other authors have proposed measures rooted on principles similar to CG, OS and US, and, therefore, aiming at evaluating the amount of misclassification in the segmentation output as well as the level of under- and over-segmentation. A representative set is summarized in Tables 2 and 3 using the notation of Table 1 and the OAM, although the formulations appearing in the original papers did not explicitly make use of this matrix. A brief discussion on them is provided in the following:

- On the one hand (Levine and Nazif, 1982) propose the *under-merging error* UM and the *over-merging error* OM. While the first one would be a synonymous for over-segmentation (i.e. the error related to merging less than it should be is over-segmentation), the second one would correspond to under-segmentation (i.e. a segmentation output which is the result of more fusion than should be leads to under-segmentation). In both measures, the authors associate to every output region $\widehat{R}_j$ the reference region R_k with the maximum spatial coincidence, and compute the error measures accumulating the resulting misclassified pixels. Therefore, they are not intended to compute the level of fragmentation of reference regions into output regions and vice versa. To combine both errors in an overall performance measure, the authors propose a sum in quadrature of UM and OM.
- Later, Huang and Dom presented two quantities very similar to the under- and over-merging errors, but not identical, which they called *missing rate* e_R^{m} and *false alarm rate* e_R^{f} (Huang and Dom, 1995). e_R^{m} rely on determining, for every reference region, the output region for which the matching is maximum, while e_R^{f} needs the correspondence between every output region and its maximum reference region. Once determined those correspondences, both measures accumulate the pixels of the non-maximum intersections between reference and output regions. As well as for Levine and Nazif, these measures are not intended to compute the level of fragmentation of reference regions into output regions and vice versa. Furthermore, if a correspondence is established between a reference region R_i and an output region $\widehat{R}_j$ in e_R^{m}, $\widehat{R}_j$ need not necessarily have the maximum intersection with R_i when computing e_R^{f}; as a consequence, what is considered misclassification by one of the measures can be considered a correct classification by the other one. OS and US avoid this difficulty by accumulating the whole region, respectively, R_i and $\widehat{R}_j$, when the over- and under-segmentation is detected, instead of accounting for the pixels not belonging to the maximum intersection, as in e_R^{m} and e_R^{f}. This fact can make e_R^{m} and e_R^{f} difficult to interpret. Finally, Huang and Dom introduce an overall performance measure p_R from the average between e_R^{m} and e_R^{f}, what, given the case depicted above, makes this measure meaningless.
- Hoover et al. (1996) classify every pair of reference R_i and output $\widehat{R}_j$ regions as *correct detections*, *over-segmentation*, *under-segmentation*, *missed* or *noise*, and then build evaluation metrics counting instances of every case. As can be observed, the percentage threshold T used during the classification plays the same role as p in CG, OS and US. Besides, as the authors show in the paper, for $0.5 < T < 1.0$, any region can contribute to up to three classifications, one each of correct detection, over- and under-segmentation. Although the measures are easily interpretable, however, unlike CG, OS and US, they depend on the number of regions in the image since they count regions not pixels. Consequently, the metrics of Hoover et al. (1996) are dependent on the number of regions of the image, which is, in general, highly variable across images. This fact thus prevents using a set of test images for getting a global measurement of the performance of the algorithm in terms of misclassification, over- and under-segmentation; that is to say, when comparing several algorithms, it will have to be done on the basis of a single image each time, instead of considering a representative set of images and computing a global performance measure for every algorithm.
- Mezaris et al. (2003) propose an overall performance measure whose cost terms take into account, apart from the spatial coincidence between reference and output regions, the distance between misclassified pixels and the corresponding reference region. The output of the evaluation process is an only measure which combines the aforementioned distance and the amount of misclassification, over- and under-segmentation in a single number. All in all, although it is clear that the measure takes lower values as the segmentation output is better and can be used to rank algorithms performance, it results difficult to interpret because a lot of information is put together in a single number.

Table 2
Other evaluation measures based on the overlapping area matrix (I)

Authors	Formulae	Measure description
Levine and Nazif (1982)	$\mathrm{UM} = \sum_{j=1}^{N_o}(n(R_k) - C_{kj}), k = \arg\max_{i=1,\ldots,N_o}\{C_{ij}\}$	Under-merging error
	$\mathrm{OM} = \sum_{j=1}^{N_o}(n(\widehat{R}_j) - C_{kj}), k = \arg\max_{i=1,\ldots,N_o}\{C_{ij}\}$	Over-merging error
	$M = \sqrt{\left(\frac{\mathrm{UM}}{n(I)}\right)^2 + \left(\frac{\mathrm{OM}}{n(I)}\right)^2}$	Overall performance
Huang and Dom (1995)	$e_R^m = \frac{\sum_{i=1}^{N_r}\sum_{j\neq \arg\max_{k=1,\ldots,N_o}\{C_{ik}\}} C_{ij}}{n(I)}$	Missing rate
	$e_R^f = \frac{\sum_{j=1}^{N_o}\sum_{i\neq \arg\max_{k=1,\ldots,N_r}\{C_{kj}\}} C_{kj}}{n(I)}$	False alarm rate
	$p_R = 1 - \frac{e_R^m + e_R^f}{2}$	Overall performance
Hoover et al. (1996)	$(R_i, \widehat{R}_j)$ is an instance of a *correct detection* if $C_{ij} \geqslant T \times n(\widehat{R}_j)$ and $C_{ij} \geqslant T \times n(R_i)$	#CC, number of correct detection instances ($0.5 < T \leqslant 1.0$ is a threshold percentage)
	$(R_i, \widehat{R}_{j_1}, \ldots, \widehat{R}_{j_x})$ is an instance of an *over-segmentation* if $C_{ij_t} \geqslant T \times n(\widehat{R}_{j_t}), \forall t$ and $\sum_{t=1}^{x} C_{ij_t} \geqslant T \times n(R_i)$	#OC, number of over-segmentation instances ($0.5 < T \leqslant 1.0$ is a threshold percentage)
	$(R_{i_1}, \ldots, R_{i_x}, \widehat{R}_j)$ is an instance of an *under-segmentation* if $\sum_{t=1}^{x} C_{i_t j} \geqslant T \times n(\widehat{R}_j)$ and $C_{i_t j} \geqslant T \times n(R_{i_t}), \forall t$	#UC, number of under-segmentation instances ($0.5 < T \leqslant 1.0$ is a threshold percentage)
	R_i is a *missed region* if it does not participate in any instance of correct detection, over-segmentation or under-segmentation	#MR, number of missed regions
	$\widehat{R}_j$ is a *noise region* if it does not participate in any instance of correct detection, over-segmentation or under-segmentation	#NR, number of noise regions

- Finally, Pignalberi et al. (2003) define another cost function combining the cost associated with erroneously segmented pixels together with handicap functions accounting for under- and over-segmentation; this function was intended as a fitness function for tuning the parameters of segmentation algorithms by means of evolutionary techniques. Contrary to the case of Mezaris et al., this time, terms accounting for misclassified pixels, over-segmentation and under-segmentation are clearly defined, although the overall performance measure suffers from the same drawback as in the case of Mezaris et al. Over- and under-segmentation level measurement takes into account whole regions, similarly to OS and US. Nevertheless, a percentage threshold is not included so that a single misclassified pixel can make a large region to be considered over- or under-segmented. Therefore, over- and under-segmentation are greatly penalized.

5. Discussion

5.1. Comparison over an illustrative example

By way of illustration of CG, OS and US, Fig. 2 shows some examples of segmentation for a very simple image, while the related performance measures appear in Table 4. The six cases considered are discussed in the following lines:

(1) This case corresponds to a perfect segmentation and, accordingly, the percentage of correctly grouped pixels is 100, while the percentages of over-segmentation and under-segmentation are 0.
(2) Now, reference region R_1 appears split into output regions $\widehat{R}_1 - \widehat{R}_4$. None of the regions $\widehat{R}_j$ mix pixels from several true regions R_i, what leads to a CG value of 100, but 25% of the image is over-segmented.
(3) This is an extreme case where all the reference regions appear split in the segmentation output, what leads to a 100% of over-segmentation. Pixels are however correctly grouped and, thus, CG = 100.
(4) In this case, two reference regions have been put together and, consequently, a certain degree of under-segmentation must be measured. Since those two regions cover 50% of the image pixels, the degree of under-segmentation is also 50%. The remaining 50% of pixels are correctly grouped.
(5) A more complex situation is presented in this example, where only pixels from regions $\widehat{R}_1$ and $\widehat{R}_4$ are

Table 3
Other evaluation measures based on the overlapping area matrix (II)

Authors	Formulae	Measure description
Mezaris et al. (2003)	Let $\mathscr{A} = \{(R_i, \widehat{R}_j)\}$ be such that, given $\widehat{R}_j$ and $i_j = \{i \mid j = \arg\max_{k=1,\ldots,N_o}\{C_{ik}\}\}$, $i = \arg\max_{k\in i_j}\{C_{kj}\}$, $\forall j = 1,\ldots,N_o$ Let $\mathscr{N}_R$ and $\mathscr{N}_S$ denote the set of, respectively, non-paired reference and non-paired output regions	
	$E_i = \sum_{p\in R_i^*} f_1(d(p,R_i)) + \sum_{p\in \widehat{R}_j^*} f_2(d(p,R_i)), \quad \forall (R_i,\widehat{R}_j)\in\mathscr{A}$	Overall performance, where: (a) $R_i^* = (R_i - R_i\cap\widehat{R}_j)$, (b) $\widehat{R}_j^* = (\widehat{R}_j - R_i\cap\widehat{R}_j)$, (c) $R(p)$ is the reference region to which p belongs, (d) $d(p,R)$ is the distance from a pixel p and the nearest boundary pixel of region R, (e) f_1 and f_2 are:
	$E_i = \sum_{p\in R_i} f_1(d(p,R_i)), \quad \forall R_i \in \mathscr{N}_R$	
	$F_j = \alpha \sum_{p\in \widehat{R}_j} f_1(d(p,R(p))), \quad \forall \widehat{R}_j \in \mathscr{N}_S$	
	$E = \sum_{i=1}^{N_r} E_i + \sum_{\widehat{R}_j\in\mathscr{N}_S} F_j$	$f_1(d) = \frac{0.001}{10}d$, $f_2(d) = \begin{cases} \frac{0.001}{10}d & \text{if } d \leqslant 10, \\ 0.001 & \text{if } d > 10, \end{cases}$ and (f) $\alpha = 2$
Pignalberi et al. (2003)	$C = \frac{\sum_{j=1}^{N_o}(n(R_{x_j}) - C_{x_j j})}{N_o}, \quad x_j = \arg\max_{i=1,\ldots,N_r}\{C_{ij}\}$	Cost of erroneously segmented pixels
	$U = n(I) - \sum_{i=1}^{N_r}\sum_{j=1}^{N_o} C_{ij}$	Cost of unlabeled pixels
	$H_u = k \sum_{j\mid\sum_{i=1}^{N_r} m_{ij}>1} n(\widehat{R}_j)$ $H_o = k \sum_{j\mid\sum_{i=1}^{N_r} m_{ij}=0} n(\widehat{R}_j)$	Handicaps of under- and over-segmentation, where (a) $m_{ij} = \begin{cases} 1 & \text{if } j = \arg\max_{k=1,\ldots,N_o}\{C_{ik}\}, \\ 0 & \text{otherwise,} \end{cases}$ and (b) k is a constant to enlarge the variability range
	$F = w_1 C + w_2 H_u + w_3 H_o + w_4 U$	Fitness function $(\sum_i w_i = 1)$

correctly grouped (18.75%) and the rest belong to regions under-segmenting the image. Since the area of all the reference regions is shared by more than one region $\widehat{R}_j$, the whole image is over-segmented (OS = 100).

(6) Finally, this case shows the effect of parameter p in the discrepancy measurements. As it has been depicted before, p determines the degree of tolerance with regions $\widehat{R}_j$ that all but a reduced set of their pixels fall in a region R_i. In the example, region $\widehat{R}_1$ is almost identical to region R_1 except for a single column of pixels which belong to region R_2 which have been "stolen" to region $\widehat{R}_2$. The same happens between regions $\widehat{R}_4$ and $\widehat{R}_3$. If $p = 100$, the percentage of correctly grouped pixels attains 45%, corresponding to regions $\widehat{R}_2$ and $\widehat{R}_3$, while for $p = 90$ this percentage goes up to 95% because only the pixels of $\widehat{R}_1$ which fall into R_2 and the ones of $\widehat{R}_4$ which fall into R_3 are discounted. As for the percentage of over-segmentation, $p = 100$ leads to account in OS for pixels from $\widehat{R}_2$ and $\widehat{R}_3$ as well as the misclassified pixels of $\widehat{R}_1$ and $\widehat{R}_4$, what leads to consider 50% of the image over-segmented. When $p = 90$, only the misclassified pixels are accounted for (5% of the image). Observe also that, in this case of $p = 90$, CG + US < 100.

In the light of the example, CG, OS and US can be said to produce accurate measures of the, respectively, correct grouping, over-segmentation and under-segmentation present in the segmentation output. Furthermore, parameter p of CG, OS and US provides the desired tolerance to low pixel dispersion among regions.

For comparison purposes, the measures for Huang and Dom (1995), Hoover et al. (1996) and Mezaris et al. (2003) for this example are also provided in Table 4. Several conclusions can be drawn from the resulting values:

- e_R^m and e_R^f behave exclusively as misclassification rates since they only take into account pixels out of the maximal matchings, not the whole regions involved in over- and under-segmentation. As a consequence, e_R^m and e_R^f measures tend to be more benevolent with the segmenta-

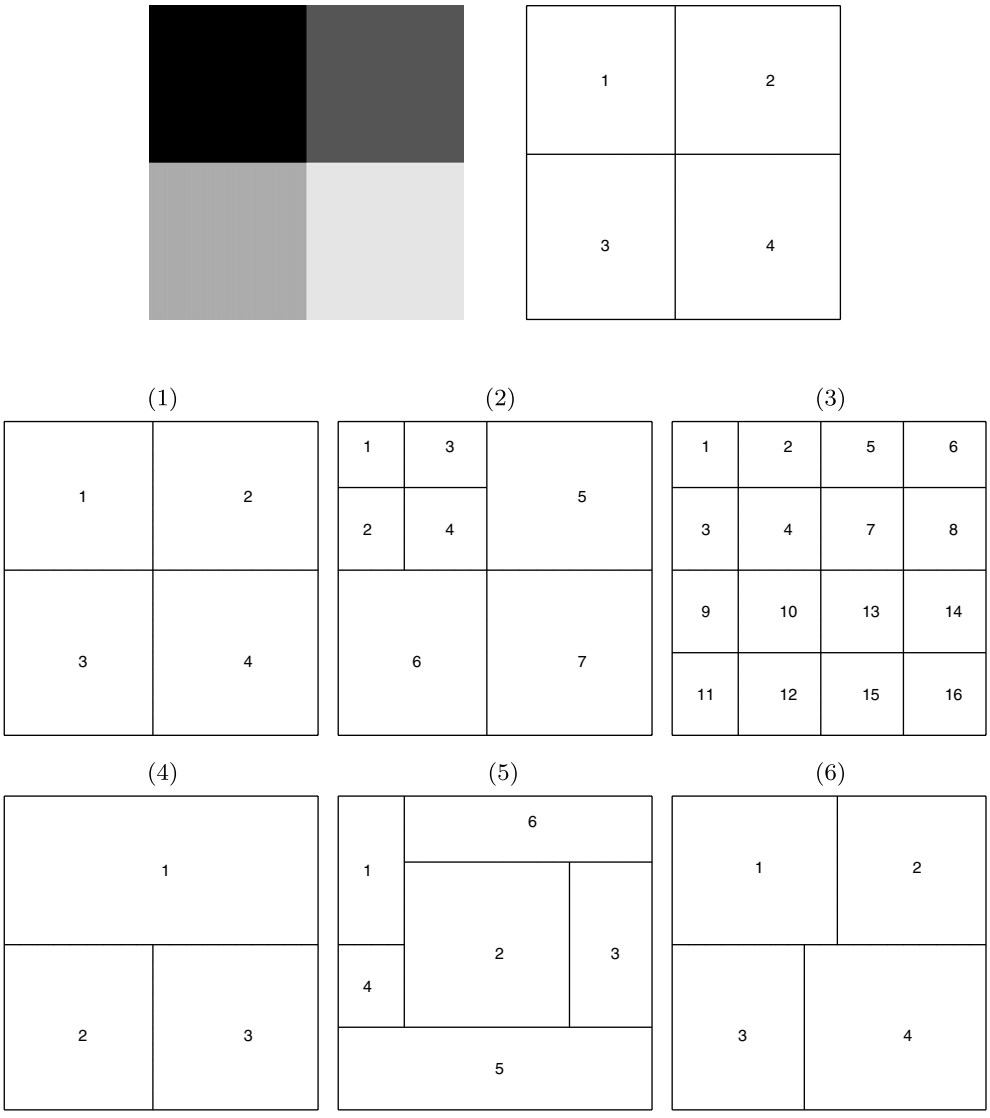


Fig. 2. Examples for illustrating CG, OS and US: [upper row, left] original image (20×20 pixels); [upper row, right] reference contour image; (1)–(6) examples of contour images.

tion output with regard to OS and US because the latter invalidate whole regions.

- The measures by Hoover et al. provide performance information closely related to CG, OS and US, but, however, no data are available about whether the detected over- and under-segmentation is significative on the particular image considered. Furthermore, it is worth discussing the results for case (6b) of Table 4. In this case, thanks to the value given to parameter T, the small dispersions of pixels are tolerated and four correct classifications are found. This is also tolerated by OS(90) and US(90), but, unlike the measures by Hoover et al., CG(90) is able to indicate that there are pixels which are not correctly labeled in the segmentation output.
- Finally, regarding the global performance measure of Mezaris et al., it is worth noting the following:
  (a) very similar values are produced for cases (2) and (4), corresponding to, respectively, over- and under- segmentation cases, although case (4) involves twice more misclassified pixels than case (2);
  (b) case (3), which corresponds to a massive over-segmentation, is ranked twice worse than case (5), which is visually and numerically less correct (look at the percentage of pixels correctly grouped, CG, of both cases).

As for (a), clearly it is due to the value given to parameter α in F_j (see Table 3), which penalizes over-segmentation (F_j) over the misclassification (E_i calculated for set $\mathscr{A}$) and under-segmentation (E_i calculated for set $\mathscr{N}_R$) measures. Regarding (b), since in case (5) both over- and under-segmentation take place simultaneously, it is more difficult to look for the cause of the misbehaviour; nevertheless, it seems to be the same as before. As was already commented, the mixture of spatial coincidence and distance information makes interpretation more difficult.

Table 4
Discrepancy measures for Fig. 2

	Case						
	(1)	(2)	(3)	(4)	(5)	(6a)	(6b)
p	100	100	100	100	100	100	90
CG	100.00	100.00	100.00	50.00	18.75	45.00	95.00
OS	0.00	25.00	100.00	0.00	100.00	50.00	0.00
US	0.00	0.00	0.00	50.00	81.25	55.00	0.00
e_R^{m}	0.00	18.75	75.00	0.00	50.00	5.00	–
e_R^{f}	0.00	0.00	0.00	25.00	43.75	5.00	–
p_R	100.00	90.63	62.50	87.50	53.13	95.00	–
T	1.00	1.00	1.00	1.00	1.00	1.00	0.90
#CC	4	3	0	2	0	0	4
#OC	0	1	4	0	0	0	0
#UC	0	0	0	1	0	0	0
#MR	0	0	0	2	4	4	0
#NR	0	0	0	0	6	4	0
E	0.0000	0.0260	0.1940	0.0330	0.0955	0.0018	–

The first group of rows is for CG, OS and US, while the rest of groups are for, respectively, (Huang and Dom, 1995; Hoover et al., 1996; Mezaris et al., 2003) measures.

5.2. Results for a real image

Fig. 3 shows results for two segmentation outputs of the same real image processed by means of the algorithm C^{3}S (Ortiz, 2005).[1] As can be observed, while case (a) shows a certain level of over-segmentation (the cup and certain areas of the flowerpot), case (b) exhibits a great amount of under-segmentation and, therefore, a deficient number of correctly grouped pixels. Both situations are clearly distinguished by means of CG, OS and US measures. The global measure of Huang and Dom (1995), however, takes almost the same value for both cases, while, clearly, case (a) is visually better than case (b); besides, only in case (b) one of the partial measures, e_R^{f}, indicates a deficient segmentation. As for Hoover et al. (1996) measures, from the 41 regions appearing in the ground truth and, respectively, 40 and 20 output regions in cases (a) and (b), almost all the reference regions are considered missed in both segmentation outputs, and almost all the output regions are classified as noise regions in both cases (a) and (b), while no indication of the relevance of this fact over the image can be obtained from the measures. Finally, the measure of Mezaris et al. (2003) correctly indicates that the segmentation output of case (b) is worse than the one of case (a), although no information about the cause is provided.

[1] C^{3}S is a segmentation algorithm insensitive to the curvature of scene objects because of being based on a physics-based image formation model. This fact allows C^{3}S to tolerate optical effects such as shading and specularities so that the segmentation output interprets better the imaged scene, in terms of the objects present in it. More precisely, C^{3}S analyzes the coupling between colour channels, which, on the basis of the study presented in (Ortiz, 2005), should always happen in uniform reflectance areas, so that when the coupling is violated is because of a reflectance transition arising there, which in turn corresponds to a surface material change and probably to an object transition.

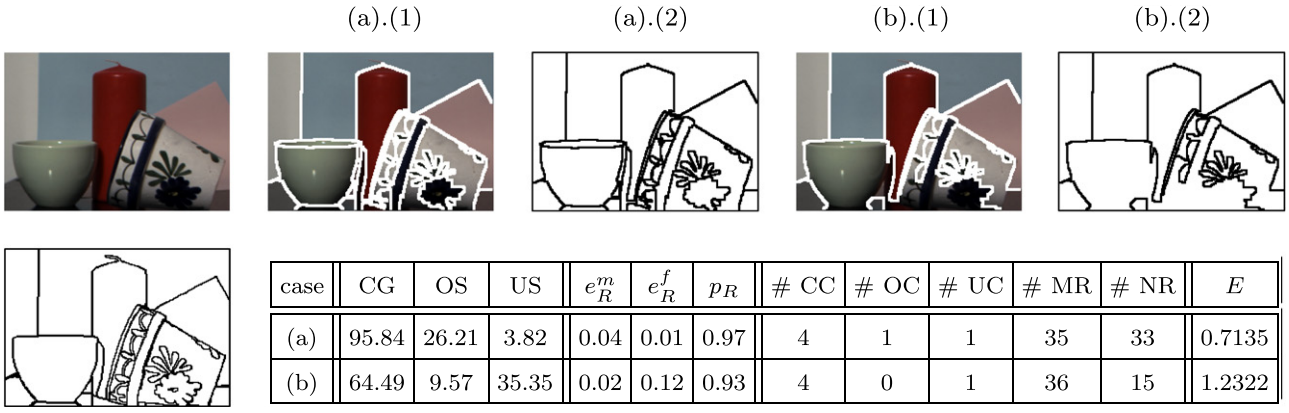
6. Conclusions

By way of conclusion, three new measures for assessing the performance of image segmentation algorithms have been provided. They focus on the classification facet of a segmentation task by means of the overlapping area matrix, and are intended to measure the level of fragmentation of the segmentation output. In order to avoid pathological cases, a tolerance percentage related with the level of pixel dispersion in the segmentation output is introduced in the definitions of the new measures.

On the other hand, from a global point of view, this paper has discussed about different aspects of other measures also based on the overlapping area matrix. Summing up:

- It is clear that global performance measures do not provide the level of detail which can be necessary to evaluate correctly the behaviour of segmentation algorithms. Besides, since the different measures are combined in a single expression, a weight has to be chosen for every term, what is clearly not trivial.
- In the same line as the previous point, mixing spatial coincidence information with distance yields performance measures which are difficult to interpret.
- Nevertheless, the previous remarks are not intended to discard the measures reviewed in this paper as well as others not appearing here. However, it seems better to use them as secondary measures providing additional information to CG, OS and US values.

Finally, it is clear that for characterizing completely the performance of a segmentation algorithm a certain measure about the location of region contours should be provided. The measures proposed in this paper do not cover this facet of the segmentation problem, although this deficiency can be alleviated by taking into account other mea-



case	CG	OS	US	e_R^m	e_R^f	p_R	# CC	# OC	# UC	# MR	# NR	E
(a)	95.84	26.21	3.82	0.04	0.01	0.97	4	1	1	35	33	0.7135
(b)	64.49	9.57	35.35	0.02	0.12	0.93	4	0	1	36	15	1.2322

Fig. 3. Discrepancy measures for two different segmentation outputs of the same real image. The ground truth and the original image appear in the leftmost part of the figure, while the two outputs are labeled with (a) and (b) in the first row, and are given in the form of (1) region contours superimposed over the real image and (2) only region contours. In the table, the first group of cells is for CG, OS and US, while the rest of groups are for, respectively, (Huang and Dom, 1995; Hoover et al., 1996; and Mezaris et al., 2003) measures. For CG, OS and US, and for the measures by Hoover et al., the parameter of tolerance was set to, respectively, $p = 90\%$ and $T = 0.90$.

sures borrowed from the evaluation of edge maps. For instance, Ortiz (2005) suggests using the Baddeley metric (Baddeley, 1992) to compute a measure of the error in the location of reflectance transitions for physics-based segmentation algorithms, given the fact it has been deemed as the best evaluation metric for edge maps in a recent survey (Fernandez-Garcia et al., 2004).

Acknowledgement

This study has been partially supported by project DPI2005-09001-C03-02 and FEDER funds.

References

Alexander, D., Buxton, B., 1997. Modelling of single mode distributions of colour data using directional statistics. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 319–324.

Baddeley, A., 1992. An error metric for binary images. In: Förstner, W., Ruwiedel, S. (Eds.), Robust Computer Vision: Quality of Vision Algorithms, Wichmann, Karlsruhe, Proc. Internat. Workshop on Robust Computer Vision, pp. 59–78.

Beauchemin, M., Thomson, K., 1997. The evaluation of segmentation results and the overlapping area matrix. Internat. J. Remote Sens. 18 (18), 3895–3899.

Borsotti, M., Campadelli, P., Schettini, R., 1998. Quantitative evaluation of color image segmentation results. Pattern Recognition Lett. 19, 741–747.

Bowyer, K.W., Kranenburg, C., Dougherty, S., 2001. Edge detector evaluation using empirical ROC curves. Computer Vision and Image Understanding 84 (1), 77–103.

Edwards, G., Taylor, C., Cootes, T., 1999. Improving identification performance by integrating evidence from sequences. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, vol. I, pp. 486–491.

Fernandez-Garcia, N., Medina-Carnicer, R., Carmona-Poyato, A., Madrid-Cuevas, F., Prieto-Villegas, M., 2004. Characterization of empirical discrepancy evaluation measures. Pattern Recognition Lett. 25 (1), 35–47.

Forbes, L., Draper, B., 2000. Inconsistencies in edge detector evaluation. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, vol. II, pp. 398–404.

Gevers, T., 2002. Adaptive image segmentation by combining photometric invariant region and edge information. IEEE Trans. Pattern Anal. Machine Intell. 24 (6), 848–852.

Gevers, T., 2004. Robust segmentation and tracking of colored objects in video. IEEE Trans. Circ. Syst. Video Technol. 14 (6), 776–781.

Gevers, T., Stokman, H., 2003a. Classifying color edges in video into shadow-geometry, highlight, or material transitions. IEEE Trans. Multimedia 5 (2), 237–243.

Gevers, T., Stokman, H., 2003b. Robust photometric invariant region detection in multispectral images. Internat. J. Comput. Vision 53 (2), 135–151.

Goumeidane, A., Khamadja, M., Belaroussi, B., Benoit-Cattin, H., Odet, C., 2003. New discrepancy measures for segmentation evaluation. In: Proc. IEEE Internat. Conf. on Image Processing, vol. II, pp. 411–414.

Haralick, R., Shapiro, L., 1985. Image segmentation techniques. Comput. Vision Graphics Image Process. 29, 100–132.

Heyden, F., 1989. Evaluation of edge detection algorithms. In: Proc. IEEE Internat. Conf. on Image Processing, pp. 618–622.

Hoover, A., Jean-Baptiste, G., Jiang, X., Flynn, P., Bunke, H., Goldgof, D., Bowyer, K., Eggert, D., Fitzgibbon, A., Fisher, R., 1996. An experimental comparison of range image segmentation algorithms. IEEE Trans. Pattern Anal. Machine Intell. 18 (7), 673–689.

Huang, Q., Dom, B., 1995. Quantitative methods of evaluating image segmentation. In: Proc. IEEE Internat. Conf. on Image Processing, vol. 3, pp. 53–56.

Levine, M., Nazif, A., 1982. An experimental rule-based system for testing low level segmentation strategies. In: Preston, K., Uhr, L. (Eds.), Multicomputers and Image Processing: Algorithms and Programs. Academic Press, New York, pp. 149–160.

Levine, M., Nazif, A., 1985. Dynamic measurement of computer generated image segmentations. IEEE Trans. Pattern Anal. Machine Intell. 7, 155–164.

Lewis, H., Brown, M., 2001. A generalized confusion matrix for assessing area estimates from remotely sensed data. Int. J. Remote Sens. 22 (16), 3223–3235.

Lim, Y., Lee, S., 1990. On the color image segmentation algorithms based on the thresholding and fuzzy *C*-means technique. Pattern Recognition 23, 935–952.

Liu, J., Yang, Y.-H., 1994. Multiresolution color image segmentation. IEEE Trans. Pattern Anal. Machine Intell. 16 (7), 689–700.

Mezaris, V., Kompatsiaris, I., Strintzis, M., 2003. Still image objective segmentation evaluation using ground truth. In: Kovář, B., Přikryl, J., Vlček, M. (Eds.), Proc. 5th COST 276 Workshop on Information and Knowledge Management for Integrated Media Communication, pp. 9–14.

Odet, C., Belaroussi, B., Cattin, H., 2002. Scalable discrepancy measures for segmentation evaluation. In: Proc. IEEE Internat. Conf. on Image Processing, vol. I, pp. 785–788.

Ortiz, A., 2005. New segmentation and edge detection methods using physics-based models of image formation. Ph.D. thesis, Department of Mathematics and Computer Science, University of the Balearic Islands (UIB), Spain.

Pal, N.R., Bhandari, D., 1993. Image thresholding: Some new techniques. Signal Process. 33, 139–158.

Pal, N.R., Pal, S.K., 1989. Entropic thresholding. Signal Process. 16, 97–108.

Pignalberi, G., Cucchiara, R., Cinque, L., Levialdi, S., 2003. Tuning range image segmentation by genetic algorithm. EURASIP J. Appl. Signal Process. 2003 (8), 780–790.

Pratt, W., 1978. Digital Image Processing. John Wiley and Sons.

Ramesh, V., Haralick, R., 1992. Performance characterization of edge detectors. In: Proc. SPIE Applications of AI X, vol. 1708, pp. 252–266.

Rees, G., Wright, W., Greenway, P., 2002. ROC method for the evaluation of multi-class segmentation/classification algorithms with infrared imagery. In: Proc. British Machine Vision Conference, Poster session.

Roman-Roldan, R., Gomez-Lopera, J., Atae-Allah, C., Martinez-Aroza, J., Luque-Escamilla, P., 2001. A measure of quality for evaluating methods of segmentation and edge detection. Pattern Recognition 34 (5), 969–980.

Sahoo, P., Soltani, S., Wong, A., Chen, Y., 1988. A survey on thresholding techniques. Comput. Vision Graphics Image Process. 41, 233–260.

Shin, M., Goldgof, D., Bowyer, K., 2001. Comparison of edge detector performance through use in an object recognition task. Computer Vision and Image Understanding 84 (1), 160–178.

Southall, B., Buxton, B., Marchant, J., Hague, T., 2000. On the performance characterisation of image segmentation algorithms: A case study. In: Proc. European Conf. on Computer Vision, vol. II, pp. 351–365.

Stehman, S., 1997. Selecting and interpreting measures of thematic classification accuracy. Remote Sens. Environ. 62 (1), 77–89.

Strasters, K., Gerbrands, J., 1991. Three-dimensional image segmentation using a split, merge and group approach. Pattern Recognition Lett. 12, 307–325.

Weszka, J., Rosenfeld, A., 1978. Threshold evaluation techniques. IEEE Trans. Systems Man Cybernet. 8, 622–629.

Yasnoff, W., Bacus, J., 1984. Scene segmentation algorithm development using error measures. Anal. Quant. Cytol. 6, 45–58.

Yasnoff, W., Mui, J., Bacus, J., 1977. Error measures for scene segmentation. Pattern Recognition 9, 217–231.

Yitzhaky, Y., Peli, E., 2003. A method for objective edge detection evaluation and detector parameter selection. IEEE Trans. Pattern Anal. Machine Intell. 25 (8), 1027–1033.

Zhang, Y., 1993. Segmentation evaluation and comparision: A study of various algorithms. In: Haskell, B.G., Hang, H.-M. (Eds.), Proc. SPIE Visual Communication and Image Processing, vol. 2094, pp. 801–812.

Zhang, Y., 1995. Influence of image segmentation over feature measurement. Pattern Recognition Lett. 16, 201–206.

Zhang, Y., 1996. A survey on evaluation methods for image segmentation. Pattern Recognition 29 (8), 1335–1346.

Zhang, Y., Gerbrands, J., 1992. Segmentation evaluation using ultimate measurement accuracy. In: Proc. SPIE, vol. 1657, pp. 449–460.

Zhang, Y., Gerbrands, J., 1994. Objective and quantitative segmentation evaluation and comparison. EURASIP J. Appl. Signal Process. 39, 43–54.