FURTHER INVESTIGATIONS ON SEGMENTATION QUALITY ASSESSMENT FOR REMOTE SENSING APPLICATIONS

Ulrike STURM, Uwe WEIDNER

Institute of Photogrammetry and Remote Sensing, University Karlsruhe (TH), Kaiserstraße 12, 76128 Karlsruhe, Germany
(ulrike.sturm, uwe.weidner)@kit.edu

KEY WORDS: segmentation, accuracy assessment, discrepancy method, segment boundary, boundary delineation, VHR, segment-based

ABSTRACT:

Object-oriented or segment-based classification approaches for remote sensing applications integrate not only the spectral signature but also shape and topological characteristics of segments. To use their shape related features, segments have to be as close as possible to the desirable object shape. The segment boundary's quality is decisive for the classification output - in order to delineate two classes and to avoid spectrally mixed segments. Despite its importance, no standard has yet been accomplished to address the segmentation's quality evaluation. This contribution presents further investigations on this topic based on boundary accuracy by using a distance dependent *weighted quality rate* and integrating other quantities yielding a *combined quality measure*. Furthermore it analyses the dependency on reference objects' quality. The reviewed quality rates are not only used for evaluating the segmentation, but are also used for the decision which segmentation level within a multi-scale segmentation should be used for classification of a certain class.

1 INTRODUCTION

Within segment-based classification approaches the segmentation step is decisive, because the resulting segments form the basis for the subsequent classification, which is based on spectral, form, topological and semantic features. Despite known investigations and approaches of quality evaluation for segmentations (see e.g. (Zhang, 2001) and (Neubert et al., 2008) for review), the question of how to access this quality with respect to remote sensing applications is not yet completely answered and no standard evaluation method has been established, which can quantitatively and thereby objectivly confirm visual assessment. Segmentation quality assessment is not only of interest at the end of the segmentation step or even of the classification, but also within the process of segmentation in order to optimise the parameters of the algorithms and to choose the appropriate segmentation level within multi-scale segmentation. Therefore, such an assessment may contribute to and improve optimisation approaches based solely on the segmentation stability within the parameter space.

In our opinion the geometry and namely the delineation of segments has a higher impact on the quality than other aspects: oversegmentation is partly acceptable, but problems may occur if geometric properties of the segments are used in the classification step. Undersegmentation definitely leads to misclassification of segments, because of the resulting mixed-segment problem due to the distortion of segment-inherent properties (Weidner and Bähr, 2007). Therefore, segmentation should provide segments that a) match the form of the objects to be classified as well as possible and b) match the form preferably as one segment if form parameters are included within the classification step. The paper discusses quantities which are supposed to check these requirements and their combination for evaluation. Emphasis lies on the improvement of our approach based on the *weighted quality rate* as a distance dependent form measure with respect to the evaluation of under- and oversegmentation (Weidner, 2008). It focuses on the positional accuracy of the segment boundaries compared to reference data and therefore belongs to empirical discrepancy indices according to (Zhang, 1996). Such reference data is obtained by manual digitising and thus underlies the influence of the clearness or ambiguity of an object class. For this reason there is a certain subjective (human) bias (Neubert et al., 2008) due to the interpretation and digitising accuracy of the operator. To investigate its influence on the evaluation results, this paper not only discusses quantities and their combination for evaluation, but also explores the influence of reference data on the quality measures. Due to different operations, there will always be at least a slight difference between automatically obtained segments and a manually obtained reference object. To overcome these differences, a buffer around the references was introduced within the *weighted quality rate*.

As an example for the evaluation, we used QuickBird data of a structurally complex rural community in Benin, West Africa, and the software package Definiens Developer. Within multi-scale segmentation, implemented in this software, segmentation levels based on the region growing approach (Baatz and Schäpe, 1999) can be combined. For every level, parameters as scale parameter, shape vs. color parameter and compactness vs. smoothness parameter have to be set. Therefore, the assessment can be used a) during parameter findings within one level and b) for choosing the best fitting segmentation level for a certain class.

2 RELATED WORK

Frameworks for quality assessments have been proposed and published in the computer vision community, e.g. (Hoover et al., 1996) for range images, (Zhang, 2001) for optical images, and (Udupa et al., 2006) for voxel data sets. (Neubert et al., 2006) and (Neubert et al., 2008) address the topic of segmentation quality for remote sensing in which context only a few investigations are published. Their quality evaluations for different segmentation approaches are based on a qualitative visual and a quantitative evaluation based on geometrical features of the segments, e.g. area, perimeter, and shape, using manually derived ground truth. Partly, the results of used quantities (e.g. average distances to ground truth) are difficult to interpret. Furthermore, some of them are correlated and the reliability of some are also dependent on the segment size, e.g. the *shape index* which is normally more reliable for larger segments. Evaluation quantities should also account for the uncertainty of the segment boundaries as e.g. proposed in (Schuster and Weidner, 2003) and for ease of interpretation should have properties of metrics as e.g. given in (Unnikrishnan et al., 2007) or be normed to a fixed range of values as e.g. in (Correira and Pereira, 2003). (Radoux and Defourny, 2008) propose an average of absolute errors on boundary position. (Neubert et al., 2008) address the need of spatially explicit outline delineation quality measures. To take a certain spatial uncertainty between reference and segmentation objects based on the data into account, the introduction of a buffer around the used reference was proposed in (Weidner, 2008). The introduction of a buffer can also be examined for landuse change analysis using different data sets (eg. in (Schöpfer and Lang, 2006)).

In (Weidner, 2008) different evaluation approaches were discussed in more detail following the categorisation used by (Zhang, 1996). Based on this discussion, a *weighed quality rate* ρ^*_{qw} was proposed, allowing to take the uncertainty of segment boundaries into account. In this contribution we extend our evaluation approach incorporating criteria like the *connectivity rate* ρ_{cc} (Cardenes et al., 2007) and ρ^*_d, which quantifies the rate of references matched by segments, into one evaluation framework. In the following section the essential formulas are compiled, followed by the section introducing the methodology. Consequently, the results and discussions are presented. A conclusion completes the paper.

3 QUANTITATIVE QUALITY ASSESSMENT

Discrepancy methods evaluate the difference between objects. In the case of segmentation evaluation that means the difference between reference objects and segments assigned to them. In order to discuss evaluation quantities, let $\mathcal{R}_{i[k]}$ denote a reference segment of class k, let $\mathcal{S}_j$ denote segments found with a segmentation and let $\mathcal{S}_{i[k]}$ denote a set of segments assigned to a reference segment $\mathcal{R}_{i[k]}$ by the *criteria of assignment*

$$\mathcal{S}_{i[k]} = \bigcup_{\mathcal{S}_j \in \mathcal{J}_k} \mathcal{S}_j \qquad (1)$$

with

$$\mathcal{J}_k = \left\{ \mathcal{S}_j \quad \text{with} \quad \frac{|\mathcal{S}_j \cap \mathcal{R}_{i[k]}|}{|\mathcal{S}_j|} > 0.5 \right\}$$

where $|\mathcal{A}|$ denotes the number of pixels of $\mathcal{A}$ or its area respectively using their overlap with the reference segment as criterion. Furthermore let $\sharp\mathcal{A}$ denote the number of segments of $\mathcal{A}$. For the ease of reading, subscipts will be omitted furtheron. Within this study, a certain number of reference objects is used for evaluation per class. In order to quantify the rate of references matched by segments fulfilling constraint (1), the *detection rate*

$$\rho^*_d = \frac{\sharp(\mathcal{S} \cap \mathcal{R})}{\sharp\mathcal{R}} \quad \text{with} \quad \rho^*_d \in [0,1] \qquad (2)$$

is considered.

One commonly used quality measure is the *quality rate*

$$\begin{aligned}\rho_q &= \frac{|\mathcal{S} \cap \mathcal{R}|}{|\mathcal{S} \cup \mathcal{R}|} = 1 - \frac{|(\mathcal{S} \setminus \mathcal{R}) \cup (\mathcal{R} \setminus \mathcal{S})|}{|\mathcal{S} \cup \mathcal{R}|} \\ &= \frac{|\mathcal{S} \cap \mathcal{R}|}{|\mathcal{S} \cap \mathcal{R}| + |(\mathcal{S} \setminus \mathcal{R})| + |(\mathcal{R} \setminus \mathcal{S})|} \qquad (3)\end{aligned}$$

with $\rho_q \in [0, 1]$. The advantages of the *quality rate* compared to other evaluation criteria like *false positive* ($\mathcal{S} \setminus \mathcal{R}$) or *false negative* ($\mathcal{R} \setminus \mathcal{S}$) is the symmetry with respect to $\mathcal{R}$ and $\mathcal{S}$ and its fixed range. The term $\delta_s = 1 - \rho_q$ has been used as similarity measure in computer vision (cf. e.g. (Keim, 1999)). Without loss of generality in the context of remote sensing applications the same class labels for the segmentation and the reference data can be assumed and therefore the *quality rate* fulfils the requirements for quantities for segmentation quality assessment defined by (Unnikrishnan et al., 2007). As example for this, $\mathcal{R} \neq \emptyset$ is assumed and three degenerated cases are considered: (a) $\mathcal{S} = \emptyset$, (b) $\mathcal{S} = \mathcal{I} \setminus \mathcal{R} = \mathcal{R}^c$ - thus being the complement of $\mathcal{R}$, and (c) $\mathcal{S} = \mathcal{I}$. For cases (a) and (b), $\rho_q = 0$, showing that the quantity is meaningful also for these degenerated cases. In case (c) $\rho_q = \frac{|\mathcal{R}|}{|\mathcal{I}|}$ and therefore directly depends on the area of $\mathcal{R}$. Thus, if $\mathcal{R} \to \emptyset$ then $\rho_q \to 0$ and if $\mathcal{R} \to \mathcal{I}$ then $\rho_q \to 1$ respectively.

Quantities like the *quality rate* ρ_q are based on binary consideration of the deviations between the two sets $\mathcal{S}$ and $\mathcal{R}$. In order to increase the influence of larger deviations between the two sets on the quality measure ρ_q, a *weighted quality rate* ρ_{qw} was introduced in (Schuster and Weidner, 2003) and a *refined weighted quality rate* ρ^*_{qw} was presented in (Weidner, 2008):

$$\rho^*_{qw} = 1 - \frac{A^*}{|\mathcal{S} \cap \mathcal{R}| + A^*} = \frac{|\mathcal{S} \cap \mathcal{R}|}{|\mathcal{S} \cap \mathcal{R}| + A^*} \quad \text{with} \quad \rho^*_{qw} \in [0,1] \qquad (4)$$

where

$$A^* = \sum_{x \in (\mathcal{S} \setminus \mathcal{R})} w(d(x, \mathcal{R})) + \sum_{x \in (\mathcal{R} \setminus \mathcal{S})} w(d(x, \mathcal{R}^c))$$

and

$$d(x, \mathcal{A}) = \inf\{\rho(x, a) : a \in \mathcal{A}\} \qquad (5)$$

$d(x, \mathcal{A})$ denotes the distance of a pixel from the reference boundary and $w(x)$ a weighting function. By defining a weight, pixels of a considered segment, which are situated further away from the reference boundary, are penalized, resulting in a lower ρ^*_{qw}. Several weighting functions and a discussion can be found in (Schuster and Weidner, 2003) or also in (Cardenes et al., 2007). Within this study linear functions like

$$w(d(x, \mathcal{A})) = \frac{1}{\Delta_d} d(x, \mathcal{A}) \qquad (6)$$

or

$$w_T(d(x, \mathcal{A})) = \begin{cases} 0 & d \leq d_T \\ \frac{1}{\Delta_d}(d(x, \mathcal{A}) - d_T) & \text{else} \end{cases} \qquad (7)$$

are used, where Δ_d denotes the *ground sampling distance* (GSD) of a pixel. The function given by (7) allows to introduce a buffer (d_T) around the reference boundary. Therefore the accuracy of the boundaries can be taken into account. As long as the segment's boundary lies within the buffer d_T, $A^* = 0$ and therefore $\rho^*_{qw} = 1$. Analysing the three cases mentioned above yields $\rho^*_{qw} = 0$ for cases (a) and (b) and for case (c) where $\mathcal{S} = \mathcal{I}$ is assumed

$$\rho^*_{qw} = \frac{|\mathcal{R}|}{|\mathcal{R}| + \sum_{x \in (\mathcal{S} \setminus \mathcal{R})} w(d(x, \mathcal{R}))}$$

Again, if $\mathcal{R} \to \emptyset$ then $\rho^*_{qw} \to 0$ and if $\mathcal{R} \to \mathcal{I}$ then $\rho^*_{qw} \to 1$ respectively. All discussed quantities can also be used for the evaluation of voxel data segmentations (c.f. (Udupa et al., 2006) for medical image processing applications).

The quantities discussed above depend on the number of segments of $\mathcal{R}$ and $\mathcal{S}$. In order to evaluate the difference, (Cardenes et al., 2007) use the *connectivity coefficient*

$$\rho_{cc} = \frac{2\min(\sharp\mathcal{S}, \sharp\mathcal{R})}{\sharp\mathcal{S} + \sharp\mathcal{R}} \qquad (8)$$

Furthermore they combine different evaluation criteria ρ_i with ranges of values from 0 to 1 using the *quadratic mean*

$$\rho_g = \sqrt{\frac{1}{n}\sum_{i}^{n}\rho_i^2} \tag{9}$$

We will also use the *connectivity coefficient* ρ_{cc}, but criteria with $[0,1]$ will be combined using the *geometric mean*

$$\rho_g^* = \left(\prod_{i}^{n}\rho_i\right)^{\frac{1}{n}} \tag{10}$$

We propose the *geometric mean*, because it combines two factors, important in our opinion: firstly, if one quantity is zero, then the *combined quality measure* is zero; secondly, the n-th root is used in order to be able to compare *combined quality measures* independently on the number of quantities ρ_i used.

4 METHODOLOY

For the presented investigation a subset of a pan-sharpened QuickBird scene with a GSD of 0.60 m from Avlekete, a structurally complex rural community in Benin, West Africa, was used (Fig. 1). The Beninese data differs significantly from German data that was used for prior tests (cf. Weidner, 2008). Other landcover and landuse classes are existent; houses are smaller and appear different due to other roof material and other surrounding terrain. For analysis on settlement processes in the coastal area of Benin, the classification of houses and other buildings is of great interest. Therefore, we focused on houses for this study. Houses show besides their spectral signature a significant shape, which should be obtained by segmentation. Due to different roof materials it is necessary to determine different house classes. Rusty metal roofs appear dark brown in band combination RGB, new metal roofs and cement asbestos roofs appear bright (see Fig. 1). Therefore one class was assigned to houses with dark rusty metal roofs and one to houses with brighter roofs. For ease of description, the first category will be called *dark houses* in the following, the second *bright houses*.

For the presented data set a segmentation was derived with scale parameters (SP) 10, 15, 20, 30, 40 and 60. For the first two segmentation levels the weight for *shape* was set to 0.3, for SP 20 and SP 30 to 0.5, for SP 40 to 0.3 and for SP 60 to 0.2. The parameter *compactness* was set to 0.5 for the first two segmentation levels and to 0.8 for the following. For SP 60 it was set to 0.5.

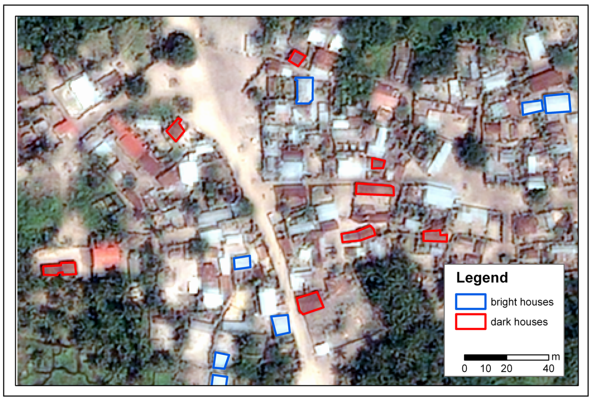

Figure 1: QuickBird data (RGB) of investigation area with reference data

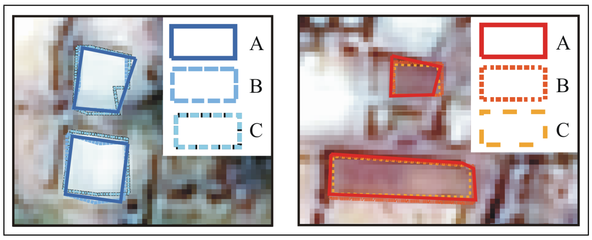

Figure 2: Examples of digitising results *A*, *B* and *C*; left: bright houses, right: dark houses

For ease of reading we will refer in the following to the different segmentation levels (*SL*) simply by their scale parameter.

For each house class, reference polygons were manually digitised based on the satellite data. To investigate the influence of the interpretation of the digitising person, two operators digitised the same houses. House boundaries might be interpreted differently depending on the visual seperability of classes. In total, three reference data sets are used: *A* and *B* are from different operators, not knowing each others digitising strategy. *C* contains the reference of the second operator knowing the other operator's digitising strategy after receiving independent samples. Fig. 2 shows two examples of the digitising results of the two operators and the reference boundary differences for *bright* and for *dark houses*.

In order to classify houses, segments should fulfill following requirements: a) all houses (reference objects) should be detected (no undersegmentation); b) the segments should be within the house boundary and should not reach too far in or out (boundary precision); c) houses should be represented preferably as one segment in order to provide the possibility of taking their shape into account for the classification. Therefore, no oversegmentation should occur. Requirement a) will be assessed by ρ_d^* (2); b) can be evaluated by ρ_{qw}^* (4); c) is assessable by ρ_{cc} (8). ρ_{qw}^* and ρ_{cc} were calculated for all matched references. Depending on the segmentation level, some reference objects might not be obtained due to undersegmentation and thereby do not fulfil the *criteria of assignment* (1). Therefore, ρ_d^*, which describes the ratio of the detected references, has to be taken into account. The three considered quantities are evaluated one by one but also combined according to the *combined quality measure* according to (10) is

$$\rho_g^* = \left(\rho_{qw}^*\rho_d^*\rho_{cc}\right)^{\frac{1}{3}} \tag{11}$$

All presented quantities can be used for the evaluation of a single object, a class or all classes together (cf. Weidner, 2008). The visual evaluation indicates, that for *bright* and *dark houses* the segmentation quality varied and different segmentation levels seem appropriate. Therefore, each class is evaluated seperately.

In addition to (Weidner, 2008) a buffer was introduced to ρ_{qw}^* for two reasons: firstly, the segment boundaries have the image rastering of 0.6 m whereas the references were smoothly digitised as vectors (see Fig. 2). Therefore, reference and segment boundary are not likely to be totally congruent. Additionally, the precision of measurement of the operator shows a second variation. Therefore, a buffer of 0.6 m respectivly 1.2 m, relating to one respectivly two pixels, was introduced and the values were compared with the original ρ_{qw}^* and ρ_g^*. The evaluation presented in this paper was conducted raster-based, but all quantities can be just as well calculated vector-based e.g. in a GIS.

5 RESULTS AND DISCUSSION

A visual comparison of the segmentations shows differences between *bright* and *dark houses* concerning an appropriate segmentation level (*SL*) and its respective segmentation quality. For *bright houses* one would choose *SL40*: all houses exist (no undersegmentation) and most of them are found as one segment. Some houses show branches of their segments, which appear yet still acceptable. In contrast to *bright houses*, *dark houses* show at *SL40* already clear undersegmentation. Therefore, a lower *SL* should be chosen. One might choose *SL15* and *SL20*, making a further distinction between house sizes: some houses appear as one segment in *SL15*, others in *SL20*; some not yet, but in *SL30* undersegmentation and strong branching of segment parts starts clearly. In *SL15* and *20* branching out of segments belonging to *dark houses* is stronger than the one for *bright houses*.

In order to quantify or control visual decision, ρ_d^*, ρ_{cc} and ρ_{qw}^* are used. Since ρ_d^*, ρ_{cc} and ρ_{qw}^* depend on reference objects, *A*, *B* and *C* were used for both house classes to check the subjective bias of the operator that digitised the reference objects. Due to the references' differences in shape and size, the set of segments fulfilling the criteria of assignment (1) varied for *A*, *B* and *C* (see Tab. 1) beeing the cause of differing ρ_d^* and ρ_{cc}. Since all values are already low for dark houses at *SL40*, *SL60* will not be considered in the following. Fig. 3 and Fig. 4 show examples for the different assignment of segments to reference objects. Fig. 3 shows examples for the assignment of segments to references of *bright houses*: for *SL10* and *15*, *B* causes the assignment of one additional segment (marked red). The other segments are assigned to either $\mathcal{R}$, also for higher *SL* (segments in yellow). Fig. 4 presents examples for the assignment of segments to refernces of *dark houses*. In this case, up to *SL30* the same references were assigned to either $\mathcal{R}$. In *SL40* no segments were assigned to *A* and *C*, yet one to *B*. The results of ρ_d^*, ρ_q, ρ_{qw}^*, $\rho_{qw(1)}^*$ ($d_T = 0.6\ m$), $\rho_{qw(2)}^*$ ($d_T = 1.2\ m$) and ρ_{cc} are presented in Tab. 2 and Tab. 3, while ρ_g^* is presented as curve in Fig. 5 and Fig. 6.

For *bright houses*, ρ_d^* has the same results for *A*, *B* and *C*. It shows that all references of *bright houses* are matched up to *SL40*, independent on the operator. The values of ρ_{cc} rise with increasing *SL* due to the reduction of the number of segments assigned to one reference object. Since the assigned number of segments differs up to *SL20* for *A*, *B* and *C*, ρ_{cc} shows different values. Furtheron it is for all $\mathcal{R}$ the same. For *A*, *B* and *C*, ρ_{cc} is relativly high with 0.88 for *SL30* (Tab. 2), but higher for *SL40* with 0.93. For *SL60* ρ_{cc} is 1, which means that all matched reference objects are represented as one segment. But ρ_d^* indicates, that some houses are already undersegmented. Therefore, *SL60* is not appropriate. ρ_{qw}^* differs slightly for all three examined $\mathcal{R}$ in its absolute values (see Tab. 2). While ρ_q shows differences of at most 0.1 for the different *SL* (Tab. 2) independent of the references, ρ_{qw}^* differs clearly for changing *SL*. For both quantities the values decline for growing *SL*. Since ρ_{qw}^* is a distance dependent measure, the

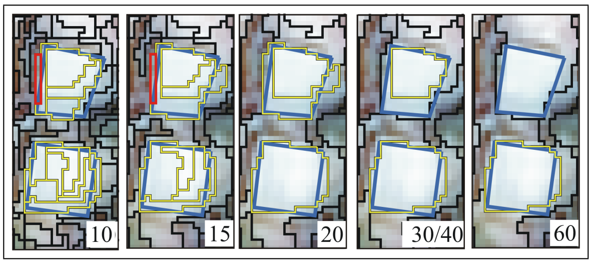

Figure 3: Segments fulfilling (1) for bright houses (red segments: only assigned to *B*, yellow segments: assigned to *A*, *B* and *C*)

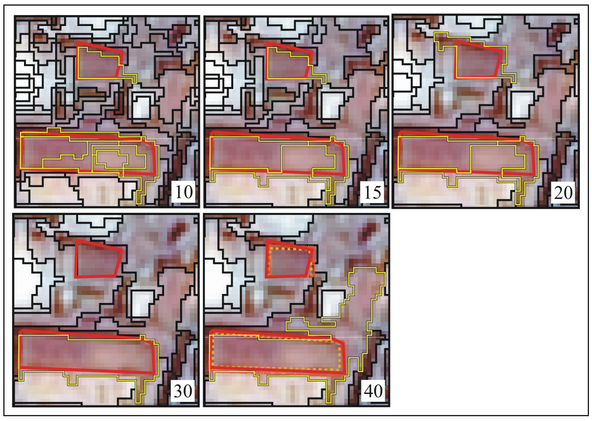

Figure 4: Segments fulfilling (1) for dark houses (red segments: only assigned to *B*, yellow segments: assigned to *A*, *B* and *C*)

bright houses			dark houses		
SL	$\sharp\mathcal{R}$	$\sharp\mathcal{S}$	SL	$\sharp\mathcal{R}$	$\sharp\mathcal{S}$
A10	7	45	A10	8	34
A15	7	26	A15	8	18
A20	7	13	A20	8	11
A30	7	9	A30	5	5
A40	7	8	A40	3	3
A60	5	5			
B10	7	57	A10	8	44
B15	7	32	A15	8	22
B20	7	14	A20	8	14
B30	7	9	A30	5	6
B40	7	8	A40	4	5
B60	5	5			
C10	7	50	A10	8	32
C15	7	28	A15	8	18
C20	7	13	A20	8	11
C30	7	9	A30	5	5
C40	7	8	A40	3	3
C60	5	5			

Table 1: Matched references and assigned segments for bright and dark houses concerning *A*, *B* and *C*

SL	ρ_d^*	ρ_q	ρ_{qw}^*	$\rho_{qw(1)}^*$	$\rho_{qw(2)}^*$	ρ_{cc}
A10	1.00	0.84	0.50	0.91	1.00	0.27
A15	1.00	0.81	0.41	0.83	0.98	0.42
A20	1.00	0.79	0.36	0.77	0.98	0.70
A30	1.00	0.78	0.33	0.72	0.96	0.88
A40	1.00	0.74	0.22	0.45	0.68	0.93
A60	0.71	0.75	0.19	0.32	0.44	1.00
B10	1.00	0.86	0.54	0.91	0.99	0.22
B15	1.00	0.85	0.48	0.84	0.98	0.36
B20	1.00	0.83	0.43	0.78	0.95	0.67
B30	1.00	0.81	0.36	0.71	0.92	0.88
B40	1.00	0.78	0.27	0.50	0.72	0.93
B60	0.71	0.81	0.25	0.39	0.52	1.00
C10	1.00	0.84	0.45	0.83	0.98	0.25
C15	1.00	0.79	0.33	0.67	0.91	0.40
C20	1.00	0.77	0.30	0.64	0.90	0.70
C30	1.00	0.76	0.29	0.62	0.90	0.88
C40	1.00	0.74	0.23	0.47	0.74	0.93
C60	0.71	0.76	0.21	0.35	0.48	1.00

Table 2: Quantities for bright houses

SL	ρ_d^*	ρ_q	ρ_{qw}^*	$\rho_{qw(1)}^*$	$\rho_{qw(2)}^*$	ρ_{cc}
A10	1.00	0.75	0.23	0.43	0.60	0.38
A15	1.00	0.69	0.14	0.24	0.36	0.62
A20	1.00	0.63	0.12	0.24	0.43	0.84
A30	0.63	0.67	0.11	0.19	0.28	1.00
A40	0.38	0.62	0.09	0.17	0.27	1.00
B10	1.00	0.80	0.31	0.57	0.78	0.31
B15	1.00	0.75	0.22	0.44	0.70	0.53
B20	1.00	0.72	0.16	0.27	0.42	0.73
B30	0.63	0.75	0.18	0.30	0.44	0.91
B40	0.50	0.66	0.05	0.07	0.08	0.89
C10	1.00	0.76	0.27	0.56	0.80	0.40
C15	1.00	0.68	0.16	0.33	0.58	0.62
C20	1.00	0.63	0.12	0.25	0.46	0.84
C30	0.63	0.66	0.12	0.21	0.34	1.00
C40	0.38	0.63	0.11	0.20	0.35	1.00

Table 3: Quantities for dark houses

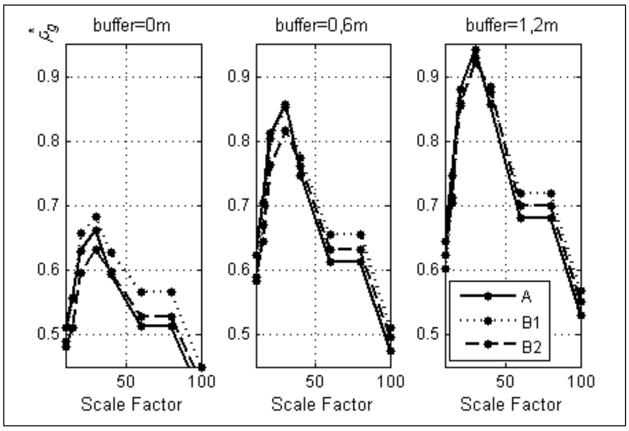

Figure 5: ρ_g^* of bright houses for references *A*, *B* and *C*

effect of a buffer around $\mathcal{R}$ was assessed. ρ_{qw}^* shows relativly low values of (4) without introducing a buffer. Already a buffer of 0.6 m (equivalent to one pixel) shows a clear increase of the values (Tab. 2). A buffer of 1.2 m results in a further yet lesser increase of ρ_{qw}^*, meaning that mostly the boundaries of $\mathcal{R}$ and $\mathcal{S}$ lie within a distance of one pixel. For the buffer of 1.2 m the quality decline between *SL30* and *SL40* is significant - independent whether *A*, *B* or *C* were used. The shape of the curves of the values of ρ_{cc} and ρ_{qw}^* is the same - independent on the reference objects and whether a buffer was used. Comparing the results of ρ_{cc} and ρ_{qw}^*, the first quantity suggests *SL40*, the second *SL30* or lower. In ρ_g^* all quantities were used in order to receive an overall result. *SL30* shows the highest value - independent whether *A*, *B* or *C* was used or whether or not a buffer (see Fig. 5) was introduced. The absolute values differ, while the shape of the either curve is the same. For *SL10* and *15*, the number of segments fulfilling (1) differs clearly. For *SL20*, *B* contains one segment more, while *A* and *C* contain the same segments. Furtheron the same segments are evaluated. Therefore, all differences of the quantities from *30* on result of the differences in shape and size of the reference objects. It can be stated, that the individual values of *bright houses* give an idea of the quality, but have to be taken with care. The plots on the other hand show unambiguously which segmentation level is the best and should be used for classification. In contrast to visual assessment, that would have chosen *SL40*, emphasing more the connectivity, the quantitative assessment suggests *SL30* by combining the quantities.

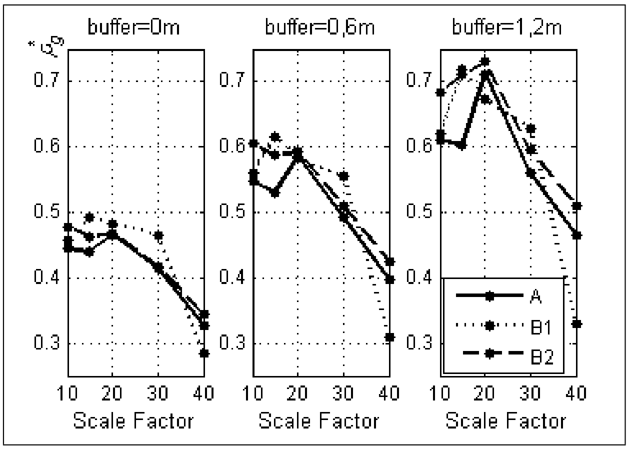

Figure 6: ρ_g^* of dark houses for references *A*, *B* and *C*

For *dark houses* visual evaluation yields different results (see above). The quantities confirm this impression. All references are matched until *SL20* (see ρ_d^* in Tab. 3). ρ_d^* is the same for the $\mathcal{R}$ of *A*, *B* and *C*, except *B*30. For the three levels *SL10*, *SL15* and *SL20*, ρ_{cc} is highest for *SL20*. By evaluating the segments' boundary delineation by ρ_{qw}^*, it is noticeable that the values are significantly lower than the ones for *bright houses*. ρ_q shows lower values as well, but the absolute difference for ρ_{qw}^* is higher, showing more obviously a quality difference, caused by branching, even after introducing the buffer. For *dark houses*, there are higher differences of ρ_{qw}^* between *A*, *B* and *C*. As a result, the shape of the curve ρ_g^* does not show the same evenness as for *bright houses*, but differs for *A*, *B* and *C*. For *A*, the highest value is found at *SL20*, for *B* at *SL15* and for *C* at *SL10* for $d_T = 0$ and $d_T = 1$, while for $d_T = 2$ the maximum is at *SL20* like *A*. The maximum for *C* at *SL10* for $d_T = 0$ and $d_T = 1$ is not very explicit and very close to*20*. According to the results for *A*, *B* and *C*, independly looking one might chose either *SL10* or *15* or *20*. But the maxima are not as significant as for *bright houses*. The results are ambiguous in contrast to the results for *bright houses*. The rise of the values for ρ_{qw}^* and therefore for ρ_g^* by introducing the buffers exists as well, but is not as high as for *bright houses*, also reflecting the further branching of the segments.

Comparing the results for *bright* and *dark houses*, it is noticable, that the values for *A* and *C* are more similar, yet not the same. Knowing the same segmentation strategy makes the result more comparable, even though the absolute values are not the same due to the different shape of the reference objects and the different fulfillment of (1). That means, the quantities are relativly independent on the operator. For the case of *dark houses* we see higher differences. Already by visual evaluation differences between the segmentation quality of *bright* and *dark houses* could be observed. *Bright houses* can be delineated better. Segments assigned to *dark houses* branch out more. The soil's spectral signature is close to the one of rusty roofs (which are often even covered with a certain dust cover). The result for *bright houses* is better. The maximum of ρ_g^* is more obvious and its value is considerably higher than the one for *dark houses*. Visually it is obvious, that the intensive reflecting new metal and asbestos roofs show a higher contrast to the surrounding terrain. If segments branch out like for *dark houses*, the assignment or non-assignment of one segment can therefore mean a higher difference for the quantity values. Comparing Tab. 1, it can be seen that for *dark houses* expecially for *A* and *B*, the number of assigned segments is almost the same, but due to the shape dissimilarities already one segment dropping out leads to relativly high differences in ρ_{qw}^*.

In this study we compared three $\mathcal{R}$. For *bright houses* we can observe a certain independence on the operator concerning the shape of the curves of quantities, for *dark houses* not. The question arises, whether or not one could appraise the certainty out of the curves using one single $\mathcal{R}$. For the presented examples it can be observed, that the maxima for *bright houses* are much more obvious for either curve. For *dark houses*, the values for *SL10*, *SL15* and *SL20* are closer to each other. That could be an indicator, that the decision about the best to be used segmentation level cannot be taken without further investigation. In this specific case it reflects also the uncertainty during visual evaluation.

6 CONCLUSIONS

In this contribution evaluation approaches for image segmentation with respect to remote sensing data are investigated. We discussed different quantities for quality evaluation and proposed a scheme for their combination into a *combined quality measure*. The single quantities reflect different requirements for a segmentation. We further investigated the influence of different reference data on the quantities focussing on house classes which constitute important classes for our application. It could be shown, that for a well segmented class like *bright houses*, the quantities lead to very similar results independent of the reference data compiled by different operators. For classes, that are more difficult to segment due to their spectral properties and their surroundings, the quantities do not only show lower values, but also a certain ambiguity for the different reference data sets. If the shape of the curve is steep, pointing well to one maximum as for *bright houses*, an operator can rely on that result. The proposed *combined quality measure* ρ_g^* provides a meaningful overall assessment. Nevertheless, the single measures ρ_d^*, $\rho_c c$ and ρ_{wq}^* give more detailed information with respect to the different requirements for the segmentation.

For further comprehensive work, a variety of classes with respect to the geometry of their boundaries have to be incorporated and larger control samples will be used. Different classes are likely to impose different requirements on the segmentation results and therefore further quality measures may be of importance, also leading to a weighting scheme for the combined measure. Furthermore, the assignment of segments to a reference object will be reevaluated and possibly a buffer will be introduced in the step.

REFERENCES

Baatz, M. and Schäpe, A., 1999. Object-oriented and multi-scale image analysis in semantic networks. In: 2nd Intl. Symposium. Operationalization of Remote Sensing, 16 - 20 August, ITC, NL, Wichmann, pp. 16 – 20.

Cardenes, R., Bach, M., Chi, Y., Marras, I., de Luis, R., Anderson, M., Cashman, P. and Bultuelle, M., 2007. Multimodal evaluation for medical image segmentation. In: CAIP2007, LNCS, Vol. 4673, pp. 229 – 236.

Correira, P. and Pereira, F., 2003. Objective evaluation of video segmentation quality. IEEE Transactions on Image Processing 12(2), pp. 186 – 200.

Hoover, A., Jean-Baptiste, G., Jiang, X., Flynn, P., Bunke, H., Goldgof, D., Bowyer, K., Eggert, D., Fitzgibbon, A. and Fisher, R., 1996. An experimental comparison of range image segmentation algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence 18(7), pp. 673–689.

Keim, D., 1999. Efficient geometry-based similarity search of 3d spatial databases. In: SIGMOD 99, Proceedings ACM SIGMOD International Conference on Management of Data, June 1-3, 1999, Philadelphia, PA, ACM Press, pp. 419–430.

Neubert, M., Herold, H. and Meinel, G., 2006. Evaluation of remote sensing image segmentation quality – further results and concepts. In: IAPRSIS, Vol. 26, Part 4/C42 - Conference on Object-based Image Analysis (OBIA 2006).

Neubert, M., Herold, H. and Meinel, G., 2008. Assessing image segmentation quality - concepts, methods and application. In: T. Blaschke, S. Lang and G. J. Hay (eds), Object-Based Imge Analysis - Spatial Concepts for Knowledge-Driven Remote Sensing Applications, Springer-Verlag, pp. 769–784.

Radoux, J. and Defourny, P., 2008. Quality assessment of segmentation results devoted to object-based classification. In: T. Baschke, S. Lang and J. G. Hay (eds), Object-Based Imge Analysis - Spatial Concepts for Knowledge-Driven Remote Sensing Applications, Springer-Verlag, pp. 257–271.

Schöpfer, E. and Lang, S., 2006. Object-fate analysis - a virtual overlay method for the categorization of object transition and object-based accuracy assessment. In: International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. XXXVI-4/C42, Salzburg, Austria.

Schuster, H.-F. and Weidner, U., 2003. A new approach towards quantitative quality evaluation of 3d building models. In: ISPRS Commission IV Joint Workshop *Challenges in Geospatial Analysis, Integration and Visualization II*, Stuttgart, pp. 156 – 163.

Udupa, J., LeBlanc, V., Zhuge, Y., Imielinska, C., Schmidt, H., Currie, L., Hirsch, B. and Woodburn, J., 2006. A framework for evaluating image segmentation algorithms. Computerized Medical Imaging and Graphics 20, pp. 75 – 87.

Unnikrishnan, R., Pantofaru, C. and Hebert, M., 2007. Toward objective evaluation of image segmentation algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence 29, pp. 929 – 944.

Weidner, U., 2008. Contribution to the assessment of segmentation quality assessment for remote sensing applications. In: IAPRSIS, Vol. 37, Part B7, pp. 1101 – 1110.

Weidner, U. and Bähr, H.-P., 2007. Vergleich von pixel- und segmentbasierter Klassifizierung am Beispiel des Kaiserstuhls. In: 27. Wissenschaftlich-Technische Jahrestagung der DGPF / Dreiländertagung SGPBF, DGPF und OVP, Muttenz, pp. 315 – 322.

Zhang, Y., 1996. A survey on evaluation methods for image segmentation. Pattern Recognition 29(8), pp. 1335 – 1346.

Zhang, Y.-J., 2001. A review of recent evaluation methods for image segmentation. In: Intl. Symposium on Signal Processing and its Applications (ISSPA), 13 - 16 August 2001, Kuala Lumpur, Malaysia, pp. 148 – 151.