Measures and Meta-Measures for the Supervised Evaluation of Image Segmentation

Jordi Pont-Tuset and Ferran Marques
Universitat Politècnica de Catalunya BarcelonaTech*
http://imatge.upc.edu

Abstract

This paper tackles the supervised evaluation of image segmentation algorithms. First, it surveys and structures the measures used to compare the segmentation results with a ground truth database; and proposes a new measure: the precision-recall for objects and parts. To compare the goodness of these measures, it defines three quantitative meta-measures involving six state of the art segmentation methods. The meta-measures consist in assuming some plausible hypotheses about the results and assessing how well each measure reflects these hypotheses. As a conclusion, this paper proposes the 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.

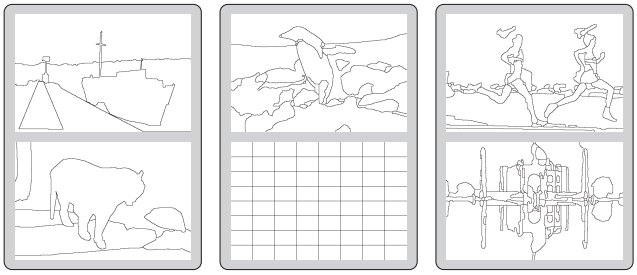
1. Introduction

Since the advent of sliding window object detectors [32], much effort has been put into providing better spatial delineation beyond sliding windows [16]. Semantic segmentation is the final objective, where detection and segmentation meet, but it is still far from being solved [8].

In this scenario, bottom-up segmentation methods often play an important role in the proposed algorithms [1, 5], and thus improving segmentation techniques would entail improvements towards better semantic segmentation [18]. In such a challenge, providing benchmarks that help researchers understand the weak and strong points of their algorithms is of paramount importance.

In this direction, in the field of object detection assessment, Hoiem *et al*. [11] stress that the results should be evaluated beyond performance summary measures in order to "help understand how one method could be improved."

*This work has been partially supported by the Spanish *Ministerio de Ciencia e Innovación*, under project TEC2010-18094 and FPU grant AP2008-01164.



(a) (b) (c)

Figure 1. Examples of the meta-measure principles: How good are the evaluation measures at distinguishing these pairs of partitions?

In other words, researchers need better feedback from the evaluation than a single number.

Back to segmentation assessment, the precision-recall curves for boundaries [20] are good examples of tools that provide richer feedback than the F-measure used as summary. Moreover, as pointed out by [2], in addition to boundary-based measures, region-oriented measures should be considered when assessing segmentations. However, the current ones are limited to summary measures [30, 22, 20, 2, 15, 13, 7, 4, 25, 24].

Our first contribution is a region-based precision-recall environment for the assessment of image segmentation. Inspired by [12, 11] and by the fact that parts of objects are important clues for object detection [9], we present the **precision-recall for objects and parts**, which is based on classifying the regions into object and parts candidates.

Summary measures also play a role in performance comparison, 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 analyze 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 particular qualitative examples [4, 30]. Extensive quantitative meta-measures, however, are desirable.

The first approach to an extensive quantitative meta-measure was proposed in [19]. 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 instances, 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. [14] presented a comparison of some measures in terms of this meta-measure.

The second contribution of this paper is to present two new meta-measures. Instead of basing our hypotheses on human-made partitions, we extend the analysis to partitions from six State-of-the-Art (SoA) segmentation algorithms.

The first assumption is that measures should be capable of distinguishing such partitions from those obtained without taking into account the content of the image. In our case, following the proposal in [2], we use a quadtree, *i.e.*, a hierarchical homogeneous rectangular grid. The meta-measure is then defined as the number of results from SoA algorithms that are judged worse than the quadtree. As a qualitative example, we assess how well a measure distinguishes between partitions like Figure 1.b.

As a second approach, we assume that any measure should be able to distinguish a partition obtained by a SoA method on an image from a partition obtained by the same method but on a different image, as the two 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.

The third contribution is to survey and structure a wide set of evaluation measures and the newly-proposed one and compare them using the three previously discussed 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, while providing richer information for researchers to interpret the results. We further interpret these two precision-recall environments by comparing six SoA segmentation algorithms.

We make the code to compute all the measures publicly available in [28], as well as all the segmentation results to make our research reproducible and to make it effortless for researchers to assess their segmentation methods.

The remainder of the paper is organized as follows. Section 2 reviews and structures the main segmentation measures available in the literature. Section 3 motivates and describes the newly proposed measure. Section 4 presents the two new meta-measures and the already available one used to compare the evaluation measures. Section 5 presents the experimental comparison of the measures using the three meta-measures. It also shows the applicability of the boundary-based and the newly proposed precision-recall curves for objects and parts in the comparison of six SoA segmentation techniques. Section 6 concludes the paper.

2. Measure Review and Structure

The state-of-the-art 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, 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. Finally, a partition can be represented as a two-class clustering of the pixel contours into boundaries and non-boundaries.

The following sections review the main measures found under each of these interpretations, keeping the notation from the original papers where possible. Table 1 shows an overview of the studied measures.

2.1. Pixel-Set Clustering

The **directional Hamming distance** from one partition S to another S' [15, 13] is defined as:

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

where R and R' are regions in S and S', respectively, and n is the number of pixels in the image. In [4] this same measure was coined as asymmetric partition distance. It is equivalent to the achievable segmentation accuracy [23] used in superpixel assessment.

A symmetric version of this measure was presented in [7] as the **van Dongen distance**:

$$d_{vD}(S, S') = D_H\left(S' \Rightarrow S\right) + D_H\left(S \Rightarrow S'\right) \qquad (2)$$

The **segmentation covering** of a partition S by a partition S' was defined in [2] as:

$$\mathcal{C}\left(S' \rightarrow S\right) = \frac{1}{n} \sum_{R \in S} |R| \cdot \max_{R' \in S'} \frac{|R \cap R'|}{|R \cup R'|} \qquad (3)$$

The intuitive step further is to measure the maximum overlap when performing a bijective matching between the regions of the two partitions. This idea was presented in [4] as symmetric partition-distance, in [14] as **bipartite-graph-matching** (*BGM*) distance, and in the context of clustering comparison, in [22] as classification error distance. It is shown in [4] 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 [19], the consistency of the BSDS300 human partitions is analyzed by means of two measures *GCE*, *LCE*, aiming at being robust against different granularities of the

Partition Interpretation	Measure Representative	References	Notation
Pixel-set clustering	Directional Hamming distance	[13, 4]	D_H
	van Dongen distance	[7]	d_{vD}
	Segmentation covering	[2]	$\mathcal{C}$
	Bipartite graph matching	[14, 4]	*BGM*
	Bidirectional consistency error	[19]	*BCE*
	Variation of information	[22]	*VoI*
Pairs-of-pixels classification	Probabilistic Rand index	[26, 30]	*PRI*
	Precision-Recall for regions	[19]	P_r, R_r
Boundary map	Precision-Recall for boundaries	[17, 19]	P_b, R_b

Table 1. Measure structure overview for the three interpretations of an image partition

scene interpretation. As the author points out, these measures are not suitable for general-purpose image segmentation evaluation. The same work proposes a measure that is not transparent to oversegmentation: the bidirectional consistency error (*BCE*), which can be rewritten as:

$$BCE(S,S') = 1 - \frac{1}{n} \sum_{\substack{R \in S \\ R' \in S'}} |R \cap R'| \min\left\{ \frac{|R \cap R'|}{|R|}, \frac{|R \cap R'|}{|R'|} \right\} \tag{4}$$

The work in [22] 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 **variation of information** is then:

$$VoI(S, S') = H(S) + H(S') - 2I(S, S') \tag{5}$$

It can be normalized by $\log N$, its maximum possible value.

2.2. Pairs-of-Pixels Classification

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 [22]:

$\mathcal{P}_{11}$: 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 [26] 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** [30] is computed as:

$$PRI(S, \{G_i\}) = \sum_i RI(S,G_i) \tag{6}$$

In this same context, the **precision-recall for regions** [19] is defined as:

$$P_r = \frac{|\mathcal{P}_{11}|}{|\mathcal{P}_{11}| + |\mathcal{P}_{10}|} \qquad R_r = \frac{|\mathcal{P}_{11}|}{|\mathcal{P}_{11}| + |\mathcal{P}_{01}|} \tag{7}$$

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

This pair of measures would be a candidate in our quest for a non-boundary-based precision-recall measure. As it will be shown in the experiments, however, this measure does not provide good meta-evaluation scores.

2.3. Boundary Map

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, [17] 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 [19, 20] leading to the well-known **precision-recall for boundaries** (P_b, R_b, and F_b).

3. Measure Proposal

In the context of image segmentation evaluation, precision-recall curves for boundaries [19, 20] 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 pointed out by [2], 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 2.

This section presents a new region benchmark that goes beyond the summary measures: the precision-recall for objects and parts. Motivated by the fact that image segmentation is increasingly being used as a preliminary step for object detection [18, 1], 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. Similarly, we interpret regions in an oversegmentation as parts of objects, if merged together can form an object of the ground truth (inspired by [12] in range image segmentation evaluation).

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

Formally, let $S=\{R_1,\dots,R_N\}$ be an image partition and $\{G_k\}$ a set of ground-truth partitions of the same image. We consider the set $G=\{R'_1,\dots,R'_M\}$ 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.

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:

$$fr(R_i)=\sum_j \left\{ O_G^{ij} \text{ s.t. } O_S^{ij} > \gamma_o \right\} \tag{8}$$

and $fr'(R'_j)$ is defined equivalently for G. The global fragmentation fr and fr' is computed adding the amount of fragmentation among all the fragmentation candidates of S and G, respectively. Figure 2 shows a toy example to illustrate the proposed classification and measures.

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

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

$$P_{op}=\frac{oc+fr+\beta\, pc}{|S|} \qquad R_{op}=\frac{oc'+fr'+\beta\, pc'}{|G|} \tag{9}$$

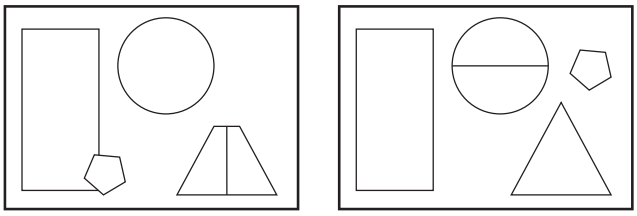


Figure 2. Classification of the regions into object and part candidates. The rectangles are classified as object candidates, despite not fully overlapping. The partition circle is a fragmentation candidate with a fragmentation of 1 (parts cover it totally), and the ground-truth half-circles are parts candidates. The opposite holds for the triangles, but in this case the fragmentation is 0.9. Both pentagons are classified as noise.

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}.

4. Meta-Measures

This section is about how to compare the goodness of the segmentation evaluation measures. The objective of this section is therefore not to tell which segmentation algorithm to use, but which evaluation measures better summarize the quality of these algorithms. To distinguish these two analyses, we will refer to the quantitative metrics to compare segmentation measures as *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 examples, 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 [30, 4].

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. 4.1) and presents two new meta-measures (Sec. 4.2 and 4.3).

4.1. 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 [21] and BSDS500 [2]) 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 measure 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 [19] 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 3 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

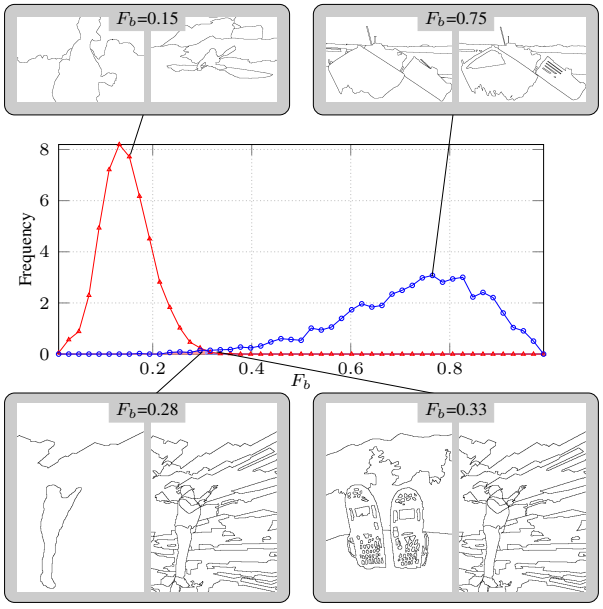


Figure 3. Distribution of F_b for the same-image pairs of partitions (—○—) and different-image pairs (—▲—). In gray rectangles, four representative pairs of partitions: a pair of correctly classified as different image (up-left) and as same image (up-right); and a pair incorrectly classified as different image (down-left) and as same image (down-right).

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, the authors reported the Bayes Risk.)

As qualitative examples, Figure 3 depicts four pairs of partitions as representatives of the type of mistakes and correct classifications using F_b.

4.2. 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 subsection 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 be able to distinguish between (i) partitions obtained by any SoA segmentation method on a given image and (ii) partitions obtained regardless of the image, that is, partitions that are created without taking into account the content of the image. These partitions are interpreted as a baseline, that is, the results that could be obtained *by chance*.

As in [2], we use a quadtree as baseline. In particular, we build the hierarchical partitions starting from the whole

image and iteratively dividing the regions into four equal rectangles. Figure 1.b shows an example of partition obtained by a SoA method and by a quadtree.

For each of the techniques considered as SoA segmentation methods, 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.

4.3. 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. To incorporate this type of comparisons to the meta-measures, we compare (i) the results created by a SoA segmentation technique with (ii) the results created by that same algorithm but on a different image.

In other words, we compare the ground-truth of a certain image with two results obtained using the same algorithm and parameterization: (i) one segmentation of that same image and (ii) 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 example of Figure 1.c, the measure should judge that the first partition is better than the second one compared both with the ground-truth of the former. 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** as the percentage of results in the database, for all the SoA methods, that the measures correctly discriminates.

5. Experimental Validation

The state of the art of segmentation is represented in this paper by the following six methods: the Ultrametric Contour Maps on the gPb contour detector (gPb-OWT-UCM) [2], the Efficient Graph-Based (EGB) image segmentation algorithm [10], the Mean Shift (MShift) algorithm [6], the Normalized Cuts (NCuts) algorithm [29], and two types of Binary Partition Trees [27]: the Normalized Weighted Euclidean distance between Models with Contour complexity (NWMC) tree [31], and the Independent Identically Distributed - Kullback Leibler (IID-KL) tree [3]. The exact parameterizations for each algorithm is detailed at [28], where we also publish the code of all measures and meta-measures used in this work. All methods are assessed at the Optimal Dataset Scale (ODS) [2] with respect to each evaluation measure.

Measure	Global Meta-Meas.	Meta-Measure SIHD	SABD	SISD
F_b	98.4	99.5	95.6	100.0
F_{op}	96.7	98.4	94.2	97.5
VoI	94.0	96.9	87.5	97.7
$\mathcal{C}(S \rightarrow \{G_i\})$	91.5	93.1	86.0	95.3
d_{vD}	90.7	95.1	86.9	90.1
$D_H(S \Rightarrow \{G_i\})$	89.5	78.5	91.3	98.8
BCE	89.2	93.3	78.9	95.4
BGM	88.1	90.7	81.6	92.0
PRI	86.7	77.7	88.8	93.7
$\mathcal{C}(\{G_i\} \rightarrow S)$	86.3	91.3	77.4	90.1
F_r	86.1	77.0	84.2	97.1
$D_H(\{G_i\} \Rightarrow S)$	80.5	73.0	92.1	76.5

Table 2. Measure comparison in terms of quantitative meta-measures. Values refer to percentages of correct results

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 [2], optimizing the global meta-measure described in the following section (See Table 2). Note that this optimization would not have been feasible without such quantitative meta-measures.

Meta-Measures Results: Table 2 shows the three meta-measure results for the test set of BSDS500, as well as a global summary meta-measure. Given that each meta-measure represents a percentage of correct results, we define the global meta-measure as the global percentage of correct results.

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, thus we propose the pair F_b-F_{op} as the measures of choice.

Regarding the computational cost of the measures, the mean time for image to compute the distances to the multiple-partition ground truth of BSDS500 is 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 limitations are tight, the authors believe that F_{op} would be the tool of choice. To provide an in-depth analysis of the final results, the tandem of precision-recall curves for boundaries and for objects-and-parts would be the most adequate option. The following section provides a thorough analysis of both frameworks on the six SoA methods used in this paper.

Precision-Recall Frameworks: Figure 4 shows the boundary and objects-and-parts precision-recall curves for

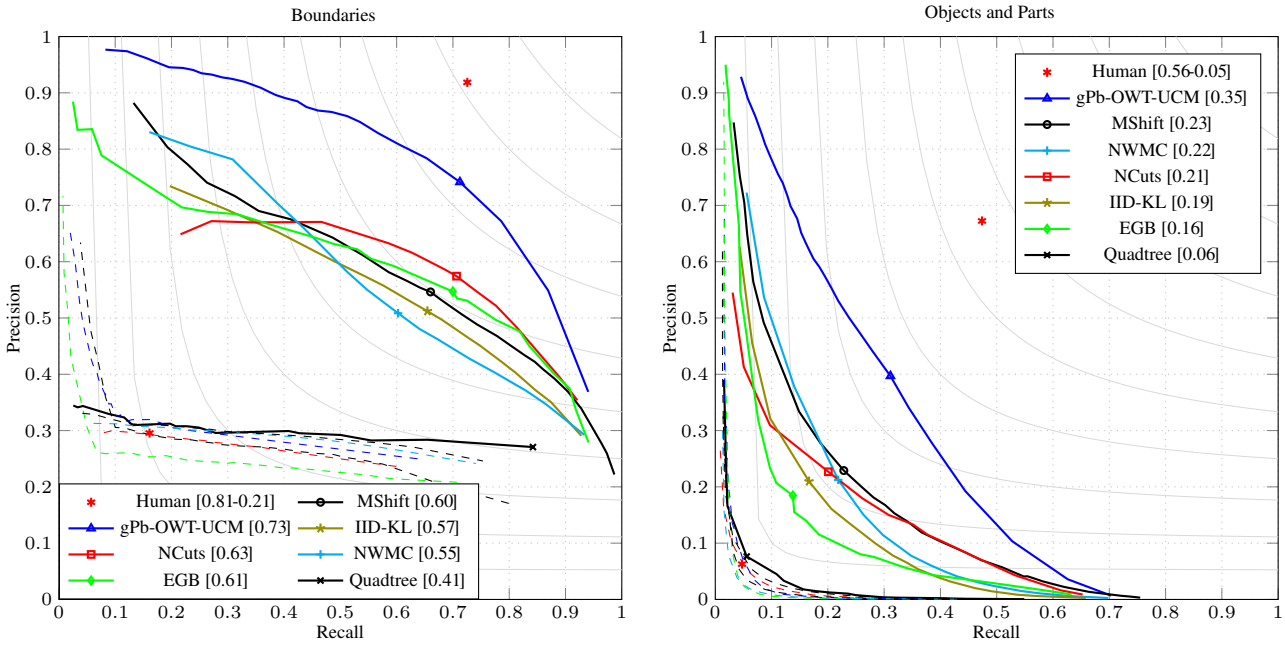


Figure 4. Precision-Recall curves for boundaries (left) and for objects and parts (right). The solid curves represent the six SoA segmentation methods and the quadtree (see legends). In dashed lines with the same color, the SoA techniques assessed on a swapped image. The marker on each curve is placed on the Optimal Dataset Scale (ODS). The isolated red asterisks refer to the human performance assessed on the same image and on a swapped image. In the legend, the F measure of the marked point on each curve is presented in brackets.

the six SoA segmentation methods studied and the human performance. Prior to the assessment of segmentation techniques, let us focus on the comparison of the two evaluation frameworks.

It is noticeable that the human baseline performance (human assessed on a different image) 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, will be 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 performance is as low as 0.56 (0.81 in F_b), which could entail that F_{op} is too strict.

Although the dynamic range is a little higher in F_b (0.60 versus 0.51), the gap between the best method and humans is much higher in F_{op} (0.08 versus 0.21). In other words, F_{op} gives more resolution at the places where improvements over the SoA would be placed.

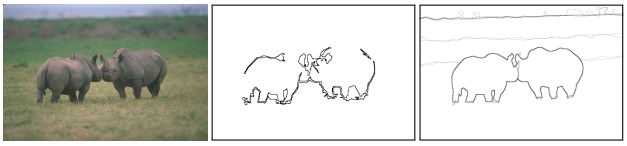
Regarding the comparison among segmentation techniques, both frameworks confirm that the gPb-OWT-UCM technique has outstanding results with respect to the rest.

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 clearly discarded but if we are interested in low recall rates it could be of interest (apart from gPb-OWT-UCM).

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 third best technique by F_b while being the worse for F_{op}. To further analyze this behavior, Figure 5 shows an image (a), an EGB result (b), and the associated ground truth (c). The EGB result consists of thin long regions that surround the object but do not close. 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.

To sum up, both measures are complementary thus we propose them in tandem as the tool of choice for image segmentation evaluation.



(a) (b) (c)

Figure 5. EGB result correctly penalized by F_{op} but not by F_b

6. Conclusions

This paper reviews an extensive set of segmentation evaluation measures 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, their precision-recall curves provide rich knowledge about the results. By making our code and datasets publicly available we allow researchers to easily assess their results and gain deeper understanding of their algorithms.

References

[1] P. Arbeláez, B. Hariharan, C. Gu, S. Gupta, L. Bourdev, and J. Malik. Semantic segmentation using regions and parts. In *CVPR*, 2012. 1, 4

[2] P. Arbeláez, M. Maire, C. C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *IEEE TPAMI*, 33(5):898–916, 2011. 1, 2, 3, 4, 5, 6

[3] F. Calderero and F. Marques. Region merging techniques using information theory statistical measures. *IEEE TIP*, 19(6):1567–1586, 2010. 6

[4] J. S. Cardoso and L. Corte-Real. Toward a generic evaluation of image segmentation. *IEEE TIP*, 14(11):1773–1782, 2005. 1, 2, 3, 5

[5] J. Carreira and C. Sminchisescu. Constrained parametric min-cuts for automatic object segmentation. In *CVPR*, 2010. 1

[6] D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *IEEE TPAMI*, 24(5):603 –619, 2002. 6

[7] S. Dongen. Performance criteria for graph clustering and markov cluster experiments. Technical Report INS-R0012, Centrum voor Wiskunde en Informatica (CWI), Amsterdam, The Nederlands, 2000. 1, 2, 3

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html. 1

[9] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE TPAMI*, 32(9):1627–1645, 2010. 1

[10] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 59:2004, 2004. 6

[11] D. Hoiem, Y. Chodpathumwan, and Q. Dai. Diagnosing error in object detectors. In *ECCV*, 2012. 1

[12] A. Hoover, G. Jean-Baptiste, X. Jiang, P. Flynn, H. Bunke, D. Goldgof, K. Bowyer, D. Eggert, A. Fitzgibbon, and R. Fisher. An experimental comparison of range image segmentation algorithms. *IEEE TPAMI*, 18:673–689, 1996. 1, 4

[13] Q. Huang and B. Dom. Quantitative methods of evaluating image segmentation. In *ICIP*, 1995. 1, 2, 3

[14] X. Jiang, C. Marti, C. Irniger, and H. Bunke. Distance measures for image segmentation evaluation. *EURASIP J. Appl. Signal Process.*, 2006:1–10, 2006. 2, 3

[15] T. Kanungo, B. Dom, W. Niblack, and D. Steele. A fast algorithm for MDL-based multi-band image segmentation. Technical report, IBM Reasearch Division, RJ 9754 (84640), 1994. 1, 2

[16] C. Lampert, M. Blaschko, and T. Hofmann. Beyond sliding windows: Object localization by efficient subwindow search. In *CVPR*, 2008. 1

[17] G. Liu and R. Haralick. Assignment problem in edge detection performance evaluation. In *CVPR*, 2000. 3, 4

[18] T. Malisiewicz and A. A. Efros. Improving spatial support for objects via multiple segmentations. In *BMVC*, 2007. 1, 4

[19] D. Martin. *An Empirical Approach to Grouping and Segmentation*. PhD thesis, EECS Department, University of California, Berkeley, Aug 2003. 2, 3, 4, 5

[20] D. Martin, C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE TPAMI*, 26(5):530–549, 2004. 1, 4

[21] 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. 5

[22] M. Meilă. Comparing clusterings: an axiomatic view. In *ICML*, 2005. 1, 2, 3

[23] S. Nowozin, P. Gehler, and C. Lampert. On parameter learning in crf-based approaches to object class image segmentation. In *ECCV*, 2010. 2

[24] B. Peng and L. Zhang. Evaluation of image segmentation quality by adaptive ground truth composition. In *ECCV*, 2012. 1

[25] J. Pont-Tuset and F. Marques. Supervised assessment of segmentation hierarchies. In *ECCV*, 2012. 1

[26] W. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971. 3

[27] P. Salembier and L. Garrido. Binary partition tree as an efficient representation for image processing, segmentation, and information retrieval. *IEEE TIP*, 9(4):561–576, 2000. 6

[28] Segmentation Evaluation Code. https://imatge.upc.edu/web/resources/supervised-evaluation-image-segmentation. 2, 6

[29] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE TPAMI*, 22(8):888–905, 2000. 6

[30] R. Unnikrishnan, C. Pantofaru, and M. Hebert. Toward objective evaluation of image segmentation algorithms. *IEEE TPAMI*, 29(6):929–944, 2007. 1, 3, 5

[31] V. Vilaplana, F. Marques, and P. Salembier. Binary partition trees for object detection. *IEEE TIP*, 17(11):2201–2216, 2008. 6

[32] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001. 1