What is a good evaluation measure for semantic segmentation?

Gabriela Csurka
gabriela.csurka@xrce.xerox.com
Diane Larlus
diane.larlus@xrce.xerox.com
Florent Perronnin
florent.perronnin@xrce.xerox.com

Computer Vision Group
Xerox Research Centre Europe
Meylan, France

Abstract

In this work, we consider the evaluation of the semantic segmentation task. We discuss the strengths and limitations of the few existing measures, and propose new ways to evaluate semantic segmentation. First, we argue that a per-image score instead of one computed over the entire dataset brings a lot more insight. Second, we propose to take contours more carefully into account. Based on the conducted experiments, we suggest best practices for the evaluation. Finally, we present a user study we conducted to better understand how the quality of image segmentations is perceived by humans.

1 Introduction

The goal of semantic segmentation is to assign each pixel of a photograph to one of several semantic class labels (or to none of them). This is a supervised learning problem which requires training a set of classifiers from data labelled at the pixel level. This is in contrast with unsupervised image segmentation, which consists in partitioning an image into coherent regions according to class-independent low-level cues, such as pixel colour or proximity. Semantic segmentation has many potential applications including scene understanding [13], removing undesired objects from photographs [21], copy-pasting objects from one photograph to another, or local class-based image enhancement. We note that these diverse applications have different requirements when it comes to judging whether the semantic segmentation algorithm has made "a good job". For instance, for the first application it might be sufficient to segment the scene into rough blobs. On the other hand, Kohli *et al*. argued [7] that in computer graphics applications where the goal of the segmentation process is to produce an alpha channel, having a precise delineation of the contours is important.

Ideally, the success of the segmentation algorithm should be measured by the success of the end application. As this is generally too difficult to evaluate, especially in graphics applications where the evaluation often requires a user study, the computer vision community has resorted to application-independent measures of accuracy. The most common strategy is to consider segmentation as a pixel-level classification problem and to evaluate it using a confusion matrix at the pixel level. Overall and per-class accuracies over the entire dataset

were the first ones to be reported [13]. The intersection-over-union segmentation measure also known as Jaccard index [4], which counts the total number of mislabeled pixels in the image, is now defacto standard. The successive PASCAL VOC competitions have used it since 2008. One of the only alternatives is the Trimap of [7] that focuses on the boundary regions. To the best of our knowledge, only few works [6, 7, 10] have used this measure.

In this paper, we raise the following question: "what is a good semantic segmentation measure?", and show that the answer is not as trivial as it sounds. Our contribution is threefold. First, we draw the attention of the community to this question that, in our opinion, has been largely overlooked, and review existing semantic segmentation measures. We show that different segmentation algorithms might be optimal for different segmentation measures. Also, for the same segmentation algorithm, it is important to optimise parameters on the target measure. Then, we propose new measures, looking at two directions: i) we propose to measure segmentation accuracy on a per-image basis rather than on the dataset as a whole, and ii) we propose a new measure based on contours, adapting a segmentation measure initially introduced for unsupervised segmentation. We show that the latter is complementary with the Jaccard index. Finally, we perform a user-study – the first we are aware of – to understand how semantic segmentation measures correlate with human preference and we use it to explore a possible combination between region-based and contour-based measures.

The rest of this paper is organised as follows. First, section 2 presents the segmentation methods we used to support our study. Then, section 3 discusses existing segmentation measures and the proposed ones. Experiments including our user study are reported in section 4, and section 5 summarises a list of recommendations.

2 Semantic Segmentation Algorithms

Semantic segmentation is formulated as a discrete labelling problem that assigns each pixel $x_i \in I\!R^3$ of an image to a label y_i from a fixed set Ψ. Given the N observations $x = \{x_1, ..., x_N\}$, the task is to predict the set of labels $y = \{y_1, ..., y_N\}$ taking values in Ψ^N. To guarantee a consistent labelling of an image, semantic segmentation algorithms [7, 10, 12, 18, 23] generally rely on conditional random fields (CRFs). Global consistency is enforced through a global conditioning that takes into account the context of the whole image. Local consistency is enforced through pairwise potentials to encourage neighbouring pixels to share the same label. Recently, there is a trend to combine recognition (local appearance) with low-level segmentation methods, which produce super-pixels. To enforce super-pixel consistency, the CRF [7, 11] can consider higher order potentials or alternatively, hard constraints can ensure that all pixels in a super-pixel have the same label [3]. Object detectors [12, 19, 22] can enhance state-of-the-art semantic segmentation models. This could have been integrated in the models we consider below, but is beyond the scope of this paper.

We consider five methods in our evaluation. We start from a simple model that only enforces a global but no local consistency. We then add more and more sophisticated consistency terms, which tend to produce more and more precise contours. This enables evaluating the impact of gradually more complex models on the different segmentation measures. We review the algorithms below (for more details, the reader is referred to the original papers).

Patch-based classification (P). The first model is directly inspired by [3] and combines two components: (i) local evidences and (ii) global context. For the first component, we use a patch-based representation. Low-level descriptors (here SIFT) are computed for each patch and transformed into Fisher Vectors (FV) [3, 16]. During training, we learn a linear SVM

classifier per class on the per-patch FVs using the corresponding labels. At test time, given a new image, we score the per-patch FVs for each class. The scores are transformed into probabilities at the pixel level yielding probability maps. For the global context, we train image-level classifiers for each class. Such classifiers output the probability that a given object is present in the image. If the score of a class is above a given threshold, then we compute the probability maps for that class. Otherwise, this class is not considered in the rest of the pipeline. In the end, each pixel is assigned to the class with the highest probability.

Patch-based classification+Mean Shift (P+MS). The second model is also inspired by [3] and relies on a post-processing of P. The probability maps obtained with the previous model generally do not fit well the object boundaries. Hence, we combine the probability maps with a low level unsupervised segmentation, and segment images into super pixels using [2, 15], to improve the segmentation boundaries. Class probabilities are averaged over the regions, and each region $\mathcal{R}$ as a whole is assigned the most likely label $c_{\mathcal{R}}$.

The following three models correspond to CRF models[1], and their energy is defined as $E(y) = E_{unary}(y) + \lambda E_{pair}(y)$ [17]. The unary term models the class appearance. The pairwise term regularises the pixel labelling by encouraging neighbouring pixels to share labels. The maximum a posteriori (MAP) labelling y^* is defined as $y^* = \text{argmin}_{y \in \Psi^N} E(y)$.

Grid based CRF (GCRF). The third model is inspired by the Grab-Cut model [17]. In the GCRF model, the unary term is $E_{unary} = \alpha_1 E_1 + \alpha_2 E_2 + \alpha_3 E_3$, where E_1 corresponds to the probability for each pixel to belong to a class computed in the same manner as for the first two models (*i.e.* probability maps), E_2 is a global prior obtained by the image-based classifiers, and E_3 captures the colour appearance of each object instance using a mixture model. We use a 4-neighbour system and a contrast sensitive Potts model for the pairwise terms. Model Inference is an iterative process that alternates i) energy minimisation (using α-expansion [1, 8]), and ii) estimation of the appearance model of each object instance.

Dense CRF (DCRF). The fourth model was directly inspired by the work of [10]. The main difference with the third model is that the pairwise potential is not limited to the four direct neighbours, but include longer range connections. In this model, we have $E_{unary} = E_1$ where E_1 is the same as in the GCRF model, and we ignore E_2 that made little difference and E_3 that is not needed thanks to the longer range dependencies.

Dense CRF+Mean Shift (DCRF+MS). Finally, the fifth model we consider [6] combines the benefits of the low-level segmentation of the P+MS model with the long-term dependencies of the DCRF model. In a nutshell, it consists in adding to the DCRF pairwise potential a term that encourages pixels in the same region $\mathcal{R}$ to have the same label $c_{\mathcal{R}}$.

3 Evaluation measures

In this section, we first review the most common evaluation measures for semantic segmentation, which we refer to as *region-based accuracies*. We then introduce our *contour-based score*. We finally discuss the issue of computing a single accuracy measure for a whole dataset, and propose a novel way to measure segmentation accuracy on a per-image basis.

While there is a significant body of work around suitable evaluation measures for foreground-background segmentation [9, 14, 20], we do not review them in this paper, as we focus on semantic segmentation whose evaluation has been less studied by far.

[1] The two previous models can also be understood as trivial CRF models with no pairwise potentials. In the P model the nodes correspond to the pixels and in the P+MS" model to the super-pixels. Because of the absence of pairwise potentials, the MAP solution is trivially found by optimising each unary term independently.

3.1 Region-based accuracies

Most semantic segmentation measures evaluate a pixel-level classification accuracy. Consequently, these measures use the pixel-level confusion matrix $\mathbf{C}$, which aggregates predictions for the whole dataset $\mathcal{D}$:

$$\mathbf{C}_{ij} = \sum_{I \in \mathcal{D}} |\{z \in I \text{ such that } S^I_{gt}(z) = i \text{ and } S^I_{ps}(z) = j\}| \tag{1}$$

where $S^I_{gt}(z)$ is the ground-truth label of pixel z in image I, S^I_{ps} is the predicted label, and $|A|$ is the cardinality of the set A. In other words, $\mathbf{C}_{ij}$ is the number of pixels having ground-truth label i and whose prediction is j. We denote by $\mathbf{G}_i = \sum_{j=1}^{L} \mathbf{C}_{ij}$, the total number of pixels labelled with i, where L is the number of classes, and by $\mathbf{P}_j = \sum_i \mathbf{C}_{ij}$ the total number of pixels whose prediction is j. We can compute the three following measures from $\mathbf{C}$:

$$OP = \frac{\sum_{i=1}^{L} \mathbf{C}_{ii}}{\sum_{i=1}^{L} \mathbf{G}_i}, \qquad PC = \frac{1}{L} \sum_{i=1}^{L} \frac{\mathbf{C}_{ii}}{\mathbf{G}_i} \qquad \text{and} \qquad JI = \frac{1}{L} \sum_{i=1}^{L} \frac{\mathbf{C}_{ii}}{\mathbf{G}_i + \mathbf{P}_i - \mathbf{C}_{ii}} \tag{2}$$

The **Overall Pixel (OP)** accuracy measures the proportion of correctly labelled pixels. This score is common on the MSRC dataset [18]. One significant limitation of this measure is its bias in the presence of very imbalanced classes. This is the case for the background class on PASCAL VOC datasets, which covers 70-75% of all pixels.

The **Per-Class (PC)** accuracy measures the proportion of correctly labelled pixels for each class and then averages over the classes. Therefore, the background region absorbs all false alarms without affecting the object class accuracies. This measure is suitable for datasets with no background class (and is common on MSRC [18]), but has a strong drawback for datasets with a large background class. As an example, if one always predicts the classes of objects present in an image, and never predicts the background class, then one artificially improves all object class accuracies, at the cost of down-scaling only a single number in the average (the background class score) which is unacceptable. Moving away from this limitation, the PASCAL VOC challenge has shift from PC to the Jaccard index.

The **Jaccard Index (JI)** measures the intersection over the union of the labelled segments for each class and reports the average. The JI thus takes into account both the false alarms and the missed values for each class. As it solves the issues mentioned for OA and PC, it is nowadays standard to evaluate the PASCAL VOC challenge since 2008. Nevertheless, one limitation is that it evaluates the amount of pixels correctly labelled, but not necessarily how accurate the segmentation boundaries are. Therefore, JI is not sufficient to compare different segmentation methods.

The **Trimap** was proposed by [7] as a complement to JI to evaluate segmentation accuracy around segment boundaries. The idea is to define a narrow band around each contour and to compute pixel accuracies in the given band ($r = 5$ in our experiments). We consider the Trimap accuracy as a region-based measure, because it measures a region-based accuracy in a boundary region. While [7] initially proposed to compute OP in the border (denoted by TO), we also propose to use the JI variant, that we will denote by TJ.

The Trimap has a strong limitation: it only evaluates the accuracy in a given band. If it is too narrow, it ignores important object/background information. If it is too large, it will converge to the OP or JI measures and disregard information about the boundary (see *e.g.* in Fig. 1). To overcome this limitation, [7] proposes to plot the accuracy as a function of the

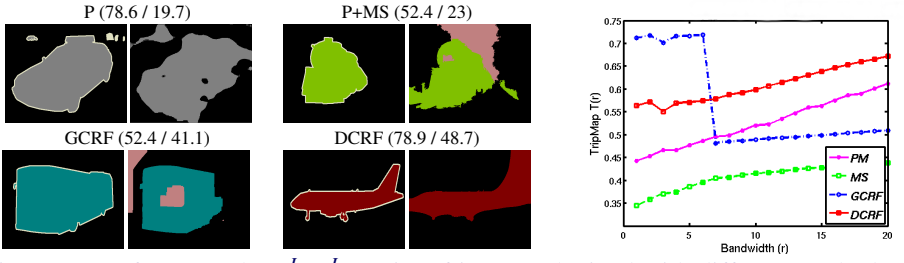


Figure 1: (Left) Example (S^I_{gt}, S^I_{ps}) pairs of images obtained with different methods (these are random samples and we are not comparing the methods here). In parenthesis, we show (JI / BF) accuracies. (Right) we show the corresponding Trimap plots (TO) varying r.

bandwidth (r). This however makes the cost of the evaluation much higher. Furthermore, having multiple accuracy measures makes it harder to pick the best model[2].

3.2 A novel semantic contour-based score

For some applications, *e.g.* in the graphics domain, the contour quality significantly contributes to the perceived segmentation quality [7]. This is not captured by the OP, PC and JI measures, and only partially by the TO and TJ measures. Therefore, we propose a novel semantic contour-based accuracy directly inspired by standard contour metrics for unsupervised segmentation. Popular contour-based measures for unsupervised segmentation include the Berkeley contour matching score [14] and the Boundary Displacement Error [5] (BDE). Both measures are based on the closest match between boundary points in the source and target segmentation maps. While the BDE measures pixel distances and hence depends on the image size, [14] computes the F1-measure from precision and recall values with a distance error tolerance θ to decide whether a boundary point has a match or not.

For this reason we focus on the F1-measure and extend [14] to semantic segmentation: we make it class-dependent by computing one value per class between the corresponding binarized segmentation maps. Let B^c_{gt} be the boundary map of the binarized ground truth segmentation map for class c, S^c_{gt}, with $S^c_{gt}(z) = [\![S_{gt}(z) == c]\!]$ and $[\![z]\!]$ is the Iversons bracket notation, *i.e.* $[\![z]\!] = 1$ if $z =$true and 0 otherwise. Similarly B^c_{ps} is the contour map for the binarized predicted segmentation map S^c_{ps}. If θ is the distance error tolerance (we used 0.75% of the image diagonal) the precision and recall for each class are:

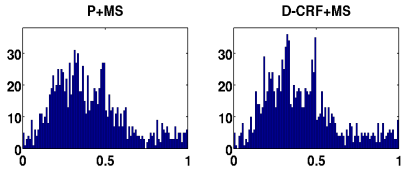
$$P^c = \frac{1}{|B_{ps}|} \sum_{z \in B^c_{ps}} [\![d(z, B^c_{gt}) < \theta]\!] \qquad \text{and} \qquad R^c = \frac{1}{|B_{gt}|} \sum_{z \in B^c_{gt}} [\![d(z, B^c_{ps}) < \theta]\!] \tag{3}$$

where $d()$ is the Euclidean distance. The F1-measure for class c is defined as:

$$F^c_1 = \frac{2 \cdot P^c \cdot R^c}{R^c + P^c}$$

Finally, to obtain the per-image F1 score (denoted BF), we average the F^c_1 scores over all classes present either in the ground-truth (S_{gt}) or in the predicted segmentation (S_{ps}), and we obtain the result on the whole dataset by averaging the per-image BF's. In contrast to

[2] We could obtain a single value by averaging over r, but as shown in Fig. 1, it might not tell the whole story.



	th	P+MS	DCRF+MS	$s_1 < s_2$	p-value
OA	.7	66	**67**	53	**2e-3**
PC	.5	**51**	48	47	**4e-2**
JI	.5	**28**	27	49	6e-2
TO	.5	64	**66**	49	**1e-2**
TJ	.5	19	**20**	46	3e-1
BF	.2	**42**	39	40	**3e-2**

Table 1: Left, histograms of per-image JI scores. Right, comparisons between P+MS and DCRF+MS: Col. 2 and 3 show the percentage of images whose score is above the threshold (th) shown in Col. 1. Col. 4 shows the percentage of images where DCRF+MS scores higher than P+MS. Col. 5 shows the *p*-value for the paired t-test, with a 5% confidence level.

the Trimap plots, the newly defined BF provides a single value per image. Also, it captures complementary information with the JI. This is illustrated in Fig. 1. While according to the JI, P and DCRF segmentation examples have similar accuracies, the BF shows that DCRF captures boundaries much more accurately than P. The same is illustrated by comparing P+MS and GCRF.

3.3 Measuring a per-image score

We believe that measuring per-image scores is important for several reasons. (1) Measures computed over the whole dataset do not enable to distinguish an algorithm that delivers a medium score on all images from an algorithm that performs very well on some images and very poorly on others. Plotting the histogram of per-image scores enables to make such a distinction. (2) We can use the set of per-image scores to evaluate the percentage of images with a performance higher than a threshold, to compare the percentage of images where one method performs better than another or to analyse the statistical difference of two segmentation algorithms with t-test (see examples in Table 1). (3) Per-image scores reduce the bias *w.r.t.* large objects, as missing or incorrectly segmented small objects have a lower impact on the global confusion matrix. (4) If one wants to ask users to compare the output of different segmentation algorithms and correlate such judgements with scores, as in our user study, we need to be able to measure per-image scores.

In the literature, semantic segmentation is always evaluated on the confusion matrix computed over the whole dataset. This is different from the proposed semantic BF which measures a per-image score. Per-image measures are uncommon for semantic segmentation probably because not all classes are present in each image. This makes the evaluation more complex as different classes may be present in the ground-truth and predicted segmentations. To deal with missing/additional objects in the predicted segmentation with respect to the ground-truth, we propose to compute the per-image score as the average of the class scores, where we consider the union of classes which are present in the ground-truth and/or predicted segmentations. If a class is absent from the ground-truth but is predicted in the image, then the corresponding class score is set to zero.

4 Experiments

We consider two widely used datasets: *PASCAL VOC 2007* and *2011* [4]. PASCAL VOC 2007 contains 20 object classes and a background class. For training, we use the 422 training

	Pascal 2007					Pascal 2011				
	P	P+MS	GCRF	DCRF	DCRFMS	P	P+MS	GCRF	DCRF	DCRFMS
global measures over all the dataset										
OP	68.9	72.3	74.3 (.1)	**74.6** (.5)	**74.6** (.2)	70.7	74.6	72.6 (2.4)	**75.0** (.4)	**75.2** (.2)
PC	41.1	41.4	42.8 (1)	**43.2** (.2)	**43.4** (.1)	43.3	**43.9**	**43.7** (0)	42.1 (.3)	41.9 (.7)
JI	25.0	26.7	28.3 (0)	**28.8** (0)	**28.9** (0)	25.4	**28.1**	26.6 (0)	27.2 (0)	**27.3** (0)
TO	45.0	52.6	52.4 (0)	**53.5** (.8)	**54.2** (.1)	45.4	**54.9**	51.9 (1.3)	54.2 (1)	**55.3** (0)
TJ	23.7	26.1	25.4 (.6)	**26.5** (.2)	**27.0** (.1)	24.5	**28.1**	26.6 (0)	26.7 (0)	**27.2** (0)
average of the per-image measures										
OP	68.6	72.0	73.9 (.2)	**74.2** (.5)	**74.2** (.2)	70.5	74.2	72.3 (2.3)	**74.7** (.5)	**74.9** (.2)
PC	46.6	**48.2**	47.5 (.6)	47.8 (.4)	**48.3** (.2)	49.0	50.1	50.9 (.7)	**51.0** (.1)	**51.3** (0)
JI	34.7	37.5	38.9 (1.5)	**39.1** (.2)	**39.3** (.3)	37.2	40.9	40.2 (1.2)	**41.1** (.3)	**41.5** (0)
TO	45.7	53.6	53.1 (0)	**54.1** (1)	**54.9** (.1)	46.3	**56.0**	53.4 (1.6)	55.7 (.9)	**56.8** (0)
TJ	23.5	**30.4**	28.4 (.1)	29.8 (.8)	**30.7** (.3)	24.4	**33.2**	31.0 (.4)	32.3 (.9)	**33.6** (0)
BF	8.4	**16.4**	14.7 (1.2)	16.3 (0)	**17.7** (0)	8.3	**20.2**	17.7 (.6)	17.6 (.6)	**19.4** (0)

Table 2: Pascal 2007 and 2011 segmentations obtained with different methods and evaluated with different measures. We selected the parameters of the CRF-based methods using the standard JI. Green means best, blue second best. If a method becomes best for a measure with the parameters tuned with it, we display the gain (shown in bracket) in green.

images annotated at the pixel level (*i.e.* with segmentation masks) and the 5,011 training images annotated with bounding boxes. The test set contains 210 images.

PASCAL VOC 2011 contains the same 21 classes. We again use all training data available: the 5,017 images with bounding boxes and the 1,112 images with segmentations. The 1,111 images of the validation set constitutes our test set [3].

4.1 Evaluation

We evaluate the segmentation methods we selected for our study using the evaluation measures presented in the previous section with both versions: the global score over the dataset and averaging the per image scores. We reported the results in Table 2.

Influence of the measures on the algorithm ranking. Each column in Table 2 corresponds to a single segmentation result (*i.e.* the exact same set of images has been considered). This means that the set of parameters for the three CRF methods is fixed across the entire table. We optimised them for the global Jaccard Index, the standard measure used in the literature for both datasets. The best result is displayed in green, the second best in blue. The main observation we can make is that the ranking of the algorithms depends on the evaluation measure, which emphasises that they capture different and complementary information.

Global vs per-image scores. Comparing the top and bottom parts of Table 2, we can also observe changes between the rankings with the global version (over the entire dataset) of a measure and its per-image version (average of the per-image scores). The per-image score brings the advantages listed in section 3. One of them is that we can use statistical significance tests, as reported in Table 1.

Comparison using per-measure optimised parameters. Optimising the parameters of the three CRF models with JI, when we use an other measure to evaluate is unfair. Therefore, we also optimised the set of parameters for each measure individually, and report the gain in Table 2 in brackets. In such a configuration, different segmentations are evaluated for different lines. If one of the methods becomes ranked first after this per-measure parameter

[3] As our study needs a large number of parameter evaluations, the validation set is more appropriate than the test set that would require many evaluations on the PASCAL server.

	PASCAL VOC 2007					
	OP	PC	JI	TO	TJ	BF
OP		.73	**.80**	.60	.54	.41
PC	64		**.80**	.63	.63	.45
JI	62	63		.60	.65	.53
TO	41	41	39		**.83**	.54
TJ	35	40	46	66		.64
BF	28	29	33	31	40	

	PASCAL VOC 2011					
	OP	PC	JI	TO	TJ	BF
OP		.78	**.83**	.63	.57	.39
PC	69		.77	.60	.59	.40
JI	69	63		.56	.63	.50
TO	42	40	36		**.82**	.50
TJ	36	37	40	67		.59
BF	30	30	34	36	39	

Table 3: Ranking correlation between different measures on the two datasets. Above diagonal is the mean of the Spearman's correlation coefficients, while below diagonal is the percentage of images where the ranking is exactly the same ($\rho = 1$).

optimisation, the number in bracket is displayed in green. We observe that the best parameter set from one evaluation measure to another varies, and can change the results significantly (up to 2.4%). The ranking can change as well.

Comparison of segmentation methods. Although the comparison of segmentation algorithms is not the primary purpose of this paper, we can make some observations from Table 2. First super-pixel based methods generally perform best (P+MS, DCRF+MS). In particular, P+MS is surprisingly competitive, although it does only a voting within super-pixels (no CRF). Overall, it seems that DCRF+MS, which uses a CRF model with longer-range dependencies and super-pixel information, generally outperforms other methods.

Correlation between measures. We have seen that the ranking of the segmentation methods differs from one evaluation measure to another. Here we would like to understand and quantify the complementarity between measures through correlation. We measure correlation using Spearman's rank correlation coefficient (SRC), which can handle ties (rank ties or value duplicates). For a sample of size n, the n raw scores A_i and B_i are converted to ranks a_i, b_i, and SRC is computed as:

$$\rho = \frac{\sum_{i=1}^{n}(a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_{i=1}^{n}(a_i - \bar{a})^2 \sum_{i=1}^{n}(b_i - \bar{b})^2}} \in [-1, 1] \tag{4}$$

where $\bar{a}$ and $\bar{b}$ are the mean of the ranks a_i and b_i respectively. The sign of ρ indicates the direction of association between the score sets A_i and B_i, where 1 means that the two ranks perfectly match, 0 no correlation, and -1 that they are anti-correlated. Table 3 shows the correlation between all measures. As expected, we observe low correlations between BF and other measures (see for example JI and BF values in Fig. 1). Our aim is to complement the Jaccard index, which is standard on this dataset. Therefore, in what follows we will consider only JI and BF.

4.2 User study

This section aims at understanding the relation between different measures and the quality of segmentations, as perceived by humans, through a user study.

We conducted a preliminary user study that showed that humans were unable to rank segmentations for which the accuracies were really low (too different from the ground truth). This seems to indicate that i) the scores have very little meaning when they are too low, and ii) a user study makes sense only in the context of segmentations that are good enough. Consequently, we selected only images where at least one of the methods had a JI value higher than 0.5. Furthermore, being interested in the complementarity of JI and BF (that exhibits

	CA	SRC
JI	70.8	.54
BF	65.6	.34
JI & BF	**72.6**	**.56**
std	.54	1.7

Table 4: Classification (CA) and correlation (SRC) with human rank.

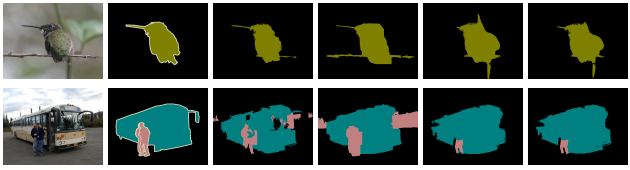


Figure 2: Example images considered in our user test. Users are shown the original image, the ground-truth segmentation, and the result obtained by P+MS, GCRF, DCRF, DCRF+MS.

low correlation in Table 3), we choose images amongst the ones for which the correlation between JI and BF is below 1 (they do not yield to the exact same ranking). We selected a set of 48 images. These images and their segmentations, obtained with MS, GCRF, DCRF and DCRF+MS, were shown (together with the ground-truth) to 14 different users, who were asked to rank segmentations of each image according to their quality. Users were also allowed to use ties when they could not make a decision.

Correlation between human rankings. To understand the correlation between the rankings given by different users, we randomly split the users into two groups and compute the mean of the SRC over the 48 image ranks. This was repeated 50 times. The average correlation over the 50 splits is 0.67 with a standard deviation of 0.04. This shows that while there are obvious variations between the different human rankings, overall groups of users tend to agree. Therefore the average over the 14 users will be used as a ground truth in what follows.

Correlation between average human ranking and JI/BF measures. First, we can compute the correlation between the ground-truth ranking and the ranking obtained with the JI and the BF. The average of the correlation values over the 48 images are respectively .54 and .34 (SRC reported in Table 4). We can see that the JI is more correlated with the human ranking than the BF. This is consistent with an observation we made during the study: accurate contours are less important in a first place than having the right categories. However, for relatively good and similar segmentations (such as the top image in Fig. 2), the BF becomes more relevant to rank segmentations. Therefore, in which follows our aim is to combine these measures to better predict human ranking.

Learning a combined measure. We consider the following two-class classification problem: for each image we consider 12 ordered pairs, each corresponding to a pair of segmentation algorithms, and we label each pair with 1 if the order matches the human ranking, and zero otherwise. The percentage of correctly ranked pairs (class accuracy) over all images (576 pairs) according to JI and BF are shown in Table 4 (class accuracy shown as CA).

To learn the combination of JI and BF, we consider the 576 two-dimensional data points obtained with JI and BF differences computed on method pairs such that a data point corresponds to *e.g.* $(JI(GCRF) - JI(DCRF), BF(GCRF) - BF(DCRF))$. The ground truth label of each data point is deducted from the human rankings as above.

For evaluation, the image set is split into 6 distinct folds (8 images per fold) and we perform a leave-one-fold out evaluation protocol: 4 folds are used for training, one for validation and one for testing. The percentage of correctly ranked pairs is computed on the union of the independently predicted six test folds (class accuracy). We repeat this operation 10 times and report the average and standard deviation (std) over the 10 random splits in Table 4. We also report the average of the corresponding SRC correlations over the 10 splits in Table 4. We can see, both from classification and correlation results, that JI&BF better predicts the human ranking than JI or BF taken individually.

5 Conclusion

In this paper we have considered different ways to evaluate semantic segmentation (through existing and newly proposed measures). From our study, we can deduce a list of recommendations. First, we would like to stress that to allow a fair comparison, parameters should be optimised for the measure that is considered for evaluation. Second, we advocate for the use of per-image scores that allow a more detailed comparisons of methods. Using per-image scores allow comparison to a threshold which can be useful in real applications where the user expects a minimum level of quality. Third, the importance of images with really low scores should be down-scaled, as such segmentations appeared meaningless during our user-study. We have also shown that the proposed BF score is complementary with JI as it more carefully takes the contours into account, and both measures should be considered simultaneously. We showed on our user study that their combination is a promising way to predict the human ranking. Possible directions for future work include a larger-scale user-study that would take into account one or several end-applications, as we believe that the evaluation of segmentation is task-dependent.

References

[1] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE TPAMI*, 26(9):1124–1137, 2004.

[2] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *PAMI*, 24(5):603–619, May 2002.

[3] G. Csurka and F. Perronnin. An efficient approach to semantic segmentation. *IJCV*, 2011.

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge. http://www.pascal-network.org/challenges/VOC.

[5] J. Freixenet, X. Mu, D. Raba, J. Mart, and X. Cuf. Yet Another Survey on Image Segmentation : Region and Boundary Information Integration. In *ECCV*, 2002.

[6] R. Hu, D. Larlus, and G. Csurka. On the use of regions for semantic image segmentation. In *ICVGIP*, 2012.

[7] P. Kohli, L. Ladický, and P. H S Torr. Robust higher order potentials for enforcing label consistency. *IJCV*, 82(1):302–324, 2009.

[8] V. Kolmogorov and R. Zabih. What energy functions can be minimized via Graph Cuts? *IEEE TPAMI*, 26(2):147–159, 2004.

[9] P. Koniusz and K. Mikolajczyk. Segmentation based interest points and evaluation of unsupervised image segmentation methods. In *BMVC*, pages 1–11, 2009.

[10] P. Krahenbuhl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, 2011.

[11] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr. Associative hierarchical crfs for object class image segmentation. In *ICCV*, 2009.

[12] D. Larlus, J. Verbeek, and F. Jurie. Category level object segmentation by combining bag-of-words models with Dirichlet processes and random fields. *IJCV*, 88(2):238–253, 2010.

[13] L-J. Li, R. Socher, and L. Fei-Fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *CVPR*, 2009.

[14] D. R. Martin, C. C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE PAMI*, 26(1), 2004.

[15] P. Meer and B. Georgescu. Edge detection with embedded confidence. *PAMI*, 23(12): 1351–1365, 2001.

[16] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, 2007.

[17] C. Rother, V. Kolmogorov, and A. Blake. “GrabCut": interactive foreground extraction using iterated graphcuts. *SIGGRAPH*, 2004.

[18] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *ECCV*, 2006.

[19] P. Sturgess, K. Alahari, C. Russell, and P. H. S. Torr. What, Where and How Many? Combining object detectors and CRFs. In *ECCV*, 2010.

[20] R. Unnikrishnan, C. Pantofaru, and M. Hebert. Toward objective evaluation of image segmentation algorithms. *PAMI*, 29(6):929–944, 2007.

[21] O. Whyte, J. Sivic, and A. Zisserman. Get out of my picture! Internet-based inpainting. In *BMVC*, 2009.

[22] W. Xia, Z. Song, J. Feng, L. F. Cheong, and S. Yan. Segmentation over detection by coupled global and local sparse representations. In *ECCV*, pages 662–675, 2012.

[23] Y. Zhang and T. Chen. Efficient inference for fully-connected CRFs with stationarity. In *CVPR*, 2012.