Comparing Hierarchies of Segmentations: Humans, Nomalized Cut, and Minimum Spanning Tree [1)]

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

Abstract:
Minimum spanning tree and normalized cut based segmentation method are evaluated with respect to segmentations made by humans. The segmentation error is determined by two discrepancy measures, as a sum of miss-counted segmented pixels. Gray value images are used for the evaluation. This benchmarking is well suited in finding the class of images best suited for a particular method.

1 Introduction

The segmentation process results in 'homogeneous' regions with respect to the low-level cues using some similarity measures. Problems emerge because:

- homogeneity of low-level cues will not map to the semantics [7], and
- the degree of homogeneity of a region is in general quantified by threshold(s) for a given measure [2].

Also some the cues can contradict each other. 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 [11]. Low-level cue image segmentation cannot produce a complete final 'good' segmentation [10], therefore the segmentation is studied only in the context of a task, as well as the evaluation of the segmentation methods. However, segmentation methods can be a valuable tools in image analysis in the same sense as edge detectors are, without taking the context into consideration. Therefore, in [9] the segmentation is evaluated 'purely' as segmentation

[1)] Supported by the Austrian Science Fund under grant FSP-S9103-N04 and P18716-N13.

by comparing the segmentation done by humans with those done by a particular method. There is a consistency of segmentation done by humans (see Fig. 1 and [9]), even thought humans segment images at different granularity (refinement or coarsening). This refinement or coarsening could be thought as hierarchical structure on the image, i.e. the pyramid. Therefore in [9] a segmentation consistency measure that does not penalize this granularity is used (see Sec. 3).

The evaluation is made having these ideas in mind:

- real world images should be used, because it is difficult to extrapolate conclusion based on synthetic images to real images [14], and
- the human should be the the final evaluator.

Evaluation of the segmentation algorithms is difficult because it depends on many factors [6] among them: the segmentation algorithm; the parameters of the algorithm ; the type(s) of images used in the evaluation; the method for evaluation of the segmentation algorithms, etc. There are two general methods to evaluate segmentations: (i) qualitative and (ii) quantitative methods. Qualitative methods are evaluated by humans, meaning that different observers would give different opinions about the segmentations (e.g. already encountered in edge detection evaluation [6], or in image segmentation [9]). On the other hand the quantitative methods are classified into analytic methods and empirical methods [13]. Analytical methods study the principles and properties of the algorithm, like processing complexity, efficiency and so on. The empirical methods study properties of the segmentations by measuring how 'good' a segmentation is close to an 'ideal' one, by measuring 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[1)] and the second one defining the function. The difference between the segmented image and the (ideal) reference one is used to asses the performance of the algorithm [13]. The (ideal) reference image could be a synthetic image or manually segmented by humans. These discrepancy method measures the difference between the segmented image and reference images. Higher value of the discrepancy means bigger error, signaling poor performance of the segmentation method. In [13], it is concluded that evaluation methods based on "mis-segmented pixels should be more powerful that other methods using other measures". In [9] the error measures used for segmentation evaluation 'count' the mis-segmented pixels.

In this paper, we evaluate two graph-based segmentation methods, the normalized cut [11](NCutSeg) and the method based on the minimum spanning tree [5](MSTBorůvSeg)(Sec. 2). We compare these two methods as in [9] i.e. comparing the segmentation result of the two graph-based

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

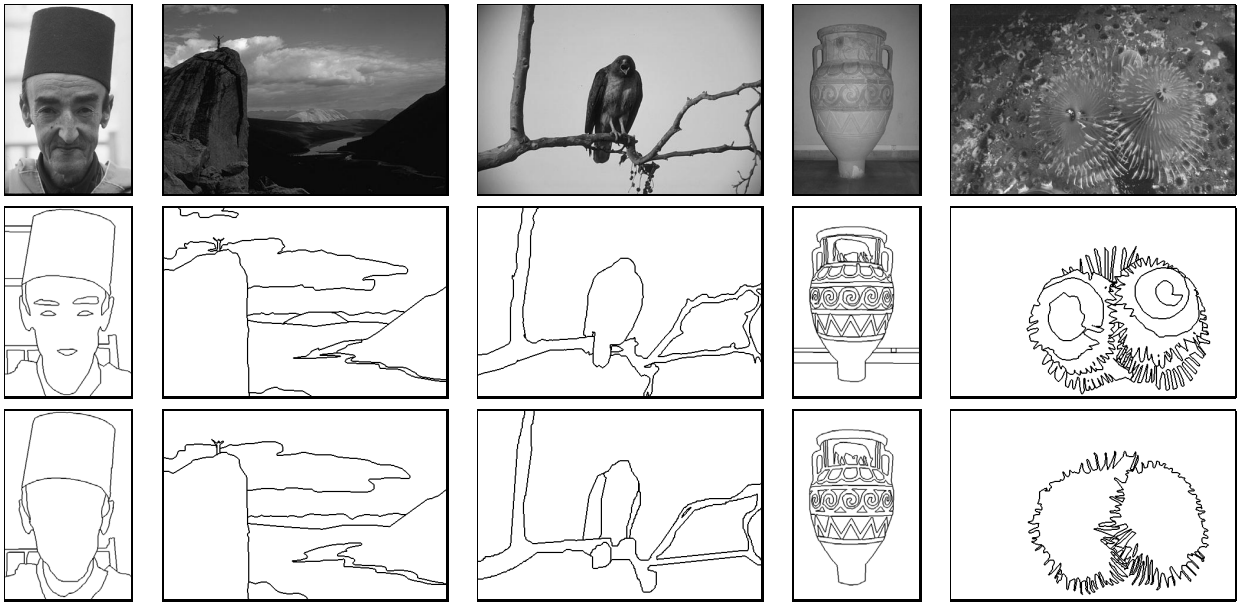

Figure 1: Images from the Berkley image database with segmentation done by humans [9].

methods with the human segmentations[2)]. The preliminary results are reported in Sec. 3.

2 Hierarchical Segmentation Methods

In the subsections below normalized cuts and minimum spanning tree based image partitioning methods used in the evaluation are summarized.

2.1 Normalized Cuts Segmentation Method (NCutSeg)

The normalized graph cut [11](NCutSeg) approach starts by representing the problem as a weighted undirected graph $G = (V, E, w_{i,j})$, where the vertices represent points in the feature space, and an edge is formed between every pair of vertices. The weight on edges is a function of the similarity between vertices that are joint by edges. The problem is posed as finding the partition of V into the sets of vertices such that the similarity within the same set is hight and across different sets is low. The solution in measuring the goodness of the image partitioning is the minimization of the normalized cut as a generalized eigenvalue problem. The graph $G = (V, E, w)$ can be partitioned into two disjoint sets by simply cutting (deleting) edges connecting vertices of these

[2)] We confirm the results in [9] of the evaluation of NcutSeg with respect to humans.

sets. The sum of the weights of deleted edges can be used to measure the dissimilarity between these two sets. In [12] this measures is minimized to produce a clustering.The clustering is done by recursively minimizing the cut criterion in the resulted segments. In [12] this global optimal criterion is used to produce 'good' segmentation on images, but the method was biased toward cutting small sets (mostly containing a single vertex). To overcome this problem, [11] propose a normalized cut criterion. Note that NCutSeg is a region splitting method. The weights of the edges in [11] are set to $w_{i,j} = \exp(-\frac{\|I(i)-I(j)\|}{\sigma_I}) \cdot \exp(-\frac{\|X(i)-X(j)\|}{\sigma_X})$ if $\|X(i) - X(j)\| < r$, otherwise 0. where $X(i)$ is the spatial location of the vertex i, $I(i)$ the intensity value. $w_{i,j} = 0$ if i and j are r pixels apart and σ_I and σ_X control the influence of intensity and/or spatial position on the overall weight. The method is instructed to give a particular number of regions[3)]. The detailed description of the algorithm is presented in [11].

2.2 Minimum Spanning Tree Segmentation Method

The minimum spanning tree [5](MSTBorůSeg) approach starts, as previously, by transforming the image to a attributed graph representation $G(V, E, w)$, where vertices represent pixels and edges their neighborhood. But instead of cutting edges, in this method the edges are added to connected components based on the minimum spanning tree principle. This method, as the previous one, is able to produce a stack of graphs, i.e. a hierarchy of partitions. In order to decide which component to merge a function is defined that measures the difference along the boundary of two components relative to a measure of differences of components' internal differences. This definition tries to encapsulate the intuitive notion of contrast: a contrasted zone is a region containing two connected components whose inner differences (internal contrast) are less than differences within it's context (external contrast). The pairwise comparison of neighboring vertices, i.e. partitions is used to check for similarities [1]. This function judges whether or not there is evidence for a boundary between two partitions. This boolean function is true i.e. the border exists, if the external contrast difference between two partitions is greater than the internal contrast differences within a partition. If the border does not exist the partitions are merged, therefore MSTBorůSeg is a region growing method. The resulted segmentation of both methods are the so called *crisp* segmentation. The detailed presentation of the algorithm is given in [5]. The attributes of edges are set as difference between pixel intensities i.e. $w(u_i, u_j) = |I(u_i) - I(u_j)|$. The parameter α is set for the function $\tau(CC) = \alpha/|CC|$, where $\alpha = const$[4)] and $|CC|$ is the

[3)] The source code is found in `http://www.cis.upenn.edu/ jshi/software/` and the parameters are not changed ($r = 10$, $\sigma_X = 30$, and $\sigma_I = 0.1$).

[4)] $\alpha = 500$.

size of region CC. A larger constant α sets the preference for larger components. Note that as size of $|CC|$ gets larger, which happens as the algorithms proceeds toward the top of the pyramid, the function $\alpha \rightarrow 0$, which means that the influence of the parameter decreases.

Note that both of the methods are capable of producing a hierarchy of images. The segmentation results of applying these methods to gray value images are shown in Fig. 2. The methods use only local contrast based on pixel intensity values. As it is expected, and can be seen from the Fig. 2, segmentation methods which are based only on low-level local cues can not create results as good as humans. Even thought it looks like, the NCutSeg method produces more regions, actually the overall number of regions in rows 1 and 3 (respectively 2 and 4) in each column of Fig. 2, are almost the same, but the MSTBorůSeg produces a bigger number of small regions. Anyway both of the methods (see Fig. 2) were capable of segmenting the face of a man satisfactory (image #35). The MSTBorůSeg 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 (Fig. 1). Both methods have problems segmenting the see creatures (image #12). Note that the size of the regions in one image produced by NCutSeg have similar sizes.

3 Evaluating Segmentations

In [9] segmentations made by humans are used as a reference and basis for benchmarking segmentations produced by different methods. Different people produce different segmentations for the same image, thus the obtained segmentations differ, only in the local refinement of certain regions. This concept has been studied on the human segmentation database (Fig. 1) [9] and used as a basis for defining two error measures, which do not penalize a segmentation if it is coarser or more refined than another. We give these measures for the sake of clarity of the presentation. In this sense, a pixel error measure $E(S_1, S_2, p)$, 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 region in segmentation S that contains pixel p. Using the local refinement error $E(S_1, S_2, p)$ the following error measures are defined [9]: the *global consistency error* (GCE), which forces all local refinements:

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

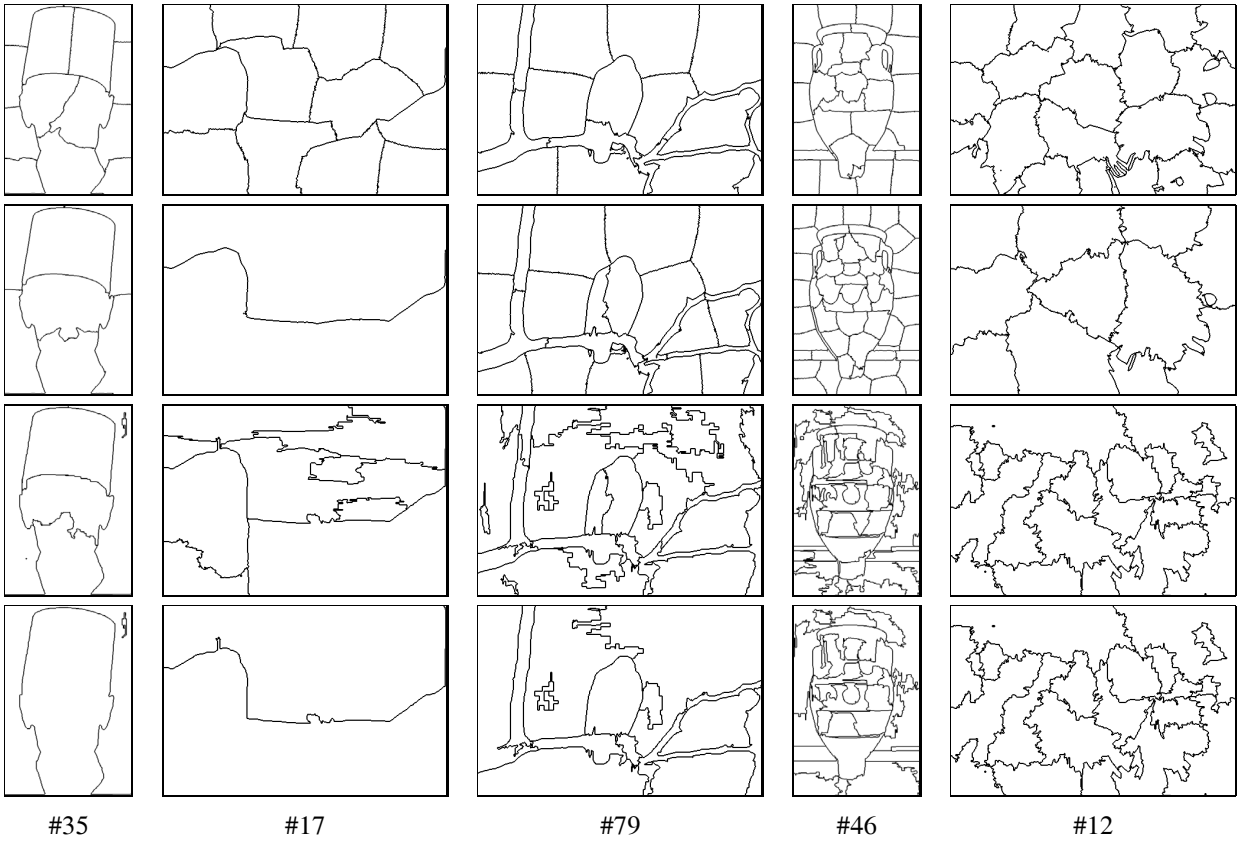

Figure 2: Segmentation produced by NCutSeg (row 1 and 2) and MSTBorůSeg (row 3 and 4)

and the *local consistency error* (LCE):

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

where I is the image size. Notice that the GCE is a stronger measure then the LCE, and both of the criteria count mis-segmented pixels. These two criteria are used to evaluate the quality of segmentation of NCutSeg [11] and MSTBorůSeg [5] against the human segmentations

4 Evaluation Results

For the evaluation, we use 32 gray level images from Berkley Image Database[5)]. For segmentation, we have used the normalized cuts Matlab implementation, from the authors [11] [6)] and for the MSTBorůSeg an implementation based on combinatorial pyramids [4].

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

[6)] http://www.cis.upenn.edu/∼jshi/software/

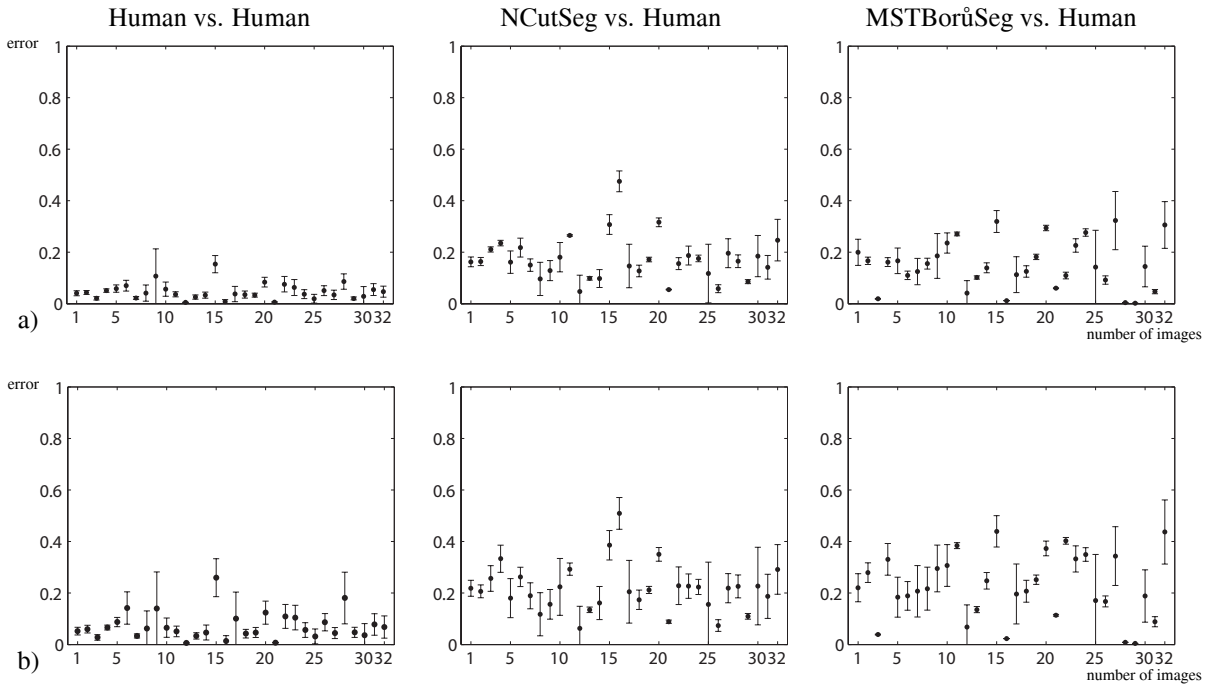

Figure 3: The a) LCE and b) GCE.

As mentioned in [9] a segmentation made of only 1 region and a segmentation where each pixel is a region can be the coarsening and refinement of any 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. In this line of ideas, taking into consideration that both methods can produce segmentations with different number of regions, we have taken for each image as a region count reference number, the average number of regions from the human segmentations available for that image. We have instructed the NCutSeg to produce the same number of regions and for the MSTBorůSeg we used the level of the pyramid with 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 2 methods and all the human segmentations available for that image. Having more then one pair of GCE and LCE for each method and image, we have calculated the mean and the standard deviation. The preliminary results on 32 gray value real world images are depicted in Fig 3 and Fig 4. There is a big similarity between the values of GCE and LCE for both methods for the same image. As a reference point, in the same figure, the results for calculating the GCE and LCE values for pairwise two segmentations made by humans, for the same image. The humans did very good and proved to be consistent when segmenting the same image (high peak in histograms near zero and $\widehat{\mu}_{LCE} = 0.0429$; $\widehat{\mu}_{GCE} = 0.0662$), and that both methods produced

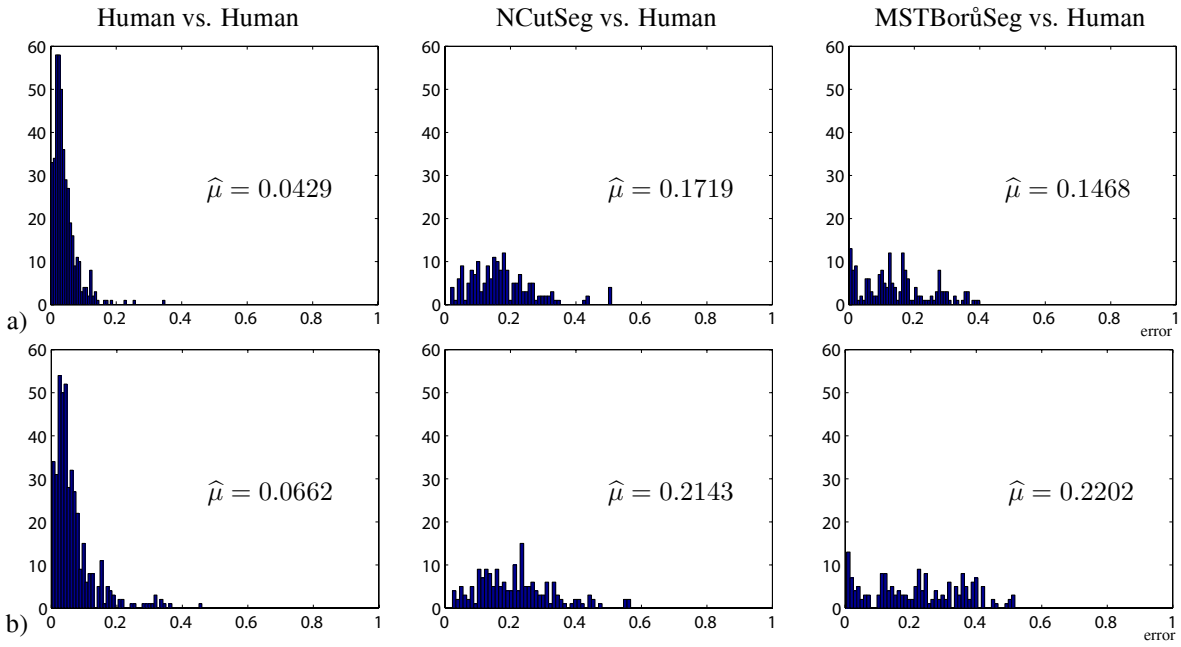

Figure 4: Histograms of a) LCE and b) GCE.

segmentations that obtained higher values for the GCE and LCE error measures ($\widehat{\mu}_{LCE} = 0.1719$; $\widehat{\mu}_{GCE} = 0.2143$ and $\widehat{\mu}_{LCE} = 0.1468$; $\widehat{\mu}_{GCE} = 0.2202$). Note that, none of the two segmentation methods have proved to outperform the other one on all images of the used image database.

5 Conclusion and Outlook

In this paper we have evaluated segmentation results of two graph-based methods; the well known method based on the normalized cuts and the method based on the minimal spanning tree principle, and compared with human segmentations. The evaluation is done by using discrepancy measures, that do not penalize segmentations that are coarser or more refined in certain regions. We used only gray images to evaluate the quality of results on one feature. In Fig. 4, histograms of the GCE and LCE values obtained, Humans vs. Humans, NCutSeg vs. Humans, and MSTBorůSeg vs. Humans. Note, that the humans were consistent in segmenting the images and the humans vs. humans histogram shows a peak very close to 0. The two segmentation methods have not proved to be as efficient as the humans, since for both the error measure results are concentrated in the lower half of the output domain and that the mean of the GCE measure is for both around the value of 0.2. This evaluation can be used to find classes of images for which NCutSeg and/or MSTBorůSeg algorithms are best suited. We plan to use a larger image database to confirm the quality of results obtained, and do the evaluation with additional low level cues.

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 retrival system. *IEEE Trans. on IP*, 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 $3d$ vector images. *PRL*, 15(5):469–475, 1994.

[4] Y. Haxhimusa, A. Ion, W. G. Kropatsch, and L. Brun. Hierarchical Image Partitioning using Combinatorial Maps. *Joint Hungarian-Austrian Conference on IP and PR*, Hungary, p:179–186, 2005.

[5] Y. Haxhimusa and W. G. Kropatsch. Hierarchy of Partitions with Dual Graph Contraction. In B. Milaelis and G. Krell, editors, *25th DAGM Symposium*, LNCS 278:338–345, 2003.

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

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

[8] W.G. Kropatsch, A. Leonardis, and H. Bischof. Hierarchical, adaptive and robust methods for image understanding. *Surveys on Math. for Ind.*, (9):1-47, 1999.

[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. *ICCV*, (2):416–423, 2001.

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

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

[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] Y. Zhang. A Survey on Evaluation Methods for Image Segmentation. *PR*, 29(8):1335–1346, 1996.

[14] Y. T. Zhou, V. Venkateshwar, and R. Chellapa. Edge Detection and Linear Feature Extraction Using 2D Random Field Mode. *IEEE PAMI*, 11(1):84–95, 1989.