Computers & Geosciences 35 (2009) 1574–1581

Contents lists available at ScienceDirect

Computers & Geosciences

journal homepage: www.elsevier.com/locate/cageo

A method for multi-spectral image segmentation evaluation based on synthetic images

André R.S. Marçal *, Arlete S. Rodrigues

Centro de Investigação em Ciências Geo-Espaciais, Faculdade de Ciências, Universidade do Porto, DMA, Rua do Campo Alegre, 687, 4169-007 Porto, Portugal

ARTICLE INFO

Article history:
Received 13 December 2007
Received in revised form
24 July 2008
Accepted 18 November 2008

Keywords:
Image segmentation
Synthetic images
Similarity indices
Segmentation metrics
Segmentation evaluation

ABSTRACT

A general framework for testing the quality of the segmentation of a multi-spectral satellite image is proposed. The method is based on the production of synthetic images with the spectral characteristics of the image pixels extracted from a signature multi-spectral image. The knowledge of the location of objects in the synthetic image provides a reference segmentation, which allows for a quantitative evaluation of the quality provided by a segmentation algorithm. The Hammoude metric and three external similarity indices (Rand, Corrected Rand, and Jaccard) were chosen to perform this evaluation, but other metrics can also be used. The proposed methodology can be used for any type of satellite image (or multi-spectral image), set of land cover types, and segmentation algorithms.

A practical application was carried out to illustrate the value of the proposed method. A SPOT satellite image was used to extract the spectral signature of 8 land cover types. Three test images were produced using the 8 land cover classes and two different 5 class sub-sets. The segmentation results provided by a standard algorithm were compared with the reference or expected segmentation. The results clearly indicate that the quality of a segmentation obtained from a multi-spectral image not only depends on the geometric properties of the objects present in the image, but also on their spectral characteristics. The results suggest that a specific evaluation should be carried out for each particular experiment, as the segmentation results are very dependent on the choice of land cover types.

1. Introduction

Multi-spectral images acquired by Earth Observation Satellite (EOS) are increasingly being used on a variety of applications and scientific fields. Over the last decade, an alternative approach to the standard per-pixel analysis has evolved to extract meaningful information from EOS images. Instead of focusing on individual image pixels, the object-based image analysis approach consists of partitioning an EOS image into meaningful image-objects, and assessing their characteristics through spatial, spectral, and temporal scale (Hay and Castilla, 2006). One of the reasons for the development of object-based methods has been the dramatic increase in commercially available high-resolution digital remote sensing imagery, with spatial resolutions of 5.0 m and finer (Hay et al., 2005). Also it has been recognised that the image pixel is not a “natural” element of an image scene. The object-based approach has been used in several remote sensing application areas, such as forest monitoring (Dorren et al., 2003; Hay et al., 2005 and Desclee et al., 2006), land cover mapping (Marçal et al., 2005 and Clevers et al., 2007), and change detection (Im et al., 2008).

A common element of all object-based image analysis systems is the segmentation stage, where the image is partitioned in a number of objects (or segments), which is clearly a critical stage of the whole process. If the segmentation fails to identify as an object a given element present in the image, the subsequent stages will generally be unable to recognise or to classify this element. It is generally accepted that only a few of the various image segmentation methods that have been used in image analysis lead to qualitatively convincing results, which are robust and applicable under operational settings (Baatz and Schape, 2000). Although the correct evaluation of the segmentation algorithm used is an important aspect of any object-based image analysis system, there is no established standard procedure for the evaluation of segmentation results (Neubert et al., 2008). The most common approaches are based on discrepancy measures between the segmentation result and a reference segmentation (Carleer et al., 2005). There are some examples of EOS image segmentation evaluations, using reference objects identified manually (Meinel and Neubert, 2004 and Moller et al., 2007), but it is usually difficult to have abundant reference segmentation data for EOS images.

The purpose of this work is to present a framework for the evaluation of image segmentation algorithms based on the production of synthetic multi-spectral images. These images are

---

* Corresponding author. Tel.: +351 220402273; fax: +351 220402209.
E-mail address: andre.marcal@fc.up.pt (A.R.S. Marçal).

0098-3004/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cageo.2008.11.008

produced with controlled characteristics, both spatially and spectrally, simulating any satellite sensor and set of land cover types.

2. Segmentation evaluation

Image segmentation is the process of partitioning an image into a set of non-overlapping regions, whose union is the entire image. The objective of the image segmentation stage is usually to identify the object of interest in the image, separating it from the background, or alternatively to divide the image in a small number of segments or objects. However, in the case of EOS images, the number of objects present in the image is normally very large. This brings a problem of evaluating the segmentation result, as the commonly used metrics, such as the true detection rate, average distance or the Hausdorff distance are mostly suitable for the case of a single interest object (Chalana and Kim, 1997). Furthermore, there is usually no reference segmentation to be used for comparison. The proposed methodology based on synthetic images provides a reference (expected) segmentation, and makes use of the Hammoude metric, which is a standard metric that does not favour under-segmentation or over-segmentation, and the external similarity indices Rand, Corrected Rand, and Jaccard to evaluate quantitatively the segmentation result. These three indices were chosen because they are all based on the same variables that can be computed efficiently, even for very large images, as it is often the case of EOS images.

2.1. Hammoude metric

The Hammoude metric allows for an evaluation of the similarity between two segmentations proposed for an object (X and Y) by comparing the number of common and non-common pixels in the two segmentations (Hammoude, 1988). The Hammoude metric is computed by

$$H = [\#(X \cup Y) - \#(X \cap Y)] / \#(X \cup Y) \tag{1}$$

where X and Y are two binary representations of the segmented object, and the operator # returns the "number of pixels ON" of a binary image. The Hammoude metric has values between 0 and 1, with a value of 1 occurring when there is no intersection between segmentations (completely dissimilar), and a value of 0 when the two segmentations are equal. In order to have consistency between the various parameters used to compare segmentations, the inverted Hammoude index (H^0) is also used, where $H^0 = 1 - H$ ($H^0 = 1$ for a perfect match and $H^0 = 0$ for two non-intersecting objects).

2.2. External similarity indices

The problem of comparing two segmented images can also be seen as a problem of comparing two classifications (data partitions), where each object (or segment) is assigned a class label. The similarity between two segmented images can thus be evaluated using external indices of cluster validity, which access the degree to which two classifications of the data agree (Dubes, 1987).

The external indices used here are the Rand coefficient (Rand, 1971), the Corrected Rand coefficient (Hubert and Arabie, 1985), and the Jaccard coefficient (Dubes, 1987). The computation of these indices is based on four variables (a, b, c, and d) that are obtained by inspecting the labels assigned to all pairs of patterns (pixels) in the dataset (image). Each image pixel (i) has two class labels, $X(i)$ and $Y(i)$, corresponding to the object numbers assigned by segmentations X and Y. The variable a is computed by counting the number of pixel pairs that have the same class in X and the same class in Y Eq. (2a). The variable d corresponds to the number of pixel pairs with different labels in both X and Y Eq. (2d), while b and c refer to pixel pairs that are assigned the same label in one of the classified images and different labels in the other Eqs. (2b) and (2c). The four values (a, b, c, and d) sum is the number of ordered pairs of pixels: $a+b+c+d = N(N-1)/2$, where N is the total number of image pixels.

$$a = \#\{(i,j): i \neq j; \quad X(i) = X(j); \quad Y(i) = Y(j)\} \tag{2a}$$

$$b = \#\{(i,j): i \neq j; \quad X(i) = X(j); \quad Y(i) \neq Y(j)\} \tag{2b}$$

$$c = \#\{(i,j): i \neq j; \quad X(i) \neq X(j); \quad Y(i) = Y(j)\} \tag{2c}$$

$$d = \#\{(i,j): i \neq j; \quad X(i) \neq X(j); \quad Y(i) \neq Y(j)\} \tag{2d}$$

The Rand coefficient (R) is the relative number of pixel pairs that are treated in the same way under both classifications (Dubes, 1987), and is computed by

$$R = (a + d)/(a + b + c + d) \tag{3}$$

The Jaccard coefficient (J) is also based on the relative number of pixel pairs that are treated in the same way under both segmentations, but discounting the situations where both classifications assign the two pixels with different classes (d).

$$J = a/(a + b + c) \tag{4}$$

Both the Rand and Jaccard indices vary between 0 and 1, with 1 corresponding to perfectly matched classifications. The Corrected Rand coefficient (CR) is a modified version of the standard Rand index, proposed by Hubert and Arabie (1985), where the range of values (0–1) is adjusted so that a fully random classification would correspond to a value of 0. The Corrected Rand coefficient is computed by Eq. (5), where R_{max} is the maximum possible value for R ($R_{max} = 1$) and R_{exp} is the expected value of R under randomness (Hubert and Arabie, 1985). In order to have a range between 0 and 1 for CR, negative values can be truncated to 0.

$$CR = (R - R_{exp})/(R_{max} - R_{exp}) \tag{5}$$

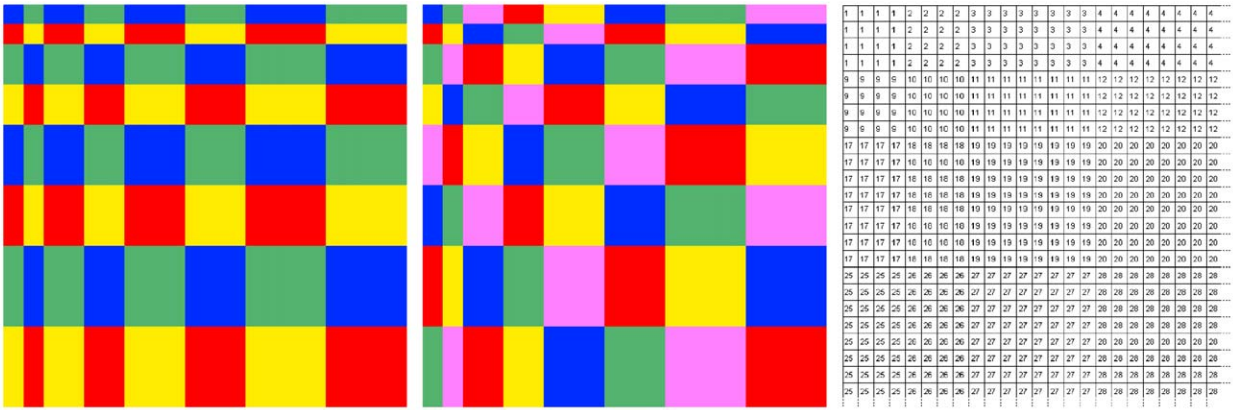
As the patterns considered are the image pixels (N), which is usually a large number, the very large number of pixel pairs ($\approx N^2/2$) could cause computational problems. However, Hubert and Arabie (1985) presented a combinatorial approach to the computation of variables a, b, c, and d that is computationally very efficient, making these external similarity indices suitable for large images with any number of classes, such as the labelled images produced by segmentation of EOS images.

3. Synthetic image production

A methodology to produce synthetic test images was developed to simulate a scenario with land parcels of various sizes, belonging to a number of cover types. Initially, a base image is produced with rectangular areas (parcels) assigned to different class labels. A multi-spectral signature image is used to obtain multi-spectral pixels belonging to each land cover type, and together with the base image produce the final multi-spectral synthetic image.

3.1. Base image

A base image is produced given the following parameters: the number of land cover types (t), the size of the smallest unit (u), the range of sizes (s) and a repetition parameter (r). The base image has a total of $r^2 s^2$ parcels of rectangular shape, with sizes from $1*1$ to $s*s$ units. There are r^2 single unit parcels (of u by u pixels), and



1	1	1	1	2	2	2	2	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4
1	1	1	1	2	2	2	2	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4
1	1	1	1	2	2	2	2	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4
1	1	1	1	2	2	2	2	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4
9	9	9	9	10	10	10	10	11	11	11	11	11	11	11	11	12	12	12	12	12	12	12	12
9	9	9	9	10	10	10	10	11	11	11	11	11	11	11	11	12	12	12	12	12	12	12	12
9	9	9	9	10	10	10	10	11	11	11	11	11	11	11	11	12	12	12	12	12	12	12	12
9	9	9	9	10	10	10	10	11	11	11	11	11	11	11	11	12	12	12	12	12	12	12	12
17	17	17	17	18	18	18	18	19	19	19	19	19	19	19	19	20	20	20	20	20	20	20	20
17	17	17	17	18	18	18	18	19	19	19	19	19	19	19	19	20	20	20	20	20	20	20	20
17	17	17	17	18	18	18	18	19	19	19	19	19	19	19	19	20	20	20	20	20	20	20	20
17	17	17	17	18	18	18	18	19	19	19	19	19	19	19	19	20	20	20	20	20	20	20	20
17	17	17	17	18	18	18	18	19	19	19	19	19	19	19	19	20	20	20	20	20	20	20	20
17	17	17	17	18	18	18	18	19	19	19	19	19	19	19	19	20	20	20	20	20	20	20	20
17	17	17	17	18	18	18	18	19	19	19	19	19	19	19	19	20	20	20	20	20	20	20	20
17	17	17	17	18	18	18	18	19	19	19	19	19	19	19	19	20	20	20	20	20	20	20	20
25	25	25	25	26	26	26	26	27	27	27	27	27	27	27	27	28	28	28	28	28	28	28	28
25	25	25	25	26	26	26	26	27	27	27	27	27	27	27	27	28	28	28	28	28	28	28	28
25	25	25	25	26	26	26	26	27	27	27	27	27	27	27	27	28	28	28	28	28	28	28	28
25	25	25	25	26	26	26	26	27	27	27	27	27	27	27	27	28	28	28	28	28	28	28	28
25	25	25	25	26	26	26	26	27	27	27	27	27	27	27	27	28	28	28	28	28	28	28	28
25	25	25	25	26	26	26	26	27	27	27	27	27	27	27	27	28	28	28	28	28	28	28	28
25	25	25	25	26	26	26	26	27	27	27	27	27	27	27	27	28	28	28	28	28	28	28	28
25	25	25	25	26	26	26	26	27	27	27	27	27	27	27	27	28	28	28	28	28	28	28	28

Fig. 1. Synthetic base images (u ¼ 4, s ¼ 4, and r ¼ 2) with 4 and 5 classes, t ¼ 4 (left) and t ¼ 5 (centre), and top left section of corresponding reference segmentation image (right).

generally r^2 parcels of $i*j$ units, with i, j ¼ 1, y , s. It is worth noting that two neighbouring parcels always belong to different classes.

As an example, Fig. 1 shows two base images, with u ¼ 4, s ¼ 4, and r ¼ 2, one with 4 classes (left) and the other with 5 classes (centre). The smallest squares on the top left section of this example have 4 by 4 pixels (1 unit), while the largest ones on the lower right part of the image have 16 by 16 pixels (4 by 4 units, s ¼ 4). In this case the repetition parameter (r) is 2, which means that there are 4 parcels of each size (r^2). The whole images are 80 by 80 pixels, with a total of 64 different parcels. There are 4 single unit parcels (4 by 4 pixels), and generally 4 parcels of $i*j$ units, with i, j ¼ 1, y , 4.

3.2. *Multi-spectral synthetic image*

To produce a multi-spectral synthetic image from the base image, a signature image is required with training areas identified for each land cover type. The multi-spectral synthetic image is produced, with the same number of bands as the signature image, by replacing each pixel in the base image by a pixel vector selected randomly from the corresponding training area. An example is presented in Fig. 2, where four land cover types were identified in a 3 band (RGB true colour) signature image. The synthetic multi-spectral image was produced using the base image of Fig. 1, with u ¼ 4, s ¼ 4, r ¼ 2, and t ¼ 4.

3.3. *Evaluation of the segmentation results for the synthetic images*

The purpose multi-spectral synthetic image is to evaluate the performance of a segmentation algorithm. The expected or ideal segmentation consists of identifying each parcel as an object. A reference segmentation image is created by assigning a different object label to each parcel in the base image. A section of the reference segmentation for the base image used in the previous sections (u ¼ 4, s ¼ 4, and r ¼ 2) is presented in Fig. 1 (right). In this example, the reference segmentation has 64 objects. The segmentation result (Y) obtained from the application of a segmentation algorithm to the multi-spectral synthetic image is compared with the reference segmentation (X).

For the Hammoude metric, a value of Hm is computed for each object m of X, by comparing it with the object n of Y that contains the central pixel of m. Average values of H are computed for all parcels of the same size and for all objects belonging to the same class, as this information is available in the base image.

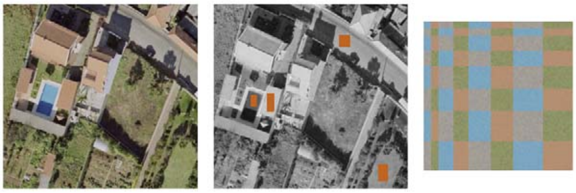


Fig. 2. Signature image (left) with 4 land cover types (centre) and resulting synthetic image (right).

For the external similarity indices, the segmentation images are divided in sub-images, corresponding to the areas where the objects have the same size in X. The Rand, Corrected Rand and Jaccard indices are computed for each sub-image.

As there is normally no difference between the horizontal and vertical directions, the average values are obtained for parcels of size $i*j$ and $j*i$ (e.g. no distinction is made between parcels of 2 by 1 and 1 by 2 units). The average indices computed are based on r^2 objects for square parcels, and $2*r^2$ objects for the other parcel sizes.

4. Experimental setup

A practical experiment was carried out to illustrate the usefulness of the proposed method. Three multi-spectral synthetic images were initially created, and then segmented using the multi-resolution algorithm in eCognition 2.1 software (Baatz et al., 2001). The results provided by the segmentation software were compared with the expected segmentation and analysed using the proposed indices.

4.1. *Synthetic image characteristics*

Two base images were created with u ¼ 3, s ¼ 8, and r ¼ 5, one with t ¼ 5 and another with t ¼ 8. These images are 540 by 540 pixels, with a total of 1600 parcels. The smallest parcels of 1 by 1 units are 3 by 3 pixels, and the biggest, of 8 by 8 units, are 24 by 24 pixels. There are 25 parcels for each of eight square sizes (between 1 by 1 to 8 by 8 units) and 50 rectangular parcels of i by j units (with i ¼ 2 to 8, j ¼ 1–7, and i4 j). In total there are 36 different parcel sizes.

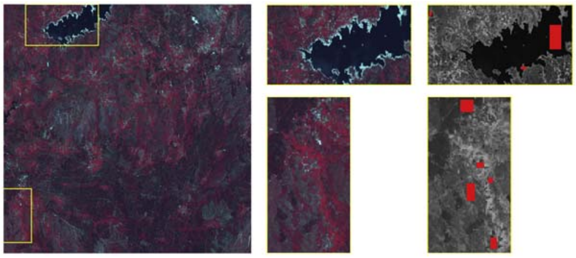


Fig. 3. Signature satellite image with training areas for 8 land cover types.

Table 1
Land cover classes used for each test image (TI8—8 classes, TI5d—5 most different classes, and TI5s—5 most similar classes).

Class ID	Class name	TI8	TI5d	TI5s
1	Irrigated permanent semi-natural mountain meadows	Yes	Yes	Yes
2	Non-irrigated permanent semi-natural mountain meadows	Yes	No	Yes
3	Evergreen forest	Yes	Yes	No
4	Deciduous forest	Yes	Yes	Yes
5	Communitarian pastures	Yes	No	Yes
6	Annual crops	Yes	No	Yes
7	Sand	Yes	Yes	No
8	Water	Yes	Yes	No

A SPOT 5 HRG satellite image covering a mountainous area around Montalegre (Portugal) was used as signature image. The SPOT image has 4 spectral bands in the visible and near infrared parts of the electromagnetic spectrum, and a spatial resolution of 10 ms.[1] A total of 8 land cover types were used, with training areas identified manually in the signature image. The section of the SPOT satellite image is presented in Fig. 3 as a RGB colour composite of bands 3, 2, 1 (near infrared, red, and green). In this figure, two sub-windows where the training areas were identified are enlarged, and the training areas used are overlaid on the greyscale versions of these sub-windows. The details of the land cover classes used are presented in Table 1. Two sub-sets with 5 out of the 8 classes were also considered, one where the 5 most different classes were selected (5d) and another with the 5 most similar classes (5s). The three sets of classes were used, together with the base image, to produce three synthetic test images (named TI8, TI5d, and TI5s). The test images have 4 bands (as the signature satellite image) and are presented in Fig. 4 as RGB colour composites of bands 3, 2, 1.

4.2. *Image segmentation*

The 3 test images were segmented using the multi-resolution segmentation algorithm, available in eCognition 2.1 software, with the default parameters: scale—10, color—0.8, shape—0.2, smoothness—0.9, and compactness—0.1 (Baatz et al., 2001). Image TI5s was also segmented with different values for the scale parameter (4, 6, 8, 10, 12, 14, 16, 18, and 20), with the other parameters kept fixed at the default values. The resulting segmentations were converted to raster, and the image pixels labelled with the corresponding object number.

5. Results

Using the default parameters, the eCognition software segmented the test images in 1563 (TI8), 1596 (TI5d), and 1545 (TI5s) objects, less than the expected (ideal) number of 1600 objects. Fig. 5 shows a section of the segmented images produced, TI8 (left), TI5d (centre), and TI5s (right), corresponding to an area of the base image with parcels of 6 by 1, 6 by 2, and 6 by 3 units (25 parcels of each size). In this section, the results obtained for TI5d perfectly match the reference segmentation image, unlike the other two results. For TI8, about half of the parcels are segmented exactly, while for the other half there are a few pixel differences between the two segmentations. For TI5s, there are considerable differences between the segmentation produced by the software and the reference segmentation. In two instances two parcels were segmented into a single object.

5.1. *Hammoude metric*

The Hammoude metric (H) was computed for each object in the reference segmentation, comparing it with the conjugate object in the segmentations produced by eCognition. The average values of H were: 0.0430 for TI8, 0.0025 for TI5d, and 0.0820 for TI5s.

The colours assigned to each object in Fig. 5 are the RGB average of its pixels, with a specific RGB colour composite and histogram enhancement assigned to each image. An inspection of this figure not only reveals that the set of classes selected to produce the synthetic image influences the results, but also that within each single image there are classes that are more likely to be correctly segmented than others. The average value of H was, therefore, computed for each land cover class, for the three test images, with the results presented in Table 2. There are two classes that have a perfect score, water and sand, which are spectrally very different than the remaining classes. For the other classes, the value of H is higher in the presence of more spectrally different classes, which results in generally better scores for all classes in TI5d, followed by TI8 and by TI5s.

The values of H for individual objects were also averaged in terms of the parcel size. The results are presented in Table 3 for TI8 and in Table 4 for TI5s. For TI5d, the average values of H were null for all parcel sizes, except the smallest sized parcels (1 by 1 unit) that had a value of H ¼ 0.1600.

5.2. *External similarity indices*

The Rand (R), Corrected Rand (CR), and Jaccard (J) indices were computed for the whole test images and also for all sub-images corresponding to the various parcel sizes.

As previously stated, the segmentation of TI5d was perfect in all but the smallest parcels. For the sub-window corresponding to these parcels (of 1 by 1 units), the similarity indices between the two segmentation images were R ¼ 0.9872, CR ¼ 0.8555, and J ¼ 0.7576. All other sub-windows had a perfect score (1), thus the overall values for TI5d were very high: R ¼ 0.9998, CR ¼ 0.9977, and J ¼ 0.9962.

For TI8, the overall values for the external similarity indices were: R ¼ 0.9977, CR ¼ 0.9767, and J ¼ 0.9623. The values of R are all above 0.99 except for parcels of 1 by 1 (R ¼ 0.9488) and 2 by 1 units (R ¼ 0.9827). For most sizes the value of R is even above 0.999. The results for the various sub-windows of TI8 are presented in Table 5 for CR and in Table 6 for J.

For TI5s, the overall values for the external similarity indices were: R ¼ 0.9941, CR ¼ 0.9549, and J ¼ 0.9348. Again the values of R are all above 0.99, except for the sub-windows with the smallest

---

[1] SPOT Image. http://www.spot.com.

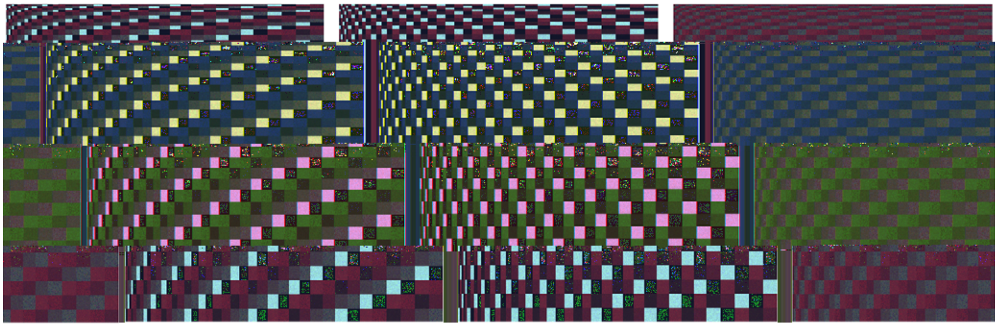


Fig. 4. Synthetic test images TI8 (left), TI5d (centre), and TI5s (right), RGB colour composites of bands 321.

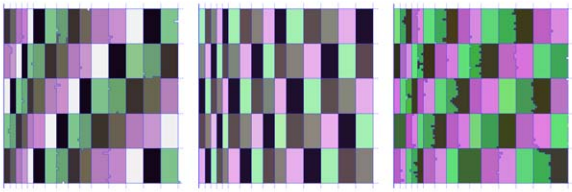


Fig. 5. Detail of images produced by segmentation of TI8 (left), TI5d (centre), and TI5s (right).

Table 2
Average Hammoude metric for each land cover class and image tested.

Class ID	8	5d	5c
1	0.0279	0.0047	0.0445
2	0.1142	[a]	0.1504
3	0.0080	0.0016	[a]
4	0.0944	0.0063	0.1083
5	0.0526	[a]	0.0522
6	0.0468	[a]	0.0546
7	0.0000	0.0000	[a]
8	0.0000	0.0000	[a]

[a] Class not used.

Table 3
Average Hammoude metric for parcels sized *i* by *j* units, for TI8.

Size *ij*	1	2	3	4	5	6	7	8
1	0.4653							
2	0.2523	0.1231						
3	0.1400	0.0261	0.0174					
4	0.1397	0.0224	0.0152	0.0131				
5	0.0588	0.0213	0.0153	0.0105	0.0181			
6	0.0563	0.0222	0.0138	0.0115	0.0080	0.0081		
7	0.0562	0.0180	0.0095	0.0087	0.0075	0.0086	0.0134	
8	0.0353	0.0264	0.0143	0.0149	0.0117	0.0081	0.0117	0.0041

parcels: 0.8816 (for parcels of 1 by 1 unit), 0.9276 (2 by 1), 0.9797 (3 by 1), and 0.9874 (4 by 1). The results for the various sub-windows of TI5s are presented in Table 7 for *CR* and in Table 8 for *J*.

5.3. *Evaluation of a segmentation parameter*

The evaluation of one parameter of the segmentation algorithm (the parameter scale) was done using the proposed method for TI5s. The number of objects produced by the segmentation process greatly varies with values assigned to the scale parameter, from 1087 (scale ¼ 20) to 25,990 objects (scale ¼ 4). The closest to the reference value occurs for the default value of 10 (1545 objects). The global indices also vary considerably with this parameter. The best value of *H* is 0.0821 for scale ¼ 10 and of *CR* is 0.9910 for scale ¼ 8.

The average value of *H* for image sub-windows with parcels of 2 2, 4 4, 6 6 and 8 8 units is presented in Fig. 6. The dotted line is the global value of *H*. The plot shows that the best choice for the scale parameter varies with the parcel size: 8 for parcels of 2 2 units, 10 or 12 for parcels of 4 4 and 16–20 for larger parcels. These results clearly show a strong relationship between

Table 4
Average Hammoude metric for parcels sized *i* by *j* units, for TI5s.

Size *ij*	1	2	3	4	5	6	7	8
1	0.7147							
2	0.5857	0.1959						
3	0.3183	0.0853	0.0306					
4	0.1972	0.0582	0.0315	0.0328				
5	0.0635	0.0364	0.0220	0.0190	0.0210			
6	0.0819	0.0507	0.0214	0.0203	0.0188	0.0155		
7	0.0773	0.0477	0.0173	0.0247	0.0135	0.0555	0.0485	
8	0.0784	0.0254	0.0223	0.0260	0.0162	0.0165	0.0396	0.0466

Table 5
Average Corrected Rand coefficient for parcels sized *i* by *j* units, for TI8.

Size *ij*	1	2	3	4	5	6	7	8
1	0.5868							
2	0.8142	0.9194						
3	0.8906	0.9988	0.9992					
4	0.9187	0.9988	0.9994	0.9997				
5	0.9768	0.9988	0.9993	0.9998	0.9992			
6	0.9589	0.9986	0.9995	0.9998	0.9999	0.9999		
7	0.9596	0.9994	0.9998	0.9997	0.9999	0.9998	0.9989	
8	0.9963	0.9990	0.9993	0.9995	0.9993	0.9997	0.9995	1.0000

Table 6
Average Jaccard coefficient for parcels sized *i* by *j* units, for TI8.

Size *ij*	1	2	3	4	5	6	7	8
1	0.4386							
2	0.6993	0.8568						
3	0.8119	0.9976	0.9985					
4	0.8556	0.9978	0.9988	0.9993				
5	0.9573	0.9977	0.9987	0.9996	0.9985			
6	0.9270	0.9974	0.9991	0.9996	0.9997	0.9998		
7	0.9282	0.9988	0.9996	0.9994	0.9998	0.9997	0.9978	
8	0.9929	0.9980	0.9987	0.9991	0.9987	0.9994	0.9991	1.0000

these plots, the values of H', R, CR, and J are presented for squared parcels as a function of their size (plot on the left) and for rectangular parcels of x by 1 units as a function of x (right), both for TI5s. The H', CR, and J values follow similar trends, with much less variability exhibited by R. One parameter that can characterise the shape of a parcel is the ratio between interior and border pixels ($r_{i/b}$). For example, for a single unit parcel of 3 by 3 pixels, there is only 1 interior pixel and 8 pixels on the border, thus the value of $r_{i/b}$ is 0.125. In the plots of Fig. 7, not only the area increases with the scale parameter x (x by x or x by 1 units), but also the values of $r_{i/b}$. Comparing groups of parcels with approximately the same area, for example, 8 by 1, 4 by 2, and 3 by 3 or 8 by 2, 5 by 3, and 4 by 4. This analysis reveals that the $r_{i/b}$ ratio seems to be more important than the parcel area in the segmentation performance. This is again not a surprise, as the segmentation errors occur mostly in the edges of objects due to neighbouring objects with close spectral signatures.

The Rand index is not very useful, as almost all sub-windows are rated highly, reducing the capability to distinguish between the various results. All other indices (H, CR, and J) provide a valuable quantification of the segmentation results, which permit to evaluate the performance of a segmentation algorithm in terms of objects shape and size. The Hammoude metric has the advantage of being computed for each individual object in the reference segmentation, allowing for an overall evaluation of the segmentation performance for each land cover type. It is possible to modify algorithms used to compute the external similarity indices to do a similar type of analysis by creating multiple sub-windows using the class information from the base image.

One of the main advantages of the similarity indices as a tool to evaluate segmentations is the fact that they are easy to compute, even for images with a very large number of objects. However, they are based on raster data, which can cause some limitations as some segmentation algorithms produce vector data that smooth (interpolate) the object outlines. A direct vector to raster conversion in these cases would result in a loss of information. This problem can nevertheless be prevented by using a higher resolution reference image in the vector to raster conversion.

6. Conclusions

A methodology to evaluate the performance of a segmentation algorithm applied to a multi-spectral satellite image was proposed. The approach is based on the production of multi-spectral synthetic images, reproducing the spectral properties present in a real satellite image, but with controlled spatial characteristics of the image-objects. This overcomes one of the main limitations in the evaluation of segmented satellite images, which is the lack of a reference segmentation.

The methods used to compare segmentations were the Hammoude metric and three external similarity indices (Rand, Corrected Rand, and Jaccard). The information provided by all these indices was consistent, but the Rand index was found to have a very limited range on the tests performed. The Hammoude metric is more convenient to evaluate the performance for distinct land cover types, while the similarity indices are more effective for parcel size evaluation. The Jaccard index has the advantage of being the easiest to compute.

The experiences carried out clearly revealed that the segmentation result obtained for a multi-spectral image not only depend on the geometric properties of objects present in the image, but also on their spectral characteristics. In fact, the selection of land cover types was found to be the single most important factor that affects the segmentation results. From the geometric point of view, the ratio between the number of interior and border pixels of objects was found to be the most meaningful parameter affecting the segmentation performance. The experiment with different values assigned to the scale parameter illustrates how the proposed method can be used to select the optimum values for the segmentation algorithm parameters, for a specific EOS image dataset. The tests carried out were simply done to illustrate the potential of the proposed method, and cannot be seen as an evaluation of a specific segmentation algorithm.

The methodology proposed can be used for any type of satellite image (or other type of multi-spectral image), set of land cover classes, and segmentation algorithm. It can also be used to fine tune the configuration parameters present in most segmentation algorithms to achieve the best results for a specific objective. Although the synthetic images reproduce the spectral characteristics of land cover types on real EOS images, there will be always limitations due to the fact that it is not possible to predict all land cover types and mixtures present on real EOS images. The addition of other metrics in the future can improve the potential of the proposed framework for the evaluation of multi-spectral image segmentation algorithms.

The software developed to create synthetic images and to compute the Hammoude metric and external similarity indices is freely available at www.fc.up.pt/sitef. The images used in this paper are also available in that web site.

Acknowledgments

The authors would like to thank the OASIS programme (No. 307) for providing the SPOT satellite image and Isabel Poças for the selection and identification of land cover classes.

References

Baatz, M., Benz, U., Dehghani, S., Heynen, M., Höltje, A., Hofmann, P., Lingenfelder, I., Mimler, M., Sohlbach, M., Weber, M., Willhauck, G., 2001. eCognition object oriented image analysis, user guide. Definiens Imaging GmbH, Munich, Germany, p. 426.

Baatz, M., Schape, A., 2000. Multiresolution segmentation: an optimization approach for high-quality multi-scale image segmentation. In: Strobl, J., Blaschke, T. (Eds.), Angewandte Geographische Informationsverarbeitung, vol. 12. Wichmann-Verlag, Heidelberg, pp. 12–23.

Carleer, A.P., Debeir, O., Wolff, E., 2005. Assessment of very high spatial resolution satellite image segmentations. Photogrammetric Engineering and Remote Sensing 71, 1285–1294.

Chalana, V., Kim, Y., 1997. A methodology for evaluation of boundary detection algorithms on medical images. IEEE Transactions on Medical Imaging 16, 642–652.

Clevers, J.G.P.W., Schaepman, M.E., Mucher, C.A., De Wit, A.J.W., Zurita-Milla, R., Bartholomeus, H.M., 2007. Using MERIS on Envisat for land cover mapping in the Netherlands. International Journal of Remote Sensing 28, 637–652.

Desclee, B., Bogaert, P., Defourny, P., 2006. Forest change detection by statistical object-based method. Remote Sensing of Environment 102, 1–11.

Dorren, L.K.A., Maier, B., Seijmonsbergen, A.C., 2003. Improved Landsat-based forest mapping in steep mountainous terrain using object-based classification. Forest Ecology and Management 183, 31–46.

Dubes, R.C., 1987. How many clusters are best?—an experiment. Pattern Recognition 20, 645–663.

Hammoude, A., 1988. Computer-assisted endocardial border identification from a sequence of two-dimensional echocardiographic images. Ph.D. Dissertation, University of Washington, Seattle, WA, p. 274.

Hay, G.J., Castilla, G., 2006. Object-based image analysis: strengths, weaknesses, opportunities and threats (SWOT). In: Proceedings of the Object-Based Image Analysis (OBIA) 2006, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) Commission V1, WG V1/4, Salzburg, Austria.

Hay, G.J., Castilla, G., Wulder, M.A., Ruiz, J.R., 2005. An automated object-based approach for the multiscale image segmentation of forest scenes. International Journal of Applied Earth Observation 7, 339–359.

Hubert, L., Arabie, P., 1985. Comparing partitions. Journal of Classification 2, 193–218.

Im, J., Jensen, J.R., Tullis, J.A., 2008. Object-based change detection using correlation image analysis and image segmentation. International Journal of Remote Sensing 29, 399–423.

Marçal, A.R.S., Borges, J.S., Gomes, J.A., Da Costa, J.F.P., 2005. Land cover update by supervised classification of segmented ASTER images. International Journal of Remote Sensing 26, 1347–1362.

Meinel, G., Neubert, M., 2004. A comparision of segmentation programs for high-resolution remote sensing data. In: Proceedings of the XXth International Society for Photogrammetry and Remote Sensing Congress, Istanbul, Turkey, The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences XXXV-B4, pp. 1097–1102.

Moller, M., Lymburner, L., Volk, M., 2007. The comparison index: a tool for assessing the accuracy of image segmentation. International Journal of Applied Earth Observation 9, 311–321.

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

Rand, W.M., 1971. Objective criteria for evaluation of clustering methods. Journal of the American Statistical Association 66, 846–850.