Evaluating Hierarchical Graph-based Segmentation

Yll Haxhimusa, Adrian Ion and Walter G. Kropatsch
Vienna University of Technology, Faculty of Informatics
Pattern Recognition and Image Processing Group 183/2,
{yll,ion,krw}@prip.tuwien.ac.at

Abstract

Using real world images, two hierarchical graph-based segmentation methods are evaluated with respect to segmentations produced by humans. Global and local consistency measures do not show big differences between the two representative methods although human visual inspection of the results show advantages for one method. To a certain extent this subjective impression is captured by the new criteria of 'region size variation'.

1. Introduction

The authors in [6] suggest to bridge and not to eliminate the representational gap, and to focus efforts on *region segmentation*, perceptual grouping, and image abstraction. Hence, evaluation of segmentations by different algorithms is also an effort worthy of concentrating. The segmentation process results in 'homogeneous' regions w.r.t low-level cues using some similarity measures. Problems emerge because i) homogeneity of low-level cues will not map to the semantics [6] and ii) the degree of homogeneity of a region is in general quantified by threshold(s) for a given measure [2]. Some cues can contradict each other, however complex grouping phenomena can emerge from simple computation on these local cues [7]. The union of regions forming the group is again a region with both internal and external properties and relations. The low-level coherence of brightness, color, texture or motion attributes should be used to come up sequentially with hierarchical partitions [10]. It is important that a grouping method has following properties [1]: i) capture perceptually important groupings or regions, which reflect global aspects of the image, ii) be highly efficient, running in time linear in the number of pixels, and iii) creates hierarchical image partitions. In general, there is no 'good' segmentation based on low-level cues [9], in many cases it is the task using the segmentation result which determine its quality. However, following the work in [8] we evaluate segmentation purely by comparing the results of human segmentation with those of a particular method. This is justified because humans tend to produce consistent segmentations (see Fig. 1 and [8]), even thought humans segment images at different granularity (refinement or coarsening). This refinement or coarsening could be thought of as a hierarchical structure on the image, i.e. a pyramid or a dendogram. Therefore in [8] a segmentation consistency measure that does not penalize this granularity is defined (see Sec. 3).

In this paper, we evaluate two segmentation methods, the state of the art in graph-based approaches, the normalized cut based [10](NCutSeg) and the method based on the minimum spanning tree [4](MSTBorůvSeg) (Sec. 2). An overview of graph-based segmentation methods (Sec. 2) is followed by a short presentation of different techniques on evaluating segmentations (Sec. 3). Segmentation results of these methods are compared with human segmentations (Sec. 3). Qualitative inspection of the produced segmentations showed differences between the two methods which the pixel-based consistency measures did not show. We therefore introduce a new criteria to quantify these qualitative differences.

2. Graph-based Segmentation Methods

A graph-theoretical clustering algorithm consists in searching for a certain combinatorial structure in the edge weighted graph, such as a minimum spanning tree (MST) [1, 3], a minimum cut [12, 10], or a search for a complete subgraph i.e. the maximal clique. Early graph-based methods [13] use fixed thresholds and local measures in finding a segmentation, i.e the MST is computed. The segmentation criterion is to break the MST edges with the largest weight, which reflect the low-cost connection between two elements. To overcome the problem of a fixed threshold, [11] normalizes the weight of an edge using the smallest weight incident on the vertices touching that edge. Note that, for the MST problem there are deterministic solutions. The methods in [1, 3] use an adaptive criterion that depends on local properties rather than global ones. The methods based on minimum cuts in a graph are designed to minimize the similarity between pixels that are being split

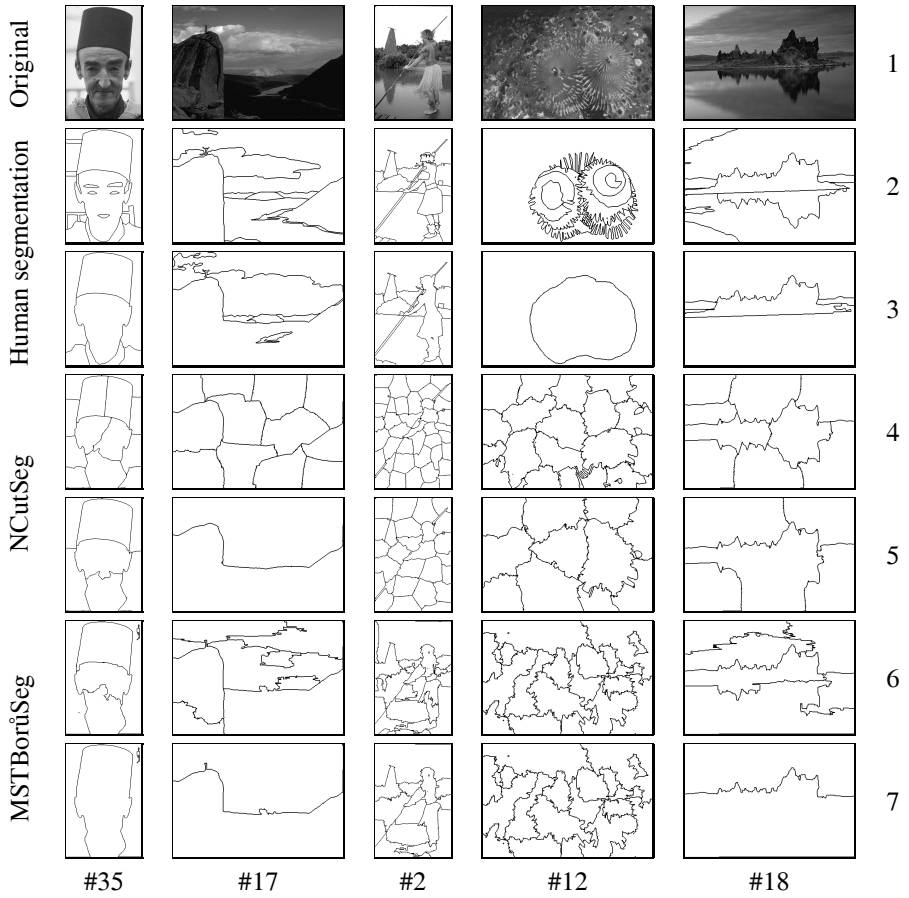

Figure 1. Segmentation of Humans [8], and NCutSeg and MSTBorůvSeg methods.

in two segments [12, 10]. A cut criterion in [12] is biased toward finding small components. The normalized cut criterion [10] is defined to avoid this problem, which takes into consideration self-similarity of regions, and it produces a divisive hierarchical tree, the dendogram. However, they provide only a characterization of such a cut rather than of a final segmentation as provided in [1]. A minimum normalized cut approximation method [10] is computationally expensive and the error in these approximations is not well understood.

Both methods considered (NCutSeg and MSTBorůvSeg) are capable of producing a hierarchy of image partitioning. The segmentation results of these methods to gray value images are shown in Fig. 1. The methods use only local contrast based on pixel intensity values. As expected, and seen on the Fig. 1, segmentation methods based only on low-level local cues cannot create results as good as humans. Even thought it looks like NCutSeg produces more regions, actually the overall number of regions in rows 4 and 6 (respectively 5 and 7) in each column of Fig. 1, is almost the same, but the MSTBorůvSeg produces a large diversity of regions sizes. The face of the man (#35) is satisfactory segmented by both methods. The MSTBorůvSeg method did not merge the statue on the top of the mountain with the sky (#17), similarly to humans. Both methods have problems segmenting the sea creatures (#12). The segmentation done by humans on the image of rocks (#18), contains the symmetry axis, even thought there is no 'big' change in the local contrast, hence both methods fail in this respect.

3. Evaluating Segmentations

Evaluating segmentation algorithms is difficult because it depends on many factors [5] e.g: the segmentation algorithm; the parameters of the algorithm; the classes of images tested; the method for evaluation of the segmentation algorithms, etc. Our evaluation copes with these facts: (i) real world images are used, since it is difficult to extrapolate conclusions based on synthetic images to real images [15], and (ii) the human should be the final evaluator.

There are two general ways to evaluate segmentations: (i) qualitative and (ii) quantitative methods. Qualitative methods involve humans, such that a variety of different opinions about the segmentations is captured (e.g. similarly to edge detection evaluation [5], or in image segmentation [8]). Quantitative methods are classified into analytic methods and empirical methods [14]. Analytical methods

study the principles and properties of the algorithm, like processing complexity, efficiency etc. The empirical methods study properties of the segmentations by measuring how close a segmentation is to an 'ideal' one, by measuring this 'goodness' with some function of the parameters. Both of the approaches depend on the subjects, the first one in coming up with the reference (perfect) segmentation and the second one in defining the function. The difference between the segmented image and the (ideal) reference one assesses the performance of the algorithm [14]. The reference image could be a synthetic image or manually segmented by humans. These discrepancy methods measure the difference between the segmented image and reference images. Higher value of the discrepancy measure signals poor performance of the segmentation method. In [14], it is concluded that evaluation methods based on "mis-segmented pixels should be more powerful that other methods using other measures".

In [8] segmentations made by humans are used as a reference and basis for benchmarking segmentations produced by different methods. The error measures used for evaluation 'count' the mis-segmented pixels. On the same image different people produce different segmentations. The obtained segmentations differ, mostly, only in the local refinement of certain regions. This observation is a core for defining two error measures in [8] , which do not penalize a segmentation if it is coarser or more refined than another. In this sense, a pixel error measure $E(S_1, S_2, p_i)$, called the local refinement error, is defined as $E(S_1, S_2, p) = \frac{|R(S_1,p) \setminus R(S_2,p)|}{|R(S_1,p)|}$ where $\setminus$ denotes set difference, $|x|$ the cardinality of a set x, and $R(S, p)$ is the set of pixels corresponding to the region in segmentation S that contains pixel $p \in I$. Using the local refinement error $E(S_1, S_2, p)$ the following error measures are defined [8]: the global consistency error (GCE), is defined as:

$$\text{GCE}(S_1, S_2) = \frac{1}{|I|} \min \left\{ \sum_{p \in I} E(S_1, S_2, p), \sum_{p \in I} E(S_2, S_1, p) \right\}$$

and the local consistency error (LCE), is defined as:

$$\text{LCE}(S_1, S_2) = \frac{1}{|I|} \sum_{p \in I} \min \{E(S_1, S_2, p), E(S_2, S_1, p)\}$$

Notice that GCE is a stronger measure than LCE (GCE$\geq$LCE). The plausibility of using these two measures for evaluation of segmentation methods is discussed in [8].

Experimental Evaluation

For the experiments, we use 100 gray level images from the Berkley Image Database[1] [8]. Segmentation results are produced by the code offered by the authors [2]. As mentioned in [8] a segmentation made of only one region and a segmentation where each pixel is a region can be the coarsening or refinement of any segmentation. Thus, the LCE and GCE measures should not be used when the number of regions in the two segmentations differs a lot. Since both methods can produce segmentations with a different number of regions, we define for each image a region count reference number, which is the average number of regions from the human segmentations. We set NCutSeg to produce the same number of regions and for the MSTBorůSeg, we take the level of the pyramid that has the region number closest to the same region count reference number.

For each of the images in the test, we have calculated the GCE and LCE using the results produced by the two methods and all the human segmentations available for that image. Having more than one pair of GCE and LCE for each method and image, we have calculated the mean. Fig. 2a) shows the histograms of the GCE and LCE values obtained, NCutSeg vs. humans, and MSTBorůSeg vs. humans. There is a big similarity between the values of GCE and LCE for both methods. In the same figure, the results of the GCE and LCE for pairwise two segmentations made by humans are shown. The humans did very well and proved to be consistent when segmenting the same image (a peak near 0), and that both methods produced segmentations that obtained higher values for the GCE and LCE error measures.

Variation in Region Sizes

To test how region sizes vary we calculated the standard deviation (σ_S) of the normalized region sizes for each segmentation (normalization was done relative to the image size). For humans, the mean of the calculated σ_S for the same image was taken. Fig. 2b) shows the resulting σ_S for 79 images (a majority for which the σ_S order Humans>MSTBorůSeg>NCutSeg existed). Results are shown sorted by the sum of the 3 σ_S for each image. The average region size variation for the whole dataset is: Humans 0.1537 , MSTBorůSeg 0.0893 and NCutSeg 0.0392. Note, that the size variation is smallest and almost content independent for the NCutSeg and largest for Humans. This shows that, the NCutSeg method is biased toward large region, since it is defined to avoid the bias of small components of cut criterion in [12].

4. Conclusion

We have evaluated segmentation results of two mostly used graph-based methods: the normalized cut (NCutSeg) and the minimum spanning tree based (MSTBorůSeg), and compared them with human segmentations. Note that, NCutSeg uses an approximation algorithm to produce segmentation results, whereas MSTBorůSeg a deterministic process. The evaluation is done by discrepancy measures, that do not penalize segmentations that are coarser or more refined in certain regions. We use only gray scale images to evaluate the quality of results on one single feature. The

[1] www.cs.berkeley.edu/projects/vision/grouping/segbench/

[2] www.cis.upenn.edu/∼jshi/software/ and www.prip.tuwien.ac.at/Research/FSPCogVis/Software

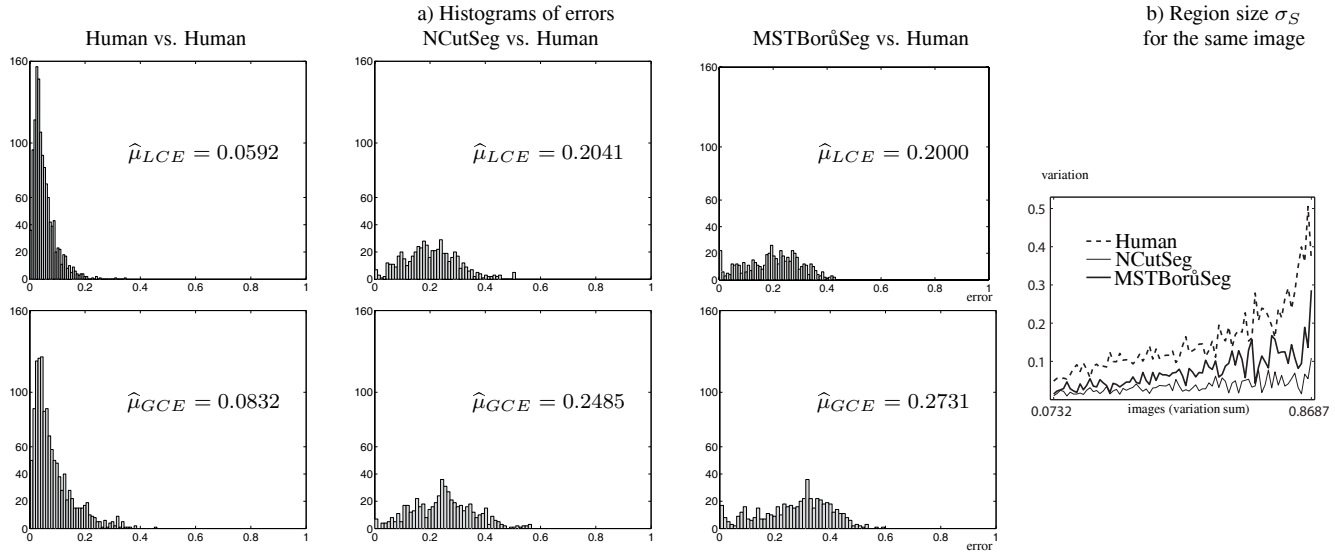

Figure 2. a) The LCE (upper) and GCE (lower) error histograms and b) region size variation (right).

two segmentation methods did not prove to be as efficient as the humans, but we showed that, for both the error measure results are concentrated in the lower half of the output domain and that the mean of the GCE measure, which is stronger than LCE , is for both around the value of 0.25. Thus both of the methods perform similar if compared with the consistency measures LCE and GCE. In the experiment with region sizes we show that humans have the biggest variation of the produced region sizes, followed by MST-BorůvSeg, and NCutSeg. This evaluation can be used to find classes of images for which the algorithms performs well.

Acknowledgment. Supported by the Austrian Science Fund under grant P18716-N13 and FSP-S9103-N04.

References

[1] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 59(2):167–181, 2004.

[2] C.-S. Fu, W. Cho, S, and K. Essig. Hierarchical color image region segmentation for content-based image system. *IEEE Trans. on IP.*, 9(1):156–162, 2000.

[3] L. Guigues, L. M. Herve, and J.-P. Cocquerez. The hierarchy of the cocoons of a graph and its application to image segmentation. *PRL*, 24(8):1059–1066, 2003.

[4] Y. Haxhimusa and W. G. Kropatsch. Hierarchy of partitions with dual graph contraction. *In Proc. Patt. Recog. Symp.*, LNCS 278:338–345, 2003.

[5] M. Heath, S. Sarkar, T. Sanocki, and K. Bowyer. A robust visual methods for assessing the relative performance of edge-detection algorithms. *IEEE PAMI*, 19(12):1338–1359, 1997.

[6] Y. Kesselman and S. Dickinson. Generic model abstraction from examples. *IEEE PAMI* 27(7):1141-1156, 2005.

[7] J. Malik, S. Belongie, T. Leung, and J. Shi. Contour and texture analysis for image segmentation. *IJCV*, 43(1):7–27, 2001.

[8] 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. *ICCV*, (2):416–423, 2001.

[9] Borra S. and S. Sarkar. A framework for performance characterization of intermediate-level grouping modules. *PR and Im. Anal.*, 19(11):1306–1312, 1997.

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

[11] R. Urquhart. Graph theoretical clustering based on limited neighborhood sets. *PR*, 13:3:173–187, 1982.

[12] Z. Wu and R. Leahy. An optimal graph theoretic approach to data clustering: theory and its application to image segmentation. *IEEE PAMI*, 15(11):1101–1113, 1993.

[13] C. Zahn. Graph-theoretical methods for detecting and describing gestalt clusters. In *IEEE Trans. Comput.*, 20, 68–86, 1971.

[14] Y. Zhang. A survey on evaluation methods for image segmentation. *PR*, 29(8):1335–1346, 1996.

[15] Y. T. Zhou, V. Venkateshwar, and R. Chellapa. Edge detection and linear feature extraction using the directional derivatives of a 2D random field model. *IEEE PAMI*, 11(1):84–95, 1989.