Quality Measures for Image Segmentation Using Generated Images

Theo E. Schouten, Maurice S. klein Gebbinck

Katholieke Universiteit Nijmegen
Computing Science Institute
Toernooiveld 1, 6525 ED Nijmegen
The Netherlands
Email ths@cs.kun.nl
tel +31-80-653175, fax +31-80-553450

ABSTRACT

To provide a quantitative measure of the quality of a segmentation of an image a "true" segmentation must be known and the differences between the two segmentations must be transformed into one or more quality values. A method is described to generate a realistic satellite image and its true segmentation to sub-pixel level using ground truth data and a real image. Quality measures are described which evaluate two kinds of errors: the splitting of a real field into more than one segment and the merging of pixels from different fields into a segment. Results for various segmentation methods are discussed.

1 INTRODUCTION

Segmentation is an important step in the processing of images. There are however few quantitative methods (see[11] for an overview) for determining the quality of a segmentation of an image. Usually only inspection by a human expert is used to provide a qualitative evaluation of the result of a segmentation process.

In[4] and[3] statistical quality measures are used based on the number and size of segments, the intensity variations inside segments and the variations between segments. For the measure in[4] it was shown in[8] that it was not useful for satellite images. In[11] artificial images are generated and features measured on objects in the segmented images are compared to their known values from the image generation process. It is our belief that the above methods will work on images with one (as used in[11]) or a few objects on a background but not for satellite images.

Our quality measures will be based on the generation of a "true" segmentation and the transformation of the differences between a segmentation and the "true" segmentation to one or more numbers indicating the quality of that segmentation.

For remote sensing applications the pixel size is often relatively large compared to the size of fields and objects. Therefore a large percentage of the pixels are "mixed": they receive intensity contributions from more than 1 field or object. The true position of the fields or objects must be known with an accuracy better than the pixel size in order to study the handling of those "mixed" pixels by the segmentation process. The "true" segmentation can be obtained in two ways:

- segmentation (by programs and careful inspection and correction by human experts) of images with higher resolution than the image under study.
- generation of an artificial image.

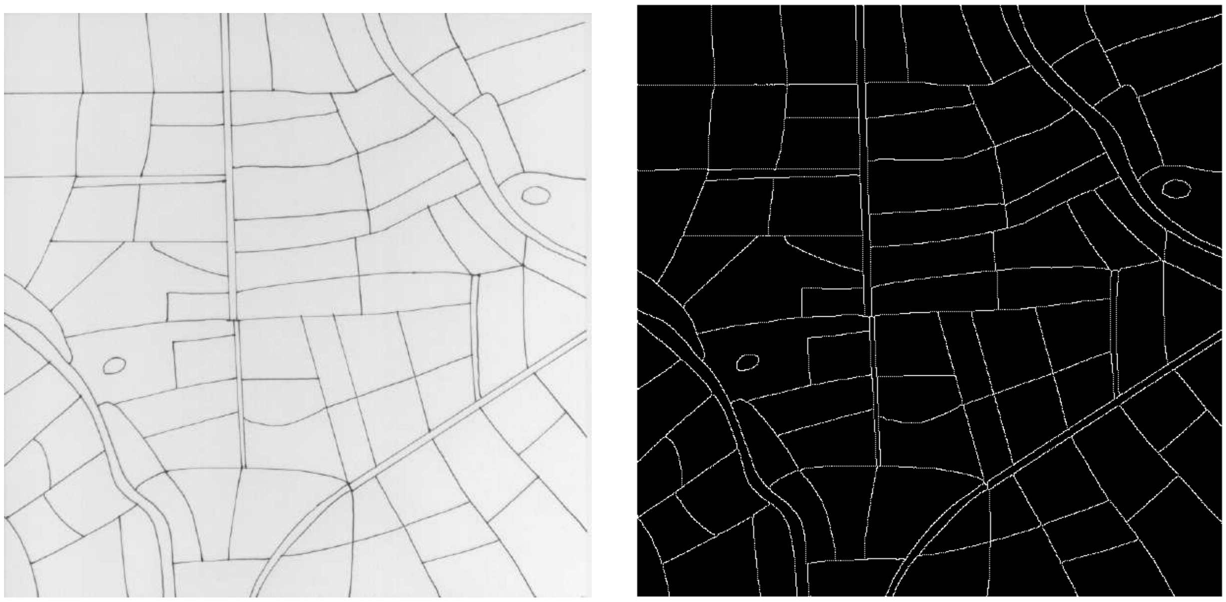
The former has the advantage of providing realistic intensity distributions of the fields, but it is difficult to obtain and the positional accuracy is limited. The latter provides a better positional accuracy but it is more difficult to obtain realistic intensity distributions. In section 2 a method is described to obtain an artificial generated satellite image with realistic intensity distributions using a ground truth data set and a real (Landsat) satellite image. In addition a "true" segmentation down to sub-pixel level is obtained.

In section 3 the segmentation methods used in this paper which are based on best merge region growing, are described.

The developed two quality measures are described in section 4. They quantify the two kinds of segmentation errors: the splitting of (the pure pixels of) a real field into more than one segment and the merging of pixels from different fields or of pure and mixed pixels into a segment. These quality measures can be used to compare different segmentation methods and also to determine the optimal parameter or threshold settings for a segmentation method.

In section 5 results of 4 segmentation methods on a generated image will be shown and evaluated.

2 ARTIFICIAL IMAGE GENERATION



(a) Scanned drawing

(b) Fields surrounded by closed polygons

Figure 1: Scanned drawing and Fields surrounded by closed polygons

Field boundaries were drawn by hand on white paper and this drawing was scanned with a laser scanner (HP Scanjet IIcx) at 100 dpi with 256 gray levels (see figure 1(a)). Thresholding was then used to obtain field (black) and field boundary (white) pixels and the image size was somewhat reduced to ensure that the field boundaries reach the outside of the image. The boundaries were thinned after filling in gaps of 1 pixel, to obtain 1 pixel wide closed polygons as field boundaries. The resulting image, see figure 1(b), has 824 lines of 816 elements and contains 90 fields.

class	*1*	*2*	*3*	*4*	*5*	*6*
sand	149	91	143	130	236	145
soil	92	45	63	64	120	64
water	66	27	28	20	20	9
wheat	91	46	69	92	121	55
barley	92	48	77	101	161	71
maize	80	37	43	125	110	43
rice	70	31	34	47	34	14
vine	104	54	75	107	161	84
grass	94	48	75	89	149	65
forest	76	34	44	81	98	41

Table 1: Used classes with mean Landsat intensities per channel

Blob coloring of the black pixels was used to label each pixel with its field number, providing an image which was used to relate each field in the drawing to its unique field number. Manually a text file was then produced relating each of these field numbers to a certain ground cover class. For example the 4 fields starting from the left top were given the following ground covers: vine, wheat, grass and soil (the small road running downwards to the middle of the image).

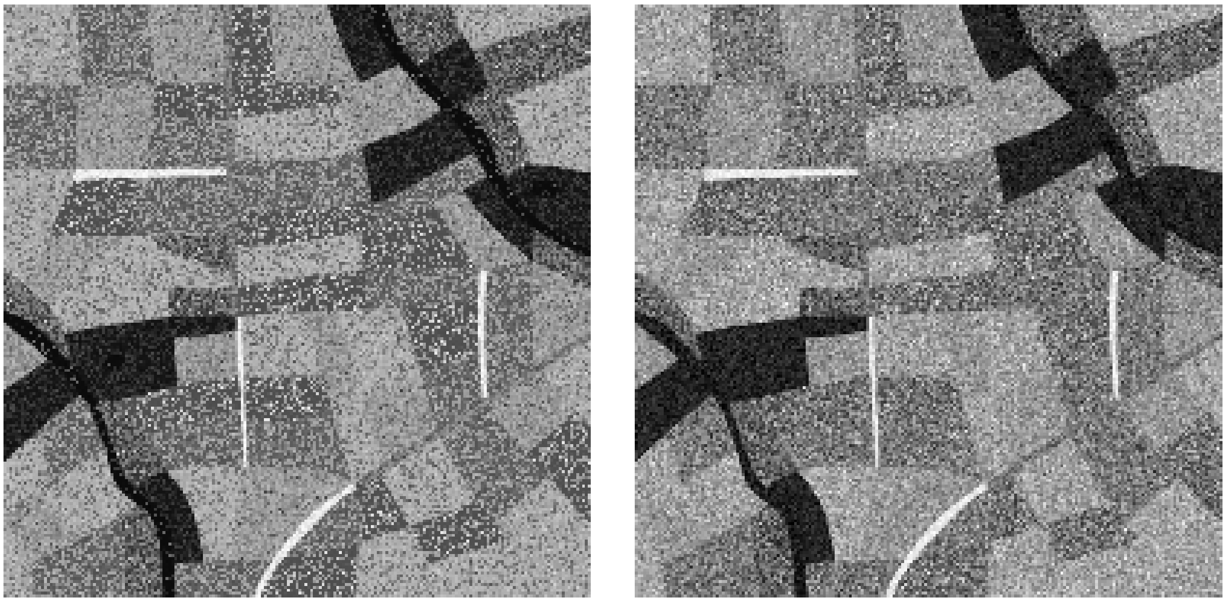
The image was then divided into blocks of n by n pixels (not taking the outer 2 pixels of the image, they were set to boundary pixels in the thinning process), a value of 4 for n was chosen for the results presented in this paper. Each block was used to create 1 pixel of the artificial satellite image, which thus may cover a number of ground cover classes. The white, field boundary pixels were taken for this paper to be of the soil ground cover class. The intensity values of each output pixel were taken as a linear mixture of the intensity values of the ground cover classes contained in it, weighted according the number of "sub-pixels" covered by each class.

For the intensities of ground cover classes a ground truth data set of an area in Portugal[7] is used. In table 1 the used classes with their mean intensities in a Landsat-TM image are shown.

First an artificial image was generated using for each output pixel a randomly selected pixel from the ground truth data set for each class covered. Then random values were used drawn according the mean values and covariance matrices for each class in the ground truth data set. The resulting images are shown in figures 2(a) and 2(b).

The variations in each field are much larger and more randomly than in the actual LANDSAT image (not shown) used for generating the ground truth data. A way to quantify this noise behavior is to determine for each pixel the difference between its intensity and the average intensity of its 8 neighbors. The mean and rms of the resulting distribution (all channels taken together) are given in table 2, showing that the rms values for the generated images are a factor 3 higher than for the real Landsat image.

In the real Landsat image the pixel to pixel variations in a field are usually smaller than the variations in average intensities between fields of the same ground cover. This is due to various reasons like different soil quality, moisture content, field inclination, crop variety or maturity. To generate a more realistic image the texture patterns of real fields must be used. If for each class c a realistic image of size $xd(c)$ by $yd(c)$ is available,



(a) Using random examples (b) Using random values

Figure 2: Channel 5 of generated Landsat-TM like images

	examples	*random*	*Landsat*	*texture*
mean	-0.009	-0.003	-0.006	0.002
rms	16.46	15.49	5.29	5.31

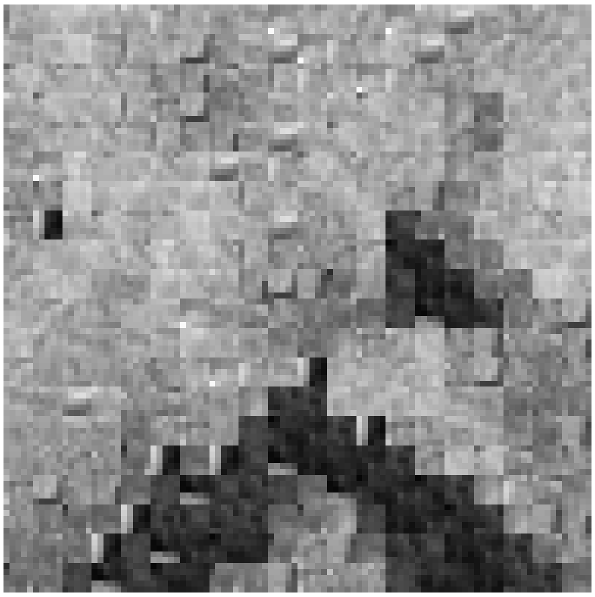
Table 2: Noise distribution in generated and real images

an output pixel can be generated as:

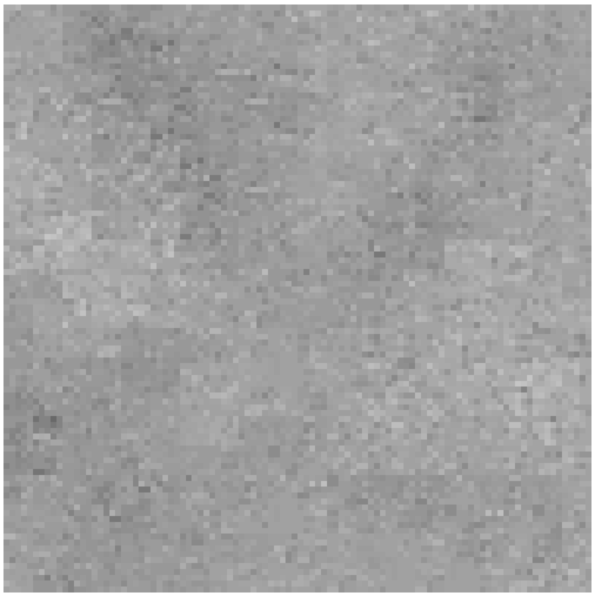
$$int_{x,y} = \sum_c fr_{x,y}(c) int_{x \ mod \ xd(c), y \ mod \ yd(c)}(c)$$

with $fr_{x,y}(c)$ the fraction of pixel x, y covered by class c and $int_{a,b}(c)$ the intensity of pixel a, b in image of class c. The sizes $xd(c)$ and $yd(c)$ (a value of 100 was used in this paper for all sizes) should be rather large to reduce the effects of discontinuities over the boundaries where the *mod* operation passes through 0.

Using the ground truth values, pixels in the real Landsat image were classified using a nearest neighbor method and requiring that the difference in intensity (summing the absolute difference in the 6 channels) was less than 18. It appeared that the image did not contain sufficient large fields, so smaller fields had to be joined. For each class blocks of 5 by 5 pixels were extracted from the Landsat image. The block closest in intensity to the mean intensity of all blocks was placed in the middle of the image to be generated. Then blocks were added choosing blocks closest in intensity to neighboring blocks, reusing already placed blocks when less than 20% remained to choose from. For the class "deciduous forest" channel 5 of the resulting image is shown in figure 3(a).



(a) Using all blocks



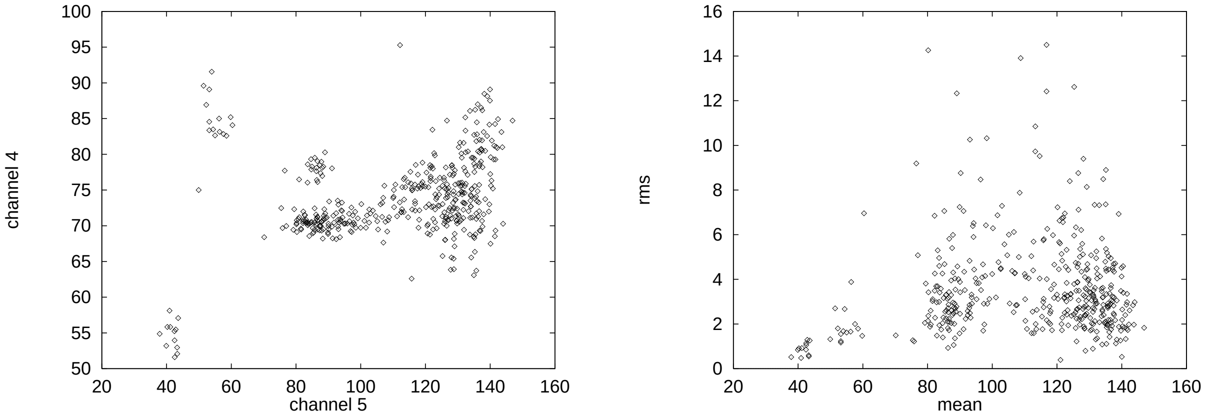
(b) Using selected blocks

Figure 3: Channel 5 of generated forest images, intensities 65-170 mapped to 0-255

Some of the blocks show a large internal variation and the differences between blocks can be large, creating artificial boundaries. This can also be judged from figures 4(a) and 4(b) were some statistical distributions of the blocks are shown.

To reduce these variations and differences the following selections on the blocks were made:

- replacing single pixels in a block deviating more than 1.5 or 2.0 times the rms from the mean value in any of the channels by a neighboring pixel.
- removing 20% to 40% of the blocks with the lowest signal to noise ratio, defined as: $\frac{1}{N_{ch}} \sum_{ch} \frac{mean_{ch}}{rms_{ch}}$, where ch runs over the number of channels.
- remove blocks not in the main peak of the intensity distribution in channel 5.
- scale all intensities of each remaining block by a factor to come as close as possible (in rms sense) to the mean intensity vector of all blocks.



(a) Mean intensity of blocks, channel 4 vs channel 5 (b) Channel 5, rms vs mean intensity

Figure 4: Some intensity distributions of forest blocks

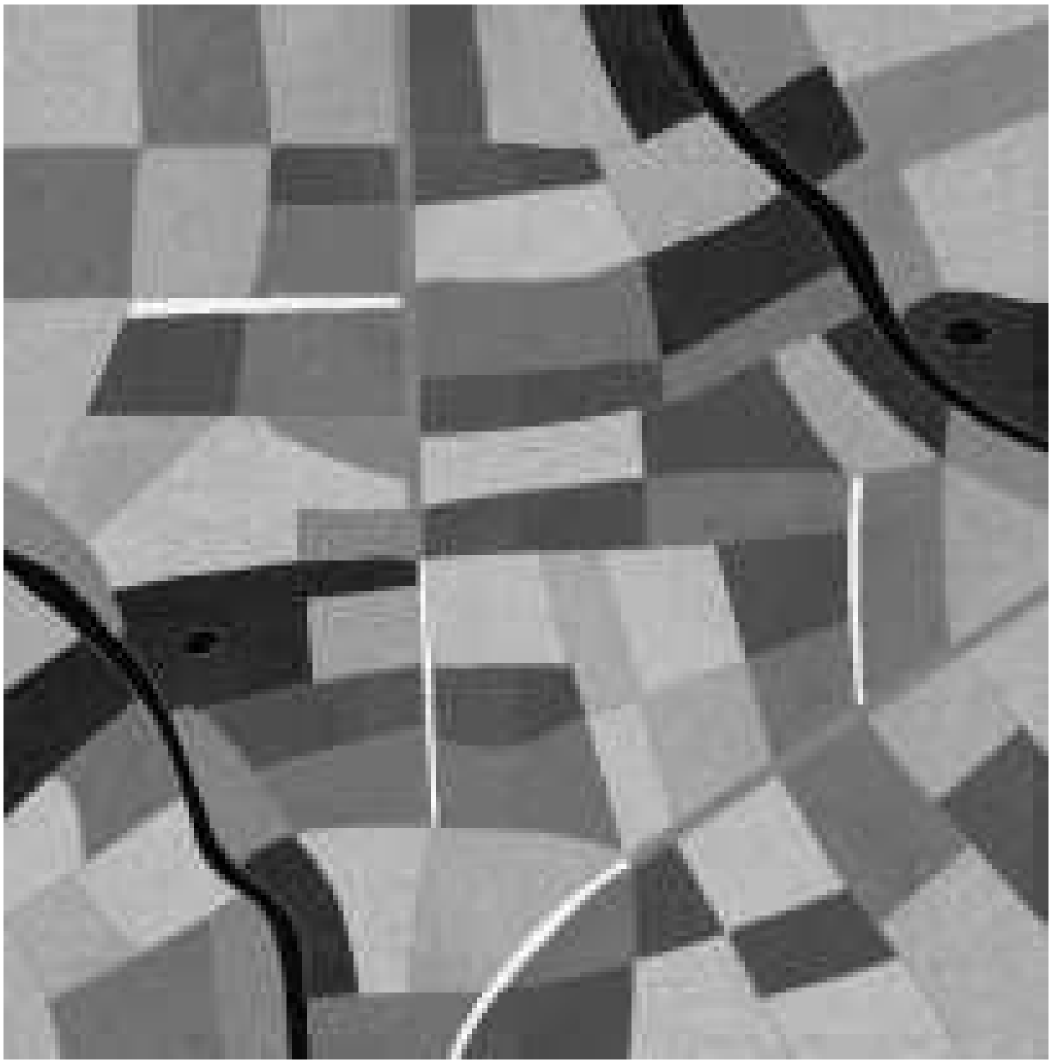
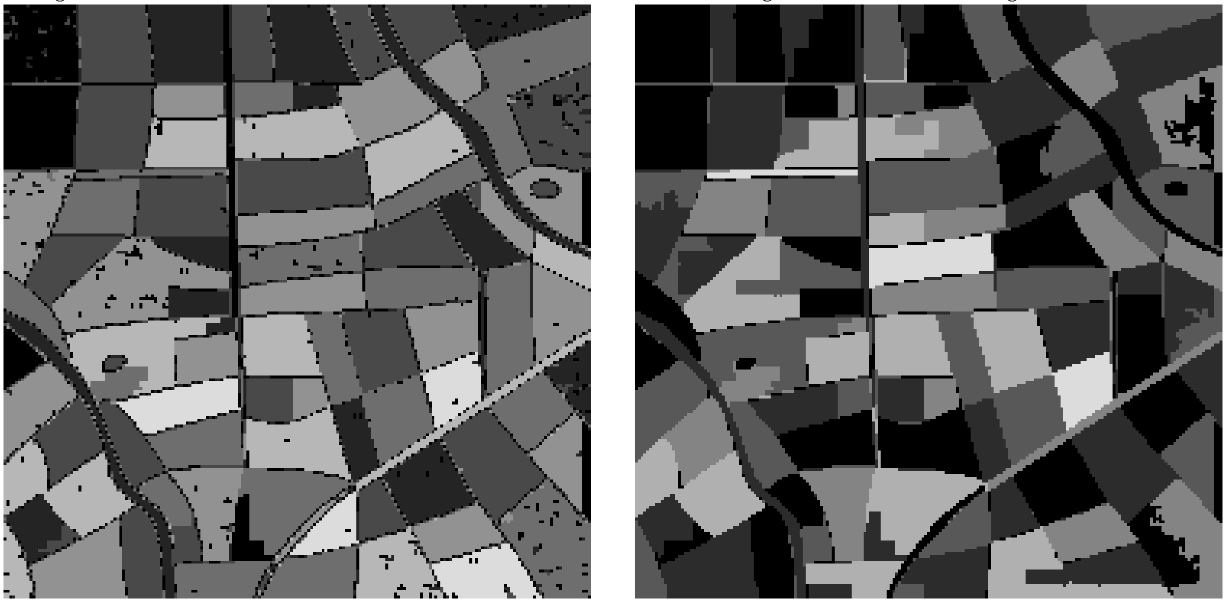


Figure 5: Channel 5 of generated Landsat-like image using texture images for each class

For the class "deciduous forest" channel 5 of the resulting image is shown in figure 3(b). No artificial boundaries are left and the texture pattern closely resembles the patterns seen in individual fields.

The generated Landsat-line image using texture images for each class, is shown in figure 5. As shown in table 2 its noise mean and rms are nearly the same as for the real Landsat image. The noise distribution of the generated image has a somewhat higher peak and lower tails than the one for the real image. Partly this can be due to the presence of more deviating mixed pixels in the real image caused by man-made structures or field separating objects which are broader or are of ground cover like water or sand which differ more in intensity from crops than soil cover as used in the artificial image. Partly this is also due to the removal of single deviating pixels and of blocks with a low signal to noise ratio. Various ways exists to take these issues into account when generating images like relaxing the selection criteria and making the boundaries of variable width and ground cover. Some discontinuities over row and column values of 100 and 200 are vaguely visible. This could be improved by joining for each class 4 properly mirrored versions of the texture image before calculating the output image.

Note that the variations between fields of a crop as seen in real images, is not present in the generated image. This has hardly any consequences for the region growing based segmentation methods as used in this paper. It could be important for segmentation methods which use (global) thresholds on intensity values in some way. Also here various ways exists to bring these field to field variations back into the generated images like scaling fields according to the distribution of the mean values of the blocks or using several texture images for each class.

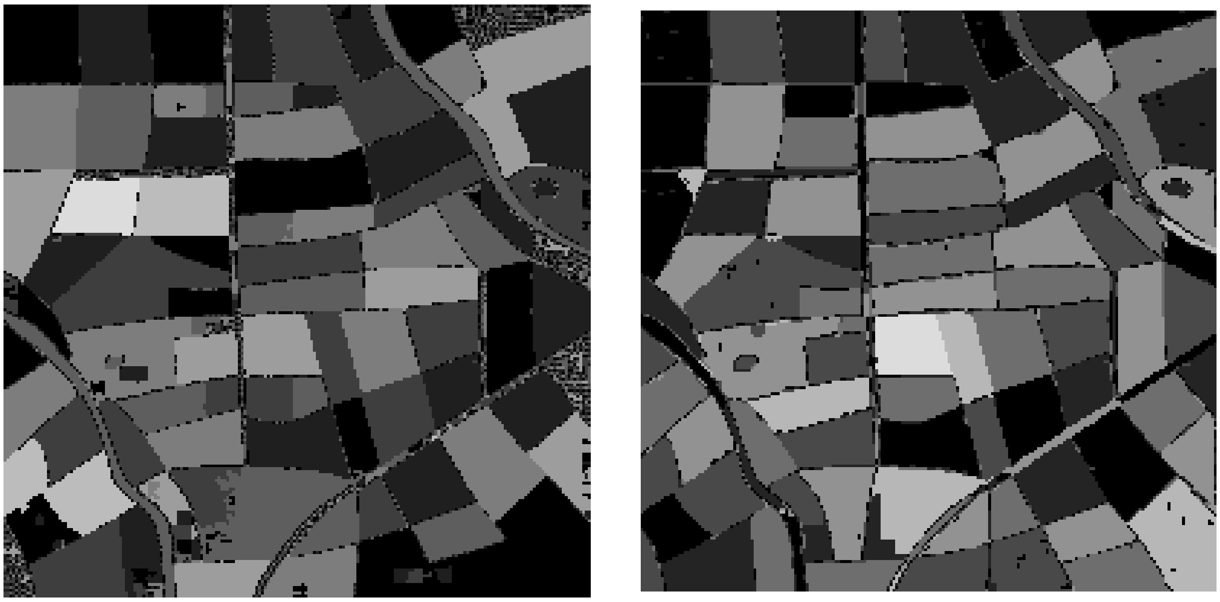


(a) "Schoen" method, threshold 5.0 (b) "B&G" method, threshold 32.5

Figure 6: Segmentation of image in figure 5

3 SEGMENTATION METHODS

The segmentation methods used in this paper are based on a best merge approach which is described in.[7] Initially each pixel is considered as a separate segment. As a first step neighboring segments having the same intensity value in all the channels are merged. Then all pairs of neighboring segments are considered and have a merging cost calculated. The two neighboring segments having the lowest merging cost are then merged into a new segment. When there is more than one group of segments with minimum cost value, they are merged in parallel. This merging of segments with the lowest merging cost is repeated until the lowest cost value is above a set threshold value. The following costs between segments i and j, each creating a different segmentation method,



(a) "SemiT" method, threshold 1.0 (b) "Pol" method, threshold 6.0

Figure 7: Segmentation of image in figure 5

are used:

Name	Cost function	Origin
Schoen	$\sum_{c=1}^{n}(v_c^i - v_c^j)^2$	as used in[7]
B&G	$\frac{N_i * N_j}{N_i + N_j} * \sum_{c=1}^{n}(v_c^i - v_c^j)^2$	as used in[1]
SemiT	$4 * \frac{N_i * N_j}{(N_i + N_j)^2} * \sum_{c=1}^{n}(v_c^i - v_c^j)^2$	derived from[9]

with n the number of channels, v_c^k the intensity value of channel c of segment k (the average value of all the pixels in segment k) and N_k the number of pixels in segment k.

In addition the method described in[7] is used, where first edge detection and linking is used to divide the image in regions with closed polygons as boundaries, then best merge using the "Schoen" cost is performed on each region and finally the edge pixels are handled by added them to neighboring segments or to other edge pixels or by keeping them as single pixel segments. This method will be named "Pol".

In figures 6(a), 6(b), 7(a) and 7(b) the resulting segments are shown for the 4 segmentation methods using thresholds as determined in section 5. So called "flag coloring" is used for these images, giving each segment a different grey value than all its neighbors. The given thresholds are squared and multiplied by the number of channels before comparing them to the cost values.

4 QUALITY MEASURES

The developed quality measures (see also[8] and[10]) try to quantify two types of segmentation errors:

- split error: a real field is split into one or more segments.

- merge error: pixels of different fields are merged into one segment.

For each field i of the true segmentation the "pure" output pixels, which are fully or nearly fully (one sub-pixel is of a different field) covered by that field, are considered. The number of those pure pixels is denoted by nin_i. Then the segments ij are determined which contain one or more (denoted by $ncov_{ij}$) pure pixels of field i. Two split quality measures are considered for field i. One determines the j (denoted by k) with largest $ncov_{ij}$ and calculates the area covered:

$$Qsplitcov_i = \frac{ncov_{ik}}{nin_i}$$

This measure is 1 if the field is fully contained in 1 segment and thus is not split into more than 1 segment. When the field is heavily split into a large number of segments, this measure approaches $\frac{1}{nin_i}$. Here only the size of the largest segment is important and not where it is located in the field or how the remaining pixels are segmented. Also the segmentation of the mixed pixels, which are partially covered by the field, are not considered.

The other measure considers all segments ij in an entropy like way:

$$Qsplitent_i = -\sum_j p_{ij} * log(p_{ij}) \ with \ p_{ij} = \frac{ncov_{ij}}{nin_i}$$

This entropy measure is 0 if the field is not split and large when the field is heavily split (approaching $log(nin_i)$).

Merge qualities are defined by reversing the role of "true" and "test" segmentations, considering for each segment the fields covering it with one or more pure pixels. All mixed pixels in the generated image are considered to belong to a single, additional field. Thus there is a requirement that the mixed pixels should not be joined with pure pixels of any field, even if that might be not obtainable because there is only a small or no difference in intensity.

Qualities for the whole image are calculated as the averages of the qualities per field or segment weighted by the number of pixels in the field (taking the pure output pixels) or segment. Other weighting schemes (no weighting, weighting with the square root of the number of pixels) were tried with slightly worse results. Split and merge qualities can be averaged to obtain a single segmentation quality.

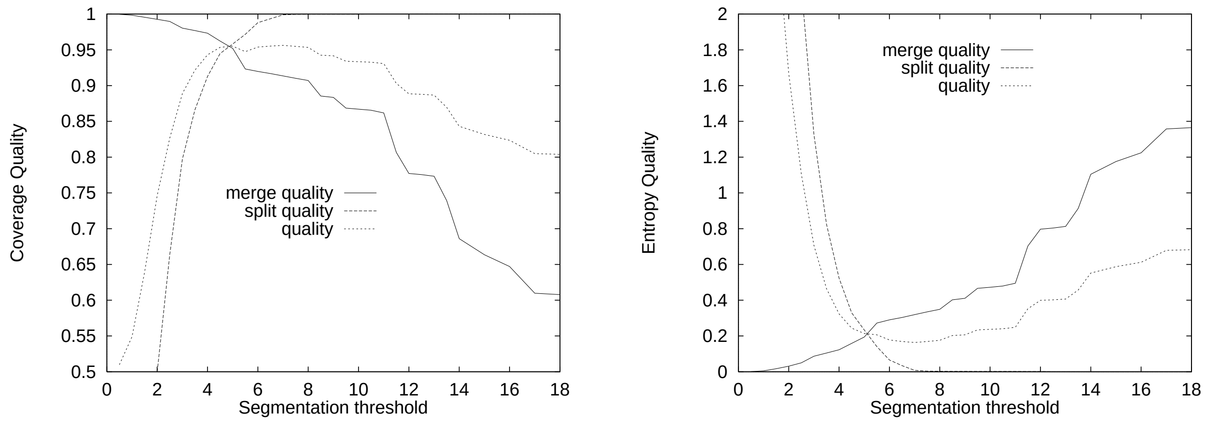
For a given segmentation method the qualities can be considered as function of the segmentation threshold or in general the parameters of the method. The best threshold is then defined as the threshold where merge and split quality are equal (the worse of the 2 qualities is there the best) or where the averaged quality is maximal (coverage quality) or minimal (entropy quality).

Different segmentation methods can be compared by looking at the qualities of the segmentations obtained at the best threshold (or parameter) settings.

5 RESULTS

In figures 8(a) to 11(b) the coverage based and entropy based quality measures are shown as function of the segmentation threshold for all segmentation methods. The split and merge quality values show a clear cross-over point, where they become equal. The overall qualities show a broad maximum and minimum, with the extrema thresholds usually higher than the cross-over thresholds. The cross-over and extrema thresholds and the corresponding quality values are shown in table 3.

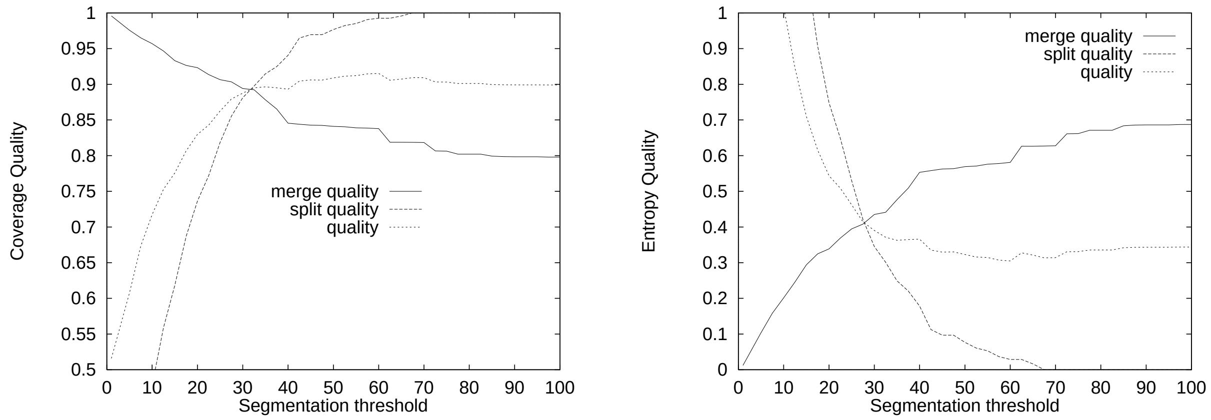
The best segmentations are clearly obtained by the "Schoen" method, followed by "Pol", "B&G" and "SemiT". The coverage and entropy based qualities give about the same threshold values. At the extrema thresholds a better



(a) Quality based on coverage by largest segment or field

(b) Quality based on entropy

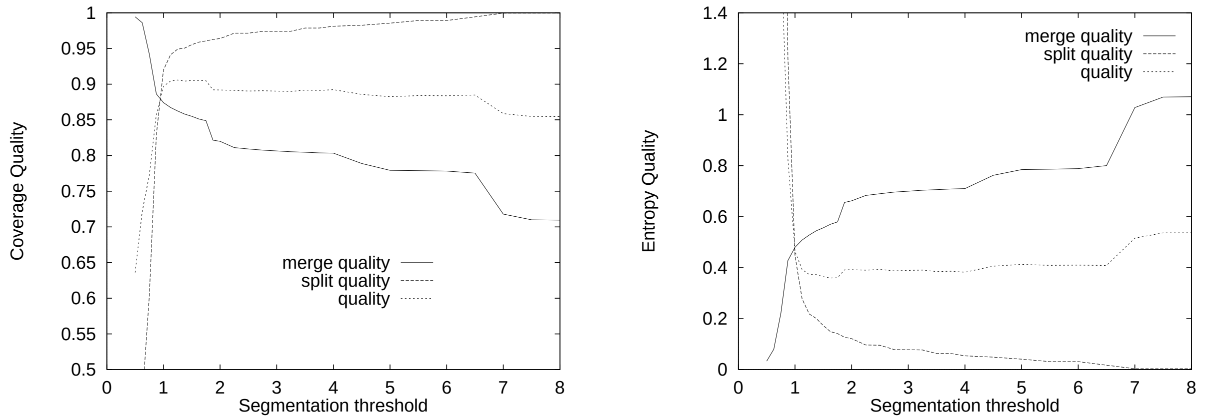
Figure 8: Quality measures as function of the segmentation threshold, "Schoen" method



(a) Quality based on coverage by largest segment or field

(b) Quality based on entropy

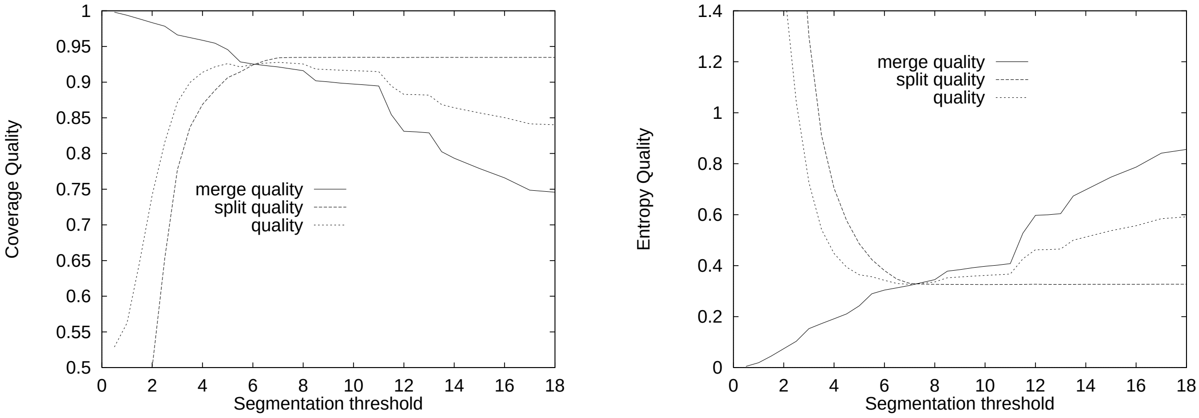
Figure 9: Quality measures as function of the segmentation threshold, "B&G" method



(a) Quality based on coverage by largest segment or field

(b) Quality based on entropy

Figure 10: Quality measures as function of the segmentation threshold, "SemiT" method



(a) Quality based on coverage by largest segment or field

(b) Quality based on entropy

Figure 11: Quality measures as function of the segmentation threshold, "Pol" method

Method	*Coverage*				*Entropy*			
	Cross-over		*Maximum*		*Cross-over*		*Minimum*	
	Thresh	*Qual*	*Thresh*	*Qual*	*Thresh*	*Qual*	*Thresh*	*Qual*
Schoen	5.0	0.954	7.0	0.956	5.0	0.214	7.0	0.163
B&G	32.5	0.894	60.0	0.915	27.5	0.410	60.0	0.305
SemiT	1.0	0.874	1.4	0.906	1.0	0.480	1.6	0.360
Pol	6.0	0.925	7.0	0.928	7.0	0.327	6.5	0.320

Table 3: Cross-over and extrema thresholds and quality values

split and a worse merge quality are obtained than at the cross-over thresholds, because the extrema thresholds are higher (the entropy based thresholds for the "Pol" method are an exception). There are less split fields at the extrema thresholds, but there are more real fields, especial neighboring fields with the same ground cover, which are merged completely. The overall qualities improve a little, but the additional joining of complete fields is probably undesirable for most applications using the segmentation results. Taking the sizes of the segments into account for the merge decision as is done for the "B&G" and "SemiT"makes the segmentation results worse. One of the reasons is that mixed pixels are too quickly merged with neighboring mixed and pure pixels, producing rather large segments on the boundary between two fields which lowers both merge and split quality.

The "Pol" method divides the image into parts using edge detection and linking into closed polygons and performing best merge segmentation on each part separately using the "Schoen" cost function. Its merge quality stays better at higher thresholds compared to the "Schoen" method, because segments can not be merged over the polygon boundaries. But the polygon detection splits some fields, especially the thin water and road fields, which causes the split quality not to reach 1 or 0 at high thresholds. Also the polygons are not exactly at the right positions, splitting off small parts of fields which are then more easy to add to pixels of neighboring fields thus causing worse split and also merge qualities at lower thresholds. The polygon detection and the post-processing of the polygon pixels could be improved and post-processing of segments near polygon boundaries could be added to improve the segmentation results. The "Pol" method, although producing a bit worse results than the "Schoen" method, has the advantage that it is faster (a factor 4 for the artificial image, a larger factor for larger images) and that it can be parallelized (see[2] for an implementation on workstations using PVM).

To improve the segmentation results the merge decision has to be based also on the geometry of the fields and on the intensity changes over the boundary between two fields.

6 CONCLUSION

A method has been developed for generating artificial satellite images using ground truth data and a real satellite image. The intensity distributions in each field of a certain ground cover closely resemble the distributions found in real images. Mixed pixels, covering area's from neighboring fields and from cover separating these fields (assumed to be of "soil" type), are modelled.

Two types of segmentation quality measures have been presented, each of them describes both split and merge errors in the segmentation. They were used to evaluate the results on an artificial image of 4 segmentation methods, all based on the best merge principle. Based on the quality measures the optimal threshold setting for each segmentation method was determined and the best segmentation method was selected.

The above methods can thus be used to select the best segmentation method for a given kind of satellite image, provided a ground truth data set is available. Also the optimal parameters or threshold settings for a segmentation method can be determined.

The artificial image generating method can be adapted to other kind of images, e.g. for medical or production control applications. The quality measures can be used whenever it is known for each pixel either to which object (the background in some images can be taken as 1 object) it belongs or that it is a mixed pixel.

7 REFERENCES

[1] J-M. Beaulieu and M. Goldberg. Hierarchy in Picture Segmentation: A Stepwise Optimization Approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 11(2):150-163, 1989

[2] Paul Janssen. Parallel Region Growing and Parallel programming Environments for Workstation Clusters. *Master thesis no. 346* Department of Informatics, Katholieke Universiteit Nijmegen, June 1995

[3] M.D. Levine and A.M. Nazif. Dynamic measurement of computer generated image segmentations. *IEEE PAMI* 7:155-164, 1985

[4] J. Liu and Y-H Yang. Multiresolution Color Image Segmentation. *IEEE PAMI* 16:689-700, 1994

[5] R.P.H.M Schoenmakers. Segmentation of Remotely Sensed Imagery. *PhD thesis* Katholieke Universiteit Nijmegen, 1995

[6] R.P.H.M Schoenmakers, G.G. Wilkinson and Th. E. Schouten. Segmentation of Remotely-Sensed Imagery: A Re-Definition for Operational Applications. *IGARSS91* pages 1087-1090, Espoo, Finland, 1991

[7] Ronald P.H.M Schoenmakers, Graeme G. Wilkinson, Theo E. Schouten. Results of a Hybrid Segmentation Method. *Image and Signal Processing for Remote Sensing, Proc SPIE 2315*, pp 113-126 Rome 1994

[8] Theo E. Schouten, Maurice S. klein Gebbinck, Ronald P.H.M Schoenmakers, Graeme G. Wilkinson. Finding thresholds for Image Segmentation *Image and Signal Processing for Remote Sensing, Proc SPIE 2315*, pp 90-101 Rome 1994

[9] J.C. Tilton. Image Segmentation by Iterative Parallel Region Growing and Splitting. *Proceedings of the International Geoscience and Remote Sensing Symposium (IGARSS89)*, pages 2235-2238, Vancouver, 1989

[10] H.F.J.M. Voogd. Error measurement for segmentation techniques. *Master thesis no. 340* Department of Informatics, Katholieke Universiteit Nijmegen, March 1995

[11] Y.J. Zhang, J.J. Gerbrands. Objective and quantitative segmentation evaluation and comparison. *Signal processing* 39(1994)43-54