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
6, chemin de Maupertuis
38920 Meylan, France
http://www.xrce.xerox.com/

Motivation. 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. Semantic segmentation has many potential applications including scene understanding, removing undesired objects from photographs, copy-pasting objects from one photograph to another, or local class-based image enhancement. These diverse applications might 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, in computer graphics applications, 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, the computer vision community has resorted to application-independent measures of accuracy. 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.

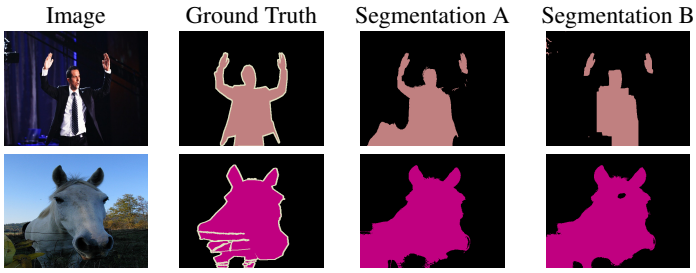

Figure 1: Which one of these two possible segmentations is better? People participating to our study found segmentation A to be more accurate (*i.e.* "closer" to the ground truth segmentation) than segmentation B. Do the segmentation evaluation measures agree ?

Contributions.

First, we draw the attention of the community to the evaluation question that has been well-studied for unsupervised segmentation, but that has been largely overlooked for semantic segmentation. We review the few existing semantic segmentation measures.

Supported by experimental evidences, we show that different segmentation algorithms can be optimal for different segmentation measures.

We propose to measure segmentation accuracy on a per-image basis rather than on the dataset as a whole. This allows visualising the distribution of scores, analysing the statistical differences between algorithms, evaluating methods on a specific image and hence performing user studies.

We propose a new measure based on contours by adapting a segmentation measure initially introduced for unsupervised segmentation [3].

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.

Existing segmentation measures. Existing measures are region-based. They consider segmentation as a pixel-level classification problem and to evaluate it using a pixel level confusion matrix accumulated over the entire dataset. Typically, overall (OA) and per-class accuracies (PC) [4] or Jaccard index (JI) [1] are computed. One of the only alternatives is the Trimap (TO) [2] that focuses on the boundary regions.

Proposed contour-based measure. We propose to extend the F1-measure of [3] to semantic segmentation: We make it class-dependent by computing one value per class by comparing the corresponding binarized segmentation maps, and we average the per-class scores over all classes present either in the ground-truth or in the predicted segmentation.

Segmentation algorithms. We consider five different segmentation algorithms 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. The two patch-based classification model are denoted by P and P+MS, and the three CRF models are denoted GCRF, DCRF and DCRFMS.

A first glance at the results. To illustrate the importance of choosing the evaluation measure correctly, we show in Table 1 results obtained for the segmentation methods mentioned above, when evaluated with the standard evaluation measures used to compare semantic segmentations (OA, PC, JI, TO). We can easily notice that the ranking of the methods is highly dependent on the evaluation measure.

	P	P+MS	GCRF	DCRF	DCRFMS
OP	70.7	**74.6**	72.6	**75.0**	**75.2**
PC	43.3	**43.9**	**43.7**	**42.1**	41.9
JI	25.4	**28.1**	26.6	**27.2**	**27.3**
TO	45.4	**54.9**	51.9	**54.2**	**55.3**

Table 1: Pascal VOC 2011. Segmentations are obtained by five different methods. The scores of the different evaluation measures are coloured according to their ranks with green for the highest, blue for the second and red for the third one.

To conclude, we suggest a **list of recommendations:**

- A single segmentation measure does not tell the whole story. Use different measures to assess the different capabilities of your algorithm.
- To allow a fair comparison, the segmentation algorithm parameters should be optimized for each measure.
- Per-image scores should be preferred as they allow a more detailed comparison of methods including statistical significance test, user study, per image reasoning.
- Below a given threshold (*e.g.* < 0.5 for JI) the segmentation scores do not seem to correlate anymore with the perceived quality of a segmentation. Hence, care should be taken when drawing conclusions from such low-scoring images.

[1] 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.

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

[3] 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.

[4] 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.