Evaluating Minimum Spanning Tree Based Segmentation Algorithms*

Yll Haxhimusa, Adrian Ion, Walter G. Kropatsch, and Thomas Illetschko

Pattern Recognition and Image Processing Group 183/2,
Institute for Computer Aided Automation, Vienna University of Technology, Austria
{yll, ion, krw, illetsch}@prip.tuwien.ac.at

Abstract. Two segmentation methods based on the minimum spanning tree principle are evaluated with respect to each other. The hierarchical minimum spanning tree method is also evaluated with respect to human segmentations. Discrepancy measure is used as best suited to compute the segmentation error between the methods. The evaluation is done using gray value images. It is shown that the segmentation results of these methods have a considerable difference.

1 Introduction

In [8] it is suggested to bridge and not to eliminate the representational gap, and to focus efforts on *region segmentation*, *perceptual grouping*, and *image abstraction*. The segmentation process results in "homogeneous" regions with respect to the low-level cues using some similarity measures. Problems emerge because i) homogeneity of low-level cues will not map to the semantics [8] and ii) the degree of homogeneity of a region is in general quantified by threshold(s) for a given measure [2]. 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 [12]. It is important that a grouping method has the 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 partitions [12].

Low-level cue image segmentation cannot produce a complete final "good" segmentation [11]. This lead researchers to look at the segmentation only in the context of a task, as well as the evaluation of the segmentation methods. However in [9] the segmentation is evaluated purely as segmentation by comparing the segmentation done by humans with those done by the normalized cuts method [12]. As can be seen in Fig. 1, there is a high degree of consistency of segmentation done by humans (already demonstrated empirically in [9]), 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. the pyramid. Therefore in [9] a segmentation consistency measure that does

* Supported by the Austrian Science Fund under grant FSP-S9103-N04.

A. Gagalowicz and W. Philips (Eds.): CAIP 2005, LNCS 3691, pp. 579–586, 2005.

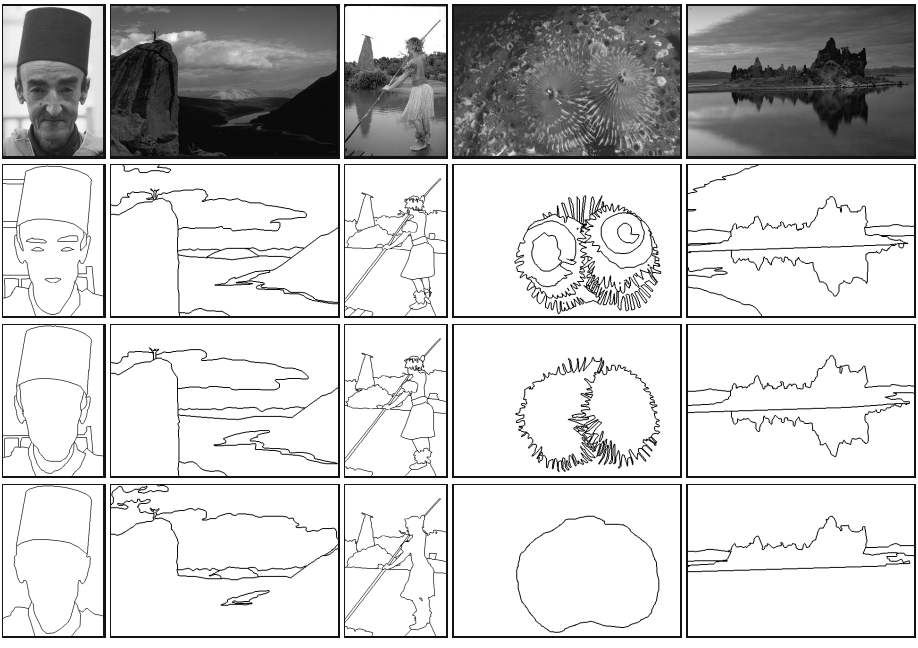

Fig. 1. Images from the Berkley image database with human segmentation [9]

not penalize this granularity difference is defined (see Sec. 4). Note that the segmented image #35 in row 2 can be coarsened to obtain the image in row 4 (and vice versa), this is called *simple refinement*; whereas to obtain image in row 3 from row 2 (or vice versa) we must coarsen in one part of the image and refine in the other (notice the chin of the man in row 3), this is called *mutual refinement*.

In this paper, we evaluate two segmentation methods based on the minimum spanning tree (MST) principle. The segmentation method based on Kruskal's algorithm [1](KrusSeg) and a parallel, hierarchical one, based on Borůvka's algorithm [6](BorůSeg) (Sec. 2). We compare these two methods following the framework of [9] i.e. comparing the segmentation results of these methods with each other. The BorůSeg is also evaluated with respect to the human segmentations. The results of the evaluation are reported in Sec. 4.

2 Segmentation Methods

A graph-theoretical clustering algorithm consists in searching for a certain combinatorial structure in the edge weighted graph, such as an MST [1,4], a minimum cut [14,12] and a search for a complete subgraph i.e. the maximal clique [10]. Early graph-based methods [15] use fixed thresholds and local measures in finding a segmentation, i.e MST is computed. The segmentation criterion is to break

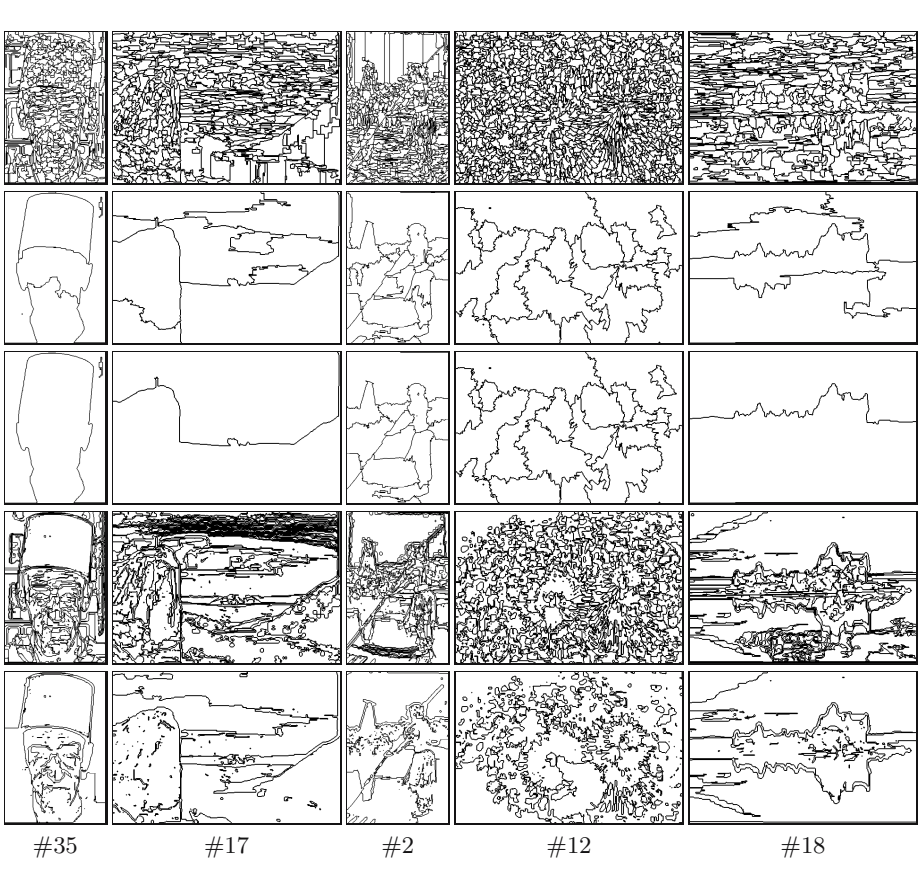

Fig. 2. Segmentation produces by BorůSeg($k = 300$) in row 1-3 (from coarser to finer segmentation), in row 4 KrusSeg($k = 300, \sigma = 1.5$) and in row 5 KrusSeg($k = 30000, \sigma = 1.5$)

the MST edges with the largest weight, which reflect the low-cost connection between two elements. To overcome the problem of a fixed threshold, Urquhart [13] normalizes the weight of an edge using the smallest weight incident on the vertices touching that edge. The methods in [1,4,6] use an adaptive criterion that depends on local properties rather than global ones.

We evaluate segmentations of the well known method [1] based on Kruskal's algorithm, with the one [6] based on Borůvka's algorithm. Since, for both methods there is a threshold dependent on the size of the connected component used ($k/|CC|$[1] see [1,6] for more details.) in the merging criteria, the segmentation inclusion trees are different, because of the way the data is processed, the first one does it in serial and the other one in parallel. Setting this threshold to zero

[1] $|CC|$ cardinality of the connected component.

both of the methods would produce the MST of the image, independent of the way the data is processed.

Some samples of the segmentation results, obtained by applying these methods to gray value images are shown in Fig. 2. The BorůSeg method is capable of producing a hierarchy of images, the pyramid (see the images in Fig. 2, where row 1 represent lower levels of the pyramid, row 2 the middle levels, and row 3 the higher levels). The methods use only local contrast based on pixel intensity values. We smoothed the images before segmenting them with the KrusSeg[2] method (Gaussian with parameter $\sigma = 1.5$), whereas BorůSeg worked with non smoothed images. As expected, and seen from Fig. 2, segmentation methods which are based only on low-level local cues can not create results as good as humans. The overall number of regions in rows 1 and 4 in each column of Fig. 2, are almost the same, and this condition is required in [9] to perform the evaluation in Sec. 4. Both of the methods are capable of segmenting the face of a man satisfactory (image #35). The BorůSeg method did not merge the statue on the top of the mountain with the sky (image #17), compared to humans which do segment this statue as a single region (see Fig. 1). Both methods have problems segmenting the sea creatures (image #12). Note that the segmentation done by humans on the image of rocks (image #18), contains the axis of symmetry, even thought there is no "big" local contrast, therefore both of the methods fail in this respect.

3 Evaluating Segmentations

There are two general methods used to evaluate segmentations: (i) qualitative and (ii) quantitative methods. Qualitative methods involve humans, meaning that different observers would give different evaluations about the segmentations (e.g. [7]). Quantitative methods are classified into analytic methods and empirical methods [16]. Analytical methods study the principles and properties of the algorithm, like processing complexity, efficiency and so on. For references on the analytic studies of methods based on minimum spanning tree see Sec. 2. The empirical methods study properties of the segmentations by measuring how "good" a segmentation is close to an "ideal" one, by determining this "goodness" with some function of parameters. Both of the approaches depend on the subjects, the first one, in coming up with the reference (perfect) segmentation[3] and the second one, in defining the "goodness" function. The difference between the segmented image and the (ideal) reference can be used to asses the performance of the algorithm [16]. The reference image could be a synthetic image or be manually segmented by humans. Higher value of the discrepancy means bigger error, signaling poor performance of the segmentation method. In [16], it is concluded that evaluation methods based on *"mis-segmented pixels should be more powerful than other methods using other measures"*. In [9] the error measures used for evaluating segmentation *counts* the mis-segmented pixels.

[2] The method is very sensitive to noise [1].

[3] Also called a gold standard [3].

In this paper we use the framework given in [9] to evaluate qualitatively the result of the KrusSeg [1] with BorůSeg [6] and of the BorůSeg with respect to humans using the discrepancy measures defined in the next section.

4 Benchmarking Segmentations

In [9] segmentations made by humans are used as a reference and basis for benchmarking segmentations produced by different methods. The concept behind this is the observation that even though different people produce different segmentations for the same image, the obtained segmentations differ, mostly, only in the local refinement of certain regions. This concept has been studied in [9] on a human segmentation database (see Fig. 1) and used as a basis for defining two error measures, which do not penalize a segmentation if it is coarser or more refined than the other. In this sense, in an image P a pixel error measure $E(S_1, S_2, p)$, between two segmentations S_1 and S_2 containing pixel $p \in P$, called the *local refinement error*, is defined as:

$$E(S_1, S_2, p) = \frac{|R(S_1, p) \backslash R(S_2, p)|}{|R(S_1, p)|} \tag{1}$$

where $\backslash$ denotes set difference, $|x|$ the cardinality of a set x, and $R(S, p)$ is the set of pixels corresponding to the connected component in segmentation S that contains pixel p. Using the local refinement error $E(S_1, S_2, p)$ the following error measures are defined in [9]: the *Global Consistency Error* (GCE), which forces all local refinements to be in the same direction, and is defined as:

$$GCE(S_1, S_2) = \frac{1}{n} \min \left\{ \sum_{p \in P} E(S_1, S_2, p), \sum_{p \in P} E(S_2, S_1, p) \right\} \tag{2}$$

and the *Local Consistency Error* (LCE), allowing refinement in different directions in different parts of the image:

$$LCE(S_1, S_2) = \frac{1}{n} \sum_{p \in P} \min \{E(S_1, S_2, p), E(S_2, S_1, p)\} \tag{3}$$

n is the number of pixels in the image. Notice that LCE $\leq$ GCE for any two segmentations. GCE is tougher measure than LCE, because GCE tolerates simple refinements, while LCE tolerates mutual refinement as well.

We have used the GCE and LCE measures presented above to evaluate the BorůSeg method [6] using the human segmented images from the Berkley humans segmented images database [9]. Also, the evaluation of BorůSeg with respect to KrusSeg is done.

4.1 Evaluation of Segmentations on the Berkley Image Database

As mentioned in [9] a segmentation consisting of a single region and a segmentation where each pixel is a region, is the coarsest and finest possible of any

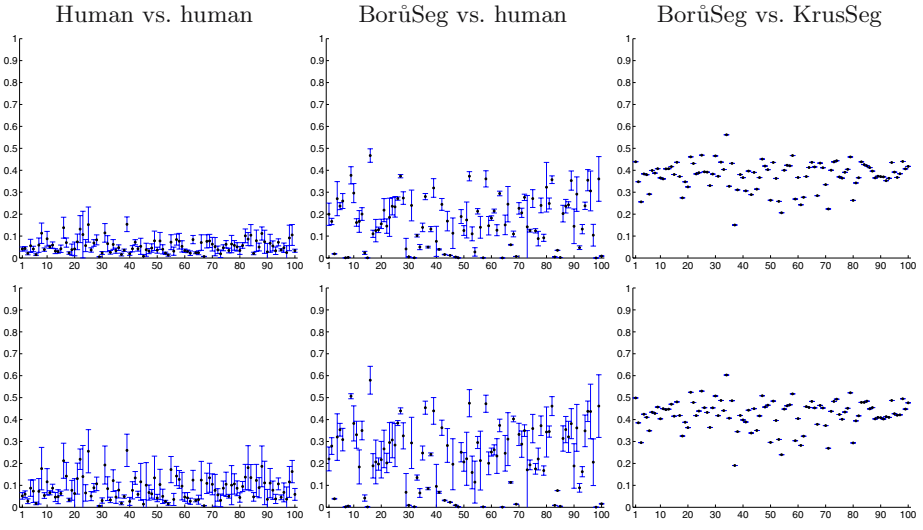

Fig. 3. The LCE (above) and GCE (below), error measure results for 100 images

segmentation. In this sense, the LCE and GCE measures should not be used when the number of regions in the two segmentation differs a lot. So, taking into consideration that the BorůSeg produces a whole hierarchy of segmentations with different number of regions (from coarser to finer), we have selected for the evaluation two levels of this pyramid. In the first case, we have taken for each image the segmentation level produced by the BorůSeg with the number of regions closest to the average number of regions produced by the humans (for the same image). When evaluating the KrusSeg we have chosen for the BorůSeg the segmentation level that had the number of regions closest to the number of regions produced by the KrusSeg method. In all the cases this meant going lower in the pyramid and taking a level which is basically a refinement of the one used when comparing to the humans. Also, as recommended by Felzenszwalb etal [1], the images given to the KrusSeg method have been smoothed with a Gaussian filter (e.g. $\sigma = 1.5$). Because the KrusSeg still produced much more regions than the human segmentations in the database have, an evaluation of the KrusSeg vs. the humans would have been unfair.

As data for the experiments, we take 100 gray level images from the Berkley Image Database[4]. For each of the images in the test, we calculate the GCE and LCE using the results produced by the KrusSeg and the corresponding level from the hierarchy produced by BorůSeg, and the human segmentations for the same image together with the corresponding level from the BorůSeg pyramid. In the case of humans and BorůSeg, having more than one pair of GCE and LCE for each image, we calculate the mean and the standard deviation. The results are summarized in Fig. 3. As a reference point, in the same figure, you can see the

[4] http://www.cs.berkeley.edu/projects/vision/grouping/segbench/

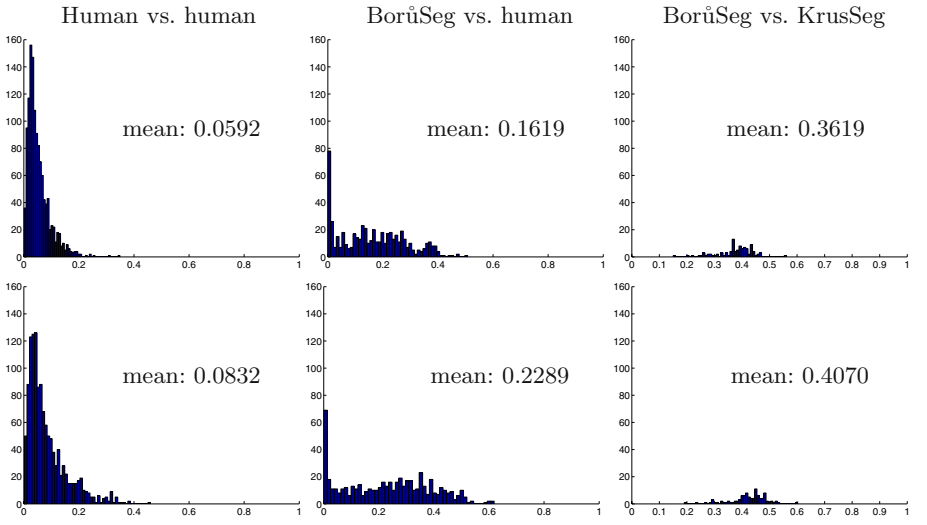

Fig. 4. Histograms of LCE (above) and GCE (below) discrepancy measure

results for calculating the GCE and LCE values for pairwise two segmentations made by humans, for the same image. We can see that the humans did very good and proved to be consistent when segmenting the same image, and that the BorůSeg produces segmentations that obtained higher values for the GCE and LCE error measures.

In Fig. 4 one can see the histograms of the GCE and LCE values obtained ($[0 \ldots 1]$, where zero means no error), humans vs. humans, BorůSeg vs. humans, and BorůSeg vs. KrusSeg. Notice that the humans are consistent in segmenting the images and the humans vs. humans histogram shows a peak very close to 0. Also, the results show that there is a considerable difference (GCE mean value 0.4) between the segmentations produced by the BorůSeg and KrusSeg methods.

5 Conclusion and Outlook

In this paper we have evaluated segmentation results of two methods based on the minimum spanning tree principle. The evaluation is done using discrepancy measures that do not penalize segmentations that are coarser or more refined in certain regions. We use gray scale images to evaluate the quality of results. In the case of BorůSeg, this evaluation can be used to find classes of images for which the algorithm has segmentation problems, corresponding to higher GCE and LCE values. We have observed that the results produced by the BorůSeg vs. KrusSeg methods have shown a considerable difference. We plan to use a larger image database to confirm the quality of the obtained results, and do the evaluation with additional low level cues (color and texture) as well as different statistical measures.

References

1. P. F. Felzenszwalb and D. P. Huttenlocher. Image Segmentation Using Local Variation. In *Proceedings of IEEE Conference on CVPR*, p:98–104, June 1998.
2. C.-S. Fu, W. Cho, S, and K. Essig. Hierarchical Color Image Region Segmentation for Content-based Image Retrieval System. *IEEE Transaction on Image Processing*, 9(1):156–162, 2000.
3. C. N. Graaf, A. S. E. Koster, K. L. Vincken, and M. A. Viergever. Validation of the Interleaved Pyramid for the Segmentation of 3*d* Vector Images. *Pattern Recognition Letters*, 15(5):469–475, 1994.
4. L. Guigues, L. M. Herve, and J.-P. Cocquerez. The Hierarchy of the Cocoons of a Graph and its Application to Image Segmentation. *Pattern Recognition Letters*, 24(8):1059–1066, 2003.
5. Y. Haxhimusa, A. Ion, W. G. Kropatsch, and L. Brun. Hierarchical Image Partitioning using Combinatorial Maps. *Joint Hungarian-Austrian Conference on Image Processing and Patt. Recog.*, p:179–186, 2005.
6. Y. Haxhimusa and W. G. Kropatsch. Segmentation Graph Hierarchies. In A. Fred, T. Caelli, R. P. Duin, A. Campilho, and D. de Ridder, editors, *Proceedings of Joint Inter. Work. on Struct., Synt., and Statis. Patt. Recog.*, LNCS 3138:343–351, 2004.
7. M. Heath, S. Sarkar, T. Sanocki, and K. Bowyer. A Robust Visual Methods for Assessing the Relative Performance of Edge-detection Algorithms. *IEEE Transactions on PAMI*, 19(12):1338–1359, 1997.
8. Y. Kesselman and S. Dickinson. Generic Model Abstraction from Examples. *IEEE Trans. on PAMI, issue on Synt. and Struct. Patt. Recog.*, 2005. to appear.
9. 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 ICCV*, (2):416–423, July 2001.
10. M. Pavan and M. Pelillo. Dominant Sets and Hierarchical Clustering. In *ICCV03*, 2003.
11. B. S. Borra and S. Sarkar. A Framework for Performance Characterization of Intermediate-level Grouping Modules. *Pattern Recognition and Image Analysis*, 19(11):1306–1312, 1997.
12. J. Shi and J. Malik. Normalized Cuts and Image Segmentation. In *Proceedings IEEE Conference CVPR*, p:731–737, 1997.
13. R. Urquhart. Graph Theoretical Clustering Based on Limited Neighborhood Sets. *Pattern Recognition*, 13:3:173–187, 1982.
14. Z. Wu and R. Leahy. An Optimal Graph Theoretic Approach to Data Clustering: Theory and Its Application to Image Segmentation. *IEEE Transactions on PAMI*, 15(11):1101–1113, 1993.
15. C. Zahn. Graph-theoretical Methods for Detecting and describing Gestalt Clusters. In *IEEE Trans. Comput.*, Vol. 20:68–86, 1971.
16. Y. Zhang. A Survey on Evaluation Methods for Image Segmentation. *Pattern Recognition*, 29(8):1335–1346, 1996.