Technical Report

An Evaluation of 3D Segmentation Algorithms using Seismic Variance Data

Klaus Köster, Xinquan Shen and Michael Spann

June 2, 1998

Contents

1 Introduction

This report describes an evaluation of two segmentation algorithms reported by Köster and Spann [1] and Shen *et al* [2] Although both algorithms are supposed to segment 2D as well as 3D data sets, this evaluation will be performed using several slices of a 3D variance filtered seismic data set. Section 2 describes the algorithms briefly.

Several evaluation and comparison attempts of algorithms in different research areas have already been published: edge detection schemes [3], global thresholding methods [4], optical flow estimation [5], stereo [6] and shape from shading [7]. Attempts for the evaluation of segmentation algorithms can be split up into two major categories: *analytical* and *empirical* methods [8]. Analytical methods directly examine and assess the segmentation techniques themselves by analysing the algorithm's principles and properties. Empirical methods indirectly judge the segmentation algorithms by applying them to test images and measuring the quality of segmentation results. Zhang [8] splits the empirical methods further up into *goodness methods* and *discrepancy methods*. Goodness methods measure some desirable properties of the segmented image (eg. intra-region uniformity [9], inter-region contrast [10], region shape [11], etc.), while discrepancy methods measure differences to a pre-defined reference segmentation (ground truth).

The most popular approaches to support the authors' argumentation in the published literature are the analytical methods. This category avoids the concrete implementation of the algorithm and the results can be exempted from the influence caused by the arrangement of evaluation experiments. The choice of this comparison method is very often the only possible due to the lack of details for a re-implementation of the published algorithms. However, evaluations of this category generating a quantitative result are very limited. Abdou and Pratt [12] analysed the performance of several edge detectors using a precisely defined and calculated detection probability ratio. This analytical method results in a quantitative measure but can only be calculated for simple edge detectors. It is in general harder to find a quantitative measure for algorithm comparison using analytical methods compared to using empirical methods.

The most obvious empirical discrepancy method is probably the evaluation using a pixel classification process. An approach proposed by Yasnoff *et al* [13] defines two error ratios:

$$M_I^{(k)} = \frac{\text{Number of pixels of class k not classified as k}}{\text{Total Number of pixels of class k}} \tag{1}$$

$$M_{II}^{(k)} = \frac{\text{Number of pixels of other classes than k classified as k}}{\text{Total Number of pixels of other classes than k}}. \tag{2}$$

This simple method has been extended by several researchers to take different quantities into account (spatial information, number of objects, object features, etc.) [8]. However, none of the proposed measures has ever been applied by other researchers and a comparison of segmentation algorithm based on these measures appears therefore to be difficult. In this report we apply an adapted version of the framework of Hoover *et al* [14] for the evaluation of range image segmentation algorithms to perform our evaluation. The framework compares the segmentation with a corresponding set of ground truth segmentations. An objective comparison tool performs the evaluation using specified metrics for correct detection, over-segmentation, under-segmentation, missed regions and noise regions. A large number of real test images, the comparison tool and results of four different research groups working on range image segmentation is publicly available and allows therefore an objective and competitive comparison of the own work with state-of-the-art algorithms. The metrics and the adaption of the comparison framework will be described in Section 3. Section 4 shows the results of the evaluation and Section 5 finishes this report with a discussion of the experimental results.

2 Description of the Algorithms

The algorithm of Shen *et al* for image segmentation is a bottom-up method successively clustering regions at each level until no further changes are possible according to the clustering conditions. The regions are graph nodes in an irregular hierarchical data structure representing the image at a number of spatial resolutions. The base level of the hierarchy is the original image data while the top level describes the result of the segmentation. The iterative calculation of the next graph hierarchy is performed by clustering regions of the lower hierarchy. Clustering at each level of the hierarchy is performed only if two adjacent regions are mutually nearest neighbours (NN). For this reason we refer to this algorithm from now on as the *NN-method*. The nearest neighbour of a region is the adjacent region which minimises a dissimilarity function and satisfies a few predicate functions at the same time. The applied dissimilarity function is the variance of the model of the two tested regions' union. The predicate function changes during the segmentation stages. At the base level, where region nodes are representing simply the pixels/voxels, the predicate function is the greylevel difference which has to be smaller than a pre-defined threshold. In the later stages the predicate functions checks the similarity of the model slopes and variances and the existence of a continuous greylevel transition between the merge candidates by exploiting Student's *t-test*. The data of a homogeneous region is represented by a linear model.

Köster and Spann's region-growing algorithm [1] also works in a bottom-up way generating an irregular graph pyramid to store neighbourhood relations between regions. Each graph node represents a region in the image and contains links to neighbour regions and the parameters of a constant data model. The algorithm works in three main stages: the initial cell tessellation, the primitive region (PR) generation stage and the iterative merging stage. The technique starts by partitioning the 3D data into a regular grid of overlapping $3{\times}3{\times}3$ voxel cells. In the next step the initial cells are shifted by at most one voxel into each direction towards a position generating a local variance minimum. Hereby cells placed on a boundary of two regions are eliminated. The variance and the region model is computed with the *least median of squares* (LMedS) estimation technique [15]. This technique allows the robust detection of inlier and outlier elements within a data set's model. The following PR generation stage generates a first PR hierarchy by recursively merging the initial cells to their neighbours if their model is mutually inlying within their neighbour's model. Starting from here the initial hierarchy is iteratively refined by merging the PRs. The applied region model similarity test is the *mutual inlier ratio* (MIR), which names this algorithm from now on as *MIR-method*. A PR A is merged to his adjacent PR B if the ratio of inliers in $\{A, B\}$ and inliers in $\{A\}$ is larger than a theoretically derived discrimination value. This value is a function of the set sizes of A and B. The test has to be satisfied mutually. The merging procedure ends when no more regions satisfy the merging criterion.

3 Evaluation Methods

In this section we describe the evaluation methods to obtain an empirical overview over the properties of the two algorithms. Section 3.1 describes the applied evaluation metrics. The generation of the necessary set of ground truth segmentation for the special case of seismic data sets will be elucidated in Section 3.2.

3.1 Evaluation Metrics

A tool objectively compares the *machine segmentation* (MS) against the corresponding GT image using a set of defined performance metrics for instances of correctly segmented, missed, and noise regions, over- and under-segmentation. The definitions for deciding classifications are based upon a tolerance T having a range of $0.5 < T \leq 1.0$ and reflecting the strictness of the definitions. Informally, Hoover *et al* [14] defined the metrics as follows:

correct detection: A pair of regions R_n in the GT image and R_m in the MS image are classified as an instance of correct detection if

1. at least T percent of the pixels in R_m are marked as pixels of R_n *and*
2. at least T percent of the pixels in R_n are marked as pixels of R_m.

over-segmentation: A region R_n in the GT image and a set of regions $R_{m1}, \ldots, R_{mx}$ in the MS image are classified as an instance of over-segmentation if

1. at least T percent of the pixels in each region R_{mi} are marked as pixels of R_n *and*
2. at least T percent of the pixels in R_n are marked as pixels of the union of $R_{m1}, \ldots, R_{mx}$.

under-segmentation: A set of regions $R_{n1}, \ldots, R_{nx}$ in the GT image and a region in the MS image R_m are classified as an instance of under-segmentation if

1. at least T percent of the pixels in region R_m are marked as pixels of the union of $R_{n1}, \ldots, R_{nx}$ *and*
2. at least T percent of the pixels in R_{ni} are marked as pixels of R_m.

missed: A region *in the GT image* not being found as an instance of correct detection, over-segmentation or under-segmentation is classified as missed.

noise: A region *in the MS image* not being found as an instance of correct detection, over-segmentation or under-segmentation is classified as noise.

3.2 Generation of Ground Truth

A set of corresponding ground truth segmentations is necessary to apply the empirical evaluation framework of Hoover *et al* [14]. The generation of a ground truth segmentation for 3D seismic images is for several reasons more difficult compared to the reported segmentation of range images. Firstly, the interpretation of the images is currently very subjective and can differ even from one expert to another. Expert knowledge in geophysics and geology is very often necessary to obtain a reasonable result. Shapes and geometry of the objects in the data can differ considerably from objects to which the human eye is usually exposed to. And finally the interpretation of 3D images is more complicated compared to 2D images because of the larger number of adjacent pixels. Hence to obtain a correct segmentation slices in front of and behind the current 2D slice have to be evaluated as well.

However, an evaluation of the algorithms is important to express strengths and weaknesses of each technique in a quantitative way. For this reason we decided to evaluate our techniques using several 2D slices of a segmented 3D set. The size of the chosen data set was $110 \times 90 \times 30$ voxels. The ground truth segmentations for the slices were obtained by computing an average

```
Split(IMAGE_A, IMAGE_B, IMAGE_OUT)
begin
  Initialise TABLE with -1;
  REG_LABEL := 0;

  PIX_A   := first pixel of IMAGE_A;
  PIX_B   := first pixel of IMAGE_B;
  PIX_OUT := first pixel of IMAGE_OUT;

  while(PIX_A)
    {
      if(TABLE(PIX_A,PIX_B) == -1)
        {
          TABLE(PIX_A,PIX_B) := REG_LABEL;
          REG_LABEL := REG_LABEL+1;
        }

      ValueOf(PIX_OUT) := TABLE(PIX_A,PIX_B);

      PIX_A   := next pixel of IMAGE_A;
      PIX_B   := next pixel of IMAGE_B;
      PIX_OUT := next pixel of IMAGE_OUT;
    }
end
```

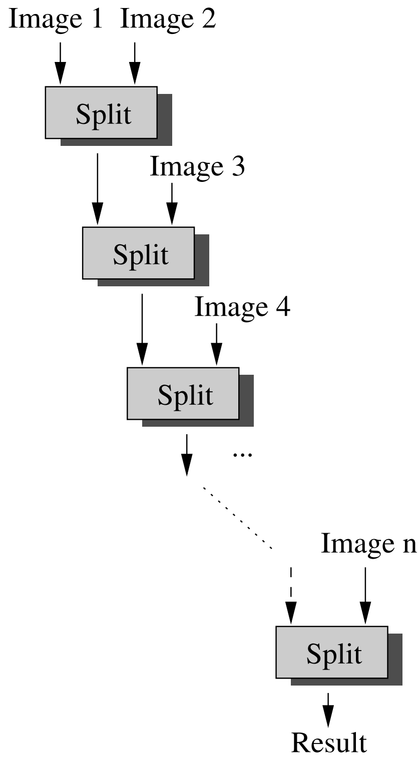


Figure 1: The *split* stage of the algorithm and its application for multiple input images.

segmentation of multiple independently obtained manual segmentations for each slice. The averaging process was introduced to avoid the influence of errors in the individual segmentations and to come to a more objective result. This averaging process will be described in detail in Section 3.2.1. To obtain another objective measurement-based ground truth set we also generated a simple threshold segmentation of the set. This approach will be elucidated in Section 3.2.2.

3.2.1 Averaging Manual Segmentations

The ground truth images were calculated by averaging three independently obtained manual segmentations. The averaging process was performed by a tool which works in two stages. An initial *split*-stage splits the data set into patches having homogeneous labels in all input segmentations. Figure 1 shows the split algorithm and the application for multiple segmentations of the same image. Before applying the algorithm every connected region has to obtain its on unique label.

The split stage results in small pixel groups that belong together according to all input segmentations. The second stage merges these groups to larger regions. Two adjacent regions are merged whenever the regions have the same label in 50% or more of the input segmentations. Given two adjacent regions A and B the decision criterion is found by determining the most frequent labels M_A and M_B in both of the regions. With these labels the ratios $r_A(n)$ and $r_B(n)$ are determined for each of the $1 \ldots n \ldots N$ input segmentations:

$$r_A(n) = \frac{\text{Number of elements with Label } M_B \text{in Region } A}{\text{Total Number of elements in Region } A} \tag{3}$$

$$r_B(n) = \frac{\text{Number of elements with Label } M_A \text{in Region } B}{\text{Total Number of elements in Region } B}. \tag{4}$$

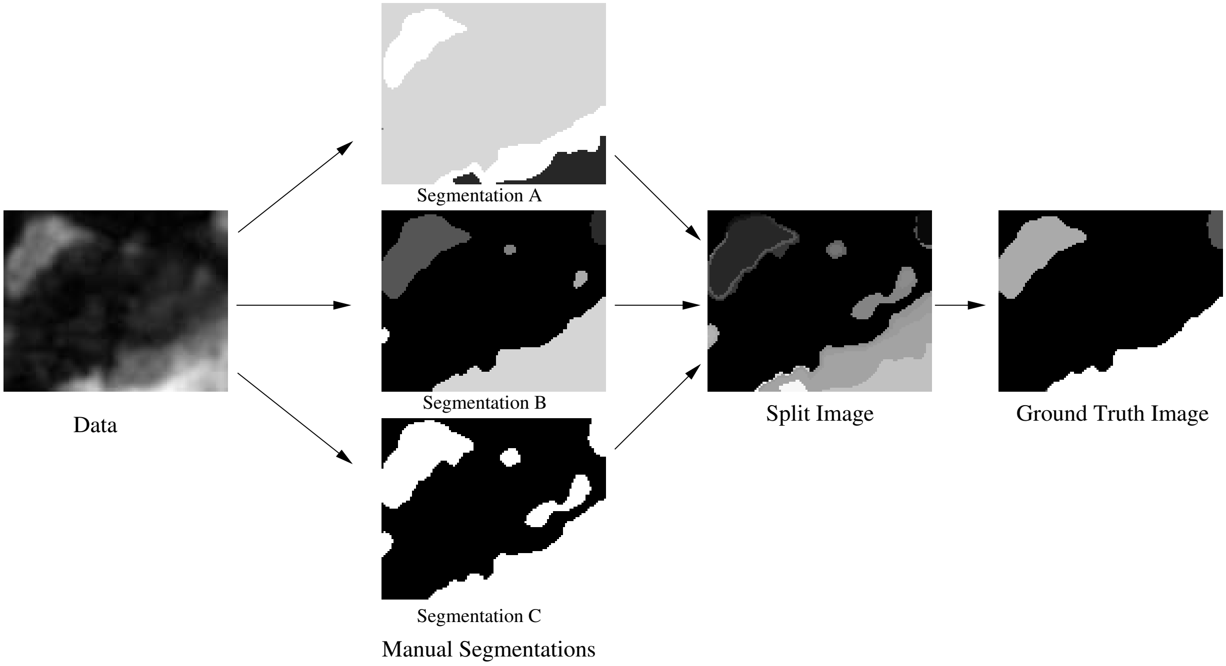


Figure 2: The ground truth generation process for a slice of a 3D data set.

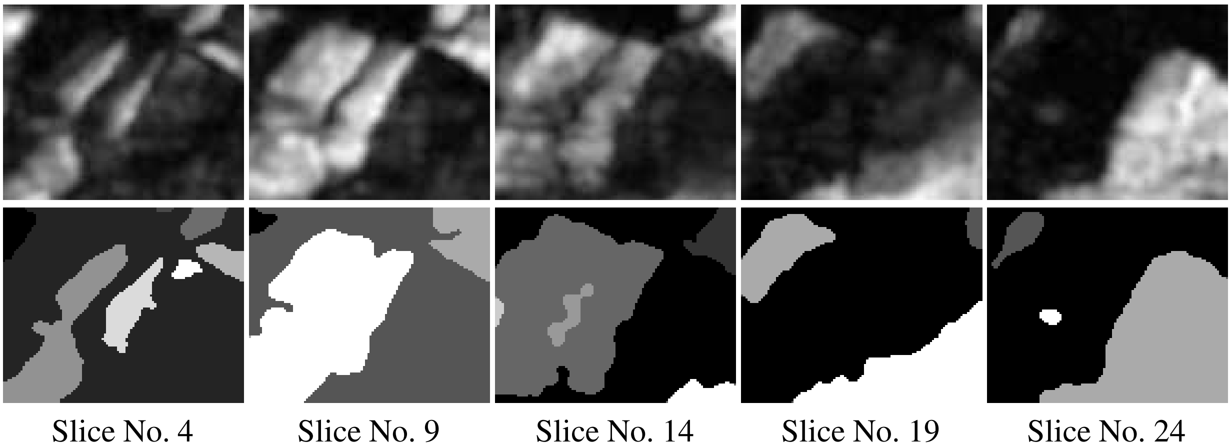


Figure 3: 5 equi-distant test slices and their corresponding average segmentations.

The ratios for the individual input images are averaged to obtain the final decision relevant ratio:

$$r = \frac{1}{2N}\sum_{N}^{n=1} r_A(n) + r_B(n). \tag{5}$$

The two regions are merged if the final ratio r is larger than 0.5. The neighbour with the highest ratio will be chosen for merging if multiple neighbours satisfy this criterion. The merging stage is iteratively repeated until no further regions can be merged. Figure 2 shows the ground truth generations process for a single slice of a data set. Figure 3 shows the five test slices of a 3D data set and their corresponding ground truth segmentations which have been chosen for this evaluation.

3.2.2 Thresholding Approach

Figure 4 shows the greylevel histogram of the variance filtered seismic data set. The graph shows spikes for greylevels below and above ≈ 66. For this reason we decided to use 66 as a threshold

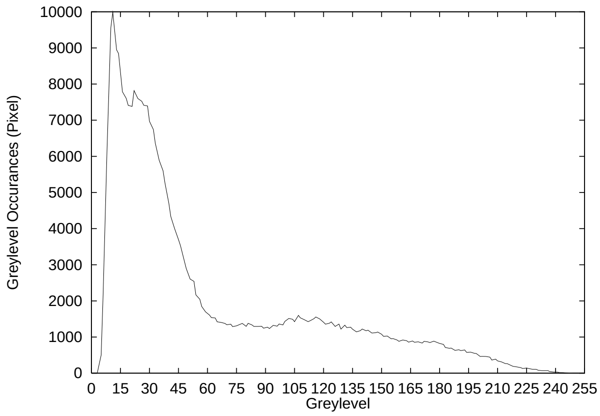
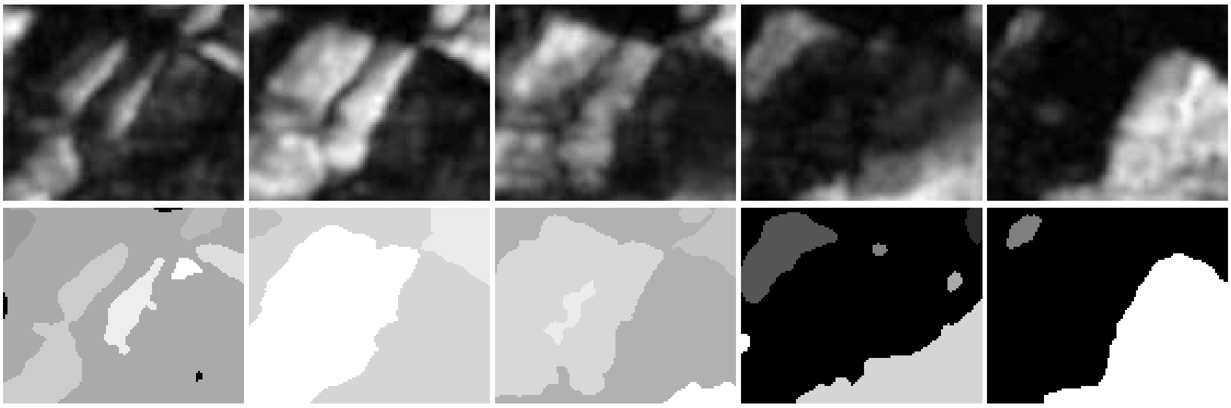


Figure 4: Greylevel histogram of a variance filtered seismic data set.



Slice No. 4	Slice No. 9	Slice No. 14	Slice No. 19	Slice No. 24

Figure 5: The 5 equi-distant test slices and the results of a simple thresholding segmentation with a greylevel threshold of 66.

to separate the regions in the 3D data set. Figure 5 shows the obtained segmentation after each connected region has been marked by a unique label. The thresholding result underlines the remarkable smoothness of the data set as all regions are reasonably large and connected. In the typical result of a thresholding approach without further post-processing we would expect a large amount of single non-connected pixels.

4 Evaluation Results

The evaluation results are split up into two parts: Section 4.1 and Section 4.2 describe results being specific to the individual segmentation techniques, while Section 4.3 illustrates a comparison of the two methods' results.

4.1 Evaluation Specific to NN-Method

The NN-method resulted in particular for larger significance levels into a very detailed segmentation with a large number of regions compared to the calculated ground truth images. Due to the slice-wise evaluation this number was growing as connected 3D regions were not necessarily connected in the 2D slices. Figures 6 to 10 show two graphs for each test slice displaying the number of region being smaller than a given size (left graph) and the percentage of the image

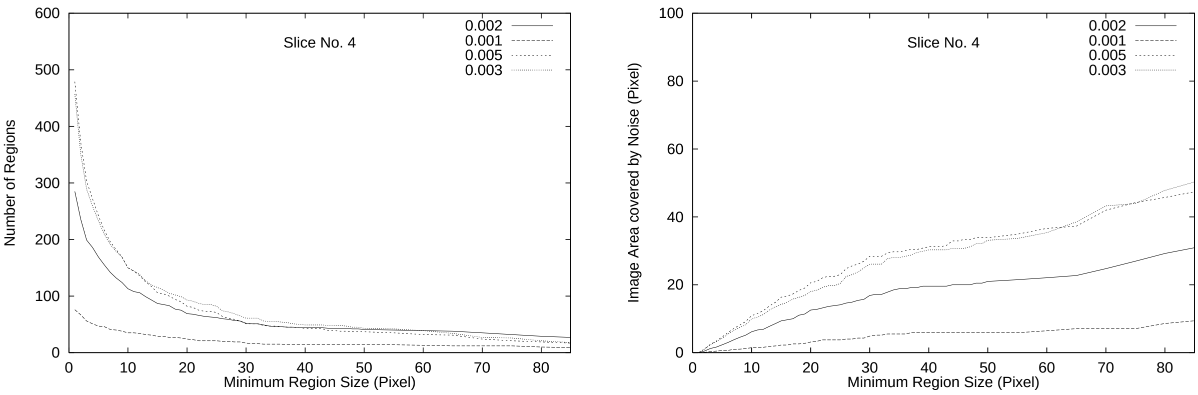


Figure 6: Number of regions being smaller than a minimum-region thresholds and the percentage of the covered image area in slice no. 4 for different levels of significance.

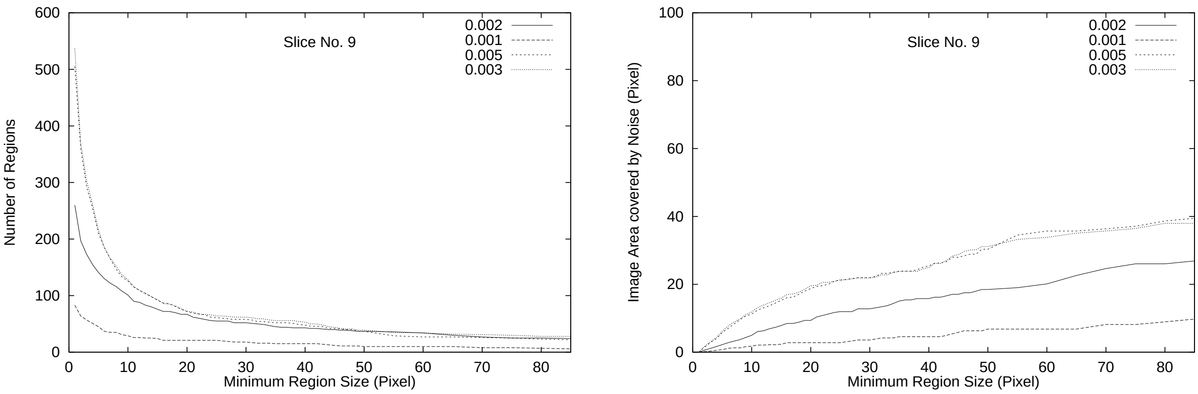


Figure 7: Number of regions being smaller than a minimum-region thresholds and the percentage of the covered image area in slice no. 9 for different levels of significance.

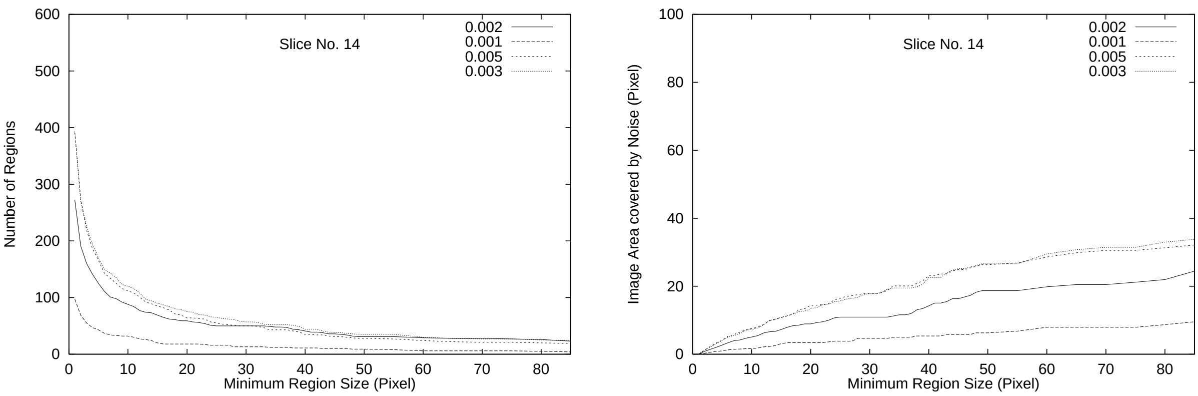


Figure 8: Number of regions being smaller than a minimum-region thresholds and the percentage of the covered image area in slice no. 14 for different levels of significance.

area these regions cover. The evaluation of such a segmentation with the described metrics results in a large number of over-segmented regions and would not reflect the true quality of the segmentation but the slice-wise evaluation. For this reason we decided to label regions smaller than 20 pixels in each slice as noise. The comparison tool reserves a few extra labels for this case. The chosen threshold marks the point from where on the graphs tend to become flatter in

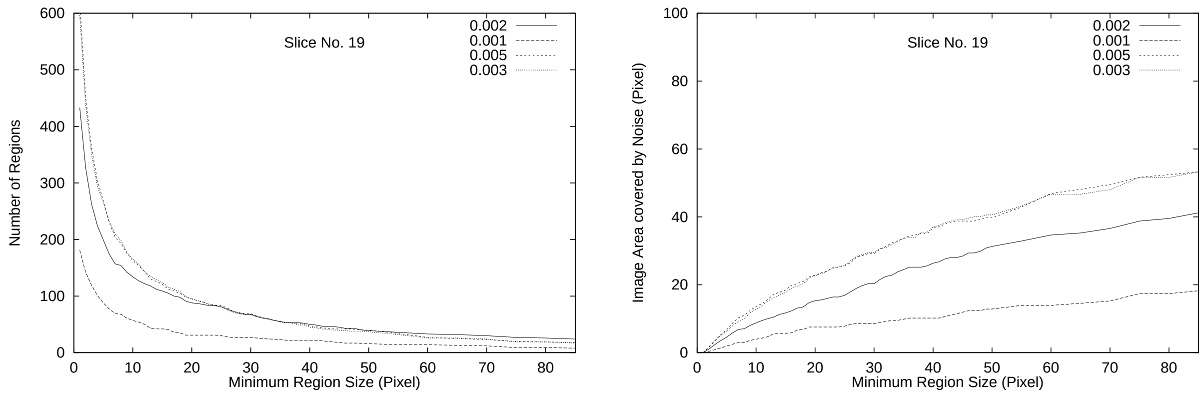


Figure 9: Number of regions being smaller than a minimum-region thresholds and the percentage of the covered image area in slice no. 19 for different levels of significance.

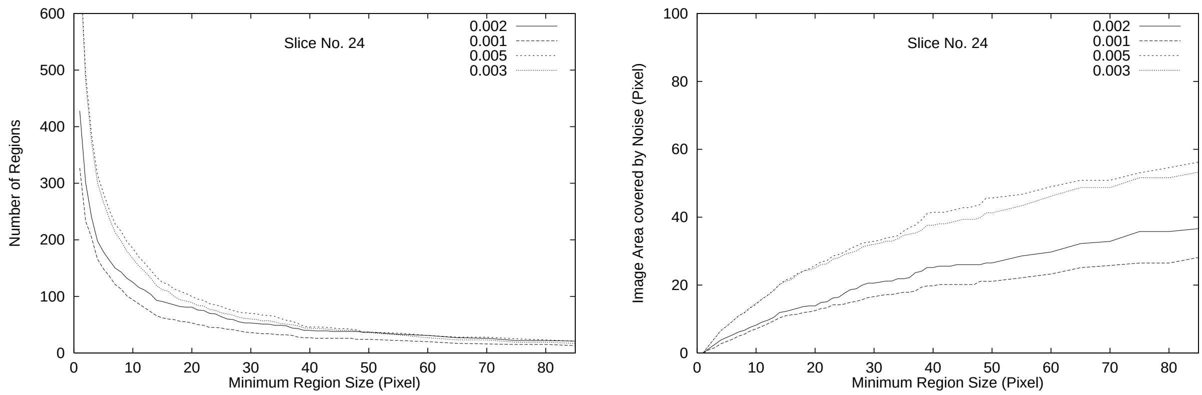
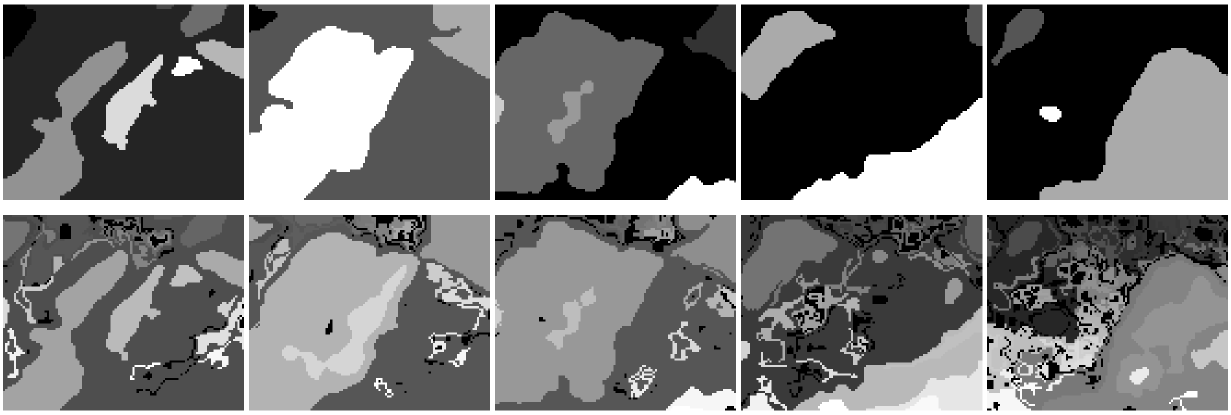


Figure 10: Number of regions being smaller than a minimum-region thresholds and the percentage of the covered image area in slice no. 24 for different levels of significance.

figures 6 to 10 and the number of regions changes slower with a rising threshold. However, using this threshold we were not able to eliminate all the small regions. Observing the slices of the segmentation one notices that all small regions have a very long and thin, almost co-planar shape. In this case the computed parameters of the linear model are imposed by a high uncertainty in one coordinate direction due to the lack of data. This suggest that the regions do not get merged to their adjacent regions because of a break-down of the estimation procedure for some of the model parameters.

Figure 11 shows a visual comparison of the obtained segmentation result and the ground truth calculated by averaging manual segmentations. For the display we chose the result obtained with a significance level of 0.001 as it provided the best test results. All major regions of the ground truth have been detected in the segmentation with a good correspondence in shape and size. The result slices also display a large number of small regions which have not been labelled as noise due to their size. These regions let us expect a large value for the metric of over-segmented regions. Regions labelled as noise are displayed with black color in figure 11.

Figure 12 shows the number of correctly detected regions for different compare tool tolerances and different levels of significance. The outstanding result has been obtained for a significance level of 0.001, where 66.6% of the regions have been correctly detected at a compare tool tolerance of 50% according to the calculated ground truth. For higher compare tool tolerances the performance drops.



Slice 4 Slice 9 Slice 14 Slice 19 Slice 24

Figure 11: Visual comparison of slices of the segmentation result of the NN algorithm (bottom row) and the ground truth images (top row).

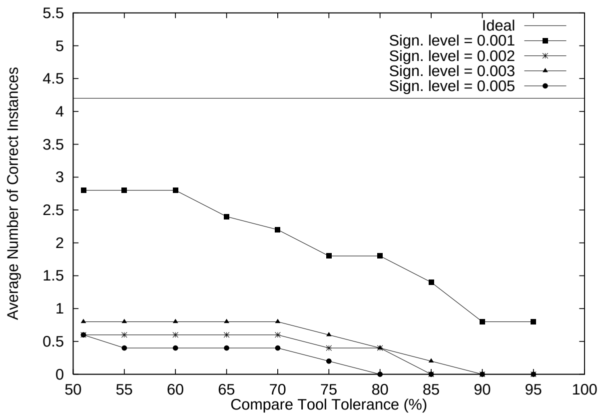


Figure 12: Number of correctly detected regions for different significance levels α of the t-test.

4.2 Evaluation Specific to the MIR-Method

The MIR-method compares the calculated inlier ratio and the result of a discrimination function to obtain a merge decision. The discrimination function result is obtained using a value α which determines the value towards which the function converges for large test set sizes [1]. Figure 13 displays the number of correctly classified regions for different values of α. The figure documents the best result for a value of $\alpha = 0.28$. In this case the algorithm classifies 71% of the regions correctly at a compare tool tolerance of 50%. For larger tolerances or values of α the performance of the algorithm drops.

Figure 14 shows a visual comparison of the test slice segmentations with the corresponding ground truth slices. For this comparison we chose a value of $\alpha = 0.28$. The major regions in the ground truth segmentation have all been detected. The shape of the segmented regions does not correspond to the ground truth shape as well a in the result of the NN-method. The region boundaries do not appear to be very smooth. However, the background (displayed dark) of the MIR-result does correspond better to the ground truth result than the over-segmented background of the NN-method.

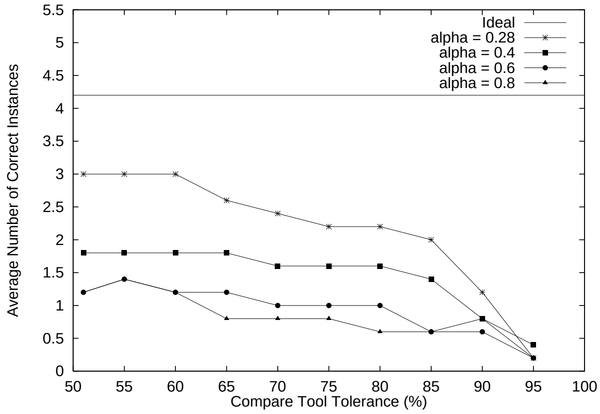


Figure 13: Number of correctly detected regions for different values for the applied α in the MIR-test.

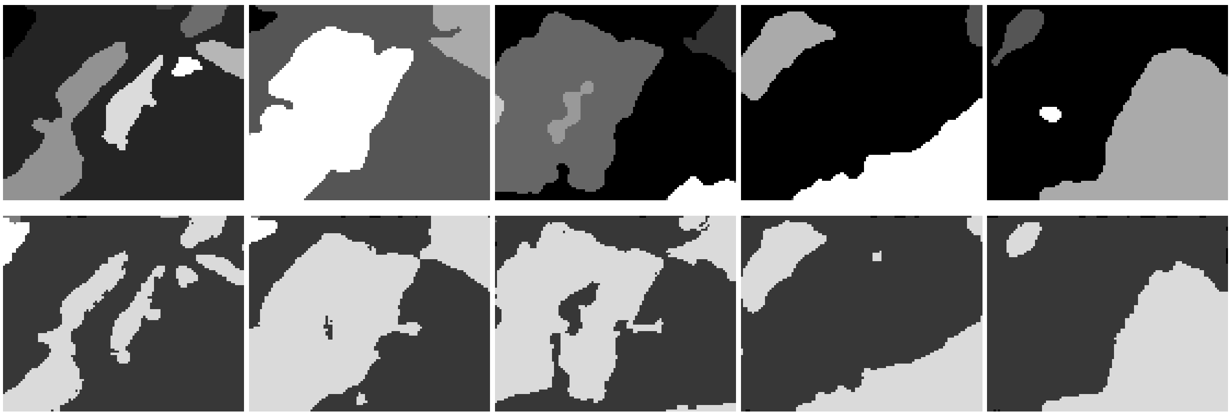


Figure 14: Visual comparison of slices of the segmentation result of the MIR algorithm (bottom row) and the ground truth images (top row).

4.3 Comparison of both Methods

A good evaluation result of a segmentation algorithm would obtain a high score for correctly classified regions and low values in the error measures of over- and under-segmentation, missed and noise regions. The results would not deteriorate for larger compare tool tolerances.

Figure 15 shows the comparison of the graphs of correctly classified regions for different compare tool tolerances. Both methods perform similarly with a slight advance for the MIR-method. Both segmenters' performances drop significantly for compare tool tolerances larger than 80%. The tests results obtained using different ground truth segmentations show similar tendencies for both segmenters.

Figure 16 shows the error metric for over-segmented instances. As expected previously, the MIR-method performs significantly better in this test due to the over-segmented background in the NN-method's result.

The error metric measuring instances classified as under-segmented returns the ideal result for the NN-algorithm in both tests in figure 17. The result states that there is no under-segmentation and the trade-off between under- and over-segmentation is not yet completely balanced by the chosen level of significance (0.001). The choice of a lower level seems to be promising to result in improved evaluation measures. The MIR-method returns a low measure of under-segmented instances.

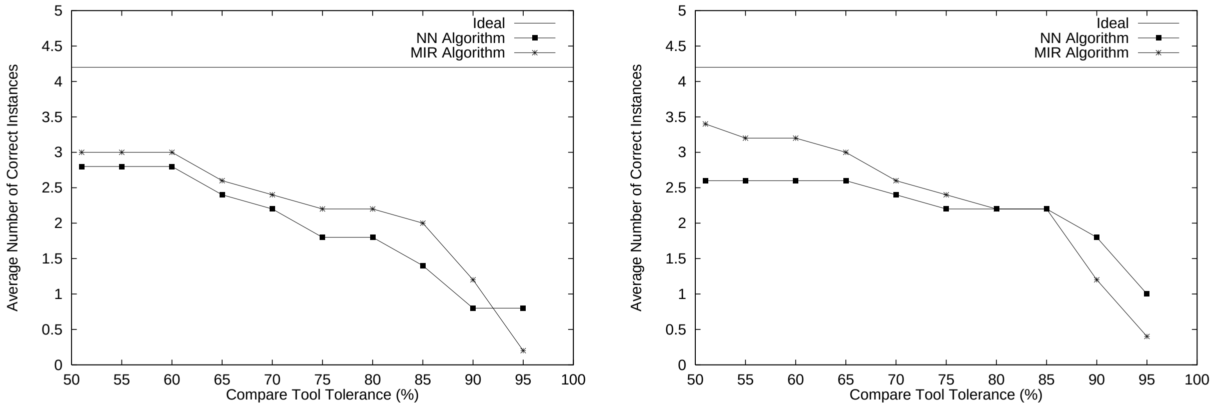


Figure 15: Average number of correctly detected regions for NN-method (significance level 0.001) and MIR method (α=0.28).

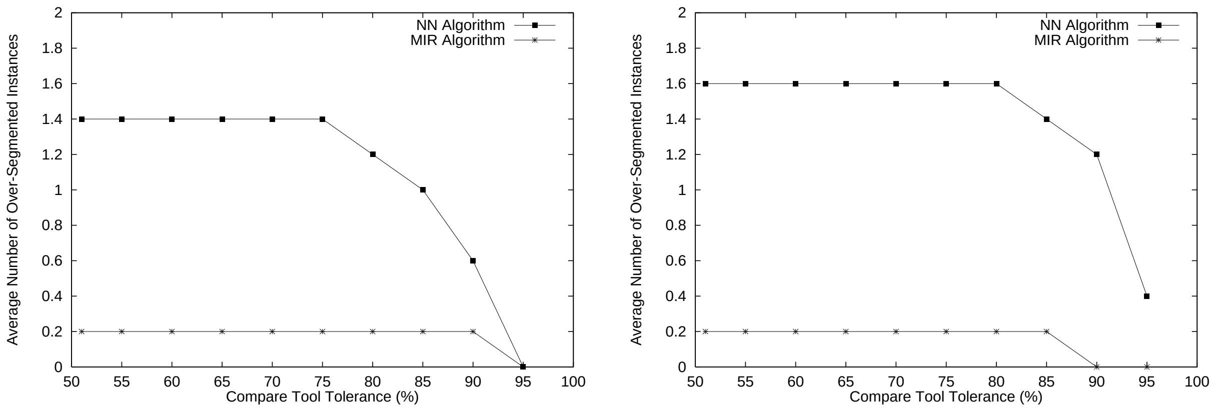


Figure 16: Average number of over-segmented instances for NN-method (significance level 0.001) and MIR method (α=0.28).

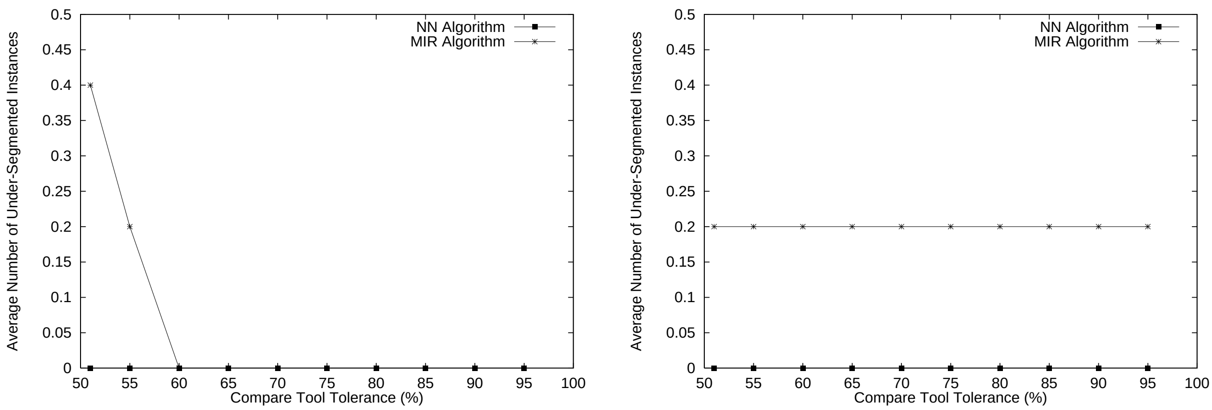


Figure 17: Average number of under-segmented instances for NN-method (significance level 0.001) and MIR method (α=0.28).

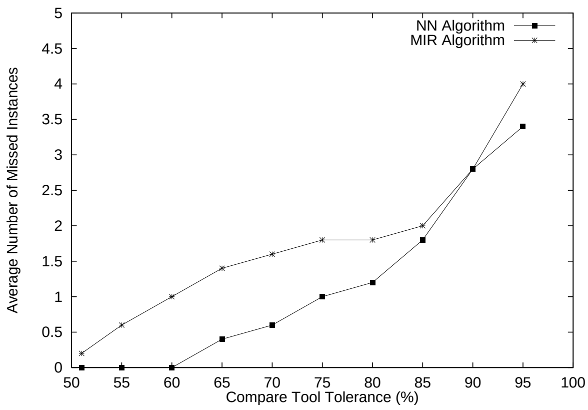


Figure 18: Average number of missed instances for NN-method (significance level 0.001) and MIR method (α=0.28).

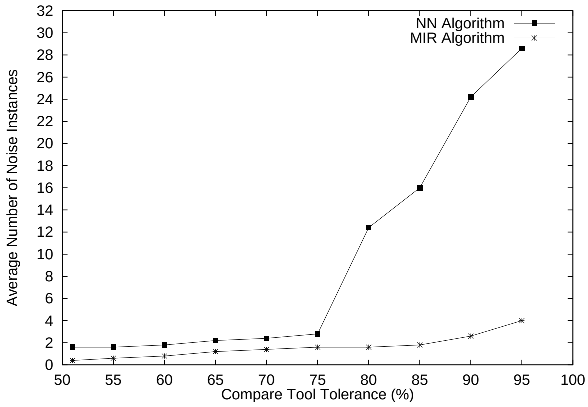


Figure 19: Average number of noise instances for NN-method (significance level 0.001) and MIR-method (α=0.28).

The average number of missed regions is displayed in figure 18. The NN-method missed in average one region more than the MIR-method. Both methods display a significant rise of missed regions for tolerance levels larger than 80%. At this point the formerly as correct classified regions do not satisfy this criterion anymore and get classified as missed.

Figure 19 shows the graph with the average number of noise instances. The NN-method returns a significantly higher error due to the large number of small regions. In particular for compare tool tolerances above 80% the measured value rises significantly as regions, which for smaller tolerances get classified as over-segmented, do not satisfy this criterion anymore and get therefore classified as noise.

5 Discussion and Conclusions

The discussion of the algorithms' evaluation and its result has two parts. Section 5.1 presents an analytical discussion of the algorithms, where we derive their properties from the examination of the applied principles and techniques. We will give our conclusions in Section 5.2.

5.1 Analytical Properties of the Methods

In this report we have compared two image segmentation algorithms for 3D data sets described in [2] and [1]. Both methods are region growing methods using different statistical hypothesis tests to find a decision to merge the regions. Both methods store the region data and the neighbour relations in a hierarchical graph description. However, there are some significant differences in the applied data models, their estimation techniques and the clustering methods.

The examined algorithm use region data models of different degrees. Shen's method (NN-method) [2] applies a linear model while in Köster and Spann's algorithm (MIR-method) [1] a simpler constant model is used to represent the data of a region. The chosen test data is the result of a 3D variance filter which high-lights image regions of high variance and applies finally a smoothing filter. The assumption of a luminance linearly dependent on the spatial image position has only limited validity in this case and a constant model is sufficient to describe the data. However, a linear model incorporates the constant model, but the additional flexibility is computational more demanding and offers also space for instabilities in the estimation procedure.

The algorithms apply two different model estimation procedures. While the NN-method applies a least-mean of squares technique (LM) the MIR-method uses the more robust LMedS estimator. However, due to the smoothness of the test data the number of outliers is very low and there is no necessity for a robust estimation procedure. We exchanged the robust LMedS estimator against a simple LM estimator and obtained better results with considerably lower computational demands.

The initial set of regions is obtained in different way in both algorithms. While the MIR-Method groups overlapping $3{\times}3{\times}3$ voxel cells, the NN-Method starts by clustering adjacent voxels with similar greylevel. The MIR-method applies a shift procedure to the cells to avoid a bad precision in the final segmentation due to the size of the calls. However, a final refinement stage towards pixel level is necessary as the segmented regions consist of $3{\times}3{\times}3$ voxel cells only. The shown segmentation results have been obtained without such a stage. The initial pixel clustering method of the NN-method has to avoid co-planar regions as the design matrix $\mathbf{A}$ will not be full ranked in this case and the estimation procedure will not work. The NN-method recognises such regions and merges them with an adjacent region in this case. More difficult to handle are regions which result in a full ranked design matrix, but are very thin in one or two coordinate directions. In this case the estimated model parameters are imposed with a high uncertainty and a miss-estimation cannot be detected by the method. In the worst case this region cannot be merged to another adjacent region which means the introduction of small thin regions into the result due to a failure of the estimation procedure. The observed large number of small regions in the shown segmentation results and the high score for over-segmented regions might support this thought.

The NN-method uses Student's t-test to obtain a merge decision while the MIR-method paper [1] proposes the MIR-test to solve this problem. The t-test is not sensitive to variance differences, which involves an additional test considering the region variances in the NN-method. To examine the similarity of two regions represented by a linear model the test procedure is applied individually to each model parameter. An combined application of the test to the complete model is not considered. The MIR test obtains a test result using the inlier/outlier information of the regions elements only. Therefore any model of any vectorial dimension can be tested in a combined way with only a single pass through the tested set. However, the scale estimation, which is important for the detection of outliers, becomes a very crucial part of the test method.

The algorithm use a hierarchical graph to describe the neighbour relationships of the regions. Both methods compute a "best neighbour for merging" for each region according to their specific criteria and merge only mutually best neighbours. This strategy is on the one hand very safe and prevents agglomerated over-merging, but on the other hand the convergence rate towards the final segmentation is very low resulting in long computation times. Hierarchical structures are memory-wise demanding due to the need to store parent, child and neighbour relations for each region. For this reason none of the examined methods stores more than two graph hierarchies at the same time.

5.2 Conclusions

Sections 3 and 4 presented an empirical evaluation of two segmentation algorithms for 3D data. The evaluation involves a comparison of the obtained segmentation with a set of corresponding ground truth images. Section 3.2 described the difficulties to generate ground truth for the seismic test data and proposed two different methods to generate ground truth segmentations. The methods for ground truth generation and the suitability of the test data for the algorithms' principles and techniques (described in Section 5.1) might not allow a final verdict on the quality of the algorithms. However, the coinciding results using two different types of ground truth images let this evaluation appear to be an indicator of the algorithms' properties.

The measurement of correctly classified regions resulted in a good score and agrees with the subjective impression of a visual comparison of the segmentation and the ground truth. Both methods detect the major regions correctly. However, the NN-method returns a large number of very small regions and returns therefore high values for the error measures of over-segmented and noise instances. Taking the low error score for under-segmented regions into account the choice of a lower level of significance might improve the total result.

Inspecting the segmentation results of the MIR-method visually the very fringed region boundaries are obvious. To obtain smoother boundaries and a better performance in the evaluation measures a further refinement stage is necessary.

References

[1] K. Köster and M. Spann, "MIR: An Approach to Robust Clustering - Application to Range Segmentation." submitted for publication in IEEE Trans. on PAMI, April 1998.

[2] X. Shen, M. Spann, and P. Nacken, "Segmentation of 2D and 3D images through a hierarchical clustering based on region modelling," *Pattern Recognition*, vol. 31, p. ??, September 1998.

[3] J. R. Fram and E. S. Deutsch, "On the quantitative evaluation of edge detection schemes and their comparison with human performance," *IEEE Trans. C*, vol. 24, pp. 616–628, 1975.

[4] S. U. Lee, S. Y. Chung, and R. H. Park, "A comparative performance study of several global thresholding techniques for segmentation," *Computer Vision, Graphics, and Image Processing*, vol. 52, pp. 171–190, 1990.

[5] J. L. Barron, D. J. Fleet, and S. S. Beauchemin, "Performance of optical-flow techniques," *International J. Computer Vision*, vol. 12, pp. 43–77, 1994.

[6] R. C. Bolles, H. H. Baker, and J. M. Hannah, "The JISCT stereo evaluation," in *Proc. Image Understanding Workshop*, (Washington, D.C.), pp. 263–274, 1993.

[7] R. Zhang, P. Tsai, J. Cryer, and M. Shah, "Analysis of shape from shading techniques," in *Computer Vision and Pattern Recognition*, (Seattle), pp. 377–384, 1994.

[8] Y. J. Zhang, “A survey on evaluation methods for image segmentation,” *Pattern Recognition*, vol. 29, pp. 1335–1346, 1996.

[9] J. S. Weszka and A. Rosenfeld, “Threshold evaluation techniques,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 8, pp. 622–629, 1978.

[10] M. D. Levine and A. Nazif, “Dynamic measurement of computer generated measurement,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 7, pp. 155–164, 1985.

[11] P. K. Sahoo, S. Soltani, A. K. C. Wong, and Y. C. Chen, “A survey of thresholding techniques,” *Computer Vision, Graphics, and Image Processing*, vol. 41, pp. 233–260, 1988.

[12] I. E. Abdou and W. K. Pratt, “Quantitative design and evaluation of enhancement/thresholding edge detectors,” *Proceedings of the IEEE*, vol. 67, pp. 753–763, 1979.

[13] W. A. Yasnoff, J. K. Mui, and J. W. Bacus, “Error measures for scene segmentation,” *Pattern Recognition*, vol. 9, pp. 217–231, 1977.

[14] A. Hoover, G. Jean-Baptiste, X. Jiang, P. J. Flynn, H. Bunke, D. B. Goldgof, K. Bowyer, D. W. Eggert, A. Fitzgibbon, and R. B. Fisher, “An experimental comparison of range image segmentation algorithms,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, pp. 673–689, July 1996.

[15] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*. New York: Wiley, 1987.