Toward a Generic Evaluation of Image Segmentation

Jaime S. Cardoso, *Student Member, IEEE,* and Luís Corte-Real, *Member, IEEE*

***Abstract*—Image segmentation plays a major role in a broad range of applications. Evaluating the adequacy of a segmentation algorithm for a given application is a requisite both to allow the appropriate selection of segmentation algorithms as well as to tune their parameters for optimal performance. However, objective segmentation quality evaluation is far from being a solved problem. In this paper, a generic framework for segmentation evaluation is introduced after a brief review of previous work. A metric based on the distance between segmentation partitions is proposed to overcome some of the limitations of existing approaches. Symmetric and asymmetric distance metric alternatives are presented to meet the specificities of a wide class of applications. Experimental results confirm the potential of the proposed measures.**

***Index Terms*—Image segmentation, objective segmentation assessment, segmentation quality evaluation.**

I. Introduction

AUTHORS currently working in the field of low-level image segmentation frequently point out the need for a standard quality measure that would allow both the evaluation and comparison of all segmentation procedures available. This need arises from the ill posedness (in the sense of Hadamard, [1]) of the image segmentation problem: For the same image, the optimum segmentation can be different, depending on the application.

Automatic segmentation is, therefore, a problem without a general solution, at least at the current state-of-the-art. A standard quality measure, if available, could be applied to automatically provide a ranking among different segmentation algorithms or to optimally set the parameters of a given algorithm, under a predefined framework.

Several methods have been proposed to evaluate the quality of segmentation algorithms. Next, we will present the main ideas underlying these methods.

A. Evaluation Methods for Image Segmentation

In the often-cited article by Zhang [2], evaluation methods are broadly divided into two categories: **analytical methods** and **empirical methods**. "The analytical methods directly examine and assess the segmentation algorithms themselves by analyzing their principles and properties. The empirical methods indirectly judge the segmentation algorithm by applying them to test images and measuring the quality of segmentation results."

Manuscript received October 24, 2003; revised September 6, 2004. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Joachim M. Buhmann.

The authors are with Departamento de Engenharia Electrotécnica e de Computadores da Faculdade de Engenharia da Universidade do Porto/INESC Porto, Porto, Portugal (e-mail: jaime.cardoso@inescporto.pt; lreal@inescporto.pt).

Digital Object Identifier 10.1109/TIP.2005.854491

Although using analytical methods to evaluate segmentation algorithms avoids the implementation of these algorithms (and so they do not suffer from influences caused by the arrangement of evaluation experiments as the empirical methods do), they have not received much attention mainly because of the difficulty to compare algorithms solely by analytical studies. The analytical methods in the literature work only with some particular models or properties (see Liedtke [3] and Abdou [4]).

Empirical methods are further classified into two types: **goodness methods** and **discrepancy methods**.

In the empirical goodness methods some desirable properties of segmented images, often established according to human intuition, about what conditions should be satisfied by an "ideal segmentation," are measured by goodness parameters. The performance of the segmentation algorithms under study is judged by the values of goodness measures. These methods evaluate and rate different algorithms by simply computing some chosen goodness measure based on the segmented image, without requiring the *a priori* knowledge of the reference segmentation. Different types of goodness measures have been proposed. Color uniformity [3], entropy [4], intraregion uniformity [5], [6], inter-region contrast [7], [8], region shape [9], etc., are some of the measures that have been proposed in the literature.

Empirical discrepancy methods are based on the availability of a **reference segmentation**, also called **gold standard** or **ground truth**. The disparity between an actually segmented image and a correctly/ideally segmented image (the gold standard, which is the best expected result) can be used to assess the algorithm's performance. Both images (actually segmented and reference) are obtained from the same input image. The methods in this group take the difference (measured by various discrepancy parameters) between the actually segmented image and the reference one into account, i.e., these methods try to determine how far the actually segmented image is from the reference image. In Section II, we will cover the early proposed methods in this group.

The distinction between empirical discrepancy methods and empirical goodness methods is not so clear cut when we think about the real meaning of selecting a goodness method with the corresponding goodness parameter(s). There is (at least) one segmentation partition that maximizes the adopted goodness measure—call it implicit gold standard. By choosing an appropriate discrepancy measure for all other possible segmentations—a rather artificial measure—we can always mimic the goodness method with the implicit discrepancy measure.

So, the difference is in how we model the reference segmentation and in what point of view seems most useful, rather than in any intrinsic difference between the methods themselves. Probably, a more meaningful name for goodness methods is *empirical with implicit reference*, contrasting with *empirical with*

explicit reference that are the so called empirical discrepancy methods.

Although conceptually similar to discrepancy methods, goodness methods have the advantage of being well suited to integrate unsupervised tools—there is no need to feed the method with any data. Also, our perception of a good segmentation might be easier to convey using these methods.

However, goodness methods also have some drawbacks. By first defining what is going to be measured—the goodness parameters—we can always construct an algorithm that will outperform all the others under the selected evaluation measure. This algorithm would generate the implicit gold standard partition. This may invalidate any assessment at all, this being especially true when similar criteria are used to design the segmentation algorithms as well as to assess their performance—in fact, goodness measures have been used to design segmentation algorithms.

B. Paper Organization

The outline of this paper is as follows. Section II presents a brief review of related work. Section III describes the details of the proposed measure. Section IV addresses some implementation issues and provides experimental results. Finally, the conclusions are drawn in Section V.

II. On the Discrepancy Methods—A Review

Taking a quick snapshot of what have been proposed so far, it is easy to conclude that current discrepancy evaluation methods lack a general and consistent approach.

Yasnoff [10] proposed to take the number of misclassified pixels and their positions into account for computing two measures: The percentage of area misclassified and the pixel distance error. However, this has only been applied to foreground/background segmentation.

A similar approach appears in [11] with figure of merit (FOM) for edge-detection evaluation. This method, applied to image segmentation, looks at the segmentation process as an edge map extractor, being only suitable for these binary edge map images. It also does not give a good general response [12].

In [13] and [14], Zhang suggests the use of the so called "ultimate measurement accuracy": "If the goal of image segmentation is to obtain measurements of object features, the accuracy of these measurements obtained from the segmented images can be used as a quantitative evaluation criteria." Mattana [15] and Huo [16] have followed a similar approach. Although this assumption may be valid in the context of image analysis, more and more applications make use of the regions created in the segmentation process, of which the new object-based compressing standards are just an example.

Chalana's proposal [17] works only for "…a single object from an image."

Betanzos [18] defines an accuracy measure for images with multiple types of objects. However, it only works when not all types of objects are present in the image. It also has to be able to count the correct and false results separately for each type of object.

Hoover [19] uses a region-based method for assessment. Nevertheless, he does not avoid unintuitive ad hoc measures that involve user defined thresholds; [20] continues the work of [19] using the same performance evaluation method; [21] proposes an adapted version of the same measure.

Roldan [12] has introduced a hybrid measure of empirical discrepancy and empirical goodness. This measure is only intended for the evaluation of low error segmentation results using the binary edge map of a segmentation.

Belaroussi [22] proposes a set of localization measures that can be used on a binary image under the knowledge of a binary reference image to evaluate the quality of the segmented edges. Although it was adapted to segmentation region maps in [23], that was only done with background/foreground segmentations.

Everingham [24], more than defining a new measure, attempts to aggregate fitness functions using the Pareto front. Measures such as ours could be used as fitness functions in the proposed methodology.

In [25], and more thoroughly in [26], Martin proposes a very interesting set of measures. Most of these measures—GCE, LCE, and BCE measures—compute the overall distance between two segmentations as the sum of the local inconsistency at each pixel. A novel methodology for judging the quality of a boundary map is also presented. The correspondence procedure, tolerant to small localization errors, resorts to the bipartite matching of "little pieces of boundary, or edgels." All measures are general enough to work with images with several objects and they all achieve excellent results in the collection of test images. However, their behavior is not always the expected, as illustrated later—see Section III-B. This is probably due to their local definition, making it also difficult to predict the performance for complex segmentations.

III. On the Discrepancy Methods—A Generic Approach

As Section II shows, only a few methods actually explore the segments (clusters) obtained from the segmentation process. Most measures are best suited to evaluate edge detection, working directly on the binary image of the regions' contours. Although we can always treat a segmentation as a boundary map, the problem lurks in the simplified use of the edge map, as simply counting the misclassified pixels, on an edge/nonedge basis. But pixels on different sides of an edge are different in the sense that they belong to different regions—that is why it may be more reasonable to use the segmentation partition itself. Realizing this, some authors have introduced "artificial corrections" to improve measures, notably counting the misclassified pixels and weighting the erred pixels according to their distance to the reference.

Most of previously proposed methods, working directly on the segments suffer from several limitations, ranging from the number of objects in the image (foreground/background segmentation; see [17] and [10] to simplifications introduced in order to be able to tackle the problem [18]–[20]). A clear exception is the work of Martin in [25] and [26].

To our knowledge, none of the proposed methods tries to define a reasonable discrepancy measure from the definition of image segmentation.

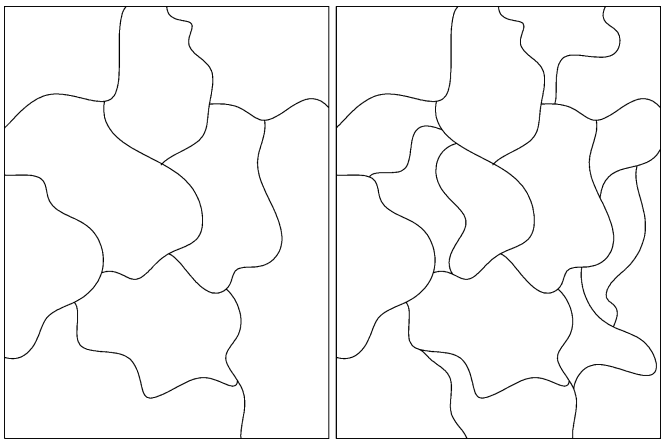


Fig. 1. Right partition is a refinement of the left partition.

Image segmentation is traditionally viewed as a process that partitions the entire image region R into n sub regions, $r_1, r_2, r_3, \ldots, r_n$, as follows.

1) Every pixel belongs to a region—$r_1 \cup r_2 \cup r_3 \cup \cdots \cup r_n = R$.
2) Every region is spatially connected.
3) All regions are disjoint—$r_i \cap r_j = 0, i \neq j$.
4) All pixels in a region satisfy a specified similarity predicate—$P(r_i) =$ true.
5) For any two adjacent regions, r_i and r_j, $P(r_i \cup r_j) =$ false, where $\boldsymbol{P}$ is the mentioned similarity predicate.

Since an image segmentation is defined as a partition, when comparing the gold standard with the segmentation under evaluation, we are, in fact, comparing two partitions. So, how do we compare two partitions? At the core of the problem are distance metrics, which define the notion of similarity between two partitions. In general terms, having a set of N elements and two different partitions defined on this set makes it possible to compare the two partitions in many ways—no single metric is useful in all circumstances. Nevertheless, one has already found applicability in other areas and, as we propose, it can also be useful in the context of evaluation of segmentation algorithms.

A. Partition Distance, d_{sym}

Before introducing the concept of partition distance, some helpful notions need to be visited. Let S be a set of N elements. A cluster is a **nonempty** subset of S. A partition of S is a set of mutually exclusive clusters, whose union is S. Two partitions P and Q of S are **identical** if and only if every cluster in P is a cluster in Q. A partition P is a **refinement** of a partition R (or P is **finer** than R) if and only if each cluster in P is contained in some cluster of R—see Fig. 1. Note that then, by definition, any partition is a refinement of itself.

The **intersection of two partitions** P and Q is a partition R so that every nonempty intersection of a cluster S_i from P and a cluster S_j from Q is an element of R—see Fig. 2. Note that R is a refinement of P and Q.

The **null partition** is the partition with only one cluster (the cluster has N elements). The **infinite partition** is the partition with N clusters (each cluster has one element).

We can now proceed to the idea of **partition distance** as it was first presented in [27]. Several alternative (but equivalent) definitions can be given (each more enlightening than the other for some background conditions).

Definition 1: "Given two partitions P and Q of S, the partition distance is the minimum number of elements that must be deleted from S, so that the two induced partitions (P and Q restricted to the remaining elements) are identical" [28].

Definition 2: "The partition distance is equal to the minimum number of elements that must be moved between clusters in P, so that the resulting partition equals Q (by definition, any set that becomes empty is no longer a cluster)" [28].

Proof That Definition 1 is Equivalent to Definition 2: Let D_1 be the set of dist_1 elements given by definition 1 and D_2 be the set of dist_2 elements given by definition 2.

a) From definition 1, P equals Q in $S \backslash D_1$. By moving the elements of D_1 in P to the same cluster as in Q, we can set $P = Q$ in S. This implies $\mathrm{dist}_2 \leq \mathrm{dist}_1$.
b) From definition 2, P equals Q in the set of unmoved elements $S \backslash D_2$. This implies $\mathrm{dist}_1 \leq \mathrm{dist}_2$.

From a) and b), we conclude the equivalence of both definitions. ■

From this definition, a useful set of properties can be deduced.

B. Properties of the Partition Distance, d_{sym}

Let P, Q, R be partitions defined in a set S of N elements. Then

1) $d_{\mathrm{sym}}(Q, P) \geq 0$.
2) $d_{\mathrm{sym}}(Q, P) = 0$, if and only if $Q = P$.
3) $d_{\mathrm{sym}}(Q, P) = d_{\mathrm{sym}}(P, Q)$.
4) $d_{\mathrm{sym}}(Q, P) + d_{\mathrm{sym}}(P, R) \geq d_{\mathrm{sym}}(Q, R)$.
5)

$$d_{\mathrm{sym}}(Q, \text{null partition}) = N - (\text{maximal cluster size in } Q).$$

6)

$$d_{\mathrm{sym}}(Q, \text{infinite partition}) = N - (\text{number of clusters in } Q).$$

7)

$$\begin{aligned} &d_{\mathrm{sym}}(\text{null partition infinite partition}) \\ &\quad = N - 1 \\ &\quad = \text{maximal distance between any two partitions.} \end{aligned}$$

8) The normalized distance $d_{\mathrm{sym}}/(N-1)$ ranges from 0 to 1.
9) Let S_1 and S_2 be two disjoint sets, P_1 and Q_1 be partitions of S_1, P_2 and Q_2 be partitions of S_2, and $P = P_1 \cup P_2$ and $Q = Q_1 \cup Q_2$ be the resulting partitions defined in $S = S_1 \cup S_2$. Then, $d_{\mathrm{sym}}(P, Q) = d_{\mathrm{sym}}(P_1, Q_1) + d_{\mathrm{sym}}(P_2, Q_2)$.

Any function with properties 1)–4) is called a **metric**.

Proof of Property 1: Follows directly from definition. ■

Proof of Property 2:

a) If $Q = P$, no points need to be removed from S to make the partitions equal. Then, $d_{\mathrm{sym}}(Q, P) = 0$.

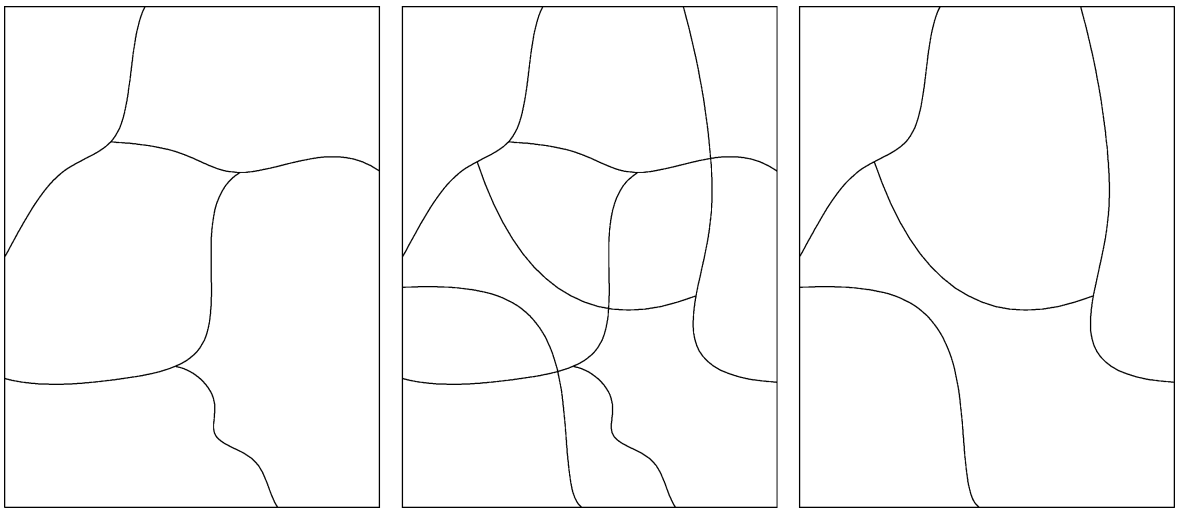


Fig. 2. Middle partition is the intersection of the left and right partitions.

b) If $d_{\mathrm{sym}}(Q,P)=0$, the number of points that had to be removed from S to make the partitions equal was 0. That is, the partitions are already equal in S.

■

Proof of Property 3: Follows directly from definition 1 of partition distance. ■

Proof of Property 4: Let D_1 be the set of $d_{\mathrm{sym}}(Q,P)$ elements to be removed in order to equal Q to P and D_2 be the set of $d_{\mathrm{sym}}(P,R)$ elements to be removed in order to equal P to R. Simultaneously, remove from S the elements of D_1 and D_2 (they may have common elements). Then, in the reduced set, we also have $Q=R$. So, removing $d_{\mathrm{sym}}(Q,P)+d_{\mathrm{sym}}(P,R)$ is enough to make Q and R equal partitions. That implies $d_{\mathrm{sym}}(Q,R)\leq d_{\mathrm{sym}}(Q,P)+d_{\mathrm{sym}}(P,R)$. ■

Proof of Property 5: Because two identical partitions have the same number of clusters, we can only keep elements from one cluster of Q in the reduced set. Then, it is easy to see that removing the elements of all clusters of Q, with the exception of those in the biggest cluster, gives the minimum number of elements that need to be removed to equal Q to the null partition. ■

Proof of Property 6: Because two identical partitions have the same number of clusters and the same number of elements in each cluster, we can only keep one element from each cluster of Q in the reduced set—otherwise, they would belong to different clusters in the infinite partition. It is easy to see that keeping only one element of each cluster of Q (anyone in fact) equals Q to the infinite partition. ■

Proof of Property 7: Making $Q=$ null partition in 6 or $Q=$ infinite partition in 5, we get the desired equality. Because it is always possible to keep at least one element of S (anyone, if fact), $(N-1)$ is the maximal possible value that d_{sym} can attain. ■

Proof of Property 8: By prop 1 and prop 7, $0\leq d_{\mathrm{sym}}\leq N-1$. Then, $0\leq d_{\mathrm{sym}}/(N-1)\leq 1$. ■

Proof of Property 9:

a) Remove from S_1 $d_{\mathrm{sym}}(P_1,Q_1)$ points to make $P_1=Q_1$ and from S_2 $d_{\mathrm{sym}}(P_2,Q_2)$ points to have $P_2=Q_2$. Then, $P_1\cup P_2=Q_1\cup Q_2$ in the set S restricted to the remaining elements. So, $d_{\mathrm{sym}}(P,Q)\leq d_{\mathrm{sym}}(P_1,Q_1)+d_{\mathrm{sym}}(P_2,Q_2)$.

b) Remove from S $d_{\mathrm{sym}}(P,Q)$ points to equal P to Q. Be n_1 the points removed from S_1. Then, $P_1=Q_1$ in S_1 excluded of the n_1 points. Then, $d_{\mathrm{sym}}(P_1,Q_1)\leq n_1$. In the same way, being n_2 the number of points removed from S_2, $d_{\mathrm{sym}}(P_2,Q_2)\leq n_2$. Then, $d_{\mathrm{sym}}(P_1,Q_1)+d_{\mathrm{sym}}(P_2,Q_2)\leq(n_1+n_2)=d_{\mathrm{sym}}(P,Q)$.

From a) and b), $d_{\mathrm{sym}}(P,Q)=d_{\mathrm{sym}}(P_1,Q_1)+d_{\mathrm{sym}}(P_2,Q_2)$. ■

We propose to apply the distance defined above to measure the discrepancy between the reference segmentation (nothing more than a partition of an image) and the segmentation under evaluation. This distance should be applied directly to the **segmentation partition** (with a different color representing each region) rather than to the edge map.

For instance, consider the two partitions of the same 8×8 image, represented in Fig. 3.

According to the distance defined above, these partitions are ten pixels away from each other. The pixels that had to be removed are highlighted in the middle image (unique solution in this particular case). Later, it will be shown how to efficiently compute this distance.

It is also interesting to compare the d_{sym} measure with the proposals in [26]. In Fig. 4(b), the BCE and d_{sym} measures are presented for two trivial segmentations. Note the nonmonotonous evolution of the BCE measure, where a monotonous behavior (not necessarily linear) presents as the most natural. In Fig. 4(c), the evolution of the measure based on mutual information from [26] is displayed when the two segmentations being compared correspond exactly. Contrast the nonconstant value of this measure, opposed to the constant value of the proposed partition distance.

C. Distance d_{sym} Applied to Binary Partitions

What do we get if we apply d_{sym} to the edge maps? These are nothing more than binary partitions of an image in edge/non edge pixels. It is easy to prove that the value given by d_{sym}

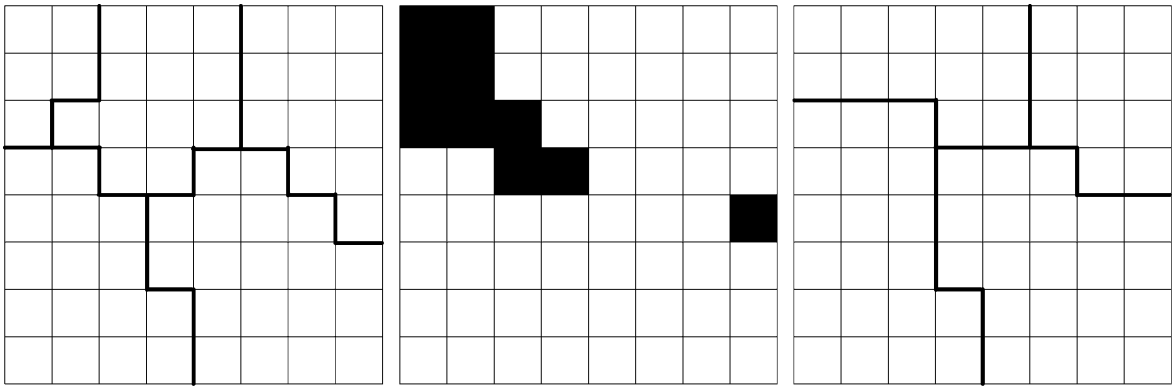


Fig. 3. Two different partitions of the same imag—the middle image highlights the points to be removed.

equals the number of misclassified pixels—this is the measure used in many of the earlier proposed methods.

Proof: Let us call $C_{ne(\mathrm{ref})}$ and $C_{ne(\mathrm{eval})}$ the cluster with the nonedge pixels in the reference and under evaluation edge map, respectively; $C_{e(\mathrm{ref})}$ and $C_{e(\mathrm{eval})}$ the cluster with the edge pixels in the reference and under evaluation edge map, respectively. To equal both partitions, we must either remove the points belonging to $C_{ne(\mathrm{ref})} \cap C_{ne(\mathrm{eval})}$ and $C_{e(\mathrm{ref})} \cap C_{e(\mathrm{eval})}$ or the points belonging to $C_{ne(\mathrm{ref})} \cap C_{e(\mathrm{eval})}$ and $C_{e(\mathrm{ref})} \cap C_{ne(\mathrm{eval})}$, that is, see the equation shown at the bottom of the page.

Clearly, for real segmentations, the number of elements in C_{ne}, both for the reference and the under evaluation edge maps, is larger than 75% of the image's total elements. Then, their intersection must have at least 50% of the elements ($|A \cap B| = |A| + |B| - |A \cup B| \geq 75\% + 75\% - 100\% = 50\%$).

So, the minimum number of points to remove is $C_{ne(\mathrm{ref})} \cap C_{e(\mathrm{eval})} + C_{e(\mathrm{ref})} \cap C_{ne(\mathrm{eval})}$, that is, the misclassified pixels. ■

Some authors have introduced pixels distance to cope with the position of misclassified pixels in the edge map. With the proposed metric, when applied to the segmentation partition, boundaries further away from their true location imply more pixels contributing to the distance between partitions.

D. Asymmetric Discrepancy Measure

In many applications, under segmentation is considered as a much more serious problem than over segmentation. This is so because it is easier to recover true segments through a merging process after over segmentation rather than trying to split a heterogeneous region. For those environments, it would be sensible to define an asymmetric distance between two partitions in such a way that the distance between a partition R and any partition Q finer than R is zero. Proceeding from the theoretical foundations already built, such a measure could be tentatively defined as follows.

Asymmetric Partition Distance, $d_{\mathrm{asy}}(R, Q)$: Given two partitions R and Q defined in a set S of N elements, the asymmetric partition distance is the minimum number of elements that must be deleted from S, so that the induced partition Q is finer than the induced partition R. Under this asymmetric distance, any partition finer than the R partition will be at zero distance from it. Notice also that, in general, $d_{\mathrm{asy}}(R, Q) \neq d_{\mathrm{asy}}(Q, R)$.

Recognizing that:

a) Q is finer than R if and only if the intersection of R and Q is equal to Q;
b) $d_{\mathrm{sym}}(Q, (R \cap Q)) = 0$, if and only if Q is finer that R.

A more *ad hoc* path could be followed to define an asymmetric distance between two partitions. In fact, $d_{\mathrm{sym}}(Q, (R \cap Q))$ should, then, convey a measure of the distance from Q to a finer partition of R. But, as it is easily verified, both definitions are equivalent.

Proof That $d_{\mathrm{sym}}(Q, (R \cap Q)) = d_{\mathrm{asy}}(R, Q)$:

a) Remove from S the d_{asy} elements needed to equal Q to a finer partition than R. Then, in the reduced set, $Q = (R \cap Q)$. That implies $d_{\mathrm{sym}}(Q, (R \cap Q)) \leq d_{\mathrm{asy}}(R, Q)$.
b) Remove from S the d_{sym} elements needed to equal Q to $(R \cap Q)$ in the reduced set. Then, Q is a finer partition of R in the reduced set. This implies $d_{\mathrm{asy}}(R, Q) \leq d_{\mathrm{sym}}(Q, (R \cap Q))$.

From a) and b), we conclude that $d_{\mathrm{sym}}(Q, (R \cap Q)) = d_{\mathrm{asy}}(R, Q)$. ■

The maximum value this asymmetric distance can attain is also $(N - 1)$ (for instance, for $Q =$ null partition, $R =$ infinite partition); so, to get a normalized distance, we just divide by $(N - 1)$. From the definition it also follows that $d_{\mathrm{sym}}(P, Q) \geq d_{\mathrm{asy}}(P, Q)$.

Working with the segmentation partitions already used to exemplify the symmetric partition distance, asymmetric distance attains the values (see Fig. 5)

$$d_{\mathrm{asy}}(\mathrm{left}, \mathrm{right}) = d_{\mathrm{sym}}(\mathrm{intersection}, \mathrm{right}) = 10$$
$$d_{\mathrm{asy}}(\mathrm{right}, \mathrm{left}) = d_{\mathrm{sym}}(\mathrm{intersection}, \mathrm{left}) = 6.$$

$$d_{\mathrm{sym}} = \min\left\{C_{ne(\mathrm{ref})} \cap C_{ne(\mathrm{eval})} + C_{e(\mathrm{ref})} \cap C_{e(\mathrm{eval})}; C_{ne(\mathrm{ref})} \cap C_{e(\mathrm{eval})} + C_{e(\mathrm{ref})} \cap C_{ne(\mathrm{eval})}\right\}$$

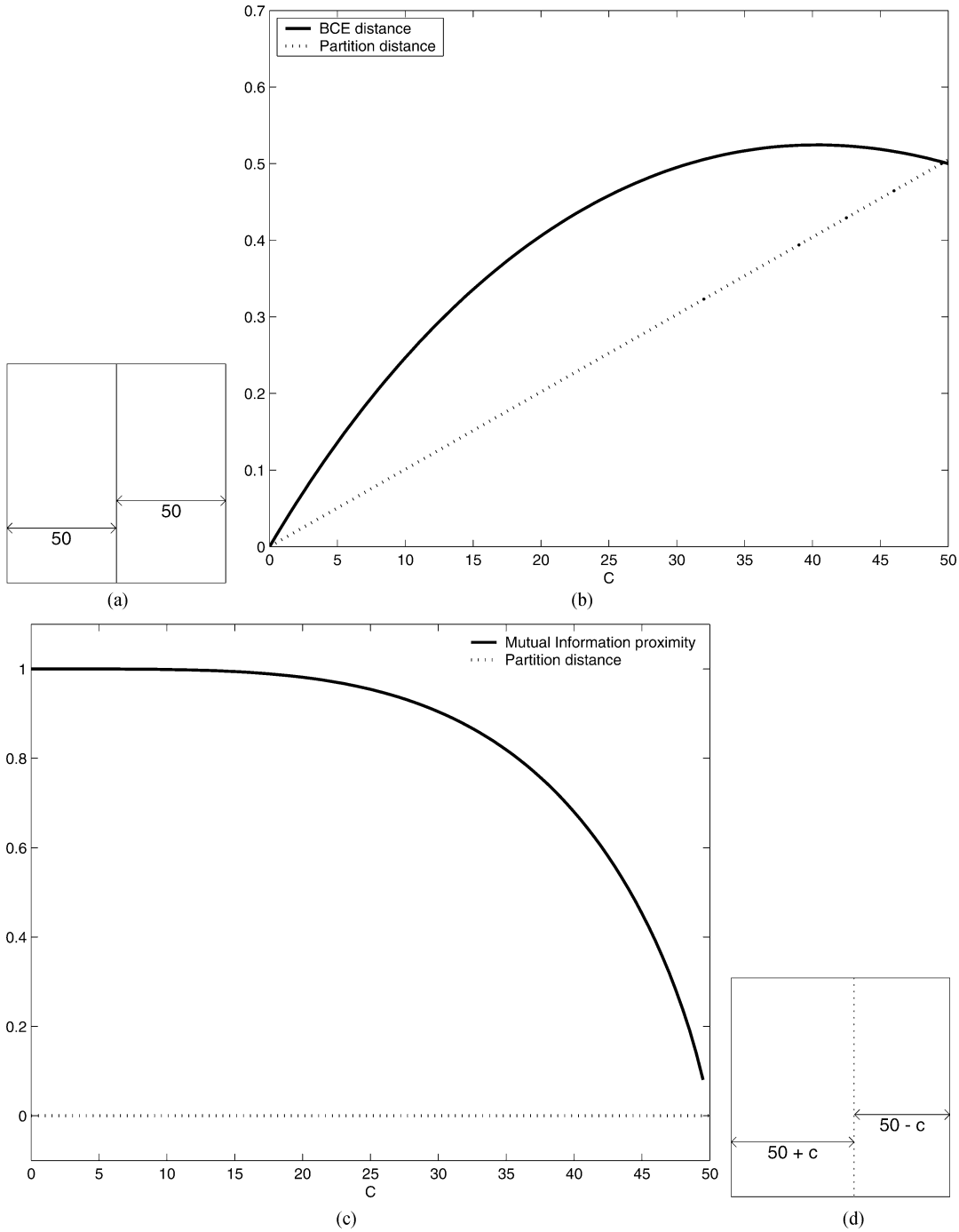


Fig. 4. Contrast in the evolution of BCE and mutual information measures from [26] and partition distance d_{sym}.

E. Mutual Refinement Measure

In some applications, it is important to have measures tolerant to mutual refinements [26]. A partition Q is said to be a mutual refinement of a partition R if the intersection of every cluster C_1 from Q with every cluster C_2 from R is either empty or equal to C_1 or C_2—see Fig. 6.

As can easily be seen, if partition Q is a mutual refinement of partition R, then R is a mutual refinement of partition Q. This concept is easily incorporated in the proposed methodology: Given two partitions R and Q defined in a set S of N elements, the mutual partition distance $d_{\mathrm{mut}}(R,Q)$ is the minimum number of elements that must be deleted from S, so that the induced partitions Q and R are mutual refinements of each other. As easily reckoned, this is a symmetric measure.

F. Proposed Discrepancy Measures

The path covered so far leads us to propose a set of different measures to evaluate the quality of an image segmentation S when comparing it to a reference segmentation R.

- Generic discrepancy measure given by the normalized partition distance between the reference segmentation and the segmentation under study: $d_{\mathrm{sym}}(R,S)/(N-1)$, where N is the number of pixels in the image.

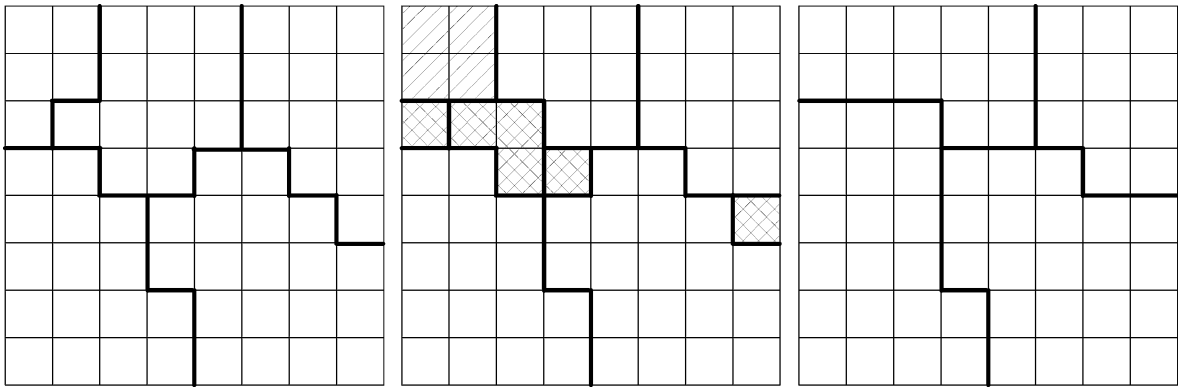


Fig. 5. Middle partition highlights the points to be removed for the asymmetric measures.

(a) (b)

Fig. 6. segmentations (a) and (b) are a mutual refinement of each other.

- Asymmetric measure for applications where over segmentation is not an issue, $d_{\text{asy}}(R,S)/(N-1)$, where R is the reference segmentation, S is the segmentation to assess and N is the number of pixels in the image.
- Asymmetric measure for applications where under segmentation is not an issue, $d_{\text{asy}}(S,R)/(N-1)$, where R is the reference segmentation, S is the segmentation to assess, and N is the number of pixels in the image.
- Mutual partition distance, $d_{\text{mut}}(S,R)$, when mutual refinements can be tolerated.

It should not be difficult to further extend this framework according to the specificities of each application.

IV. Results

To be of any practical use, the proposed measures have to be efficient to compute. It is shown in [28] that the partition distance can be computed in polynomial time, formulating the problem as an instance of the classical assignment problem. "An instance of the classical assignment problem consists of a matrix of numbers M, and an assignment is a selection of cells of M such that no row or column contains more than one selected cell. An optimal assignment is an assignment whose selected cell values have the largest sum over all possible assignments. An optimal assignment can be computed in polynomial time as a function of the size of M. To solve the partition-distance problem, create an instance $M(P,Q)$ of the assignment problem with one row i for each cluster S_i in P and one column j for each cluster S_j in Q. Associate cell (i,j) with the subset $(S_i \cap S_j)$ and write the number $|(S_i \cap S_j)|$ in cell (i,j). Next, solve the assignment problem on $M(P,Q)$ and let $A(P,Q)$ denote the value of the assignment. Then, the partition distance equals $N - A(P,Q)$. Moreover, the elements to remove from N are all those elements not associated with any selected cells of the optimal assignment" [28].

The asymmetric distance, although possible to compute using the general algorithm described above and the equivalence $d_{\text{asy}}(R,Q) = d_{\text{sym}}(Q,(R \cap Q))$, can be obtained much more efficiently, realizing that $d_{\text{asy}}(P,Q) = N - \sum_i(\max_j\ (S_i \cap S_j))$, for all S_i in P and S_j in Q follows directly from properties 5 and 9 of partition distance. This is readily obtained from matrix M, defined above as $N - \sum_i(\max_j\ M(i,j))$.

The proposed metrics were applied to a selected set of segmentation partitions outputted by segmentation algorithms and results were compared in order to assess the metrics' quality. For that end, a software application[1] was developed to implement the proposed metrics. The assignment problem was solved based on the well-known Hungarian method by Kuhn [29]. For HD images (1920×1080) with less than 256 regions, the computation takes less than one second in a regular PC (1-GHz AMD microprocessor, 256 MB RAM).

[1]The software, as well as all streams used in the tests, is available upon request to the authors.

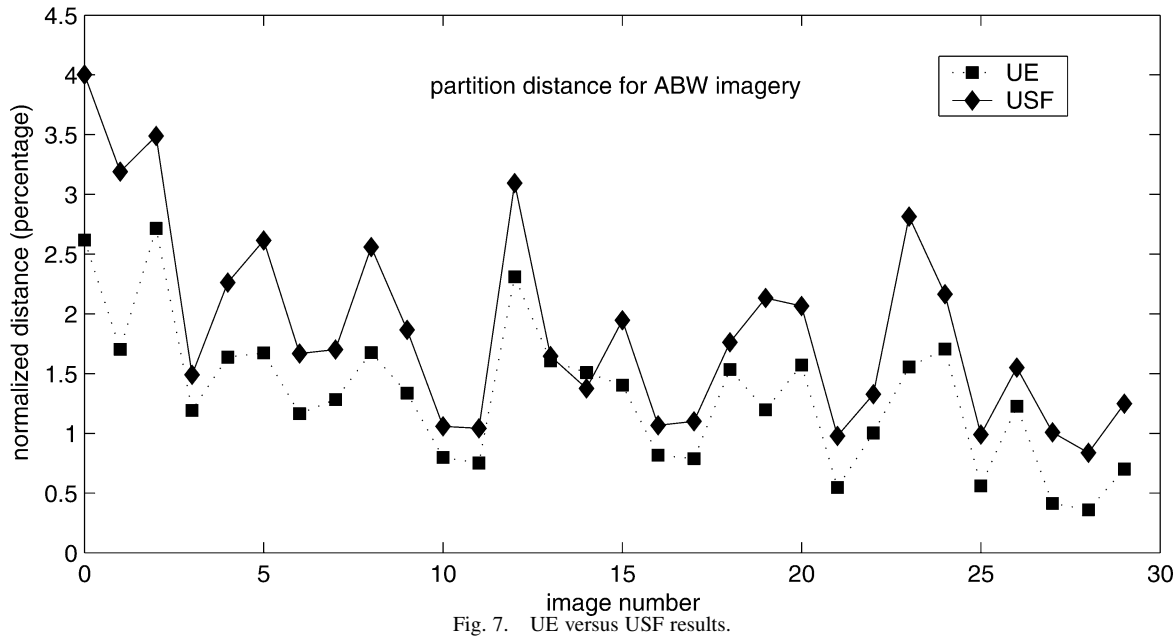


Fig. 7. UE versus USF results.

Fig. 8. Segmentation partition S_{23} on left, segmentation partition S_{226} on right, original image on center.

TABLE I
TABLES SHOWING THE SYMMETRIC AND ASYMMETRIC RESULTS (PERCENTAGE VALUES)

(a)

d_{sym}	S_{23}	S_{47}	S_{77}	S_{84}	S_{129}	S_{226}
S_{23}	0.00	14.66	30.09	30.10	34.95	51.30
S_{47}	14.66	0.00	18.84	18.86	23.92	40.73
S_{77}	30.90	18.84	0.00	0.09	12.89	30.87
S_{84}	30.10	18.86	0.09	0.00	12.88	30.85
S_{129}	34.95	23.92	12.89	12.88	0.00	25.12
S_{226}	51.30	40.73	30.87	30.85	25.12	0.00

(b)

d_{asy}	S_{23}	S_{47}	S_{77}	S_{84}	S_{129}	S_{226}
S_{23}	0.00	0.85	1.80	1.80	2.10	2.20
S_{47}	14.61	0.00	2.98	2.98	3.03	3.25
S_{77}	29.98	18.21	0.00	0.00	4.20	3.67
S_{84}	29.99	18.23	0.09	0.00	4.24	3.70
S_{129}	34.68	23.39	10.51	10.46	0.00	3.76
S_{226}	51.11	40.32	30.16	30.11	24.74	0.00

In a first test to check the adequacy and performance of the proposed solution to evaluate segmentation's quality, the symmetric metric was applied to the output of two range segmentation algorithms (UE and USF) presented in [19], using the ABW imagery, provided by the same author. The distance from the ground truth segmentation and the segmentation produced by each algorithm was calculated for each of the 30 test images on the set.

The partition distance results, presented in Fig. 7, consistently attribute better quality to UE, except for frame 15. This rating was found consistent with the subjective evaluation that a human observer would make by direct visualization of the segmentation partitions. These results are also in accordance with the average values in [19].

In a second test, the strength of the proposed asymmetric distances was also gauged. Toward this end, a segmentation algorithm that can be parametrically configured was selected. Different segmentation partitions, S_n, were produced for the same image (see Fig. 8), where n stands for the number of regions obtained for the partition. For each pair of segmentation partitions, we computed the d_{sym} and d_{asy} distances. Results are presented in Table I.

From Table I, we see that d_{sym} increases as we move away from the main diagonal. This is expected because as $|i-j|$ increases S_i and S_j become more and more different. However, for a given S_i, $d_{\text{asy}}(S_i, S_j)$ decreases while j increases until i, attaining 0 when $j = i$. It then stabilizes in very low values for $j > i$. This is so because segmentation algorithms tend to produce finer partitions as the segmentation resolution is increased.

Finally, the mutual partition distance was assessed with the Berkeley Segmentation Dataset [30]. The dataset consists of a collection of images where each image was segmented by different humans in color, grayscale, and inverted negated [26].

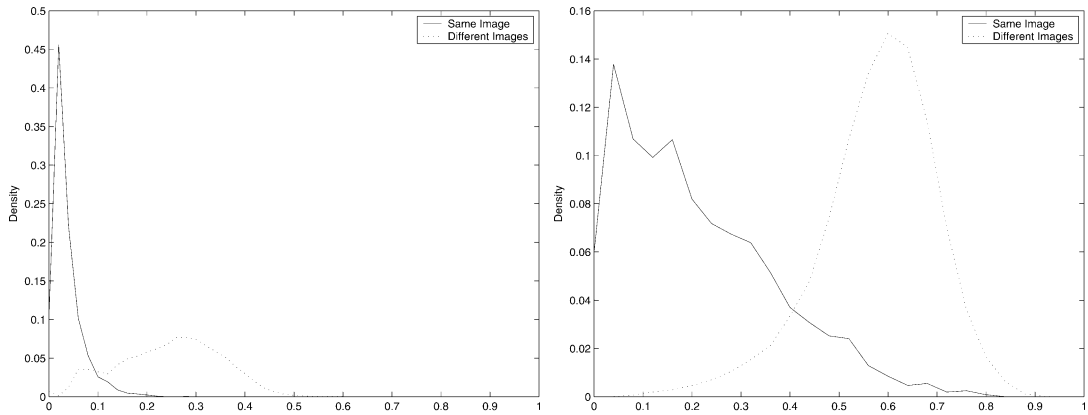


Fig. 9. Comparison of d_{mut} and d_{sym} for pairs of human segmentations.

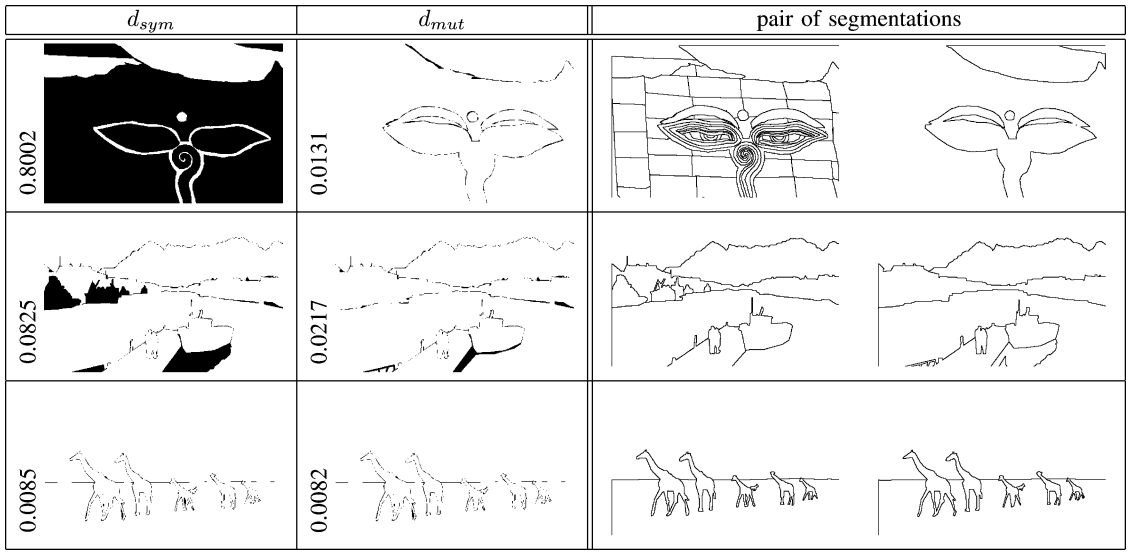


Fig. 10. Example pairs at various d_{sym} and d_{mut} values.

Humans may segment an image differently: The same scene may be perceived differently; different subjects may attend to different parts of the scene; subjects may segment an image at different granularities. However, according to [26], segmentations of the same image tend to be consistent in the sense that they are mutual refinement of each other.

Fig. 9 shows the distribution of d_{mut} and d_{sym} over the dataset for pairs of segmentations of the same image and pairs of segmentations of different images. As seen, although two segmentations of the same image may differ appreciably, as given by the d_{sym} measure, they are almost always identical, in what concerns the d_{mut} measure. For segmentations of the same image, the mutual partition distance exhibits a strong peak near zero error, given evidence of the consistency of human segmentations. It is also visible that the fraction of overlap—Bayes risk—is smaller for the mutual partition distance. Some examples of segmentation pairs, at different values of d_{sym} and d_{mut}, are shown in Fig. 10, each distance being presented both as a numerical value and as black pixels of a mask image.

These results are in accordance with the results achieved in [26].

V. Conclusion

Image segmentation quality evaluation is a key element when comparing segmentation algorithms. Segmentation quality evaluation allows the assessment of segmentation algorithms' performance for a given target application and the tuning of algorithms for optimal performance. It is believed that objective image segmentation quality evaluation is a very present-day problem, for which a satisfying solution is not yet available in the literature.

A generic framework for segmentation evaluation was presented in this paper. While some of the most recent segmentation quality evaluation methods only deal with two objects (foreground and background), the proposed methodology copes with multiple regions in the segmentation partition, using a clean, comprehensive technique.

The aim of this work is not to propose an evaluation measure incorporating perceptual or contextual information. As a low level measure, the proposed technique should rather produce valid results under all applications where a reference segmentation is available. This technique could also be used as a

building block in more complex and application specific evaluation schemes.

In the conducted experiments, the proposed segmentation quality evaluation metric showed the ability to estimate the segmentation quality according to what a human observer would do. Also, the asymmetric measure was shown to stand out in applications insensitive to over segmentations.

REFERENCES

[1] Hadamard, "Sur les problemes aux derivees partielles et leur signification physique," *Princeton Univ. Bull.*, pp. 49–52, 1902.

[2] Y. J. Zhang, "A survey on evaluation methods for image segmentation," *Pattern Recognit.*, vol. 29, no. 8, pp. 1335–1346, 1996.

[3] J. S. Weszka and A. Rosenfeld, "Threshold evaluation techniques," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-8, no. 3, pp. 622–629, Jun. 1978.

[4] N. R. Pal and S. K. Pal, "A review on image segmentation techniques," *Pattern Recognit.*, vol. 26, pp. 1277–1294, 1993.

[5] M. D. Levine and A. M. Nazif, "An experimental rule-based system for testing low level segmentation strategies," in *Multicomputers and Image Processing Algorithms and Programs*. New York: Academic, 1982, pp. 149–160.

[6] M. Borsotti, P. Campadelli, and R. Schettini, "Quantitative evaluation of color image segmentation results," *Pattern Recognit. Lett.*, vol. 19, no. 8, pp. 741–747, 1998.

[7] M. D. Levine and A. Nazif, "Dynamic measurement of computer generated image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-7, no. 2, pp. 155–164, Feb. 1985.

[8] C. Rosenberger and K. Chehdi, "Genetic fusion: Application to multicomponents image segmentation," in *Proc. IEEE Int. Conf. Acoustics Speech and Signal Processing*, vol. 4, 2000, pp. 2219–2222.

[9] P. K. Sahoo, S. Soltani, and A. K. C. Wang, "A survey of thresholding techniques," *Comput. Vis., Graph., Image Process.*, vol. 41, pp. 233–260, 1988.

[10] W. A. Yasnoff, J. K. Mui, and J. W. Bacus, "Error measures for scene segmentation," *Pattern Recognit.*, vol. 9, pp. 217–231, 1977.

[11] E. Abdou and W. K. Pratt, "Quantitative design and evaluation of enhancement/thresholding edge detectors," *Proc. IEEE*, vol. 67, no. 7, pp. 753–763, Jul. 1979.

[12] R. Ramán-Roldán, J. F. Gómez-Lopera, C. Atae-Allah, J. Martínez-Aroza, and P. L. Luque-Escamilla, "Measure of quality for evaluating methods of segmentation and edge-detection," *Pattern Recognit.*, vol. 34, pp. 969–980, 2001.

[13] Y. J. Zhang and J. J. Gerbrands, "Segmentation evaluation using ultimate measurement accuracy," in *Proc. CVPR*, vol. 1657, 1992, pp. 449–460.

[14] ——, "Objective and quantitative segmentation evaluation and comparison," *Signal Process.*, vol. 39, no. 1–2, pp. 43–54, 1994.

[15] M. F. Mattana, J. Facon, and A. S. Britto, "Evaluation by recognition of thresholding-based segmentation techniques on brazilian bankchecks," *Proc. SPIE*, vol. 3572, pp. 344–348, 1999.

[16] Z. M. Huo and M. L. Giger, "Evaluation of a computer segmentation method based on performances of an automated classification method," *Proc. SPIE*, vol. 3981, pp. 16–21, 2000.

[17] V. Chalana and Y. Kim, "A methodology for evaluation of boundary detection algorithms on medical images," *IEEE Trans. Med. Imag.*, vol. 16, no. 5, pp. 642–652, May 1997.

[18] A. A. Betanzos, B. A. Varela, and A. C. Martínez, "Analysis and evaluation of hard and fuzzy clustering segmentation techniques in burned patient images," *Image Vis. Comput.*, vol. 18, no. 13, pp. 1045–1054, 2000.

[19] A. Hoover, G. Jean-Baptiste, X. Y. Jiang, P. J. Flynn, H. Bunke, D. B. Goldgof, K. W. Bowyer, D. W. Eggert, A. W. Fitzgibbon, and R. B. Fisher, "An experimental comparison of range image segmentation algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 7, pp. 673–689, Jul. 1996.

[20] J. Min, M. Powell, and K. W. Bowyer, "Automated performance evaluation of range image segmentation algorithms," presented at the Workshop on the Application of Computer Vision, 2000.

[21] K. I. Chang, "Evaluation of texture segmentation algorithms," in *Proc. CVPR*, vol. 1, 1999, pp. 294–299.

[22] B. Belaroussi, C. Odet, and H. Benoit-Cattin, "Scalable discrepancy measures for segmentation evaluation," presented at the IEEE Int. Conf. Image Processing, 2002.

[23] A. B. Goumeidane, M. Khamadje, B. Belaroussi, H. Benoit-Cattin, and C. Odet, "New discrepancy measures for segmentation evaluation," presented at the IEEE Int. Conf. Image Processing, Barcelona, Spain, 2003.

[24] M. R. Everingham, H. Muller, and B. T. Thomas, "Evaluating image segmentation algorithms using the pareto front," in *Proc. 7th Eur. Conf. Computer Vision*, May 2002, pp. IV:34–48.

[25] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. 8th Int. Conf. Computer Vision*, vol. 2, Jul. 2001, pp. 416–423.

[26] D. Martin, "An empirical approach to grouping and segmentation," Ph.D. dissertation, Dept. Comp. Sci., Univ. California, Berkeley, 2003.

[27] A. Almudevar and C. Field, "Estimation of single generation sibling relationships based on dna markers," *J. Agricult., Biol., Environ. Stat.*, vol. 4, pp. 136–165, 1999.

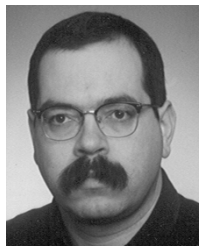
[28] D. Gusfield, "Partition distance: A problem and class of perfect graphs arising in clustering," *Inf. Process. Lett.*, vol. 82, pp. 159–164, May 2002.

[29] H. Kuhn, "The hungarian method for the assignment problem," *Nav. Res. Log. Quar.*, vol. 2, pp. 83–97, 1955.

[30] Berkeley Segmentation Dataset (2002). [Online]. Available: http://www.cs.berkeley.edu/projects/vision/bsds

Jaime S. Cardoso (S'04) was born in Póvoa de Varzim, Portugal, in 1976. He received the degree in electrical and computer engineering from the Faculdade de Engenharia, Universidade do Porto, Porto, Portugal, in 1999. He is currently pursuing the Ph.D. degree at the Departamento de Engenharia Electrotécnica e de Computadores da Faculdade de Engenharia da Universidade do Porto.

Since 1999, he has been a Researcher at INESC Porto, an R&D institute affiliated with the Universidade do Porto. His research interests include cryptography, data compression, and image/video processing.



Luís Corte-Real (M'91) was born in Vila do Conde, Portugal, in 1958. He received the degree in electrical engineering from the Faculdade de Engenharia, Universidade do Porto, Porto, Portugal, in 1981, the M.Sc. degree in electrical and computer engineering from the Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisbon, Portugal, in 1986, and the Ph.D. degree from the Faculdade de Engenharia, Universidade do Porto, in 1994.

In 1984, he joined the Universidade do Porto as a Lecturer of telecommunications. He is currently an Associate Professor with the Departamento de Engenharia Electrotécnica e de Computadores da Faculdade de Engenharia da Universidade do Porto. Since 1985, he has been a Researcher at INESC Porto, an R&D institute affiliated with Universidade do Porto. His research interests include image/video coding and processing and content-based image/video retrieval.