# Design of statistical measures for the assessment of image segmentation schemes

Marc Van Droogenbroeck and Olivier Barnich *

Department of Electricity, Electronics and Computer Science,
Institut Montefiore B-28, Sart Tilman, B-4000 Liège, Belgium

**Abstract** Image segmentation is discussed for years in numerous papers, but assessing its quality is mainly dealt with in recent works. Quality assessment is a primary concern for anyone working towards better segmentation tools. It both helps to objectively improve segmentation techniques and to compare performances with respect to other similar algorithms.
In this paper we use a statistical framework to propose statistical measures capable to describe the performances of a segmentation scheme. All the measures rely on a ground-truth segmentation map that is supposed to be known and that serves as a reference when qualifying the results of any segmentation tool. We derive the analytical expression of several transition probabilities and show how to calculate them. An important conclusion from our study, often overlooked, is that performances can be content dependent, which means that one should adapt a measure to the content of an image.

## 1 Introduction

Segmentation is one the most difficult task in automatic image analysis. It consists in partitioning an image into objects (segments) homogeneous with respect to a specific property. Many algorithms for segmentation have been proposed over the years and this number still continues to raise. One of the reasons for this proliferation of techniques is that no segmentation technique offer enough universality to meet the requirements of a broad family of applications.

While the development of new segmentation techniques has attracted significant attention, fewer efforts have been spent on their evaluation. Some could also argue that no satisfactory evaluation measure has been proposed so far and that the discipline is still in its infancy.

In [1] ZHANG reviews some methods for segmentation evaluation. He divides the family of evaluation methods into two categories:

1. the *analytical* methods, which evaluates the properties and the principles of segmentation algorithms,
2. and the *empirical* methods, that judge algorithms by applying them to test images and by measuring the results.

According to ZHANG [1], empirical methods can been further divided into two types: *goodness* methods and *discrepancy* methods. In the first category results are qualified according to human intuition and judged by the values of goodness measures. In the second category some segmentation references, called *ground-truth* maps, that represent expected results are given, and results are compared with these references by counting the difference.

In [2] a procedure for evaluating the intrinsic quality of segmentation masks in video sequences where the existence of ground-truth masks is assumed is proposed. The procedure adopted in MPEG-4 for objective and automatic evaluation is described in [3]. It assumes that ground-truth mask are available and performs two types of analysis on segmented sequences that measure the spatial accuracy and the temporal coherence. As an alternative ERDEM *et al.* [4] propose measures to evaluate the performance of video segmentation without ground-truth segmentation maps.

* This work was supported by the Belgian Walloon Region (http://www.wallonie.be), under the CINEMA project.

All in all there is a conceptual difference between the two tasks of *quality assessment* and *benchmarking*, often confused in literature. On the one hand researchers want to develop efficient segmentation algorithms. In order to tune their algorithms they need a criterion capable to measure the quality of the segmentation. On the other hand it is sometimes useful to compare the performances of competitive algorithms targeting the same applications. This second task is known as *benchmarking*. It is defined as an objective measure of the performances of a segmentation algorithm obtained by evaluating them on test data.

In practical terms both tasks are related. A developer may start by tuning an algorithm, then gradually transform it to end up with a completely new algorithm. Quality assessment then turns off to a sort of poor benchmarking. One may then object that this methodology does not prove the reliability of the new algorithm nor that the final algorithmic expression has the best performances. Like COURTNEY and THACKER [5], we believe that practical approaches neglect the important role that statistics must play in algorithm development. It is even worse to notice that most articles, like [6], simply ignore any statistical issues. To the contrary, papers dealing with image quality assessment do consider the statistical distribution of results [7]. For all these reasons our work concentrates on measures that provide statistical insights on the segmentation performances.

In the following we investigate the statistical significance of a discrepancy method. In Section 2 we develop a framework for describing a segmentation result. This model leads to statistical measures that are defined and discussed in Section 3. From our study it appears that evaluating the quality of segmentation depends on the data and that one has to adapt the measures to the size of the segmentation maps. These conclusions are presented in Section 4.

## 2 Statistical model for assessing segmentation techniques

Let $x$ be the location of a pixel inside the image that can be of any type (a 2D flat image, a volumetric 3D image, or an image flow like a video). Generally speaking image segmentation produces a region map, in which each pixel is labeled with a number designating the region to which it was assigned. In the following we restrict the number of regions to a single object, that might be composed of several disconnected parts, and a background.

There are various ways to generate segmentation references. Ideally a reference is specified on the base of a perfect segmentation process. If this last is available the need to measure the quality of segmentation techniques is rather low. One possible alternation is to use synthetic images made by the superposition of an object (the blue screen technique can help producing a realistic object) on a real background. This principle is illustrated in Figure 1. In this example, some images (a) were captured after a snow storm. Then the white color was made transparent after thresholding and superimposed on a real background (b) to produce a realistic scene with a perfectly known segmentation map (c). A more work intensive method

(a) (b) (c)

**Figure1.** Process to build ground-truth segmentation maps by superimposing a thresholded image on a natural background.

is to segment images by hand.

Ground-truth segmentation maps are not widely available and one often has to build his own. A segmentation dataset has been distributed by a research group at the Bekerley University [8]. From this large dataset one can download files which contain the probability of a boundary at each location of 200 training and 100 test images.

In the following we assume that there exists perfect segmentation maps. The known background is denoted by $b[x]$, and $f[x]$ is the image captured by the camera. If no object is superimposed on the background then $b[x] = f[x]$ for any $x$. To the contrary, when an object is added to the scene, the previous equality holds for some pixels but not all of them anymore. In this case, the background is masked by a function denoted $m[x]$. We define $m[x] = 0$ when the background is visible and $m[x] = 1$ when an object in the foreground hides the background. If per coincidence the color of the object is identical to the background color, i.e. $b[x] = f[x]$, we still consider that the background is masked and therefore that $m[x] = 1$. Note that this simple model does not discard transparent objects like windows.

With the aforementioned model and notations the segmentation algorithm has to process

$$f[x] = m[x]f[x] + (1 - m[x])b[x] \tag{1}$$

where

$$m[x]f[x] = o[x] \tag{2}$$

denotes the superimposed object. From all the functions, $b[x]$ is known and $f[x]$ is observed. If we choose a non-cooperative design scheme, the algorithm has no prior knowledge of $o[x]$, nor of $m[x]$. Despite that relation (1) holds for any $x$, there is not enough information for the algorithm to recover $o[x]$ or $m[x]$.

One of the key techniques to segment an image is *background substraction* [9,10]. A background is first estimated, by time integration for example, and then the estimated background $\widehat{b}[x]$ is compared to $f[x]$. If one assumes that noise has been filtered out, a simple decision rule states that if $f[x] \neq \widehat{b}[x]$ then $m[x] = 1$ and $f[x] = o[x]$. Clearly this technique does not suffice as $f[x] = \widehat{b}[x]$ does not imply that $m[x] = 0$. In order words background substraction produces underestimated object surfaces. Therefore background substraction is usually combined to an object tracking algorithm. This avoids the two main drawbacks of background substraction techniques : the progressive inclusion in the background of static objects and the non-detection of objects that have the same color than the background (like transparent objects).

### 2.1 A statistical interpretation of the segmentation process

Regardless of whether the background is known or not the segmentation process may be seen as a stochastic process. Let us consider image segmentation as a two states pixel classification process $M[x]$. For any location $x$, $M[x]$ is a random variable equal to

- $1$ when $x$ belongs to a foreground object $\mathcal{O}$, and
- $0$ when $x$ is included in the background $\mathcal{B}$.

When the ground-truth segmentation map is not available, $M[x]$ can only be described in terms of probabilities characterizing two possible outcomes: $M[x] = 1$ or $M[x] = 0$. Let the probabilities of these events be $p(x \in \mathcal{O})$ and $p(x \notin \mathcal{O}) = p(x \in \mathcal{B})$. For simplicity these probabilities are denoted $p(o)$ and $p(b)$ respectively. Obviously $x \in O$ or $x \in B$, so that $p(o) = 1 - p(b)$.

The role of segmentation is to estimate the masking function $M[x]$, hopefully as close as possible to the real segmentation mask. In practical terms we have to estimate the function $\widehat{M}[x]$ which should be equal to $M[x]$ almost everywhere. Since a perfect match is not achievable, we have to model the segmentation process with some probabilities. Let $p_s(o)$ and $p_s(b)$ be the probabilities for a pixel to be classified as a foreground object or as a background respectively. The probability $p_s(o)$ sums the probability of two cases: $x$ belongs to the object or, although $x$ is in the background, it has been assigned to the object.

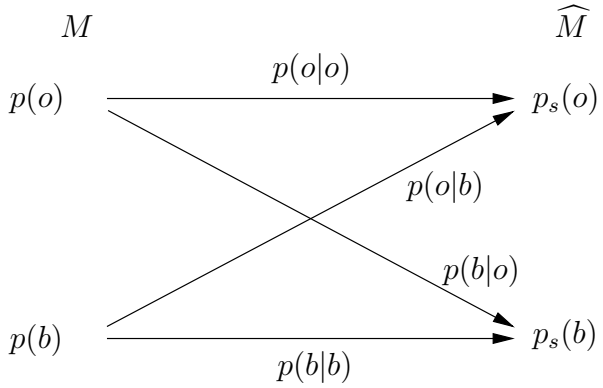


**Figure2.** Binary model for segmentation.

Figure 2 shows the model used in the following. Let us consider a given location $x$. The input, drawn on the left-hand side, represents the original two possible states (and probabilities) for the mask. A referenceless segmentation produces an estimated binary value $\widehat{M}[x]$; it is drawn on the right -hand side of Figure 2.

As suggested by MARTIN [11], we could compute the mutual information $I$ between $\widehat{M}$ and the ground-truth map $M$ which is a global measure. Our model puts the focus on several probabilities rather than on a global measure. The model is characterized by the set of possible original states, the set of possible outcomes, and a set of conditional probabilities also called transition probabilities. For example, $p(o|b)$ is the probability of an error for a background pixel to be labelled as an object.

As a consequence

$$p_s(o) = p(o|o)p(o) + p(o|b)p(b). \tag{3}$$

A similar relation yields for $p_s(b)$:

$$p_s(b) = p(b|b)p(b) + p(b|o)p(o). \tag{4}$$

Segmentation errors originate from the diagonal probabilities $p(b|o)$ and $p(o|b)$. The larger these probabilities, the larger the segmentation error rate will be. The probability of error $p_e$ for a two-class problem can be defined by [12]

$$p_e = p(b|o)p(o) + p(o|b)p(b) \tag{5}$$

where $p(o)$ and $p(b)$ are viewed as *a priori* probabilities. An extension of $p_e$ for multi-class problems can be found in [13].

## 3 Statistical discrepancy measures

The idea of computing discrepancy based on the number of misclassified pixels is also reflected in some edge-detection evaluation schemes. The most elementary way to match reference region boundaries and computer generated boundaries is to compute the percentage of boundary pixels that overlap. However corresponding boundary pixels, though typically close to one another, often do not overlap. In [14] PAGLIERONI introduces some tolerance on spatial overlap that states that patterns are said to be potentially co-occurrent as long as they are separated by less than some tolerance distance. The purpose of spatial tolerance is to take into account the fact that edges are difficult to localize exactly in continuous images. But this introduces the need for a local correlation model whose validation is difficult from a practical point of view. As we are dealing with a binary mask, we clearly would like to favour a simpler spatial model.

### 3.1 Statistical assumptions

The statistical segmentation model summarized in Figure 2 is location-dependent. Indeed classification probabilities as well as transition probabilities are related to the position of a pixel in the image. We will nevertheless suppose that $M[x]$ is spatially stationary in the wide sense, which implies that its mean is constant. This assumption is debatable but if an object moves equally over the image plane or inside a 3D volume, wide-sense stationarity is acceptable.

Additionally we assume that $M[x]$ is *mean-ergodic*. As a consequence the constant local mean theoretically equals the average over the observation volume. In order words, if $x$ is observed over $\mathcal{D} \in \mathbb{R}^n$ then

$$\mu_M = E\{M\} = \frac{1}{\sharp(\mathcal{D})} \int_{\mathcal{D}} m(x)\, dx \tag{6}$$

where $\sharp(\mathcal{D})$ is the cardinality of $\mathcal{D}$. Again this is acceptable if $\mathcal{D}$ is large enough, which is the case for a usual image size, like a $640 \times 480$ VGA image.

### 3.2 Estimation of the means

Assuming wide-sense stationarity and ergodicity in the mean it is possible to compute the means of $M$ and $\widehat{M}$. The statistical mean of $M$ is equal to

$$\mu_M = E\{M\} = 1 \times p(o) + 0 \times p(b) = p(o). \tag{7}$$

If the ground-truth segmentation map is known, the cardinality of the objects by $\sharp(o)$ is easily computed so that $\mu_M = p(o)$ is nothing but the ratio of $\sharp(o)$ to the image size $\sharp(\mathcal{D})$:

$$\mu_M = \frac{\sharp(o)}{\sharp(\mathcal{D})}. \tag{8}$$

Computing the mean of $\widehat{M}$ is not as straightforward. Analytically,

$$\mu_{\widehat{M}} = E\left\{\widehat{M}\right\} = 1 \times p_s(o) + 0 \times p_s(b) = p_s(o). \tag{9}$$

Using equation (3) this yields

$$\mu_{\widehat{M}} = p(o)p(o|o) + p(b)p(o|b). \tag{10}$$

Because of segmentation inaccuracies, $\mu_M \neq \mu_{\widehat{M}}$.

### 3.3 Probabilistic quality measures

Basically all four transition probabilities drawn on Figure 2 are interesting measures but for different reasons:

- $p(o|o)$ directly relates to the aim of segmentation,
- $p(b|b)$ is a useful measure for gauging the quality of any background detection tool, and
- $p(o|b)$ and $p(b|o)$ determine the overall segmentation errors.

Like for relation (8), the mean of $\widehat{M}$ can be estimated by counting the number of object pixels divided by the image size:

$$\mu_{\widehat{M}} = \frac{\sharp(o_s)}{\sharp(\mathcal{D})} \tag{11}$$

where $o_s$ represents the objects after segmentation. Accordingly we can easily compute a valid statistical estimate of $\mu_{\widehat{M}}$, and its value is assumed to be known hereafter.

Let us now reconsider equation (10). A substitution of $p(o)$ and $p(b)$ by their values yields

$$\mu_{\widehat{M}} = \frac{\sharp(o)}{\sharp(\mathcal{D})} p(o|o) + \left(1 - \frac{\sharp(o)}{\sharp(\mathcal{D})}\right) p(o|b) \tag{12}$$

Further simplifications are needed to isolate $p(o|o)$ and $p(o|b)$. We will first consider the case of large objects and then the case of small objects as the number of samples impact of the statistical significance of the estimates.

**First case: large objects.** If the objects occupy a large portion of the image and the segmentation performs relatively well –it would be pointless to address the performances of a poor segmentation technique!–, $p(o|b) \ll p(o|o)$. Consequently $\mu_{\widehat{M}}$ reduces to

$$\mu_{\widehat{M}} \simeq \frac{\sharp(o)}{\sharp(\mathcal{D})} p(o|o). \tag{13}$$

This provides the value of $p(o|o)$:

$$p(o|o) \simeq \mu_{\widehat{M}} \frac{\sharp(\mathcal{D})}{\sharp(o)} = \frac{\sharp(o_s)}{\sharp(o)}. \tag{14}$$

So two simple counting processes on the segmentation reference and on the real segmentation are sufficient to compute a criterion capable to estimate the object segmentation quality. Note that $p(o|o)$ might be superior to 1 which is theoretically impossible. Therefore we should use a modified criterion, like the absolute value of $1 - p(o|o)$, to evaluate the segmentation performances.

To compute $p(b|b)$ we start with the complementary probability of $p_s(o)$, $1 - p_s(b)$, and replace $p_s(b)$ by its value (see relation 4):

$$p_s(o) = 1 - p(b)p(b|b) - p(o)p(b|o), \tag{15}$$

$$= 1 - \left(1 - \frac{\sharp(o)}{\sharp(\mathcal{D})}\right) p(b|b) - \frac{\sharp(o)}{\sharp(\mathcal{D})} p(b|o). \tag{16}$$

Remember that $\mu_{\widehat{M}} = p_s(o)$ and considering that the large objects hypothesis also implies that $p(b|o) \ll p(b|b)$, we get after some simplifications,

$$p(b|b) = \frac{\sharp(\mathcal{D}) - \sharp(o_s)}{\sharp(\mathcal{D}) - \sharp(o)}. \tag{17}$$

To determine the missing diagonal transition probabilities we use the coherence relationship between probabilities originated from the same original event: $p(b|o) + p(o|o) = 1$. Therefore

$$p(b|o) = 1 - p(o|o) = \frac{\sharp(o) - \sharp(o_s)}{\sharp(o)}. \tag{18}$$

Likewise,

$$p(o|b) = 1 - p(b|b) = \frac{\sharp(o_s) - \sharp(o)}{\sharp(\mathcal{D}) - \sharp(o)}. \tag{19}$$

**Second case: small objects.** Expression (14) is inadequate when the objects occupy a negligible part of the image. More precisely, if $\sharp(o) \ll \sharp(\mathcal{D})$, then

$$\mu_{\widehat{M}} \simeq \frac{\sharp(o)}{\sharp(\mathcal{D})} p(o|o) + p(o|b). \tag{20}$$

We now consider that the segmentation is symmetric, i.e. that $p(b|o) = p(o|b)$; a non-symmetric segmentation would otherwise be biased towards the foreground or the background and would lead to unacceptable results in the case of small objects. As $p(o|o) = 1 - p(b|o)$,

$$\mu_{\widehat{M}} \simeq 1 + \left( \frac{\sharp(o)}{\sharp(\mathcal{D})} - 1 \right) p(o|o) \tag{21}$$

so that, after further simplifications,

$$p(o|o) \simeq \frac{1 - \mu_{\widehat{M}}}{\left(1 - \frac{\sharp(o)}{\sharp(\mathcal{D})}\right)} \simeq \left(1 - \mu_{\widehat{M}}\right) \left(1 + \frac{\sharp(o)}{\sharp(\mathcal{D})}\right) = 1 + \frac{\sharp(o) - \sharp(o_s)}{\sharp(\mathcal{D})} + \frac{\sharp(o)\sharp(o_s)}{\sharp(D)^2}. \tag{22}$$

$\sharp(o_s)$ and $\sharp(o)$ are small compared to $\sharp(D)$, so that the quadratic term is negligible and therefore

$$p(o|o) \simeq 1 + \frac{\sharp(o) - \sharp(o_s)}{\sharp(\mathcal{D})}. \tag{23}$$

This probability gets very close to 1 as $\sharp(o_s)$ tends to 0. We then obtain $p(b|o)$ on the spot:

$$p(b|o) = 1 - p(o|o) = \frac{\sharp(o_s) - \sharp(o)}{\sharp(\mathcal{D})}, \tag{24}$$

which is also the value of $p(o|b)$. Again, by symmetry, $p(b|b) = p(o|o)$. The probability of error is then

$$p_e = \frac{\sharp(o_s) - \sharp(o)}{\sharp(\mathcal{D})}. \tag{25}$$

**Discussion.** In [1] ZHANG concludes that evaluation methods based on discrepancy measures are more powerful than evaluation methods using other measures. Moreover he compared several discrepancy measures to rank their ability to discriminate the overall quality. While there are many discrepancy measures, it appears that $p_e$ is one of the best quality measure. Subsequently we can rely on this conclusion and do not have to validate $p_e$ as a useful measure.

In the meanwhile we have computed additional probabilities that can have their relevance for certain segmentation purposes. A possible measure could be any weighted summation of $p(o|o)$, $p(b|b)$, $p(o|b)$, and $p(b|o)$. But one has to be careful with the interpretation of such a criteria because its statistical suitability is questionable. A sounder approach consists in comparing the probabilities separately, but then one has to cope with multiple criteria.

All four transition probabilities offer different insights on the quality of the segmentation but we can notice that the transition probabilities, in particular $p(o|b)$, seem less sensitive to the object size. They are also analytically very close to $p_e$. Therefore, if one is looking for a measure independent of the size of the object, we recommend $p(o|b)$. On the other hand $p(o|o)$ and $p(b|b)$ can also be useful if one wants a measure that changes its discriminating power with the foreground size.

## 4 Conclusions

In this paper we have derived several statistical measures to assess the quality of a segmentation algorithm with respect to a ground-truth reference. The measures are expressed in terms of transition probabilities; their analytical expression are summarized in Table 1.

From these values we can conclude that:

– appropriate estimates of transition probabilities depend on the data content.

| | **Large objects** | **Small objects** |
|---|---|---|
| Assumptions: | $p(b\|o) \ll p(b\|b),\ p(o\|b) \ll p(o\|o)$, and $\sharp(o) \gg$ | $\sharp(o) \ll \sharp(\mathcal{D})$, and $p(b\|o) = p(b\|o)$ |
| $p(o\|o)$ | $\frac{\sharp(o_s)}{\sharp(o)}$ | $1 + \frac{\sharp(o)-\sharp(o_s)}{\sharp(\mathcal{D})}$ |
| $p(o\|b)$ | $\frac{\sharp(o_s)-\sharp(o)}{\sharp(\mathcal{D})-\sharp(o)}$ | $\frac{\sharp(o_s)-\sharp(o)}{\sharp(\mathcal{D})}$ |
| $p(b\|o)$ | $\frac{\sharp(o)-\sharp(o_s)}{\sharp(o)}$ | $\frac{\sharp(o_s)-\sharp(o)}{\sharp(\mathcal{D})}$ |
| $p(b\|b)$ | $\frac{\sharp(\mathcal{D})-\sharp(o_s)}{\sharp(\mathcal{D})-\sharp(o)}$ | $1 + \frac{\sharp(o)-\sharp(o_s)}{\sharp(\mathcal{D})}$ |
| $p_e$ | $p(o\|b)p(b) + p(b\|o)p(o)$ | $\frac{\sharp(o_s)-\sharp(o)}{\sharp(\mathcal{D})}$ |

**Table1.** Statistical measures for assessing the quality of a segmentation technique.

- the statistical relevance of these estimates varies with the size of the object in the foreground. Analytical expressions show that $\sharp(o)$ and $\sharp(o_s)$ appears in all the transition probabilities. Since $\sharp(o_s)$ is an estimate, it would be interesting to investigate the impact of the variability of $\sharp(o_s)$ on the probabilities.
- we recommend $p(o|b)$ as a assessment criterion insensitive to the size of the object.

Further works are needed to examine the influence of several parameters of the model. Still we have established the unsuitability to trust a single criterion all over the foreground object size range.

## References


1. Zhang, Y.: A survey on evaluation methods for image segmentation. Pattern Recognition **29** (1996) 1335–1346
2. Villegas, P., Marichal, X.: Perceptually-weighted evaluation criteria for segmentation masks in video sequences. IEEE Transactions on Image Processing **13** (2004) 1092–1103
3. Wollborn, M., Mech, R.: Refined procedure for objective evaluation of VOP generation algorithms. Doc. ISO/IEC JTC1/SC29/WG11 MPEG98/3448, Fribourg, Switzerland (1997)
4. Erdem, Ç., Sankur, B., Tekalp, M.: Performance measures for video object segmentation and tracking. IEEE Transactions on Image Processing **13** (2004) 937–951
5. Courtney, P., Thacker, N.: Performance characterisation in computer vision: The role of statistics in testing and design. In: Imaging and Vision Systems: Theory, Assessment and Applications, NOVA Science Books (2001)
6. Zhang, Y.: Evaluation and comparison of different segmentation algorithms. Pattern Recognition Letters **18** (1997) 963–974
7. Wang, Z., Bovk, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13** (2004) 600–612
8. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proc. 8th Int'l Conf. Computer Vision. Volume 2. (2001) 416–423
9. Li, L., Huang, W., Gu, Y.H., Tian, Q.: Statistical modeling of complex backgrounds for foreground object detection. IEEE Transactions on Image Processing **13** (2004) 1459–1472
10. Radke, R., Andra, S., Al-Kofahi, O., Roysam, B.: Image change detection algorithms: a systematic survey. IEEE Transactions on Image Processing **14** (2005) 294–307
11. Martin, D.: An Empirical Approach to Grouping and Segmentation. PhD thesis, University of California, Berkeley (2002)
12. Lee, S., Chung, S., Park, R.: A comparative performance study of several global thresholding techniques for segmentation. Computer Vision, Graphics, and Image Processing **52** (1990) 171–190
13. Lim, Y., Lee, S.: On the color image segmentation algorithm based on the thresholding and the fuzzy C-means techniques. Pattern Recognition **23** (1990) 935–952
14. Paglieroni, D.: Design considerations for image segmentation quality assessment measures. Pattern Recognition **37** (2004) 1607–1617