This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2481406, IEEE Transactions on Pattern Analysis and Machine Intelligence

Supervised Evaluation of Image Segmentation and Object Proposal Techniques

Jordi Pont-Tuset and Ferran Marques, *Senior Member, IEEE*

Abstract—This paper tackles the supervised evaluation of image segmentation and object proposal algorithms. It surveys, structures, and deduplicates the measures used to compare both segmentation results and object proposals with a ground truth database; and proposes a new measure: the precision-recall for objects and parts. To compare the quality of these measures, eight state-of-the-art object proposal techniques are analyzed and two quantitative meta-measures involving nine state of the art segmentation methods are presented. The meta-measures consist in assuming some plausible hypotheses about the results and assessing how well each measure reflects these hypotheses. As a conclusion of the performed experiments, this paper proposes the tandem of precision-recall curves for boundaries and for objects-and-parts as the tool of choice for the supervised evaluation of image segmentation. We make the datasets and code of all the measures publicly available.

Index Terms—Image segmentation, object proposals, supervised evaluation, meta-measures

1 Introduction

Since the advent of sliding window object detectors [1], much effort has been put into providing better spatial delineation beyond sliding windows [2], as a preprocessing step of many state-of-the-art algorithms [3], [4]. Bottom-up segmentation methods often play an important role in the proposed algorithms [5], [6], [7], [8], [9], and thus improving segmentation techniques would entail improvements towards better computer vision applications.

In such a challenge, providing benchmarks that help researchers understand the weak and strong points of their segmentation and object proposal algorithms is of paramount importance. Among these, the supervised evaluation, i.e., comparing the results with an annotated database called ground truth, is the most common approach; and the measures we use to grade the partitions are the cornerstone of the evaluation.

The first contribution of this paper is to **survey and structure** a large set of evaluation measures available in the literature. We first focus on the measures that assume a foreground-background ground truth (Section 2), which we refer to as object-based measures. Given their current relevance, we describe how to extend these measures to evaluate object proposal techniques, i.e., algorithms that propose a reduced set of locations and shapes among which it is probable to find the objects in the image (e.g. [5], [10], [11]).

To evaluate the generic image segmentation measures, which we refer to as partition-based (Section 3), we show that they can be classified depending on the interpretation of image partition they are based on. The most obvious one (*region-based* interpretation) is to interpret an image partition as a clustering of the set of pixels into regions, so any generic measure to evaluate clustering algorithms can be applied in this context. We can also cast the problem to a two-class clustering of the set of all pairs of pixels: those pairs belonging to the same region, and those coming from different regions (*pairs-of-pixels* interpretation). Finally, we can also interpret segmentation as a detection problem, aiming at telling apart the pixel contours that are true boundaries from those that are not (*boundary-based* interpretation).

Many of the most used evaluation measures, however, are limited to provide a single number, that is, given a pair of partitions (machine-generated and ground truth) they give us a single value that somehow reflects the degree of agreement between both. In the field of object detection assessment, Hoiem et al. [12] refer to these measures as *performance summary measures* and they stress that results should be evaluated beyond this type of measures in order to "help understand how one method could be improved." In other words, researchers need better feedback from the evaluation than a single number.

Back to segmentation assessment, the precision-recall curves for boundaries [13] are good examples of tools that provide richer feedback than the F measure used as summary. Moreover, as pointed out by [14], in addition to measures based on the boundary-based interpretation of a partition, region-based measures should be considered when assessing segmentations. However, the current region-based measures are limited to summary ones (e.g. [15], [16], [13], [17]).

The second contribution of this work (Section 4) is a precision-recall environment for the assessment of image segmentation that relies on the region-based interpretation of an image partition. Inspired by [18],

• *This work was mainly done at Universitat Politècnica de Catalunya, BarcelonaTech. The last revision was done while J. Pont-Tuset was at ETH Zürich. E-mail: jordi.pont@upc.edu and ferran.marques@upc.edu*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2481406, IEEE Transactions on Pattern Analysis and Machine Intelligence

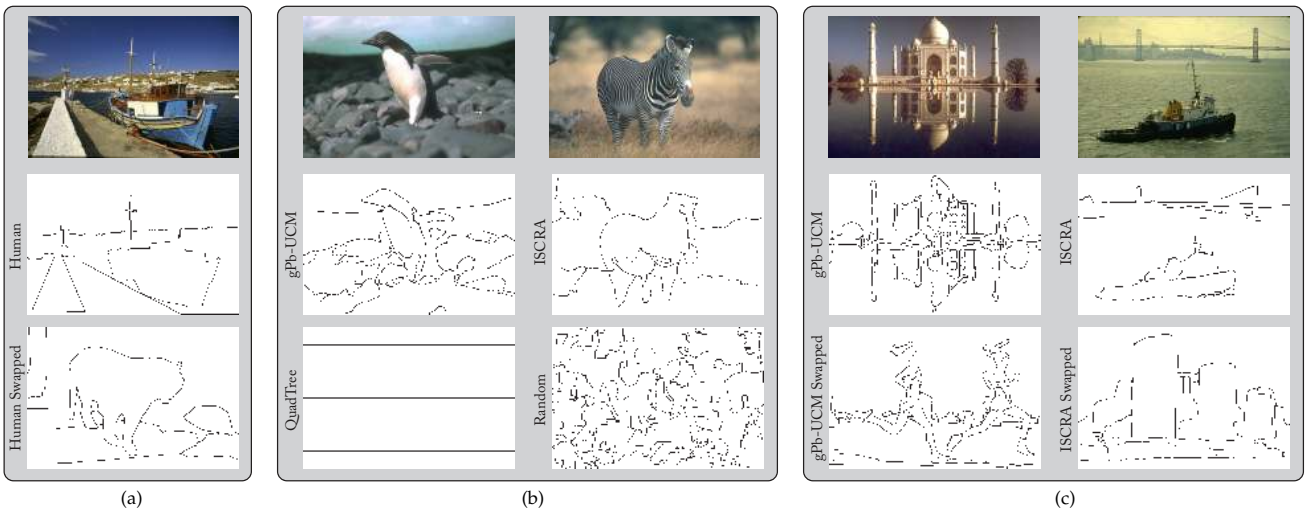

Fig. 1. **Quantitative meta-measure principles**: How good are the evaluation measures at ranking the second-row partitions better than the third-row ones?

[12] and by the fact that parts of objects are important clues for object detection [19], [20], we present the **precision-recall for objects and parts**, which is based on classifying the regions into object and part candidates.

Summary measures also play a role in performance comparison and researchers have a large list to choose from, thus the question that now arises is how to compare the goodness of an evaluation measure. In other words, we should define a *meta-measure* to compare the evaluation measures. The principle of a meta-measure is to assume a plausible hypothesis about the segmentation evaluation and assess how well measures match this hypothesis.

Some previous works based their claims on qualitative meta-measures, that is, showing the behavior of the measures on a few particular qualitative examples [21], [15]. The first approach to an extensive *quantitative* meta-measure was proposed in [13]. The hypothesis in this work was that measures should be able to discriminate between two pairs of human-marked partitions coming from different images (for instance, the two partitions in Figure 1.a). In an annotated database with multiple partitions per image, the quantitative meta-measure was defined as the number of same-image partition pairs that the measure judges as less similar than other pairs of partitions coming from different images. [22] presented a comparison of some measures in terms of this meta-measure.

The third contribution of this work (Section 5) is to present **two new quantitative meta-measures**. Moreover, instead of basing our hypotheses only on human-made partitions, we extend the analysis to partitions from nine State-of-the-Art (SoA) segmentation techniques.

The first hypothesis is that measures should rank higher SoA partitions than those obtained by means of two baseline techniques. The meta-measure is then defined as the number of results from SoA algorithms that are judged better than the baselines. As an example, we assess whether the measures score higher SoA partitions like those in the top row of Figure 1.b than the baseline ones in the lower row.

As a second meta-measure, we assume that any measure should rank higher a partition obtained by a SoA method on a given image than a partition obtained by the same method but on a different image, as the two pairs of partitions shown in Figure 1.c. The meta-measure in this case is defined as the number of cases in which the measure correctly judges the same-image partition as better.

Finally, Section 6 presents the experimental validation of this paper. For the foreground-background case (*object-based*), we analyze the boundary- and pixel-based measures, as well as three different generalization strategies to object proposals. We show qualitative results and the quantitative comparison of eight SoA object proposal techniques that show the complementarity of the proposed measures. For the *partition-based* case, we first compare all surveyed evaluation measures using the three quantitative meta-measures. We show that the two precision-recall measures (boundary- and objects-and-parts-based) have outstanding results as summary measures with respect to the rest of measures. We further analyze these two precision-recall frameworks by comparing nine SoA segmentation algorithms and show qualitative results illustrating the complementarity between the two frameworks.

Overall, the experiments show that the tandem of boundary- and region-based measures should be the choice for the supervised evaluation of both image segmentation and object proposals techniques. This work is an extended version of [23].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2481406, IEEE Transactions on Pattern Analysis and Machine Intelligence

2 Object-based Measures: Review and Improvements

This section focuses on the specific case of image segmentation where both the segmentation and the ground-truth are foreground-background partitions. Measures can focus either on evaluating how well the *pixels* of the ground truth are detected, or on how accurate the *boundaries* are represented. Sections 2.1 and 2.2 review and deduplicate the measures found in the literature under both interpretations. Then, Section 2.3 extends these measures to evaluate object proposals.

2.1 Pixel-based object measures review

Given an object detection method m, its resulting single-object detection can be written, from a *pixel perspective* as a division of the image pixel set I into two disjoint classes $I = P_m \cup N_m$, where P_m and N_m refer to positive and negative pixels, respectively, and the subscript stands for the method used. Equivalently for the ground-truth $I = P_{gt} \cup N_{gt}$.

The goal of any automatic algorithm is to achieve a perfect detection, i.e., $P_m = P_{gt}$, but if this is not the case, we define the following sets:

- **True positives**: Pixels that are detected as object and they are labeled as so in the ground truth: $TP = P_m \cap P_{gt}$.
- **False positives**: Pixels that are detected as object but they are not labeled as so in the ground truth: $FP = P_m \cap N_{gt}$.
- **False negatives**: Pixels that are classified as non-object but they are labeled as object in the ground truth: $FN = N_m \cap P_{gt}$, also known as *misses*.

The objective is, therefore, to maximize the true positives while minimizing both the false positives and the false negatives.

2.1.1 Precision, Recall, and F Measure

A widely used and accepted pair of measures to assess a detection algorithm is the following:

- **Precision**: Measures the percentage of detected pixels that are actually true:

$$Precision = \frac{|TP|}{|P_m|} = \frac{|P_m \cap P_{gt}|}{|P_m|} \leq 1$$

- **Recall**: Measures the percentage of ground-truth positives that are actually detected:

$$Recall = \frac{|TP|}{|P_{gt}|} = \frac{|P_m \cap P_{gt}|}{|P_{gt}|} \leq 1$$

Our objective is to maximize both measures, but in general there is a trade-off between them, which we can measure using the **F measure**, that is, the harmonic mean between precision and recall:

$$F = 2\,\frac{Prec \cdot Rec}{Prec + Rec} = \frac{2\,|TP|}{2\,|TP| + |FN| + |FP|} \quad (1)$$

To the knowledge of the authors, this coefficient was first reported by Czekanowski in 1913 [24], in the context of anthropology. Later, Dice used it in 1945 [25] to compare the number of species in two samples, with respect to the shared species in both. He coined it as *coincidence index*. It was also used in the context of plant sociology by Sørensen in 1948 [26]. Named after them, the coefficient is also known as *Czekanowski*, *Dice's*, or *Sørensen's coefficient*. More recently, the F measure is used as the evaluation metric in the Weizmann segmentation database [27], in the context of multi-object tracking [28], [29], or in the medical imaging context [30], where it is also referred to as *Spatial Overlap Index*.

2.1.2 Jaccard Similarity Coefficient

The Jaccard index was introduced in the context of plant sociology by Jaccard in 1901 [31], and in the context of object segmentation it is often referred to as *Intersection over Union* (*IoU*) between the machine and the ground-truth results:

$$J(P_m, P_{gt}) = \frac{|P_m \cap P_{gt}|}{|P_m \cup P_{gt}|} = \frac{|TP|}{|TP| + |FN| + |FP|} \quad (2)$$

In the PASCAL Visual Object Classes Challenge 2010 [32] the Jaccard coefficient (called area of overlap a_0) is used to assess whether a particular object has been detected ($a_0 \geq 0.5$) or not ($a_0 < 0.5$). In the context of object detection in [33], object accuracy is measured by means of the same value, denoted as $\mathcal{A}_0$. The performance measure used in the salient object extraction evaluation in [34], [35] is also J, although denoted as P. The work in [36] uses also this measure but it is denoted as *Overlap Score* (*OS*), or *spatial support score*. In [14] the Jaccard index is referred to as *overlap* and in [37], as *ratio of intersection*.

2.1.3 The Jaccard and F measures are equivalent

Comparing the expression of the F (Eq. 1) and J (Eq. 2), we can deduce the following equality:

$$\frac{F}{2-F} = \frac{\frac{2\,|TP|}{2|TP|+|FN|+|FP|}}{2 - \frac{2\,|TP|}{2|TP|+|FN|+|FP|}} = \frac{2\,|TP|}{4|TP| + 2|FN| + 2|FP| - 2|TP|} = J$$

That is, both measures are functionally related. Figure 2 plots the value of J as a function of F, in the range of interest $[0, 1]$. Given that their relationship is a monotonically increasing function, any ranking between algorithms using any of the two functions would be the same. In other words, for the purpose of segmentation algorithm comparison, both measures are equivalent. Despite this simple equivalence, there exist works in the literature [38] that report results using both measures in parallel.

In this work we will mainly use the Jaccard coefficient, since it is more used in the literature, and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2481406, IEEE Transactions on Pattern Analysis and Machine Intelligence

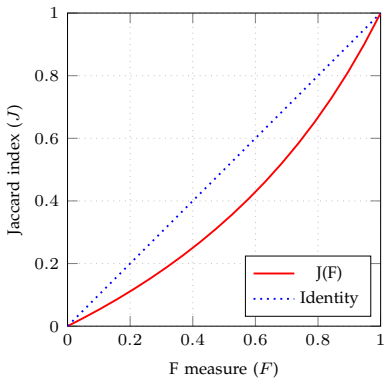

Fig. 2. **F measure versus Jaccard index**: J and F are functionally related

therefore, comparing results will be easier. We believe, however, that the main reason why this measure was selected against the F measure is aesthetic: the expression in terms of P_m and P_{gt} is more compact; although from a *detection* point of view, the F measure is theoretically more justified in our opinion.

We refer the reader to [39] for a comparison of specific measures to evaluate foreground maps, which although very similar to object segmentations, it is out of the scope of this paper. In it, the authors compare their measures using the meta-measures presented in this work (and previously in [23]).

2.2 Boundary-based object measures review

Differently to the pixel-based approach, a single-object detection result can be represented, from a *boundary perspective*, by the boundary between the foreground and the background pixels. Comparing the boundaries from the ground truth and the result we could therefore assess the quality of the detected objects.

Section 3.3 presents a review of boundary-based measures for full partitions, and highlights the well-known *precision-recall for boundaries* [13] as the measure of choice. The main idea behind this measure is to perform a *Bipartite Graph Matching* (BGM) between the pieces of boundary and then compute the precision, recall, and F measure (F_b). Given that we are evaluating simpler foreground-background masks, therefore, we could even compute more informative measures that specifically evaluate, for instance, how similar the represented shapes are [40], [41], perception-inspired losses [42], [33], etc.

When evaluating object proposals, however, we need to compare the ground truth with a large set of potential proposals (in the order of thousands, usually). The measure must be, therefore, very efficient, leaving out the majority of approaches introduced previously, which usually involve a costly BGM. To overcome this issue, we propose to do a simple morphological approximation of the *precision-recall for boundaries* [13] (F_b) that avoids the BGM: to compute precision we dilate the boundary pixels of the ground-truth shape and count the object boundary pixels that intersect the resulting mask (recall is computed the other way around). We then compute the morphological boundary F measure ($\tilde{F}_b$).

2.3 Evaluating object proposals techniques

A current trend in image and object segmentation is generating object proposals [10], [43], [11], [5], [44], [45], [6], which aims at generating a pool of region proposals (or candidates) with the objective of being as accurate as possible, while minimizing the size of the pool. From the point of view of object detection, they can be seen as a reduced set of potential locations and shapes where to look for objects, thus we would like the pool of candidates to be as small as possible (for our algorithm to be fast), while not losing set quality due to not considering all the set of possible locations and shapes.

To evaluate object proposals, therefore, we should account for two counterbalancing aspects: number of proposals versus the maximum achievable quality within the candidates in the pool. When training an algorithm to find its optimal parameterization we could perform optimization in the Pareto front of this two-dimensional space, as in [36] for generic image segmentation. In this work we focus on the evaluation at testing time, where the parameters are fixed.

For a given image in the database, we will, therefore, scan all proposals, compute an object-based metric M with respect to the ground truth, and get the maximum value. To compute the overall performance metric, we explore three different strategies. First, we could simply average the maximum measure value for all the annotated objects. Second, we could compute the median instead, to try to be more robust to outliers (e.g. missed objects with M close to 0).

In both cases, we are summarizing a large set of results into a single number so we are missing the distribution of the results. For instance, we would not distinguish a method whose proposals on half the objects are perfect ($M = 1$) and half missed ($M = 0$) from a result whose proposals are always at $M = 0.5$. We might, however, prefer one strategy against the other depending on the application.

A histogram reflects well the distribution of M values for a given number of proposals, but then we would end up having a 3-dimensional evaluation measure (number of proposals, binned M, bin counts), which is always tricky to plot. An in-between solution is to plot the percentiles of the histogram with respect to the number of proposals, that is, the percentage of objects on which the achievable M is higher than a threshold. In detection terms, these percentiles are the recall rates for different M thresholds.

We will discuss and analyze the results obtained using these three measures on eight state-of-the-art object proposal algorithms in the experiments section.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2481406, IEEE Transactions on Pattern Analysis and Machine Intelligence

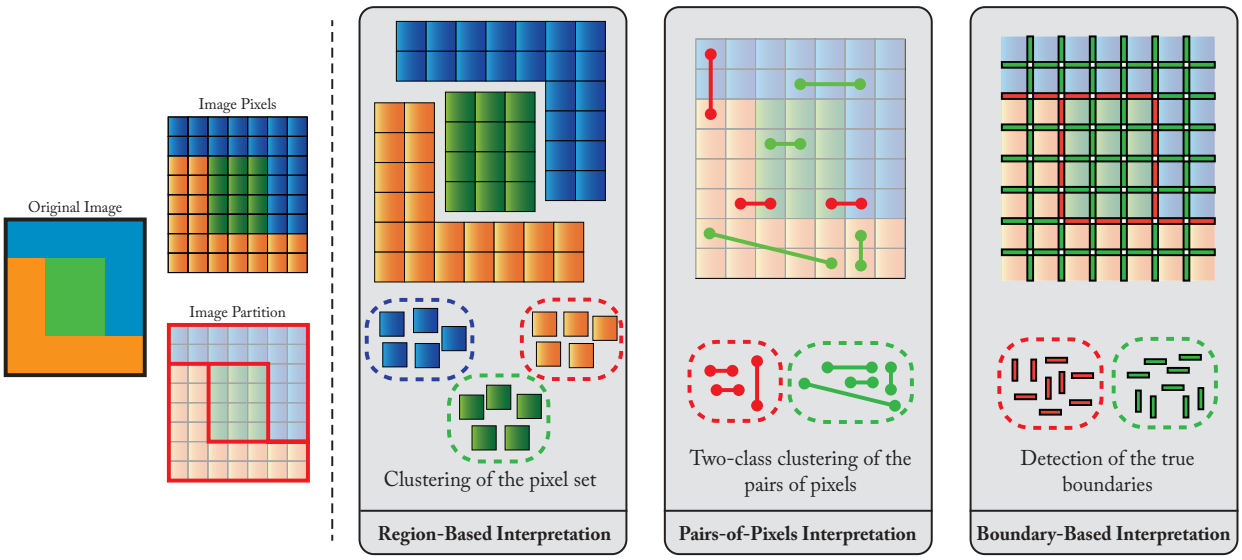

Fig. 3. **The three interpretations of image partition**: Clustering of the pixel set (region-based), two-class clustering of the pairs of pixels (pairs-of-pixels-based), and detection of true pixel contours (boundary-based)

3 Partition-Based Measures: Review and Structure

The state-of-the-art supervised evaluation measures can be classified depending on the image partition interpretation on which they are based. The most common interpretation is as a clustering of the pixel set into a number of subsets or regions, which we will refer to as *region-based* interpretation. A partition can also be interpreted as a two-class clustering of the set of pairs of pixels, with some pairs linking pixels from the same region and others linking pixels from different regions, which we will call *pairs-of-pixels* interpretation. Finally, a partition can be interpreted as a detection result, aimed at selecting the true boundaries on the image, which we will refer to as *boundary-based* interpretation. Figure 3 illustrates these three different partition interpretations.

The contributions of Sections 3.1 to 3.3 are to review, de-duplicate, and discuss about the main measures found under each of these interpretations, keeping the notation from the original papers where possible. In [50], the reader can find an interpretation of most of these measures in terms of simple measures such as the F measure, the Jaccard index, or precision-recall. Finally, Table 1 shows an overview of the measures.

3.1 Region-Based Measures

The **directional Hamming distance** from one partition S to another S' [51], [46] is defined as:

$$D_H\left(S \Rightarrow S'\right) = n - \sum_{R' \in S'} \max_{R \in S} |R' \cap R| \qquad (3)$$

where R and R' are regions in S and S', respectively, and n is the number of pixels in the image. In [21] this same measure was coined as asymmetric partition distance. Moreover, it is equivalent to the achievable segmentation accuracy [52] used in superpixel assessment.

As shown in [50], this measure is a generalization of the local measure *precision* between R and R':

$$1 - \frac{1}{n} D_H\left(S \Rightarrow S'\right) = \frac{1}{n} \sum_{R' \in S'} |R'| \cdot \max_{R \in S} \frac{|R' \cap R|}{|R'|}$$

The **segmentation covering** of a partition S by a partition S' was defined in [14], and can be interpreted as the generalization of the local measure *Jaccard index* between R and R':

$$\mathcal{C}\left(S' \rightarrow S\right) = \frac{1}{n} \sum_{R \in S} |R| \cdot \max_{R' \in S'} \frac{|R \cap R'|}{|R \cup R'|} \qquad (4)$$

A symmetric version of D_H was presented in [47] as the **van Dongen distance**:

$$d_{vD}(S, S') = D_H\left(S' \Rightarrow S\right) + D_H\left(S \Rightarrow S'\right) \qquad (5)$$

The intuitive step further is to measure the maximum overlap when performing a bijective matching between the regions of the two partitions, instead of the *local* matchings done in the measures above. This idea was presented in [21] as symmetric partition distance, in [22] as **bipartite-graph-matching** (*BGM*) distance, and in the context of clustering comparison, in [16] as classification error distance. It is shown in [21] that it is equivalent to the minimum number of pixels that must not be taken into account for the two partitions to be identical.

In [13], the consistency of the BSDS300 human partitions is analyzed by means of the **bidirectional**

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2481406, IEEE Transactions on Pattern Analysis and Machine Intelligence

Partition Interpretation	Measure Representative	References	Notation	Beyond Summary Measures
Region based	Directional Hamming distance	[46], [21]	D_H	✗
	van Dongen distance	[47]	d_{vD}	✗
	Segmentation covering	[14]	$\mathcal{C}$	✗
	Bipartite graph matching	[22], [21]	*BGM*	✗
	Bidirectional consistency error	[13]	*BCE*	✗
	Variation of information	[16]	*VoI*	✗
Pairs-of-pixels based	Probabilistic Rand index	[48], [15]	*PRI*	✗
	Precision-Recall for regions	[13]	P_r, R_r	✓
Boundary based	Precision-Recall for boundaries	[49], [13]	P_b, R_b	✓

TABLE 1
Measure structure overview: Based on the three interpretations of image partition

consistency error, which can be rewritten as:

$$BCE(S,S') = 1 - \frac{1}{n} \sum_{\substack{R \in S \\ R' \in S'}} |R \cap R'| \min\left\{ \frac{|R \cap R'|}{|R|}, \frac{|R \cap R'|}{|R'|} \right\} \tag{6}$$

The work in [16] introduced a new point of view to the measures of clustering assessment based on information-theoretic results. The author defines a discrete random variable taking N values that consists in randomly picking any pixel in the partition $S = \{R_1, \ldots, R_N\}$ and observing the region it belongs to. Assuming all the pixels equally probable to pick, the entropy $H(S)$ associated with a partition is defined as the entropy of such random variable. The mutual information $I(S,S')$ between two partitions is defined equivalently. The measure **variation of information** is then:

$$VoI(S, S') = H(S) + H(S') - 2I(S, S') \tag{7}$$

If divided by $\log N$, its maximum possible value, we get the **normalized variation of information** ($nVoI$).

3.2 Pairs-of-Pixels Measures

An image partition can be viewed as a classification of all the pairs of pixels into two classes: pairs of pixels belonging to the same region, and pairs of pixels from different regions. Formally, let $I = \{p_1, \ldots, p_n\}$ be the set of pixels of the image and consider the set of all pairs of pixels $\mathcal{P} = \{ (p_i, p_j) \in I \times I \,|\, i < j \}$. Given two partitions S and S', we divide $\mathcal{P}$ into four different sets, depending on where a pair (p_i, p_j) of pixels fall [16]:

$\mathcal{P}_{11}$: in the same region both in S and S',
$\mathcal{P}_{10}$: in the same region in S but different in S',
$\mathcal{P}_{01}$: in the same region in S' but different in S,
$\mathcal{P}_{00}$: in different regions both in S and S'.

The **Rand index**, originally defined in [48] as a clustering evaluation measure, arises naturally in this context:

$$RI(S,S') = \frac{|\mathcal{P}_{00}| + |\mathcal{P}_{11}|}{|\mathcal{P}|}$$

It *counts* the pairs of pixels that have coherent labels for the two partitions being compared, with respect to the number of possible pairs of pixels.

In the context of image segmentation and having a set $\{G_i\}$ of ground-truth partitions of the same image, the **Probabilistic Rand Index** [15] is computed as:

$$PRI(S, \{G_i\}) = \sum_i RI(S,G_i) \tag{8}$$

In this same context, the **precision-recall for regions** [13] is defined as:

$$P_r = \frac{|\mathcal{P}_{11}|}{|\mathcal{P}_{11}| + |\mathcal{P}_{10}|} \quad R_r = \frac{|\mathcal{P}_{11}|}{|\mathcal{P}_{11}| + |\mathcal{P}_{01}|} \tag{9}$$

As a summary measure, the F measure F_r is used.

3.3 Boundary-Based Measures

All measures above could be applied to any clustering algorithm, no matter the nature of the elements being classified. In fact, the majority of the indices presented come from the application of general-clustering assessment measures to image segmentation.

Image pixels, however, are spatially distributed in the image plane, and so the concept of neighborhood arises naturally. Therefore, an image partition with connected components can be unambiguously defined by their boundaries, i.e., a bijection could be made between all possible image partitions and all possible closed boundaries maps.

Recalling the definition of $\mathcal{P}$ as the set of pairs of pixels in the image, let us define the set of pairs of neighboring pixels as $\mathcal{N} \subset \mathcal{P}$. One can define a bijection between the set of boundary segments B and $\mathcal{N}$ linking each segment to the pair of pixels at each of its sides. Using this notation, boundary detection can be understood as a two-class clustering of B, dividing the segments into those being boundaries and those not. This way, comparing two partitions can be translated into comparing two clustering of B.

To be robust to unnoticeable shifts of boundary localization, [49] proposes to compute the optimal matching between the segments of boundaries of the two partitions as a maximum-weight bipartite-graph matching. The algorithm is improved in [13] leading to the well-known **precision-recall for boundaries** (P_b, R_b, and F_b).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2481406, IEEE Transactions on Pattern Analysis and Machine Intelligence

4 New Measure: F Measure for Objects and Parts

In the context of image segmentation evaluation, precision-recall curves for boundaries [13] are a boon for researchers. They statistically reflect, for instance, that an algorithm is providing too coarse segmentations (low recall, high precision) or instead its results are too fragmented (low precision, high recall).

As we will show in the experiments, and as pointed out by [14], however, region benchmarks are also needed apart from the boundary benchmarks when assessing image segmentation. Region benchmarks, however, are currently limited to summary measures as the ones reviewed in Section 3.1. (Note that in the vocabulary used in this paper, region-based measures are the ones based on the interpretation of a partition as a clustering of the set of pixels.)

This section presents a new region benchmark that goes beyond the summary measures: the precision-recall curves for objects and parts. Motivated by the fact that image segmentation is increasingly being used as a preliminary step for object detection [19], [20], we propose to assess segmentation under this perspective, that is, we interpret regions in a partition as potential object candidates, and classify them as correct or not depending on their overlap with the ground-truth regions.

Figure 4 shows a toy example (left: ground truth, and right: partition) to illustrate the proposed classification. First, we classify those regions from the partition that overlap significantly with a ground-truth region as *object candidates* (rectangle on the left and background). We then take oversegmentation into account, and define *part candidates* as those regions that can be used as a part to form a ground-truth region (triangle on the right). Undersegmentation *fragmentation candidates* are defined equivalently, as those regions that have incorrectly been merged together in the partition (circle and star).

Precision and recall are then the weighted fraction of candidates with respect to the total number of regions, that is, part candidates are only *partially counted*.

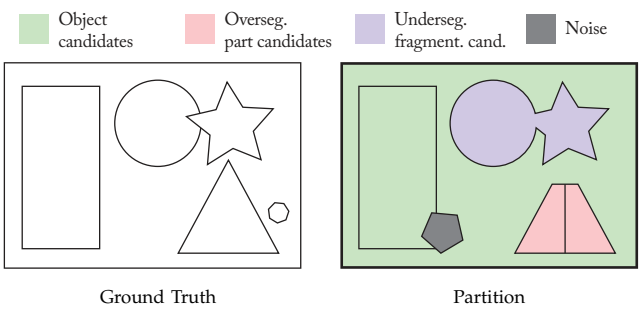

Fig. 4. **Region classification example**: Regions are classified into object, part candidates, fragmentation candidates, and noise

Formally, let $S=\{R_1,\ldots,R_N\}$ be an image partition and $\{G_k\}$ a set of ground-truth partitions of the same image. We consider the set $G=\{R'_1,\ldots,R'_M\}$ of all the regions in $\{G_k\}$. For each pair of regions $R_i \in S$, $R'_j \in G$ we compute the relative overlaps as:

$$O_S^{ij}=\frac{|R_i \cap R'_j|}{|R_i|} \qquad O_G^{ij}=\frac{|R_i \cap R'_j|}{|R'_j|}$$

We define an *object threshold* γ_o and a *part threshold* $\gamma_p < \gamma_o$ and classify the regions in both partitions as described in Algorithm 1, where "$\leftarrow$" means that a region is classified only if it previously did not have a more favorable classification.

Algorithm 1 Region candidates classification

for all $R_i \in S$, $R'_j \in G$ **do**
  if $O_S^{ij} > \gamma_o$ **and** $O_G^{ij} > \gamma_o$ **then**
    $R_i, R'_j \leftarrow$ Object candidates
  else if $O_S^{ij} > \gamma_p$ **and** $O_G^{ij} > \gamma_o$ **then**
    $R_i \leftarrow$ Fragmentation candidate
    $R'_j \leftarrow$ Part candidate
  else if $O_S^{ij} > \gamma_o$ **and** $O_G^{ij} > \gamma_p$ **then**
    $R_i \leftarrow$ Part candidate
    $R'_j \leftarrow$ Fragmentation candidate
  else
    $R_i, R'_j \leftarrow$ Noise
  end if
end for

Let oc and oc' be the number of object candidates in S and G, respectively (note that they can differ, given that G can be formed by more than one partition and thus a region in S can be matched as object with more than one region in G), and pc and pc' the number of part candidates. Regarding the fragmentation candidates, we compute the percentage of the object that could be formed from the matched parts. Formally, we define the amount of fragmentation $fr(R_i)$ of a region $R_i \in S$ as the addition of the relative overlaps of the part candidates matched to R_i:

$$fr(R_i)=\sum_j \left\{ O_G^{ij} \text{ s.t. } O_S^{ij} > \gamma_o \right\} \tag{10}$$

$fr'(R'_j)$ is defined equivalently for G. The global fragmentations fr and fr' is computed adding the amount of fragmentation among all fragmentation candidates of S and G, respectively.

We then define the **precision-recall for objects and parts** as follows:

$$P_{op}=\frac{oc+fr+\beta\, pc}{|S|} \qquad R_{op}=\frac{oc'+fr'+\beta\, pc'}{|G|} \tag{11}$$

Intuitively, in a completely oversegmented result, the recall would be high but the precision very low. Conversely, a completely undersegmented result (one single region) would entail a high precision but very low recall. As a summary measure, we propose to use the F measure (F_{op}) between P_{op} and R_{op}.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2481406, IEEE Transactions on Pattern Analysis and Machine Intelligence

5 Quantitative Meta-Measures

A meta-measure analysis must rely on accepted hypotheses about the segmentation results and assess how coherent the measures are with such hypotheses As an example, an accepted hypothesis can be the human judgment of quality of some particular examples. The meta-measure is then defined as a quantization of how coherent the evaluation measures are with this judgment, as done in works such as [15], [21].

To provide statistically significant results, however, one must go beyond a handful of examples and provide a quantitative analysis on an annotated database. The remainder of this section explains one meta-measure already published in the literature (Sec. 5.1) and presents two new meta-measures (Sec. 5.2 and 5.3).

The two new meta-measures differ significantly from the already-existing one in the sense that, instead of being based only on human-made partitions, we base our analysis on a large set of partitions made by state-of-the-art segmentation techniques. In turn, these meta-measures can be easily updated as new state-of-the-art segmentation techniques are presented.

5.1 Swapped-Image Human Discrimination

Given an image, there is no unique valid segmentation, since it depends on the perception of the scene, the level of details, etc. In order to cope with this variability, the Berkeley Segmentation Dataset (BSDS300 [53] and BSDS500 [14]) consists of a set of images each of them manually segmented by more than one individual.

The hypothesis behind the first meta-measure is that an evaluation metric should be able to tell apart the ground-truth partitions coming from two different images. In other words, given a pair of ground-truth partitions from BSDS500, a measure should be able to tell whether they come from the same image (thus differences are an acceptable refinement) or different images (unacceptable discrepancies).

As first proposed by [13] to evaluate the coherence of BSDS300, given an evaluation measure m, we compute the Probability Density Function (PDF) of the values of m for all the pairs of partitions in BSDS500, grouped in two classes: those coming from different images and those from the same one. Figure 5 shows the PDFs for these two types of pairs of partitions using the F_b measure.

A simple classifier was then defined setting a threshold on the measure to discriminate the two types of pairs. The **Swapped-Image Human Discrimination** (SIHD) meta-measure is defined as the percentage of correct classifications of that classifier, that is, the sum of the area under the curve above and below the threshold for the same-image and different-image pairs, respectively. (In the original work [13], the authors reported the Bayes Risk.)

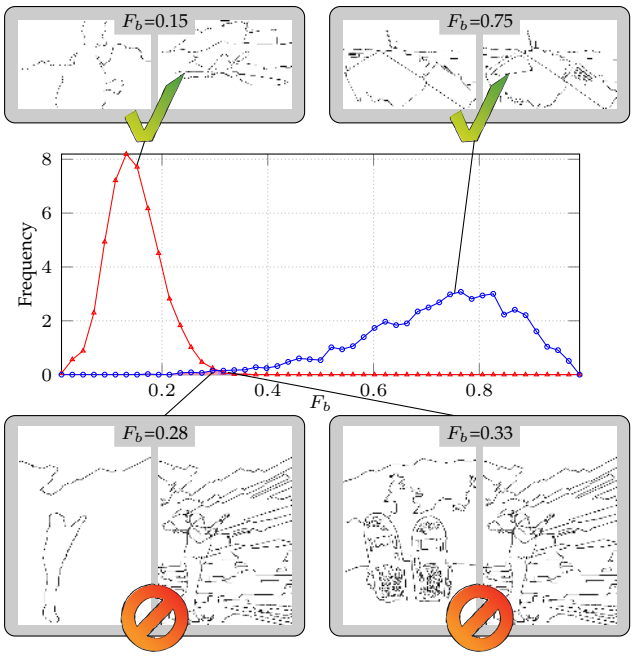

Fig. 5. **SIHD example**: Distribution of F_b values for the same-image pairs of partitions (—○—) and different-image pairs (—△—). In gray rectangles, four representative pairs of partitions: a pair of correctly classified as different image (up-left) and as same image (up-right); and a pair incorrectly classified as different image (down-left) and as same image (down-right).

As qualitative examples, Figure 5 depicts four pairs of partitions as representatives of the type of mistakes and correct classifications using F_b.

5.2 SoA-Baseline Discrimination

One of the reasons why SIHD can be criticized is the fact that it is based only on human-made partitions, that is, it does not show how measures handle the *real-world* discrepancies found between SoA segmentation methods. This section and the following are devoted to present two meta-mesures based on SoA segmentation results.

The hypothesis on which we base the meta-measure presented in this section is that evaluation measures should, for a given image, rank higher partitions obtained by any SoA segmentation method than partitions obtained by baseline methods. In particular, in this work we will use nine SoA techniques and two baseline methods.

As the first baseline technique we consider a quadtree (as in [54], [14]), which consists in hierarchical partitions starting from the whole image support and iteratively dividing the regions into four equal rectangles, regardless of the content of the image. Figure 1.b(left) shows an example of partition obtained by a SoA method and by a quadtree.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2481406, IEEE Transactions on Pattern Analysis and Machine Intelligence

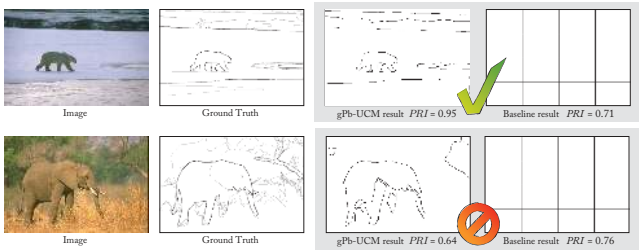

Fig. 6. **SABD example**: Correct and incorrect judgments by a segmentation evaluation measure.

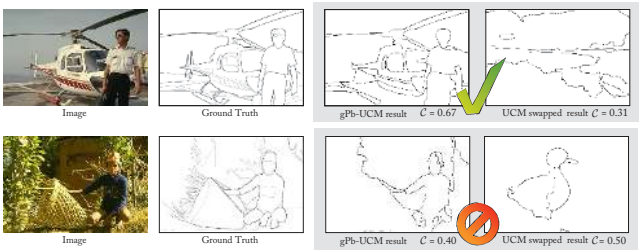

Fig. 7. **SISD example**: Correct and incorrect judgments by a segmentation evaluation measure.

As a second baseline, we use a *random* hierarchy, that is, we compute the SLIC [55], [56] superpixels of the image and then iteratively merge random pairs of neighboring regions. Figure 1.b(right) shows an example of partition obtained by a SoA method and by a random hierarchy.

As the partitions given by the baselines can be considered as obtained *by chance*, the SoA partition should be judged better than the baseline, regardless of the application we are focused on. For each of the techniques considered as SoA segmentation methods, therefore, we compute the number of images in the dataset in which an evaluation measure correctly judges that the baseline result is worse than the SoA generated partition. We refer to the resulting meta-measure as **SoA-Baseline Discrimination** (SABD), and it is defined as the global percentage of correct judgments for a given measure.

Figure 6 shows an example of a correct and an incorrect judgment by a segmentation evaluation measure of the quality of a baseline result with respect to a SoA result.

5.3 Swapped-Image SoA Discrimination

Segmentation evaluation measures are often used to adjust the parameters of a segmentation technique. They are therefore used to compare different partitions created by the same algorithm with slightly different parameterizations and we want the evaluation measures to differentiate between *good* and *better* results in order to learn the best parameters. A necessary condition, therefore, it is that a measure should be able to tell apart an *acceptable* result from a *wrong* result. Given an image, we consider a SoA partition as acceptable result and a partition done by the same technique and parameters but on a different image as a wrong one.

In other words, we compare the ground-truth partitions of a certain image with two results obtained using the same algorithm and parameterization: one segmentation of that same image and one of a different image. The hypothesis in this case is that the evaluation measures should judge that the same-image result is better than the different-image one. In the examples of Figure 1.c, the measure should judge that the first-row partitions are better than the second-row ones, when compared both with the ground-truth of the images of the first row. In this meta-measure, evaluation measures have to tackle the potential bias of the SoA methods towards their specific type of results.

For each SoA segmentation technique, we compute the number of images in the dataset in which an evaluation measure correctly judges that the same-image SoA result is better than the different-image one. We define the meta-measure **Swapped-Image SoA Discrimination** (SISD) as the percentage of results in the database, for all the SoA methods, that the measures correctly discriminate.

Figure 7 shows an example of a correct and an incorrect judgment by a segmentation evaluation measure of the quality of a SoA result judged on the same versus a different image.

6 Experimental Results

This section presents the experimental validation of the measures and meta-measures proposed in this paper. We will use the images from BSDS500 [14], with the object ground truth from [11] and partition ground truth from [14]. Section 6.1 presents a qualitative comparison of the object-based studied measures and Section 6.2 describes the experiments on object proposals, focusing on the behavior and complementarity of the proposed measures. Section 6.3 shows the comparison of all partition-based evaluation measures in terms of the proposed quantitative meta-measures. As a result of the analysis, we propose the two best performing measures F_b and F_{op} to be used in tandem. Section 6.4 analyzes the state-of-the-art segmentation techniques in terms of the precision-recall curves of these two measures, illustrating the usefulness of these two frameworks in tandem and the richness gained by going beyond summary measures. We also present some experiments to further analyze their differences and show their complementarity, reinforcing the choice of using them in tandem.

6.1 Object-based Measures

This section shows some qualitative results to highlight the differences between the pixel- and boundary-based measures from an object perspective. Figure 8

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2481406, IEEE Transactions on Pattern Analysis and Machine Intelligence

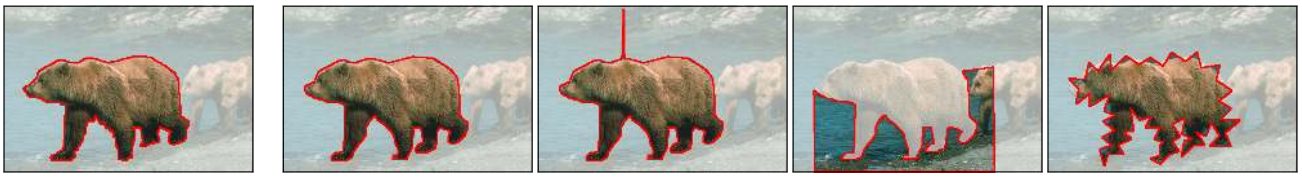

Fig. 8. Object-based J versus F_b: Complementary examples where the behavior of the two measures differs.

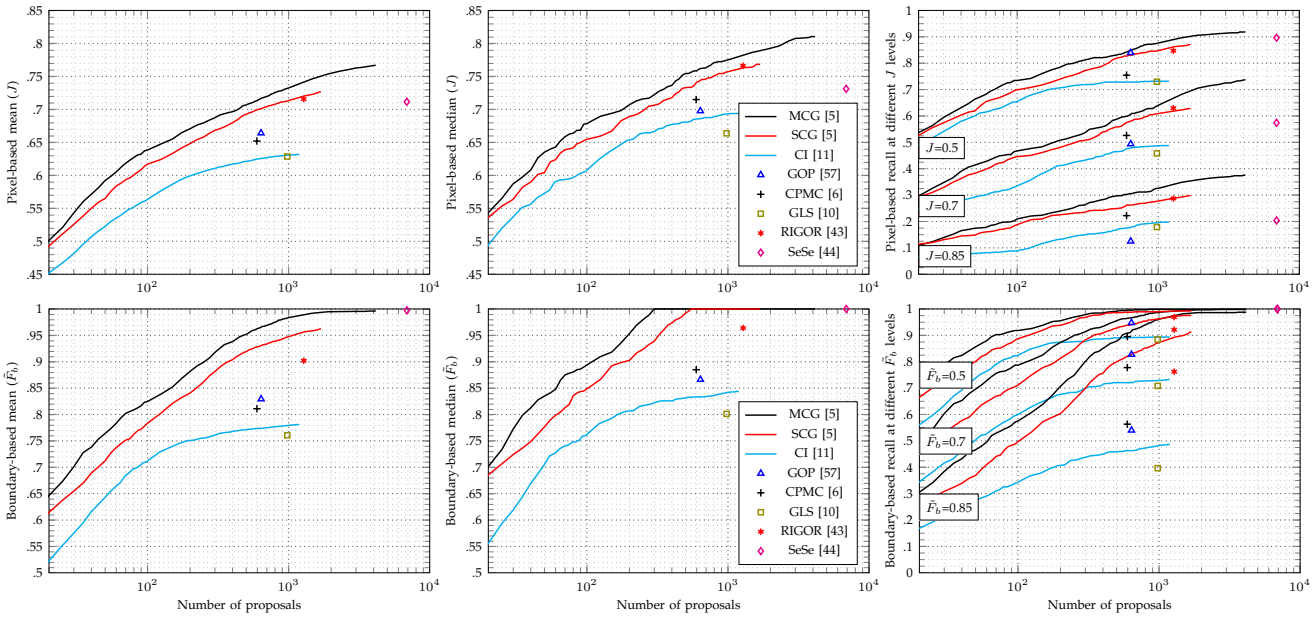

Fig. 9. **Object proposal evaluation**: Pixel- and boundary-based measures (J and $\tilde{F}_b$): mean, median and recall.

depicts a ground-truth annotated object and four different cases, each of which evaluated with the pixel-based measure Jaccard (J) and the boundary-based $\tilde{F}_b$. We also report the original F_b to intuitively check whether the morphological approximation $\tilde{F}_b$ is acceptable and to adjust its parameters.

Figure 8(a) shows a human-made partition to represent the quality upper-bound. The three measures report very high values, although not a perfect 1. Figure 8(b) shows a degenerate case where the pixel-based measure does not penalize the result, because the number of *wrong* pixels is very small. In contrast, $\tilde{F}_b$ correctly penalizes it. Figure 8(c) shows the dual result, in which the object is completely missed in terms of pixels but the boundary-based measure does not penalize it properly because there is a significant boundary overlap. Figure 8(d) shows a result with altered boundaries, which is considered worse than (c) in terms of $\tilde{F}_b$, although pixel-wise is a good result.

All examples show that $\tilde{F}_b$ is a good approximation of the original F_b. To achieve so, we adapted the boundary tolerance (8% of the image diagonal) in $\tilde{F}_b$ to be robust to degenerate cases such as (d).

Overall, we observe a dual behavior between the boundary-based and pixel-based measures, so our proposal is to use both measures in tandem for the object-based evaluation of segmentation.

6.2 Object Proposals

The state of the art in object proposals is represented in this work by the following eight methods: GOP [57], MCG [5], SCG [5], CI [11], CPMC [6], GLS [10], RIGOR [43], and SeSe [44]. Figure 9 shows the pixel- and boundary-based evaluation results for these methods using the three proposed generalization measures.

As expected, the general trend is that the more proposals, the better the achievable quality. The mean and median measures show similar overall behavior, with MCG being the best performing and slight differences such as the comparison between CPMC [+] and GOP [▵], which have inverted rankings with the two measures. Being the median consistently better than the mean suggests that there are some outliers on the lower part of the distribution, i.e., some missed objects where the achievable quality is close to zero.

Focusing on the pixel-based measure (top row), these differences are better reflected in the recall plots, on the right-most part of the plot. We depict the plots for $J = 0.5$ (very imprecise result), $J = 0.7$ (approximate result), and $J = 0.85$ (precise result). Again, it is interesting to compare the behavior between CPMC [+] and GOP [▵]. GOP has outstanding results for $J = 0.5$ but the ranking is exchanged for $J = 0.7$ and 0.85, which reflects that the majority of results by

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2481406, IEEE Transactions on Pattern Analysis and Machine Intelligence

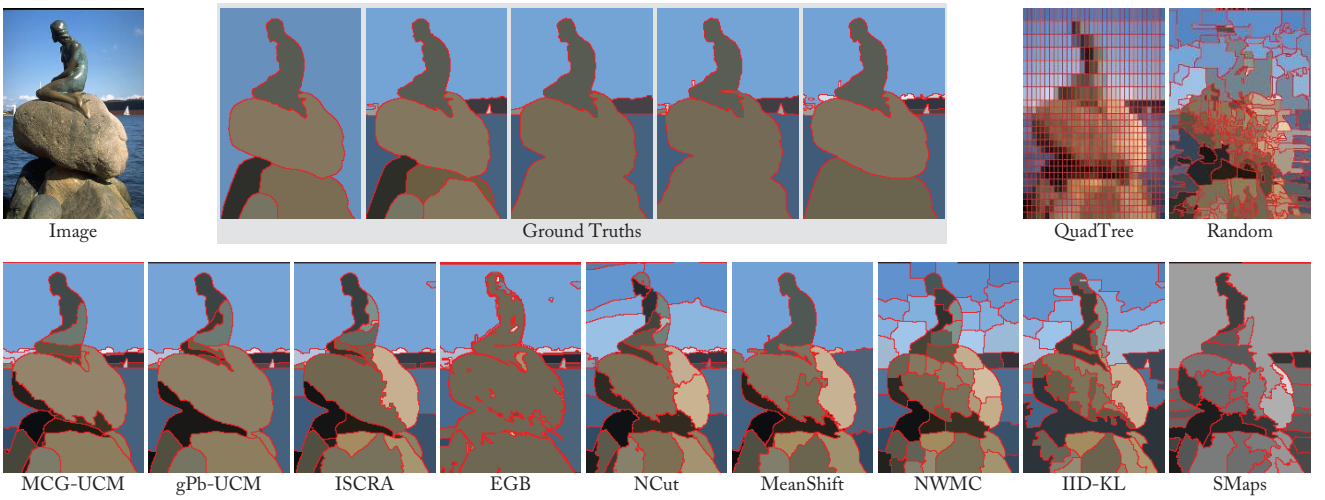

Fig. 10. **State-of-the-art examples**. Top row: An image from BSDS500, the five ground-truth partitions done by different humans, and the partitions obtained by the two baseline techniques. Bottom row: Partitions obtained by the nine representative SoA segmentation techniques, each of them at its ODS with respect to F_b

GOP has an imprecise representation but they are not missed. In contrast, CPMC provides results that are much more precise, but also many more misses.

Focusing on the boundary-based measure (bottom row), we observe that MCG and SCG are even better than the rest of SoA. This suggest that these techniques have very accurate boundaries but they might miss some parts of the objects, which is further penalized in the pixel-based measure. The saturation of the median F_b to 1 (middle plot) tells us that more than 50% of the results have perfectly accurate boundaries (within the matching tolerance) but the ones that do not are almost missed, because the mean is considerably lower than the median.

Overall, we propose the two measures and three generalization strategies together as a good representative of the quality and behavior of the object proposal methods.

6.3 Quantitative Meta-Measures

The state of the art of segmentation to compute the meta-measures is represented in this paper by MCG-UCM [5], gPb-UCM [14], ISCRA [58], EGB [59], Normalized Cuts [60], Mean Shift [61], NWMC [62], IID-KL [63], and Saliency Maps (on grayscale images) [64]. As baselines, we use two techniques: a rectangular homogeneous grid (Quadtree) and a random merging of superpixels (Random). All methods (SoA and Baselines) are assessed at the Optimal Dataset Scale (ODS) [14] with respect to each evaluation measure, that is, using the parameters that entail the best value of the measure in mean on the whole training set of BSDS500. In other words, we run the segmentation techniques sweeping their parameters (from coarse to fine partitions), and then choose the optimal parameter in terms of each evaluation measure, globally in the whole training set. Figure 10 shows an image, the various ground-truth partitions, and the baseline and SoA partitions at their ODS with respect to F_b.

The parameter values of the newly proposed measure are: $\gamma_o = 0.95$, $\gamma_p = 0.25$, and $\beta = 0.1$. They have been trained on the training set of BSDS500, by optimizing the global meta-measure described below. Note that this optimization would not have been feasible without quantitative meta-measures.

As an additional property of the measures, we analyze their definition when multiple ground-truth annotations $\{G_k\}_1^n$ are available. The most common approach to evaluate a partition P using measure m is to compute the mean over all annotations $\frac{1}{n}\sum_k^n m(P, G_k)$. In contrast, some measures have specific definitions that take further advantage of the multiple annotations. We also tested computing the maximum and median instead of the mean over annotations, with no significant differences in the results, thus we show the ones for the mean only.

Table 2 shows the quantitative meta-measure results for the test set of BSDS500, as well as which measures have a specific definition for multiple ground truths.

In global terms, F_b and F_{op} are the two top-ranked summary measures. On top of that, they both provide much richer information in form of precision-recall curves. Interestingly, the two measures are also the only ones with a specific definition for the multiple-ground-truth case (See Section 4), which reinforces the intuition that specifying the definition is a good choice. We believe, therefore, that the tandem F_b-F_{op} should be the evaluation measures of choice. Section 6.4 reinforces this choice by showing their complementarity in realistic scenarios.

Regarding the computational cost of the measures, the mean time per image to compute the distances to the multiple-partition ground truth of BSDS500 is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2481406, IEEE Transactions on Pattern Analysis and Machine Intelligence

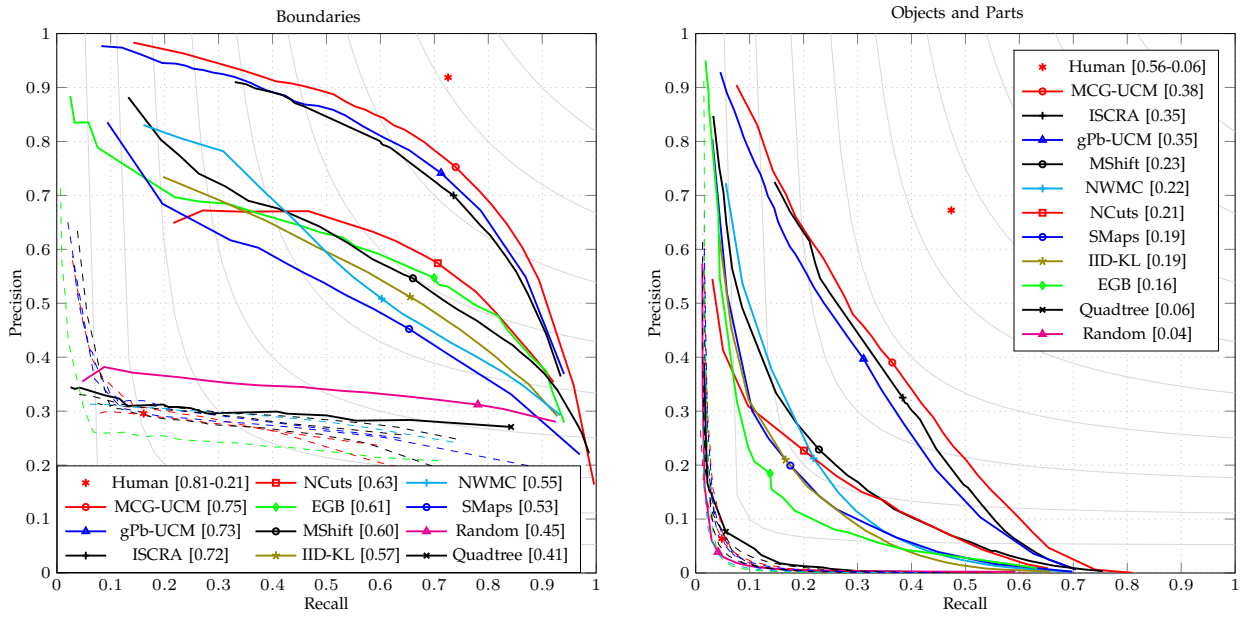

Fig. 11. **Precision-Recall curves for boundaries (left) and for objects and parts (right)**. The solid curves represent the nine SoA segmentation methods and the baselines (see legends). In dashed lines with the same color, the SoA techniques assessed on a swapped image. The marker on each curve is placed on the Optimal Dataset Scale (ODS), F measure in the legend. The isolated red asterisks refer to the human performance, i.e. ground truth partitions, assessed on the same image and on a swapped image.

Measure	Specific multiple	Quant. Meta-Meas. SIHD	SABD	SISD	Global
F_b	✓	**99.5**	93.6	**99.9**	**97.7**
F_{op}	✓	98.4	**94.8**	97.9	97.1
NVI	✗	96.7	83.3	96.8	92.3
$\mathcal{C}(S \rightarrow \{G_i\})$	✗	92.7	85.6	95.6	91.3
BCE	✗	93.1	79.6	95.7	89.5
PRI	✗	78.8	89.3	94.2	87.5
d_{vD}	✗	95.0	76.5	91.6	87.7
BGM	✗	90.2	78.8	93.2	87.4
$D_H(S \Rightarrow \{G_i\})$	✗	78.1	83.7	98.8	86.9
F_r	✗	89.3	76.4	93.3	86.3
$\mathcal{C}(\{G_i\} \rightarrow S)$	✗	91.4	72.0	90.9	84.8
$D_H(\{G_i\} \Rightarrow S)$	✗	73.8	58.1	77.1	69.7

TABLE 2
Measure comparison in terms of quant. meta-meas.

3.79 ± 2.06 s for F_b and at least one order of magnitude lower for the rest of measures. In particular, F_{op} takes 0.078 ± 0.020 s. In scenarios where the time constraints are tight, therefore, F_{op} would be the recommended measure (or the morphological approximation $\tilde{F}_b$).

6.4 Precision-Recall Frameworks

This section tests the proposed tandem of measures to compare a large set of state-of-the-art segmentation techniques, and evaluates the complementary behavior of F_b and F_{op}, which supports their use in tandem.

Figure 11 shows the boundary and objects-and-parts precision-recall curves for the nine SoA segmentation methods studied, the two baselines, and the human performance. Prior to the assessment of segmentation techniques, let us focus on the comparison of the two evaluation frameworks.

Precision-recall value ranges: The theoretical range of F_b and F_{op} values is [0,1]. To estimate the maximum *expectable* range of values of each measure in practice, we take advantage of the fact that BSDS500 contains various annotations per image. To estimate the maximum experimental value, we evaluate the ground-truth partitions against the partitions done by other individuals, in a leave-one-out way. In the other extreme, we estimate the minimum experimental value by evaluating the ground-truth partition of a given image against the ground-truth of a different image. We represent both extremes as red asterisks.

It is noticeable that the human minimum performance for F_b is 0.21, which could be interpreted as F_b being too lax. In this same direction, the baseline boundary precision for F_b is between 0.2 and 0.3, that is, any result, no matter how wrong it is, is judged as providing at least a 0.2 precision.

While in the case of F_{op} the human baseline is correctly downgraded to 0.05 (as well as the swapped-image results), then the surprising fact is that human maximum performance is as low as 0.56 (0.81 in F_b), which could entail that F_{op} is too strict.

To sum up, when judging results using both measures, one should take into account that the experimental range of values of F_b is 0.21-0.81 and that of F_{op} is 0.06-0.56, and extract the conclusions about their results with respect to these values.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2481406, IEEE Transactions on Pattern Analysis and Machine Intelligence

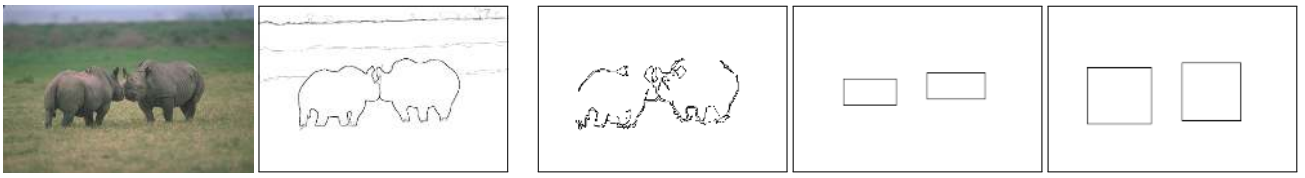

Fig. 12. F_b versus F_{op}: Complementary examples where the behavior of one of the measures is not the expected.

Analysis of the precision-recall curves: Regarding the comparison among segmentation techniques, both frameworks confirm that MCG-UCM outperforms the state of the art at all regimes under both measures. If we were to decide between gPb-UCM and ISCRA for the second place, however, gPb-UCM is consistently better in terms of boundary localization, while ISCRA outperforms gPb-UCM from the point of view of regions and parts.

The advantages of *going beyond* the summary measures are also clear on these plots. For instance, the summary F_b measure of quadtree (0.41) judges this technique close to NWMC (0.55), but in the precision-recall curves it is clear that quadtree is much worse. Similarly, judging by F_b, NWMC would be discarded with respect to NCuts for instance, but if we are interested in low recall rates it could be of interest.

As common points between the two measures, NCuts is judged as being much better at high recall rates than at low ones and, conversely, NWMC is much better at high precision rates. The measures are coherent also in the fact that human results have a better precision than recall. As one of the main discrepant points, however, EGB is judged as the fourth best technique by F_b while being the worse for F_{op}. This behavior is further analyzed below.

Qualitative results on complementary cases: Figure 12 shows an image and the associated ground truth. The EGB result (a) consists of thin long regions that surround the object but do not close to create the regions of interest. The assessment value of this result is $F_b = 0.62$ and $F_{op} = 0.05$. From a region-based point of view, this type of results is correctly penalized by F_{op} and not by F_b, since as a contour detector the result is correct.

We further compare the measures qualitatively by creating two academic examples (Figure 12 (b) and (c)) that show the complementary behavior, that is, examples where the F_{op} behavior is not intuitive. First, partition (b) is composed of two boxes completely included on the objects of interest. F_{op} interprets them as part candidates, since they are completely included in the objects and cover a significant part of them, so it does not penalize the partition significantly. On the other hand, F_b penalizes the result because the contours of the boxes do not overlap with the true boundaries. If we slightly increase the size of the boxes (Figure 12 (c)), however, making the contours overlap but having a small part of the boxes outside of the object, the situation is changed: the boxes are not considered parts anymore ($F_{op} = 0.04$) and the boundary measure does not judge the results as being very bad ($F_{op} = 0.28$).

To sum up, intutively, both measures can be *tricked* by incorrect results giving good evaluation values, but F_{op} will usually not fail when F_b does and viceversa. In other words, F_b and F_{op} are very complementary.

Qualitative results at ODS: Figure 13 shows example partitions from four SoA techniques and a baseline. For each of them we plot the ground-truth partitions (first row), and the partition at the Optimal Dataset Scale (ODS) with respect to F_b (second row) and F_{op} (third row).

We observe a general trend especially in the number of regions in the partitions. In the case of F_b, the partitions try to cover all the contour segments, even those marked only by one annotator, which usually leads to a number of small regions and over-fragmented results. On the other hand, the ODS partitions for F_{op} have less regions, increasing the probability of having a single region approximating each object but in exchange, they miss more annotated contours.

This behavior is also reflected in the baseline partitions done by Quadtree (last column), in which having many small rectangles (ODS F_b) entails a better probability to sweep annotated contours, while having only eight big rectangles (ODS F_{op}) is the best chance to overlap with an annotated object.

Experiments reproducibility: We present the package SEISM [65] (Supervised Evaluation of Image Segmentation Methods), which makes the code to compute all the measures publicly available, as well as all the segmentation results and scripts to make our research reproducible and to make it effortless for researchers to assess their segmentation methods.

Conclusions of the experiments: To sum up, both measures are complementary in terms of the properties of the partitions they evaluate, they both provide useful precision-recall curves, they achieve the best meta-measure results as summary measures, they have specific definitions for multiple ground truth, and their code is public to ensure reproducibility; thus we propose them in tandem as the tool of choice for image segmentation evaluation.

7 Conclusions

This paper reviews and structures an extensive set of segmentation evaluation measures, showing that the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2481406, IEEE Transactions on Pattern Analysis and Machine Intelligence

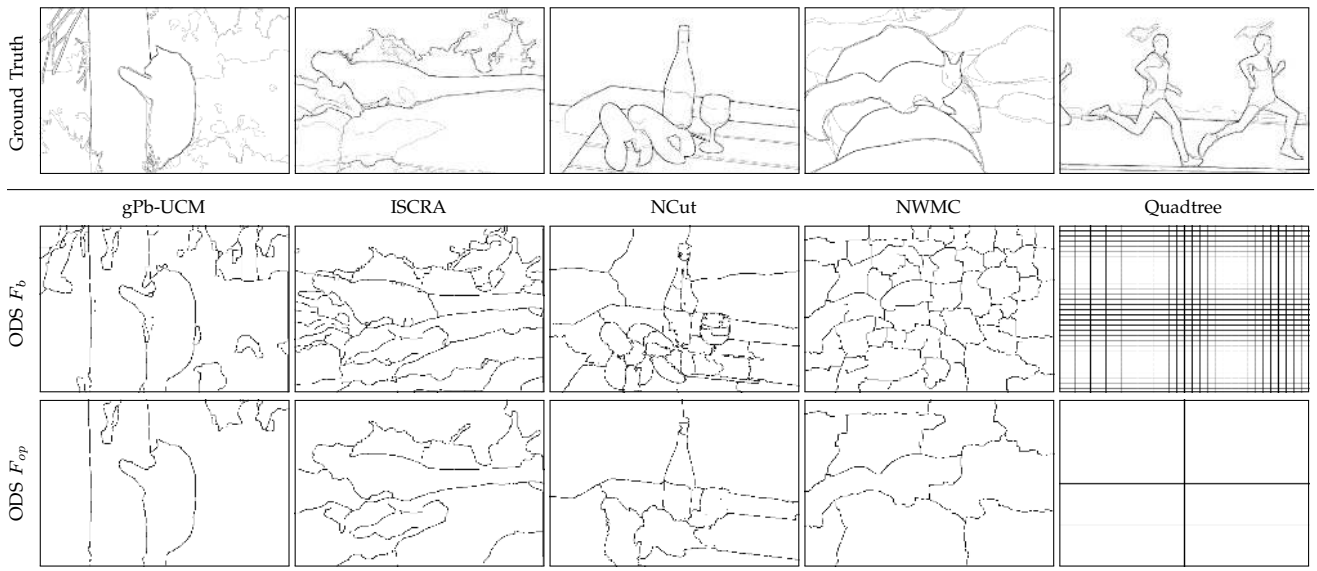

Fig. 13. **Qualitative comparison of F_b and F_{op}**: ODS partitions with respect to both measures

Jaccard index and the F measure are equivalent, and presents the new precision-recall measure for objects and parts. Three meta-measures are used (two newly proposed) to quantitatively compare the goodness of the evaluation measures. The results show that the tandem boundary and objects-and-parts precision-recall curves is a good candidate for benchmarking segmentation algorithms; since apart from obtaining the best meta-measure results as summary measures, their precision-recall curves provide rich knowledge about the results and they are very complementary in terms of the properties of the partitions they reflect. In the object-based analysis, we propose the pixel- and boundary-based pair of measures, and three generalization strategies to evaluate object proposals. We perform an extensive experimental validation on eight state-of-the-art object proposal techniques and on nine generic image segmentation techniques. By making our code and datasets publicly available we allow researchers to easily assess their results and gain deeper understanding of their algorithms.

Acknowledgements: This work was partially supported by ERDF project BIGGRAPH-TEC2013-43935-R and FPU grant AP2008-01164.

REFERENCES

[1] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *CVPR*, 2001.
[2] C. Lampert, M. Blaschko, and T. Hofmann, "Beyond sliding windows: Object localization by efficient subwindow search," in *CVPR*, 2008.
[3] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014.
[4] B. Hariharan, P. Arbelez, R. Girshick, and J. Malik, "Simultaneous detection and segmentation," in *ECCV*, 2014.
[5] P. Arbeláez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," in *CVPR*, 2014.
[6] J. Carreira and C. Sminchisescu, "CPMC: Automatic object segmentation using constrained parametric min-cuts," *IEEE TPAMI*, vol. 34, no. 7, pp. 1312–1328, 2012.
[7] P. Arbeláez, B. Hariharan, C. Gu, S. Gupta, L. Bourdev, and J. Malik, "Semantic segmentation using regions and parts," in *CVPR*, 2012.
[8] A. Ion, J. Carreira, and C. Sminchisescu, "Image segmentation by figure-ground composition into maximal cliques," in *ICCV*, 2011.
[9] A. Levinshtein, C. Sminchisescu, and S. Dickinson, "Optimal contour closure by superpixel grouping," in *ECCV*, 2010.
[10] P. Rantalankila, J. Kannala, and E. Rahtu, "Generating object segmentation proposals using global and local search," in *CVPR*, 2014.
[11] I. Endres and D. Hoiem, "Category-independent object proposals with diverse ranking," *IEEE TPAMI*, vol. 36, no. 2, pp. 222–234, 2014.
[12] D. Hoiem, Y. Chodpathumwan, and Q. Dai, "Diagnosing error in object detectors," in *ECCV*, 2012.
[13] D. Martin, "An empirical approach to grouping and segmentation," Ph.D. dissertation, EECS Department, University of California, Berkeley, Aug 2003.
[14] P. Arbeláez, M. Maire, C. C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE TPAMI*, vol. 33, no. 5, pp. 898–916, 2011.
[15] R. Unnikrishnan, C. Pantofaru, and M. Hebert, "Toward objective evaluation of image segmentation algorithms," *IEEE TPAMI*, vol. 29, no. 6, pp. 929–944, 2007.
[16] M. Meilă, "Comparing clusterings: an axiomatic view," in *ICML*, 2005.
[17] J. Pont-Tuset and F. Marques, "Supervised assessment of segmentation hierarchies," in *ECCV*, 2012.
[18] A. Hoover et al., "An experimental comparison of range image segmentation algorithms," *IEEE TPAMI*, vol. 18, pp. 673–689, 1996.
[19] C. Gu, J. Lim, P. Arbelaez, and J. Malik, "Recognition using regions," in *CVPR*, 2009.
[20] F. Li, J. Carreira, and C. Sminchisescu, "Object recognition as ranking holistic figure-ground hypotheses," in *CVPR*, june 2010, pp. 1712 –1719.
[21] J. S. Cardoso and L. Corte-Real, "Toward a generic evaluation of image segmentation," *IEEE TIP*, vol. 14, no. 11, pp. 1773–1782, 2005.
[22] X. Jiang, C. Marti, C. Irniger, and H. Bunke, "Distance measures for image segmentation evaluation," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. 1–10, 2006.
[23] J. Pont-Tuset and F. Marques, "Measures and meta-measures for the supervised evaluation of image segmentation," in *CVPR*, 2013.
[24] J. Czekanowski, "Zarys metod statystycznych w zastosowaniu do antropologii," *Prace Towarzystwa Naukowego Warszawskiego*, vol. 5, 1913.

[25] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. pp. 297–302, 1945.
[26] T. Sørensen, "A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons," *Biologiske Skrifter*, vol. 5, pp. 1–34, 1948.
[27] S. Alpert, M. Galun, A. Brandt, and R. Basri, "Image segmentation by probabilistic bottom-up aggregation and cue integration," *IEEE TPAMI*, vol. 34, no. 2, pp. 315 –327, 2012.
[28] K. Smith, D. Gatica-Perez, J.-M. Odobez, and S. Ba, "Evaluating multi-object tracking," in *CVPR*. IEEE, 2005.
[29] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler, "MOT challenge 2015: Towards a benchmark for multi-target tracking," *arXiv:1504.01942*, 2015.
[30] K. H. Zou, S. K. Warfield, A. Bharatha, C. M. Tempany, M. R. Kaus, S. J. Haker, W. M. Wells, F. A. Jolesz, and R. Kikinis, "Statistical validation of image segmentation quality based on a spatial overlap index," *Academic radiology*, vol. 11, no. 2, pp. 178–189, 2004.
[31] P. Jaccard, "Étude comparative de la distribution florale dans une portion des alpes et des jura," *Bulletin de la Société Vaudoise des Sciences Naturelles*, vol. 37, pp. 547–579, 1901.
[32] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2010 Results," http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html, 2010.
[33] K. McGuinness and N. E. O'connor, "A comparative evaluation of interactive segmentation algorithms," *Pattern Recognition*, vol. 43, pp. 434–444, 2010.
[34] F. Ge, S. Wang, and T. Liu, "Image-segmentation evaluation from the perspective of salient object extraction," in *CVPR*, 2006.
[35] ——, "New benchmark for image segmentation evaluation," *Electronic Imaging*, vol. 16, no. 3, pp. 033 011.1–16, 2007.
[36] T. Malisiewicz and A. A. Efros, "Improving spatial support for objects via multiple segmentations," in *British Machine Vision Conference*, 2007.
[37] B. Russell, W. Freeman, A. Efros, J. Sivic, and A. Zisserman, "Using multiple segmentations to discover objects and their extent in image collections," in *CVPR*, 2006.
[38] A. Suinesiaputra et al., "Left ventricular segmentation challenge from cardiac MRI: A collation study," in *International Conference on Statistical Atlases and Computational Models of the Heart*, 2012.
[39] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?" in *CVPR*, 2014.
[40] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *PAMI*, vol. 24, no. 4, pp. 509–522, 2002.
[41] R. C. Veltkamp, "Shape matching: Similarity measures and algorithms," in *Shape Modeling and Applications*. IEEE, 2001, pp. 188–197.
[42] F. J. Estrada and A. D. Jepson, "Benchmarking image segmentation algorithms," *IJCV*, vol. 85, pp. 167–181, 2009.
[43] A. Humayun, F. Li, and J. M. Rehg, "RIGOR: Recycling Inference in Graph Cuts for generating Object Regions," in *CVPR*, 2014.
[44] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *IJCV*, vol. 104, no. 2, pp. 154–171, 2013.
[45] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE TPAMI*, vol. 34, pp. 2189–2202, 2012.
[46] Q. Huang and B. Dom, "Quantitative methods of evaluating image segmentation," in *ICIP*, 1995.
[47] S. Dongen, "Performance criteria for graph clustering and markov cluster experiments," Centrum voor Wiskunde en Informatica (CWI), Amsterdam, The Nederlands, Tech. Rep. INS-R0012, 2000.
[48] W. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, 1971.
[49] G. Liu and R. Haralick, "Assignment problem in edge detection performance evaluation," in *CVPR*, 2000.
[50] J. Pont-Tuset, "Image segmentation evaluation and its application to object detection," Ph.D. dissertation, Universitat Politècnica de Catalunya, UPC BarcelonaTech, 2014. [Online]. Available: http://jponttuset.github.io
[51] T. Kanungo, B. Dom, W. Niblack, and D. Steele, "A fast algorithm for MDL-based multi-band image segmentation," IBM Reasearch Division, RJ 9754 (84640), Tech. Rep., 1994.
[52] S. Nowozin, P. Gehler, and C. Lampert, "On parameter learning in crf-based approaches to object class image segmentation," in *ECCV*, 2010.
[53] 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 *ICCV*, 2001.
[54] M. Everingham, H. Muller, and B. Thomas, "Evaluating image segmentation algorithms using the pareto front," in *European Conference on Computer Vision*, 2006, pp. 255–259.
[55] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels," EPFL, Tech. Rep. 149300, 2010.
[56] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC Superpixels Compared to State-of-the-art Superpixel Methods," *IEEE TPAMI*, vol. 34, no. 11, pp. 2274 – 2282, 2012.
[57] P. Krähenbühl and V. Koltun, "Geodesic object proposals," in *ECCV*, 2014.
[58] Z. Ren and G. Shakhnarovich, "Image segmentation by cascaded region agglomeration," in *CVPR*, 2013.
[59] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *IJCV*, vol. 59, p. 2004, 2004.
[60] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE TPAMI*, vol. 22, no. 8, pp. 888–905, 2000.
[61] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE TPAMI*, vol. 24, no. 5, pp. 603 –619, may 2002.
[62] V. Vilaplana, F. Marques, and P. Salembier, "Binary partition trees for object detection," *IEEE TIP*, vol. 17, no. 11, pp. 2201–2216, 2008.
[63] F. Calderero and F. Marques, "Region merging techniques using information theory statistical measures," *IEEE TIP*, vol. 19, no. 6, pp. 1567–1586, 2010.
[64] L. Najman and M. Schmitt, "Geodesic saliency of watershed contours and hierarchical segmentation," *IEEE TPAMI*, vol. 18, no. 12, pp. 1163–1173, 1996.
[65] J. Pont-Tuset and F. Marques, "SEISM: Supervised Evaluation of Image Segmentation Methods," 2013. [Online]. Available: http://goo.gl/OLIXZV

Jordi Pont-Tuset is a post-doctoral researcher at ETHZ, Switzerland, in Prof. Luc Van Gool's computer vision group (2015). He received the degree in Mathematics in 2008, the degree in Electrical Engineering in 2008, the M.Sc. in Research on Information and Communication Technologies in 2010, and the Ph.D in 2014; all from the Universitat Politècnica de Catalunya, BarcelonaTech (UPC). He worked at Disney Research, Zürich (2014).

Ferran Marques received the degree in Electrical Engineering and the Ph.D. from the Universitat Politècnica de Catalunya, BarcelonaTech (UPC), where he is currently Professor at the department of Signal Theory and Communications. He has been President of the European Association for Signal Processing (EURASIP), Associate Editor of the IEEE Transactions on Image Processing, and Area Editor for Signal Processing: Image Communication (Elsevier). Currently, he is Dean of the Electrical Engineering School at UPC. He has published over 150 conference and journal papers, 2 books, and holds 4 international patents.