# On Segmentation Evaluation Metrics and Region Counts*

Allan Hanbury and Julian Stöttinger
*PRIP, Institute of Computer-Aided Automation, Vienna University of Technology*
*Favoritenstraße 9/1832, A-1040 Vienna, Austria*
{*hanbury,julian*}*@prip.tuwien.ac.at*

## Abstract

*Five image segmentation algorithms are evaluated: mean shift, normalised cuts, efficient graph-based segmentation, hierarchical watershed, and waterfall. The evaluation is done using three evaluation metrics: probabilistic Rand index, global consistency error, and boundary precision-recall. We examine region-based metrics as a function of the number of regions produced by an algorithm. This allows new insights into algorithms and evaluation metrics to be gained.*

## 1. Introduction

Image segmentation is an important component in many image analysis and computer vision tasks. It is in general used in two contexts. The first is the segmentation of images from a specialised area, such as medical images, for which one can usually obtain ground truth from experts in the field, allowing an objective evaluation of segmentation algorithms. The second context in which it is applied is in the understanding of general images. An example is in [1], where an image is first segmented and then tags are automatically assigned to each region in order to describe the contents of the image. For such an application, the segmentation algorithm should produce regions which correspond well to real-world objects in the image. The problem with such a requirement is that there are many "correct" segmentations of an image, depending on the scale at which one interprets the image. A commonly accepted benchmark for evaluating segmentation algorithms for the segmentation of general images is the Berkeley Segmentation Dataset [7], of which 300 images are usually used in evaluations. During the collection of this dataset, human subjects were asked to manually segment each image. The possible difference in interpretation of the images is taken into account by having at least five manual segmentations by different people for each image. An average evaluation metric is usually calculated over all ground truth segmentations of an image.

One aspect of unsupervised segmentation that is usually not considered in the evaluation metrics is the number of regions produced. For the human segmentations of the Berkeley dataset images, the mean number of regions per image is 18. Unsupervised segmentation algorithms often produce many more regions than this. The main contribution of this paper is an investigation of three evaluation metrics and their relation to the number of regions produced by segmentation algorithms. Five unsupervised segmentation algorithms are evaluated over a range of parameters: mean shift, normalised cuts, efficient graph-based segmentation, hierarchical watershed, and waterfall.

## 2. Segmentation Evaluation

Segmentation evaluation metrics can be divided into boundary-based and region-based methods [4]. For region-based evaluation, we investigate the widely used Global Consistency Error (GCE) [7], as well as the recently introduced Probabilistic Rand Index (PRI) [10]. The GCE measures the extent to which regions in one segmentation are subsets of regions in a second segmentation (i.e. the refinement). Because the GCE is zero for the extreme cases where one image is a single region or every pixel in an image is assigned a unique label, it only makes sense to compare the GCE of images segmented into (almost) the same number of regions. For this reason, we introduce a plot of the mean GCE (over all test images) for a specific set of parameters of an algorithm against the mean number of regions produced by these parameters. One can then easily compare the algorithm performance for a given number of regions.

The PRI measures the consistency of labellings between a segmentation and its ground truth by the ratio of pairs of pixels having the same labels. The mean over

*Partly supported by the FIT-IT project OMOR (815994)

978-1-4244-2175-6/08/$25.00 ©2008 IEEE

all the human segmentations for an image takes their differences into account. As above, we plot the mean PRI (over all test images) against the number of regions to accentuate the relation between these two metrics. We choose to use the PRI instead of the Normalised PRI [10], as the latter is simply a linear scaling of the former.

For boundary-based evaluation, we use the boundary precision-recall curves of [6]. A correspondence is computed between machine boundary and human labelled boundary maps, after which the precision and recall of boundary pixels is calculated. These are plotted on a precision-recall curve, where each point represents the mean precision $P$ and recall $R$ (over all test images) for a specific set of parameters of an algorithm. Each point is also characterised by an $F$-measure, defined as $F = 2PR/(P + R)$, where a higher $F$-measure indicates a better segmentation.

## 3. Algorithms Evaluated

We evaluate five unsupervised segmentation algorithms, divided into two groups. The first group contains algorithms for which the number of regions required in the segmentation is an input parameter: normalised cuts (ncut) [9] and the hierarchical watershed with volume extinction values (wsvol) [8]. In the second group, the parameters of the algorithms are less directly linked to the number of regions obtained: mean shift (ms) [2], efficient graph-based (fz) [3] and waterfall (wfall) [5] segmentation. For all algorithms, we used implementations obtained either directly from the authors (watershed algorithms) or from their web pages. For the normalised cuts implementation, the only parameter is the number of regions. However, due to the time and memory-intensive nature of the algorithm, the images were reduced to half their size before applying the algorithm. The resulting region labelled image was then doubled in size. Before applying the watershed and waterfall algorithms, a leveling filter was applied to simplify the image, as recommended in [8]. The size of this leveling filter is one input parameter, along with the number of regions (volume extinction watershed) or number of waterfall simplification iterations. The mean shift implementation does clustering in a 5-dimensional space, with 2 spatial and 3 colour dimensions. Parameters $h_s$ and $h_r$ are the spatial and range (colour) kernel bandwidths. In addition, parameter $m$ is the size in pixels of the smallest region that can be produced. The efficient graph-based algorithm has parameters $\sigma$, the size of a Gaussian pre-filtering; $k$, a threshold on the region comparison predicate; and $m$ as for the mean shift algorithm. Table 1 summarises the abbreviations for the algorithms and the groups to which they belong. Group (Gr) 1 contains algorithms that have the number of regions specified as a parameter, while the algorithms in group 2 produce a varying number of regions.

| Algorithm | Gr | Fixed Params | Varied Params |
|---|---|---|---|
| ncut | 1 | | num. reg. = 10,20,...,90 |
| wsvol_f3 | 1 | leveling size=3 | num. reg. = 10,20,...,90 |
| wsvol_f0 | 1 | leveling size=0 | num. reg. = 10,20,...,90 |
| ms s12 rx m50 | 2 | $h_s = 12, m = 50$ | $h_r$ = 4,6,...,20 |
| fz s0.8 kx m50 | 2 | $\sigma = 0.8, m = 50$ | $k$ = 50,150,...,1050 |
| wfall_f3 | 2 | leveling size=3 | iteration=1,2,3 |

**Table 1. The evaluated segmentation algorithms and their parameters. The second column (Gr) gives the group of the algorithm, as explained in the text.**

## 4. Results

We evaluated the algorithms over a range of parameters. For the algorithms where the number of regions is specified (wsvol and ncut), this was varied from 10 to 90 in steps of 10, corresponding to the range in which the number of regions in the manually produced ground truth segmentations falls. In addition, the effect of the leveling parameter on the wsvol was tested, with no leveling (wsvol_f0) and a leveling with an alternating sequential filter (ASF) of size 3 (wsvol_f3) applied. For the efficient graph-based approach, a value of $\sigma = 0.8$ was chosen, as done in [3]. The $k$ parameter was varied from 50 to 1050 in steps of 100. For the mean shift, the spatial parameter was chosen as $h_s = 12$ based on the size of the images as recommended in [2]. The range parameter was varied from 4 to 20 in steps of 2. For both algorithms, the minimum region size $m = 50$ was used. Three iterations of the waterfall simplification are examined. A leveling with an ASF of size 3 was applied before the segmentation. These parameters are also summarised in Table 1.

Figures 1 and 2 show respectively the mean PRI and mean GCE against the mean number of regions over all 300 images in the Berkeley Segmentation Dataset. Note that the error bars in these figures represent one fifth of the standard deviation of the evaluation metric (on the $y$-axis) and number of regions (on the $x$-axis). The reduction in the length is to improve the ease of reading the graphs. Figure 3 shows the boundary precision-recall graph. The position on each curve with the highest $F$-measure (interpolated) is marked by a large point, and the $F$-measure ($F$), coordinates (@) and parameter value ($d$) of these points are given in the legend. In this plot, the parameters resulting in a higher number of regions have points plotted with higher recall.

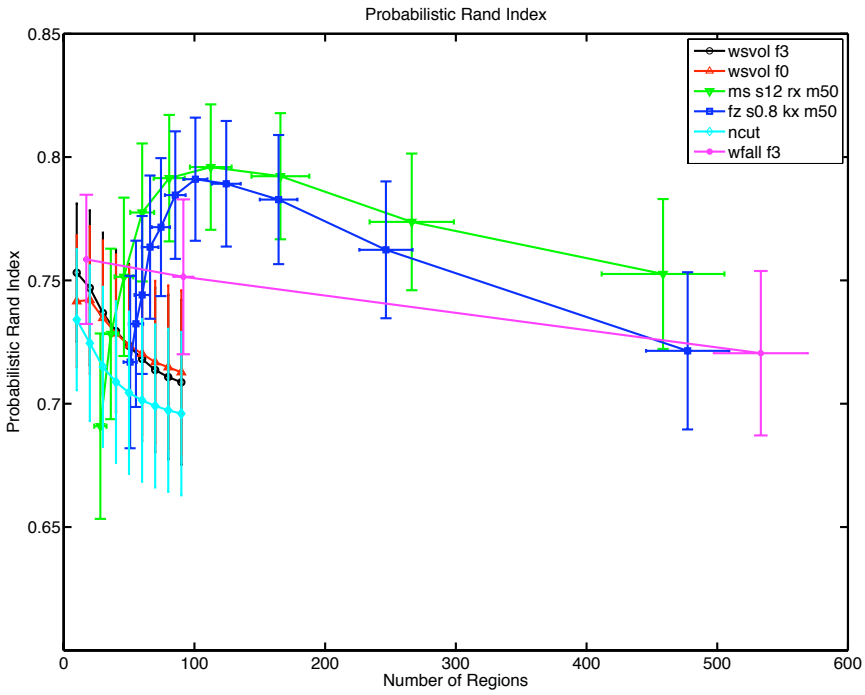


**Figure 1. PRI against Number of Regions. A larger PRI implies a better segmentation.**

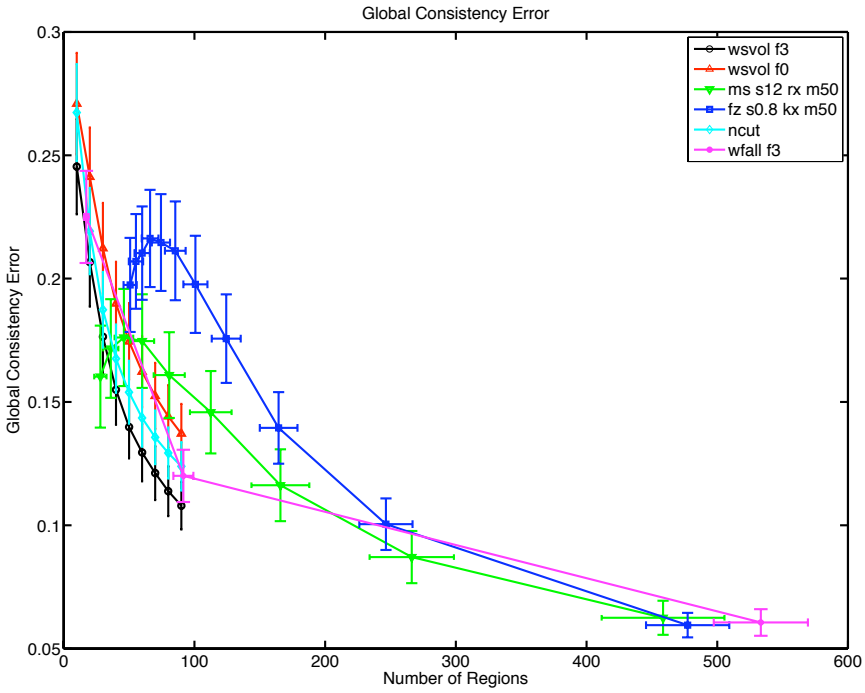


**Figure 2. GCE against Number of Regions. A smaller GCE implies a better segmentation.**

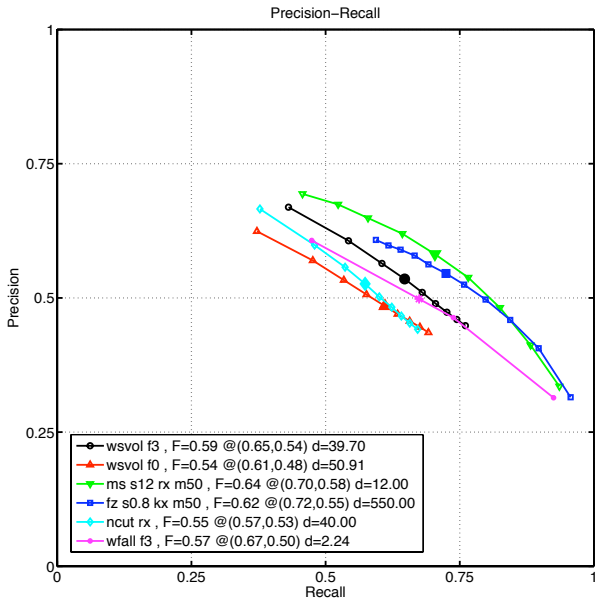


**Figure 3. Boundary Precision and Recall.**

## 5. Discussion

A striking feature of the PRI and GCE graphs is the difference in the general shapes of the curves between the two groups of algorithms. In the PRI graph, the algorithms for which the number of regions are specified as a parameter (Group 1) generally have a larger PRI as fewer regions are specified, with the wsvol curves above the ncut curve. On the other hand, the fz and ms algorithms have a maximum PRI for segmentations into close to 100 regions, and drop below the other curves for a small number of regions. The rapid decrease in the number of regions for each waterfall iteration is also visible, although this curve also rises with a decreasing number of regions. The waterfall iteration after the last one shown in the plot results in a single region for almost every image.

For the GCE plot, it is important to remember that one cannot examine trends in the GCE value as the number of regions varies, as this metric only makes sense when comparing segmentations into a similar number of regions. For this plot, the algorithms within the same groups are generally ordered as in the PRI plot, with ms better than fz, and wsvol_f3 better than ncut. There are however discrepancies in metrics between the groups, especially for a number of regions between 50 and 100, where fz and ms are better than the wsvol and ncut according to the PRI and worse according to the GCE.

For the boundary precision-recall graph, the ordering of the algorithms appears clear, based on the ordering of the curves and the best $F$-measures listed in the legend. It is interesting that there is no crossing of the fz and ms curves with the wsvol curves, as occurs in the PRI and GCE plots.

## 6. Conclusion

It is clear from the discussion that evaluating algorithms on a group of images can lead to different rankings depending on the metric chosen. The large error bars for the metrics indicate their large variation over the set of 300 images. Examining the number of regions using the proposed plots leads to some useful insights. In particular, the change in the ranking of the algorithms based on the number of regions obtained is a useful result — it allows the best algorithm to be chosen based on the number of regions required. However, given the large variations in the values of the metrics over the test images, more insight can only be gained by examining the values of the metrics on individual images.

## References


[1] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. I. Jordan. Matching words and pictures. *J. of Machine Learning Research*, 3:1107–1135, 2003.

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

[3] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *Int. Journal of Computer Vision*, 59(2):167–181, 2004.

[4] J. Freixenet, X. Munoz, D. Raba, J. Martí, and X. Cufí. Yet another survey on image segmentation: Region and boundary information integration. In *Proc. ECCV 2002*, pages 408–422, 2002.

[5] B. Marcotegui and S. Beucher. Fast implementation of waterfall based on graphs. In *Proc. ISMM'05*, pages 177–186, 2005.

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

[7] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. 8th Int. Conf. Computer Vision*, pages II: 416–423, July 2001.

[8] F. Meyer. An overview of morphological segmentation. *Int. Journal of Pattern Recognition and Artificial Intelligence*, 15(7):1089–1118, 2001.

[9] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. PAMI*, 22(8):888–905, 2000.

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