International Journal of Remote Sensing
Vol. 31, No. 3, 10 February 2010, 617–645

Taylor & Francis
Taylor & Francis Group

Segmentation performance evaluation for object-based remotely sensed image analysis

PADRAIG CORCORAN*, ADAM WINSTANLEY and PETER MOONEY
National Centre for Geocomputation, Department of Computer Science, National University of Ireland Maynooth, Co. Kildare, Ireland

(*Received 18 July 2008; in final form 1 December 2008*)

The initial step in most object-based classification methodologies is the application of a segmentation algorithm to define objects. Modelling the human visual process of object segmentation is a challenging task. Many theories in cognitive psychology propose that the human visual system (HVS) initially segments scenes into areas of uniform visual properties or primitive objects. If an accurate primitive-object segmentation algorithm is ever to be realized, a procedure must be in place to evaluate potential solutions. The most commonly used strategy to evaluate segmentation quality is a comparison against ground truth captured by human interpretation. A cognitive experiment reveals that ground truth captured in such a manner is at a larger scale than the desired primitive-object scale. To overcome this difficulty we consider the possibility of evaluating segmentation quality in an unsupervised manner without ground truth. Two requirements for any method which attempts to perform segmentation evaluation in such a manner are proposed, and the importance of these is illustrated by the poor performance of a metric which fails to meet them both. A novel metric, known as the spatial unsupervised (SU) metric, which meets both the requirements is proposed. Results demonstrate the SU metric to be a more reliable metric of segmentation quality compared to existing methods.

1. Introduction

The introduction of high resolution data from satellites such as QuickBird has opened up the possibility of capturing highly detailed land-use classifications of the Earth's surface (Aplin *et al.* 1997). Prior to their introduction, data from older satellites such as Landsat were simply at too great a spatial scale for this to be possible. Traditional pixel-based remote sensing techniques, developed for such low resolution data, cannot be successfully applied to data of a high spatial resolution (Blaschke *et al.* 2000). This is due to the fact that as the spatial resolution increases the land-use heterogeneity also increases. These difficulties have been the impetus for the development of a new form of remote sensing known as object based image analysis (OBIA). The benefit of moving from a pixel to object representation is that you permit the incorporation of more expert and spatial information. Many researchers have drawn similarities between OBIA and the human visual system (HVS) which can accurately interpret aerial imagery (Blaschke 2003). Accurate modelling of the HVS would be beneficial in many ways, none more so than in the development of accurate land-use classification systems. If such an OBIA implementation, which attempts to model aspects of the HVS, is ever to be accomplished it must draw from current theories and findings in cognitive psychology.

*Corresponding author. Email: padraigc@cs.nuim.ie

International Journal of Remote Sensing
ISSN 0143-1161 print/ISSN 1366-5901 online
http://www.tandf.co.uk/journals
DOI: 10.1080/01431160902894475

Many researchers in the field of cognitive psychology believe that object segmentation cannot be achieved in a completely bottom-up manner and that segmentation and classification in most cases are strongly coupled (Corcoran and Winstanley 2007). A bottom-up process is defined as any process which is not influenced by our knowledge (our prior knowledge about the image), our desires (what information we are aiming to extract from the image) and our expectations (what we expect to see in the image). On the other hand a top-down process is affected by these influences. Many theories in cognitive psychology propose that the HVS initially segments an image into areas of uniform visual properties or primitive-objects, which are defined as areas of uniform texture or colour, and that this is a bottom-up process. These segments are then merged and parsed by the HVS in a top-down manner to define the object-hierarchy which contains the various scales of segmentation (Corcoran and Winstanley 2007). The overall goal of our research is to model the initial primitive-object segmentation stage in the HVS. As an alternative to this approach, previous papers (Kühnert *et al.* 2006) propose that land-use patterns may be understood as a result of self-organization principles. One useful technique for analysing such patterns with self-organizing behaviours is detrended fluctuation analysis (DFA) (Varotsos *et al.* 2004, 2006, 2007, Varotsos 2005).

Segmentation is a very active research area in both the fields of remote sensing and computer vision with hundreds of papers published on the subject (Zhang 2006a). Given the vast number of existing algorithms it is important to have a procedure in place to evaluate the performance of all these potential solutions and guide future efforts (Wirth *et al.* 2006). Segmentation evaluation strategies can be broadly classified as analytical, task-based, supervised and unsupervised methods (Yang *et al.* 1995, Jiang *et al.* 2006, Zhang 2006b). Analytical methods characterize a segmentation algorithm by applying mathematical analysis without reference to any implementation or test data. Predicting the performance of a particular algorithm using analytical methods is extremely difficult and therefore is rarely performed in isolation. Task-based methods perform segmentation evaluation by measuring the overall performance of the system using the segmentation algorithm, for example an object-recognition system. However, this strategy can become unfair and, more seriously, inconsistent when evaluating algorithms that are tailored to different applications (Unnikrishnan *et al.* 2007). Supervised methods generally involve a comparison against ground truth captured by photo-interpretation (Congalton and Green 1998, Martin *et al.* 2001). If accurate ground truth can be obtained this represents an optimal objective evaluation strategy. If ground truth is not available segmentation evaluation may be performed in an unsupervised manner. Unsupervised methods generally define a metric for relative segmentation quality which is a function of low-level image features.

McCane (1997) recognized that all segmentation algorithms should be evaluated using a variety of techniques to provide a border and fairer evaluation. Despite this, there have been few studies of how best to determine the performance of segmentation algorithms when applied to remotely sensed data (Carleer *et al.* 2005, Neubert *et al.* 2007). Most segmentation methods are evaluated in a qualitative manner by visual inspection (Chen *et al.* 2006). Some attempts have been made to perform this evaluation in a quantitative manner but most have been based on the assumption that accurate ground truth can be captured (Carleer *et al.* 2005, Neubert *et al.* 2007). In §2 of this paper we show that, in the case of primitive-object segmentation, this assumption is invalid. To overcome this difficulty we propose a new metric with which it is possible to perform this evaluation in an unsupervised manner without ground truth.

This paper is structured as follows. In §2 all supervised segmentation evaluation methods are reviewed and their value assessed. In §3 a review of existing metrics for evaluating segmentation quality in an unsupervised manner is presented; and two requirements for any metric which attempts to perform this evaluation are proposed. Two new unsupervised metrics are also proposed in this section. In §4 we present experimental results which demonstrate the performance of both these metrics. Finally in §5 we draw conclusions from this work and propose future research directions.

2. Supervised segmentation evaluation

Visual assessment of segmentation results is probably the most commonly used supervised evaluation strategy. This method involves simply viewing segmentations and giving a subjective qualitative opinion on their accuracy (Chen *et al.* 2006). Rosenberger *et al.* (2006) evaluates a number of visualization techniques which allow the most effective visual comparison of segmentation results. McCane (1997) recognized that a subjective human being is not the best judge to evaluate the output of any segmentation algorithm and that performance must be quantified.

Quantitative supervised evaluation may be performed by the collection of ground truth and the use of a metric which measures the similarity between a segmentation result and this ground truth. These methods are known as supervised discrepancy methods. If accurate ground truth is available supervised discrepancy methods represent an optimal objective evaluation strategy and should be used if possible (Hoover *et al.* 1996, McCane 1997). If synthetic images are used then accurate ground truth is known. On the other hand, if real or remotely sensed images are used the capturing of accurate ground truth is not straightforward. This is due to the subjectivity involved in the human interpretation of images. Foody (2002) identified eight issues with the current methods of classification accuracy assessment that are commonly used and recommended in remote sensing literature. One of the issues identified in the work of Foody was the inherent inaccuracy of ground-truth data, stating that the common usage of the term 'truth' when describing ground data is problematic and should be avoided. A procedure is described in Usamentiaga *et al.* (2006) for the capturing of ground-truth images which utilizes a group of trained photo interpreters. Each interpreter is asked to segment a given image. All segmentations are then fused to form a single ground truth consisting of only those edges established by more than half the interpreters. In Hoover *et al.* (1996) and Zhu *et al.* (2000) ground truth is generated by a single interpreter then reviewed by another to identify any obvious errors.

If accurate ground truth can be obtained, Unnikrishnan *et al.* (2007) proposed a set of requirements for any supervised discrepancy method which attempts to perform an objective comparison between a segmentation and this ground truth. These requirements are:

1. **Non-degeneracy**: no degenerate cases where unrealistic segmentations that are not well represented by the ground truth give abnormally high values of similarity.
2. **No assumptions about data generation**: does not assume equal cardinality of labels or object sizes between segmentation and ground truth.
3. **Adaptive accommodation of refinement**: accommodate label refinement only in regions that humans find ambiguous and penalize differences in refinement elsewhere. Label refinement is defined as differences in the scale of segmentations of a given scene and is illustrated in figure 1. This requirement only applies

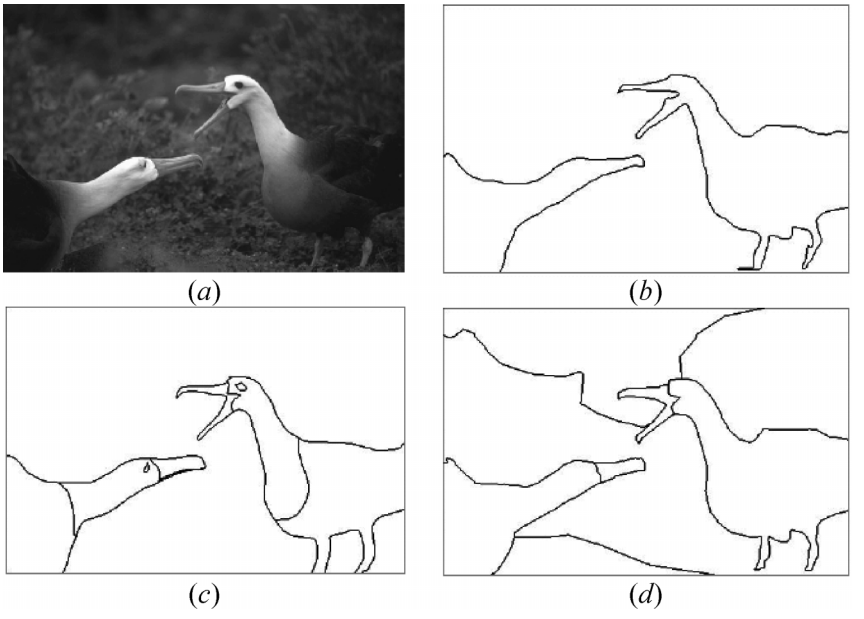


Figure 1. The motivation for making a supervised discrepancy method tolerant to refinement. The original image in (*a*) is taken from the Berkeley segmentation dataset. Three segmentations of this image are displayed in (*b*)–(*d*). (*c*) and (*d*) are simple refinements of (*b*), and are mutual refinements of each other. Taken from Martin *et al*. (2001).

when comparing segmentation against multiple ground truths captured by different image interpreters.

4. **Comparable score**: allows comparisons between segmentations of different images and segmentations of the same image.

All supervised discrepancy methods can be classified as object-discrepancy or edge-discrepancy methods (Usamentiaga *et al.* 2006). Edge-discrepancy methods use the alignment between the edges in the segmented image and ground truth as a measure of similarity (Huang and Dom 1995, Freixenet *et al.* 2002, Martin *et al.* 2004). These methods are not tolerant to refinement and therefore it is possible for two segmentations that are mutual refinements of each other to have a very low similarity score when compared.

Object-discrepancy methods use the properties of the segmented objects and generally operate on pixel labels to provide a measure of similarity. An example of an object-discrepancy method is the Rand index, which counts the pairs of pixels that have a similar label relationship in the segmentation and ground truth being compared. More specifically, given segmentation S and ground truth S' each of N points with labels $\{l_i\}$ and $\{l'_i\}$ assigned respectively, the Rand index R can be computed as the ratio of the number of pairs of pixels having the same label relationship in S and S':

$$R(S,S') = \frac{1}{\binom{N}{2}} \sum_{\substack{i,j \\ i \neq j}} [I(l_i = l_j \ \Lambda \ l'_i = l'_j) + I(l_i \neq l_j \ \Lambda \ l'_i = l'_j)], \tag{1}$$

where I is the identity function and the denominator is a binomial coefficient or the number of combinations of N data points taken in pairs. The number of unique labels

in S and S' is not restricted to being equal. It has been shown that for the comparison of a segmentation to a single ground truth, the Rand index fulfils all four requirements stated by Unnikrishnan. Numerous other supervised discrepancy methods exist in the literature, see Rosenberger *et al.* (2006), Zhang (2006b) and Unnikrishnan *et al.* (2007) for an in-depth review of these.

Most supervised discrepancy methods can only be applied to the case where a segmentation is to be compared to a single ground truth (Unnikrishnan *et al.* 2007). In cases where there are multiple ground truths to which we want to compare a segmentation, these methods cannot be applied and therefore fail to meet Unnikrishnan's third requirement. To address this concern Unnikrishnan proposed two generalizations of the Rand index, namely the Probabilistic Rand (PR) index and the Normalized Probabilistic Rand (NPR) index. In this work we do not compare segmentations to multiple ground truths; therefore we do not describe these techniques in detail.

In the following section we detail a cognitive experiment in which we attempted to assess whether accurate primitive-object ground truth can actually be captured by human interpretation.

2.1 *Capturing of ground truth by photo interpretation*

Vecera and Farah (1997) showed the process of human visual segmentation to be influenced by top-down factors. We would therefore expect any such segmentation to be of a larger scale than primitive objects. This is because the process which generates such segmentations is a high-level process built on the early-vision primitive-object segmentation and has been affected by top-down factors. Therefore these ground truths will have some similarities to the early-vision primitive-object segmentation we are attempting to model but they will also have significant differences. In §2.1.1 a cognitive experiment to test the validity of this assertion is outlined. §2.2.2 presents the results of this experiment.

2.1.1 Cognitive experiment. In a cognitive experiment to assess the similarity of segmentations captured by visual interpretation to the desired primitive-object segmentation, we asked five subjects who were unaware of the research background to segment five remotely sensed images. The images consisted of scanned aerial photography, with a 0.25 m ground sample distance, of Southampton city obtained from Ordnance Survey UK, Southampton. Each image was of size 256×256 pixels and in RGB format. Texture is a spatial property and any features used to describe it must be estimated via reference to a neighbourhood. On the other hand, multispectral features, such as RGB values, are non-spatial properties and can be estimated without reference to a neighbourhood. As a consequence fusing multispectral and texture features is a complex task (Corcoran and Winstanley 2007). To reduce this complexity we decided to convert the RGB images to grey-scale before any analysis was performed. The HVS can still accurately discriminate most primitive objects in aerial imagery even when colour information has been removed so this will not hinder performance significantly. Although all images are of urban areas the primitive objects contained within these images is very diverse. For example an image in this dataset is displayed in figure 2. The primitive objects in this image are very varied and correspond to regions such as individual tree tops, roof surfaces, shadows, road markings, etc.

The subjects in the experiment were untrained photograph interpreters but familiar with aerial photography. The instructions given to them were brief:

Figure 2. An example of an image contained in the remotely sensed dataset. Ordnance Survey Crown Copyright. All rights reserved.

(*a*) (*b*)

Figure 3. Segmentation results captured by two individuals for the image in figure 2. Segmentation object boundaries are represented by the colour white. Ordnance Survey Crown Copyright. All rights reserved.

> You will be presented with a photographic image. Divide the image into some number of segments, where the segments represent 'things' or 'parts of things' in the scene. The number of segments is up to you, as it depends on the image.

These instructions are similar to those used by Martin *et al.* (2001) to capture the Berkeley segmentation ground-truth dataset for natural scenes.

2.1.2 Cognitive experiment results. Figure 3 displays two example segmentation results captured by two individuals for the image in figure 2. From these ground

truths we see that, although large-scale conceptual objects such as trees, buildings, roads and sidewalks are represented, a large number of primitive objects, for example road markings and garden features, are not. The differences between individual segmentations are due to the fact that the top-down influences affecting segmentation will vary between individuals. Although only two segmentation results are shown here, these properties were uniform across the whole dataset captured.

After completion of the segmentation task subjects were asked to give reasons for their failure to capture certain primitive objects in the scene. All reasons given were similar to 'I did not notice that object' or 'I believed that object to be of the same visual properties as the larger scale object captured' or 'I believed that object not to be of significant importance to require representation'. These results inform us that when asked to produce ground truth, human subjects tend to produce segmentation at a larger scale than the desired primitive objects.

To overcome this issue one could collect multiple ground truths for each image using the same procedure. This would result in a large percentage of all primitive-object boundaries being represented in at least one ground truth. If a metric such as the NPR index, which fulfils all four of Unnikrishnan's requirements listed above, was then used to compare a segmentation to its corresponding set of ground truths, it would provide a relative measure of segmentation quality. This is because such a metric would tolerate label refinement only in these regions that humans find ambiguous, corresponding to boundaries which are only represented in some ground truths, and penalize differences in refinement elsewhere, corresponding to boundaries which are not represented in any of the ground truths. A drawback of this approach is that capturing multiple ground truths for each individual image is very time consuming. An alternative strategy could be to repeat the above experiment with the instructions given to the interpreters edited so that they are instructed to capture all primitive objects. Again this would require a great effort on the interpreter's behalf and would be labour intensive.

3. Unsupervised segmentation evaluation

Given the difficulties in capturing accurate primitive-object ground truth required for supervised evaluation, we now investigate the possibility of performing evaluation in an unsupervised manner. For an unsupervised method to be accurate, it must incorporate all the information required to generate the segmentation which we are attempting to implement. That is, it must know exactly what the segmentation is attempting to achieve if it is to evaluate its quality. The segmentation we are attempting to emulate is a bottom-up primitive-object segmentation which incorporates little or no top-down knowledge. Therefore the accuracy of any solution to this problem can be assessed in an almost entirely bottom-up manner. Building on this point, we propose that any metric which attempts to evaluate the quality of primitive-object segmentation in an unsupervised manner must exhibit the following properties:

1. It must operate on a feature set which is an accurate model of the low-level features used in the HVS to define primitive-objects.
2. The HVS performs segmentation in the spatial domain only seeking contrast across the boundaries it defines and not between primitive objects which do not share a common boundary. Therefore any unsupervised metric must also be implemented in the spatial domain.

Most existing unsupervised metrics in the literature attempt to optimize a metric which is a function raw intensity features (Zhang 2006b). This would be a suitable feature set if each primitive object was of uniform intensity, but this is not the case. Most primitive objects in remotely sensed data contain small scale texture and this is a valuable cue used by the HVS to determine boundaries. Therefore these methods are founded on an inaccurate model of the low-level features used by HVS and break the first of our required properties. In Rosenberger *et al.* (2006) and Chabrier *et al.* (2006) the authors propose to optimize a metric which is a function of a competing intensity and texture feature set. Within this strategy each region and boundary is defined exclusively by a single visual cue of texture or intensity, but not both. Most primitive objects exhibit both unique texture and intensity properties which are exploited simultaneously by the HVS to define segmentation. Therefore this competing feature set does not represent an accurate model of the low-level features used by HVS, again breaking the first of our required properties. Choosing a suitable cost metric to operate on a given feature set is a non-trivial problem. Most existing cost metrics operate in the feature space and neglect attributes of the spatial domain (Zhang 2006b). These methods therefore break the second of our required properties.

In this work we propose two novel unsupervised metrics. The first metric, known as the PBM metric, operates in the feature domain and is a function of a complementary intensity and texture feature set. This feature set allows all objects and object boundaries to be described in terms of both visual cues simultaneously. It therefore represents a more accurate model of the low-level features used by the HVS than previous competing feature sets proposed. The PBM therefore satisfies the first of our required properties but being implemented in the feature domain it fails to satisfy the second. The motivation for choosing a metric which deliberately fails this requirement is because this requirement has been ignored by most existing metrics and we wanted to highlight its importance. The second metric we propose is known as the spatial unsupervised (SU) metric. This metric is again a function of a complementary intensity and texture feature set but instead of operating in the feature domain it operates in the spatial domain. It therefore satisfies both of our stated requirements. In the next part of this section we describe the complementary feature set used. This is followed by in-depth details of the two unsupervised metrics we propose. The final part of this section presents the strategies we used to evaluate the accuracy of both metrics.

3.1 *Complementary feature set*

The practice of using competing features is motivated by two properties of raw intensity and texture feature images. Firstly, an edge detector applied to a raw intensity image with the aim of detecting primitive-object intensity boundaries will not only respond to such boundaries but also the intensity variation due to primitive-object texture. Secondly, texture is a spatial property and any features used to describe it must be calculated within a neighbourhood. This results in what Corcoran and Winstanley (2007) refer to as the texture boundary-response problem. That is where a unique response is observed at primitive-object boundaries due to the feature extraction algorithm responding to a mixture of textures and/or a primitive-object intensity boundary. An edge detector applied to such a raw texture feature image will not result in the desired single response to each primitive-object texture boundary. Instead two responses, corresponding to each primitive object to boundary-response edge, will result at such locations. Within most competing feature set strategies a measure of

texturedness is used to overcome these traits by modulating each feature image in all locations. Intensity features in the presence of texture are inhibited, reducing false-positives resulting from primitive-object texture intensity variation. Whereas texture features in the presence of no texture are inhibited, removing false-positives resulting from boundary responses at pure intensity boundaries. Using such a competing feature set strategy means each object and object boundary can only be described in terms of a single visual cue of either texture or intensity. These strategies therefore suffer from reduced discrimination strength and are not an accurate model of the features used by the HVS to define primitive objects.

To overcome this issue we propose to use a complementary texture and intensity feature set. The texture features are calculated using a popular Gabor filter bank implementation and the intensity features are calculated using a diffusion process. This set allows all objects and object boundaries to be described in terms of both visual cues simultaneously. Corcoran and Winstanley (2007) provide details on how this complementary feature set is computed.

3.2 *PBM metric*

Given a dataset that possesses a clustering structure the task of a clustering algorithm is to reveal this structure. All clustering algorithms are based on some form of prior knowledge about the properties of the clusters contained in the dataset, for example their number or shape. Poor prior knowledge may lead to inaccurate conclusions about the clustering structure of the data. Therefore the need for further evaluation of clustering results is apparent (Theodoridis and Koutroumbas 2006). For a given partitioning of a dataset, the task of cluster validation is to quantify how closely this structure imposed on the dataset actually fits its natural structure. It is easy to think of image segmentation in terms of clustering; as the aim is to represent the image in terms of clusters of pixels (Forsyth and Ponce 2002). In fact, many segmentation algorithms are based on a clustering of image pixels (Comaniciu and Meer 2002). This implies that techniques from the area of cluster validation may be applied to segmentation results to provide a metric of segmentation performance.

Clustering validation metrics are traditionally classified as unsupervised, supervised and relative types (Tan *et al.* 2006:533). Unsupervised cluster validation metrics measure the accuracy of a particular clustering without reference to external information. These measures typical involve a measure of cluster cohesion or separation or both and are often referred to as internal metrics because they only utilize information present in the original data (Theodoridis and Koutroumbas 2006). Supervised cluster validation metrics measure the extent to which a particular clustering matches some external information or ground truth. The goal of these metrics is similar to that of object-discrepancy segmentation evaluation methods, and therefore methods from both areas share many similarities. Supervised cluster validation metrics are often referred to as external metrics because they utilize information not present in the original data. Relative cluster validation metrics are supervised or unsupervised metrics used for the purpose of comparing clusterings. They are not a separate form of cluster validation metric but instead a specific use of existing metrics.

In this work we are interested in using an unsupervised cluster validation metric as a relative cluster validation metric to perform relative unsupervised segmentation evaluation. Pal *et al.* (2000), Acharyya and Kundu (2001), Pal and Mitra (2002) and Mitra *et al.* (2004) utilize a cluster validation index called the B metric to provide a

relative evaluation of segmentation algorithms applied to remotely sensed data. The B metric is simply a ratio of the total dataset variance to the sum of individual cluster variances, with the intention that higher values will represent better segmentations. The variance of the total dataset is a constant so this equation reduces to one over the sum of individual cluster variances. One of the requirements stated by Unnikrishnan *et al.* (2007) for any supervised discrepancy metric is that no degenerate cases should exist (see §2). This requirement is also vital for any unsupervised metric. A degenerate case occurs in the B metric when every pixel in the image is considered an individual cluster. In this case the variance of each cluster is zero; therefore the sum of cluster variances is also zero, giving a B metric of infinity. This is obviously a very poor segmentation but it has an abnormally high metric value. Chen and Lee (2001) used an alternative cluster validation index to evaluate the quality of remotely sensed image segmentations.

Within this study we evaluate the effectiveness of a cluster validation index, known as the PBM metric, at determining the relative quality of segmentation results. The PBM metric was proposed by Pakhira *et al.* (2004) and is defined as a product of three factors:

$$\mathrm{PBM}(K) = \left(\frac{1}{K} \times \frac{E_1}{E_K} \times D_K\right)^2, \tag{2}$$

where K is the number of clusters. Given

$$E_K = \sum_{k=1}^{K} E_K \tag{3}$$

where

$$E_k = \sum_{i=1}^{n} u_{kj} ||x_j - z_k|| \tag{4}$$

and

$$D_K = \max_{i,j} ||z_i - z_j|| \tag{5}$$

Here n is the total number of points in the dataset, $U(X) = [u_{kj}]_{K \times n}$ is a partition matrix and z_k is the centre of the kth cluster. The objective is to maximize this metric in order to obtain the relative best clustering of the given data. The PBM metric consists of the three factors $1/K$, E_1/E_K and D_K. The first factor decreases as K increases which in turn reduces the metric value. The second factor consists of a ratio of E_1, which is constant for a given data set, and E_K, which decreases as K increases. Consequently this factor increases as E_K increases and encourages more compacted clusters. The final factor D_K measures the maximum separation between all cluster pairs and increase with K. While the first factor is decreasing the other two are increasing with increased K. This is motivated by the fact that we want to keep the number of clusters to a minimum while increasing their compactness and separation as much as possible. Pakhira *et al.* (2004) provided an in-depth theoretical and empirical evaluation of this cluster validation metric and showed that these three factors compete with each other in a sophisticated manner to assign the highest score to the best clustering of a given dataset. They showed the PBM metric outperforms some of the most established cluster validation indices on both real and synthetic datasets.

In this work we make the PBM metric a function of our complementary feature set which consists of 12 texture features and a single intensity feature. If Euclidean distance was used to measure the distance between points in this feature space, due to their number the texture features would dominate the intensity feature. We view texture and intensity as equally important visual cues and therefore we decided to give each an equal weight when measuring distance between data points. To achieve this a weighted norm was used to measure distance:

$$||x|| = \sqrt{\sum_{i=1}^{m} (w_i x_i)^2}, \tag{6}$$

where m is the number of features, x_i is the individual feature and w_i is the individual feature weighting. In our implementation w_i takes the value 0.5 for our intensity feature and 0.5/12, which equals 0.0417, for each texture feature.

3.3 *Spatial unsupervised (SU) metric*

One of the problems with using a metric from the area of cluster validation is that these techniques are calculated in the feature domain and neglect the spatial attributes of each data point. We propose a novel metric which incorporates the spatial properties of the segmentation being evaluated. Instead of measuring the contrast between every primitive object this metric only measures the contrast between those which share a common boundary. This function, called the spatial unsupervised (SU) metric, is defined as a ratio of primitive-object separation to cohesion:

$$\mathrm{SU} = \frac{\text{separation}}{\text{cohesion}}. \tag{7}$$

Cohesion is defined as a sum of the norms of individual object feature variances weighted by individual object size:

$$\text{cohesion} = \sum_{i=1}^{k} \frac{\text{ObjectSize}(i)}{\text{TotalArea}} ||\text{Variance}(\text{Features}(i))|| \tag{8}$$

Here k is the total number of objects. The variables ObjectSize(i) and Features(i) are the size and features of the ith object respectively, and TotalArea is the total number of pixels in the image. Separation is defined as the sum of contrast for each object to all its neighbouring objects with which it shares a common boundary weighted by object size:

$$\text{separation} = \sum_{i=1}^{k} \frac{\text{ObjectSize}(i)}{\text{TotalArea}} \sum_{j=1}^{k} \text{neighbours}(i,j) \frac{\text{ObjectSize}(j)}{\text{neighboursSize}(i)} ||\text{Mean}(i) - \text{Mean}(j)||, \tag{9}$$

where

$$\text{neighbourSize}(i) = \sum_{j=1}^{k} \text{neighbours}(i,j) \times \text{ObjectSize}(j) \tag{10}$$

and

$$\text{neighbours}(i,j) = \begin{cases} 1 & \text{if } i \text{ and } j \text{ share a boundary} \\ 0 & \text{otherwise} \end{cases}. \tag{11}$$

The variables Mean(i) and Mean(j) are vectors containing the mean values of each feature for the ith and jth objects respectively. The weighted norm discussed in the previous section is also used for the calculation of distances in this metric. A good segmentation should maximize object separation and minimize object cohesion; therefore the goal of segmentation analysis is to maximize the SU metric. Rosenberger *et al.* (2006) also proposed a metric which is implemented in the spatial domain but this metric is a function of a competing texture and intensity feature set. This feature set therefore does not represent an accurate model of the low-level features used by the HVS and breaks the first of the required properties stated earlier.

One issue with all current unsupervised cluster validation metrics and unsupervised segmentation evaluation metrics, including the two presented here, is the fact that none are normalized with respect to a baseline. Therefore these techniques only offer a relative measure of performance on a particular dataset or image. They cannot specify how good a particular clustering or segmentation is in isolation or compare clusterings or segmentations across different datasets. As stated in §2, Unnikrishnan *et al.* (2007) believed this to be a desirable property for any supervised discrepancy method which attempts to perform an objective comparison of a segmentation to ground truth. To achieve this they normalized the Rand index with respect to a baseline.

3.4 *Metric evaluation strategy*

To quantitatively evaluate the accuracy of both the proposed unsupervised segmentation evaluation metrics, synthetic and remotely sensed datasets were used. When accurate ground truth is known an objective supervised discrepancy method may be used to provide an accurate metric of segmentation performance. A measure of similarity of behaviour between this metric and the unsupervised metrics can then be used to measure the precision of the unsupervised metric on this data. To measure this similarity of behaviour a number of strategies may be employed. Zhang (1996) proposed to plot curves of the supervised and unsupervised metric values as the segmentation scale was varied from under- to over-segmented. Properties of the curves such as depth are then extracted and compared to give a measure of performance. Chabrier *et al.* (2006) proposed a measure called cumulative similarity of correct comparison (SCC), which compares the orderings of a set of segmentation results in terms of metric values to measure the equivalence between supervised and unsupervised metrics. The correlation coefficient has also been used to measure the correspondence of behaviour between metrics (Huang and Dom 1995, Rosenberger *et al.* 2006). A correlation coefficient indicates the strength and direction of a linear relationship between two variables. It varies from 0 (random relationship) to 1 (perfect positive linear relationship) to –1 (perfect negative linear relationship) providing a precise measure of the linear relationship between the variables. For this reason we have chosen to use the correlation coefficient between our supervised and unsupervised metrics as a measure of similarity of behaviour, and consecutively a measure of the unsupervised metric accuracy.

To aid visualization and quantification of results we propose to also use a novel measure of similarity of metric behaviour based on histogram similarity. Segmentation is run at multiple scales and for each metric we calculate a histogram

containing a count of the number of images which achieved a metric maximum at each individual scale. For example, if at segmentation scale x a metric achieves a maximum for 10 images, then the histogram bin corresponding to scale x will contain a count of 10. This procedure produces a corresponding histogram for each metric and metrics with similar behaviour will produce similar histograms because for any image they will achieve a maximum value at a similar scale. Therefore a measure of histogram similarity provides a measure of similarity of metric behaviour. If $H = \{h_i\}$ and $K = \{k_i\}$ are histograms where h_i and k_i are the counts in the i bins of H and K respectively, then the L_1-distance between histograms is defined as:

$$d(H, K) = \sum_i |h_i - k_i|. \tag{12}$$

The metric (12) overestimates distances because neighbouring bins are not considered when there is no match between the exact corresponding bins in the two histograms (Rubner *et al.* 1998). This problem can be overcome by calculating the L_1-distance between the cumulative histograms instead (Rubner *et al.* 1998), where a cumulative histogram is defined as:

$$\widehat{\mathrm{h}}_\mathrm{i} = \sum_{\mathrm{j} \le \mathrm{i}} \mathrm{h}_\mathrm{j}. \tag{13}$$

In the remainder of this paper whenever we refer to the histogram of a particular metric we are referring to its original histogram and not its cumulative histogram. Also whenever we refer to the distance between two histograms we are referring to the L_1-distance between the corresponding cumulative histograms. In this work all multi-scale segmentation results were generated by varying the h parameter in the h-minima transform before application of the watershed transform (Corcoran and Winstanley 2007). In all cases the difference between all consecutive h values was uniform, ensuring a uniform increase or decrease in segmentation scale. Consequently histogram distance will be a linear function of the difference in the segmentation scales represented by the two histograms in question. Two histograms where the difference in scales represented by each is small will have a proportionally small histogram distance; two histograms where the difference in scales represented by each is larger will have a proportionally larger histogram distance.

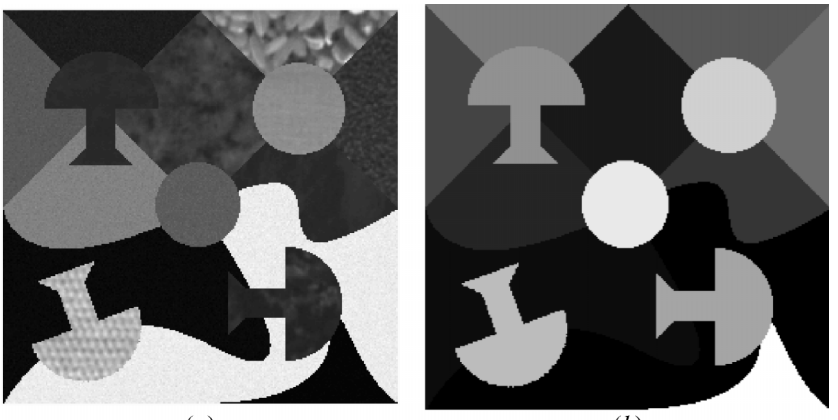
In this work all segmentation evaluation metrics are also evaluated in a qualitative manner using visual inspection. To achieve this, for a particular image a set of segmentations are generated and these are ordered from best-to-worst using visual inspection. This optimal ordering is then compared to the ordering produced by each metric to signify the accuracy of their behaviour.

4. Results

The results section is divided into two parts. In the first part we evaluate the performance of the proposed unsupervised metrics on synthetic data, while in the second part we evaluate their performance on remotely sensed data.

4.1 *Results on synthetic data*

The synthetic dataset used was originally created by Chabrier *et al.* (2006) to evaluate an alternative unsupervised metric and is freely available for download. From this



(*a*) (*b*)

Figure 4. (*a*) An example image taken from the synthetic dataset, with (*b*) corresponding ground-truth. Each colour in the ground-truth image represents a different object.

dataset 100 images of size 256 × 256 were used with each containing 15 regions or primitive objects. Within each image almost half of the primitive objects are of uniform intensity with added Gaussian noise while the remainder are textured. An example of an image taken from this dataset with corresponding ground truth is displayed in figure 4. Usamentiaga *et al.* (2006) makes the point that any synthetic images used should be a true representation of real images. The synthetic images used in this study are clearly not of the same complexity as remotely sensed images in terms of shape and size of primitive objects. Nevertheless the boundaries contained in these images are of a comparable complexity because many consist of a fusion of an intensity and texture boundary. Therefore any results achieve on this data would have a strong relationship with corresponding results achieved on our remotely sensed data.

For synthetic images accurate ground truth is known; therefore an objective supervised discrepancy metric may be used to provide an accurate metric of segmentation performance. The accurate supervised metric used in this study was the Rand index which, for this dataset with a single accurate ground truth, fulfils Unnikrishnan's four requirements for any supervised discrepancy method stated in §2. To generate segmentation at multiple scales for use in the evaluation process an implementation of the watershed transform proposed by Corcoran and Winstanley (2007) was used. Image segmentation was defined for 20 different scales ranging from over- to under-segmented and this is illustrated in figure 5. Figure 6 displays example plots for each segmentation evaluation metric versus segmentation scale for the synthetic image in figure 4(*a*). The majority of plots for each image in the dataset had the same general form as those displayed in these figures; such that the SU metric and Rand index generally first increase then decrease and the PBM metric generally continuously increases as the scale increases. The Rand index and SU metric plots are quite similar with each achieving a maximum at a smaller scale. On the other hand, the PBM metric reaches its maximum at a very large scale. Referring to the plots in figure 6 we see that the Rand index ordered the segmentation scales in figure 5 as 6, 4, 2, 20 in a best-to-worst ordering. On the same scales the SU metric produced a best-to-worst ordering of 4, 6, 2, 20, while the PBM achieved a best-to-worst ordering of 20, 6, 4, 2. Visual inspection informs us that the Rand index ordering is optimal, the SU metric

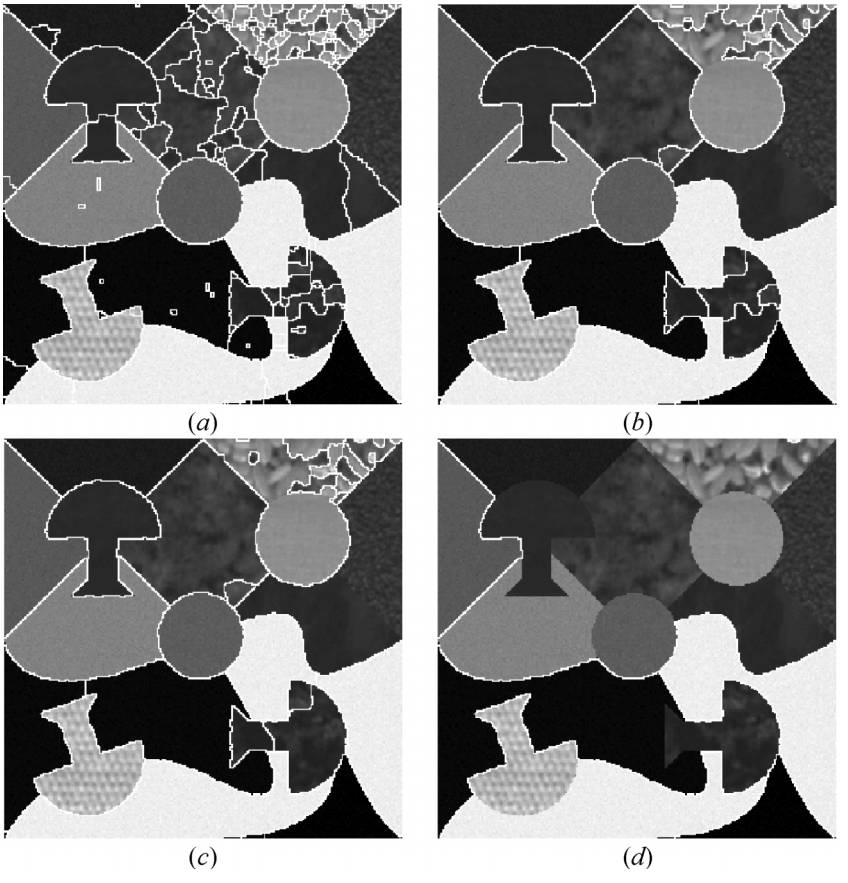


Figure 5. Segmentation results ranging from over- to under-segmented at the scales (*a*) 2, (*b*) 4, (*c*) 6 and (*d*) 20 for the image in figure 4(*a*). Segmentation boundaries are represented by the colour white.

ordering is close to optimal and BPM ordering is far from optimal. Figure 7 displays another synthetic image with corresponding multi-scale segmentation results. The Rand index ordered these scales 8, 6, 4, 10, 2 in a best-to-worst ordering. On the some scales the SU metric produced a best-to-worst ordering of 6, 4, 8, 2, 10, while the BPM achieved a best-to-worst ordering of 10, 8, 6, 4, 2. Again, visually the Rand index ordering is optimal. The SU metric ordering is slightly different but it still ranks the best result in the top two. The BPM ordering is far from optimal. We believe that the BPM metric consistently targets an over-segmented result and consequently its inaccurate behaviour is due to its failure to incorporate spatial information. If two primitive objects, for example two buildings, have similar visual properties but do not share a common boundary, then obviously these should be represented as two separate objects. A PBM metric does not consider this spatial information and therefore will promote segmentation at a very large scale where these two primitive objects are merged. These results indicate that any metric which operates in the feature domain will be inaccurate.

Each performance evaluation metric is designed so that its value should increase with segmentation accuracy; therefore the desired result is a strong positive

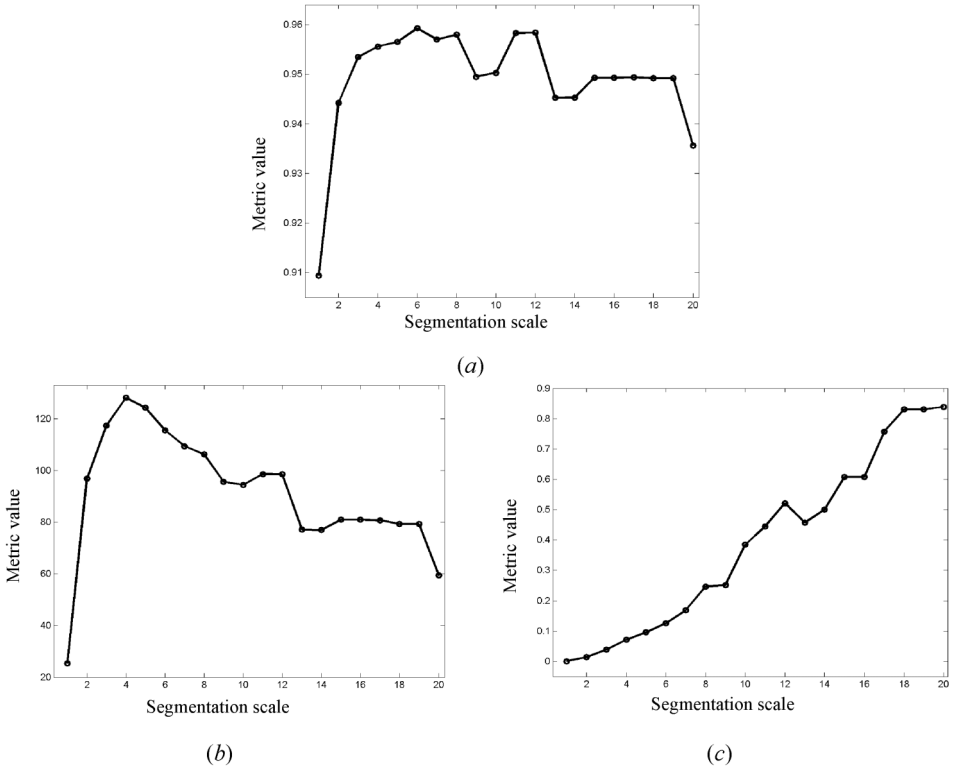


Figure 6. For the synthetic image in figure 4(*a*), each segmentation evaluation metric value on the *y*-axis is plotted versus segmentation scale on the *x*-axis. The segmentation scale increases from left to right. The Rand index is plotted in (*a*), the SU metric in (*b*) and the PBM index in (*c*). The curves of Rand index and SU metric have a strong positive correlation with a correlation coefficient of 0.87. On the other hand the curves of the Rand index and PBM index do not with a correlation coefficient of only 0.02. The Rand index achieves a maximum at the 6th scale, the SU metric achieves a maximum at the 4th segmentation scale and the PBM index achieves a maximum at the 20th scale.

correlation between the supervised and unsupervised metrics. For each image in the synthetic dataset we calculated the correlation coefficient between supervised and unsupervised metric values for the multi-scale segmentation results. The mean correlation coefficient between the PBM metric and the Rand index over the 100 synthetic images was –0.47. This negative correlation, as opposed to the desired strong positive correlation, indicates that the PBM metric does not provide a truthful metric of relative segmentation quality. This demonstrates the need to incorporate the spatial domain when performing unsupervised segmentation evaluation. On the same synthetic dataset the SU metric achieved a strong positive mean correlation coefficient of 0.72 with the Rand index. This demonstrates that the SU metric provides an accurate measure of relative segmentation performance. The metric of Rosenberger *et al.* (2006) represents the current state of the art in terms of unsupervised segmentation evaluation for images containing textured regions. On the same synthetic dataset as used in our study, this unsupervised metric only achieved a relatively poor correlation coefficient of 0.143 with a supervised metric. Huang and Dom (1995) reported a high correlation coefficient between supervised and unsupervised metrics. The data used in this study did not

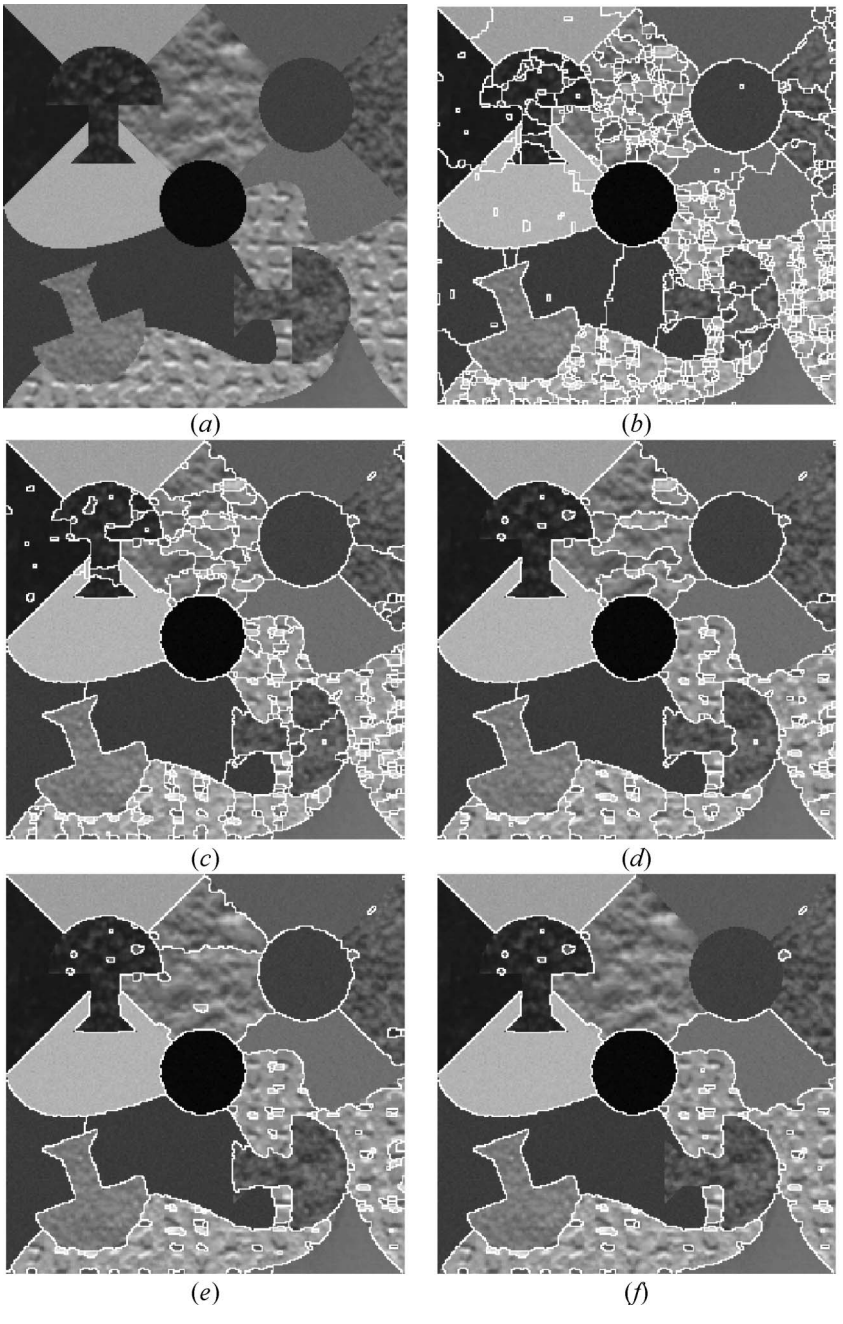


Figure 7. Segmentation results for the image in (*a*) ranging from over- to under-segmented at the scales (*b*) 2, (*c*) 4, (*d*) 6, (*e*) 8 and (*f*) 10. Segmentation boundaries are represented by the colour white.

contain any textured regions, therefore reducing the complexity of the unsupervised evaluation task. The study also calculated the correlation coefficient on a very small number of data points (30) compared to 2000 (20 scales per image for 100 images) used in our study. Zhang (1996) also achieved a high correlation coefficient between supervised and unsupervised metrics but again the data used in this study were composed of highly uniform intensity regions. These comparative results indicate that the proposed SU metric represents the current best approach to accomplishing segmentation evaluation in an unsupervised manner for images containing textured regions.

Histograms showing the segmentation scale at which each of the 100 images achieved a maximum for the Rand index, the SU and PBM metrics are displayed in figure 8. The histograms corresponding to the Rand index and SU metric are quite similar, with both centred around scales 6 and 7, and skewed to the right (the Rand index slightly more so). Also neither histogram contains a mode which is significantly greater than neighbouring values. We believe this is due to the fact, as can be seen in figure 6, that there is no clear best segmentation with neither metric assigning a maximum value which is significant greater than other values. The shape of the histogram corresponding to the PBM metric is significantly different to that corresponding to the Rand index. It has a single significant mode located at the much larger scale 20 and its spread is much less. This signifies that the PBM metric targets a segmentation of significantly larger scale compared to the Rand index. Segmentation

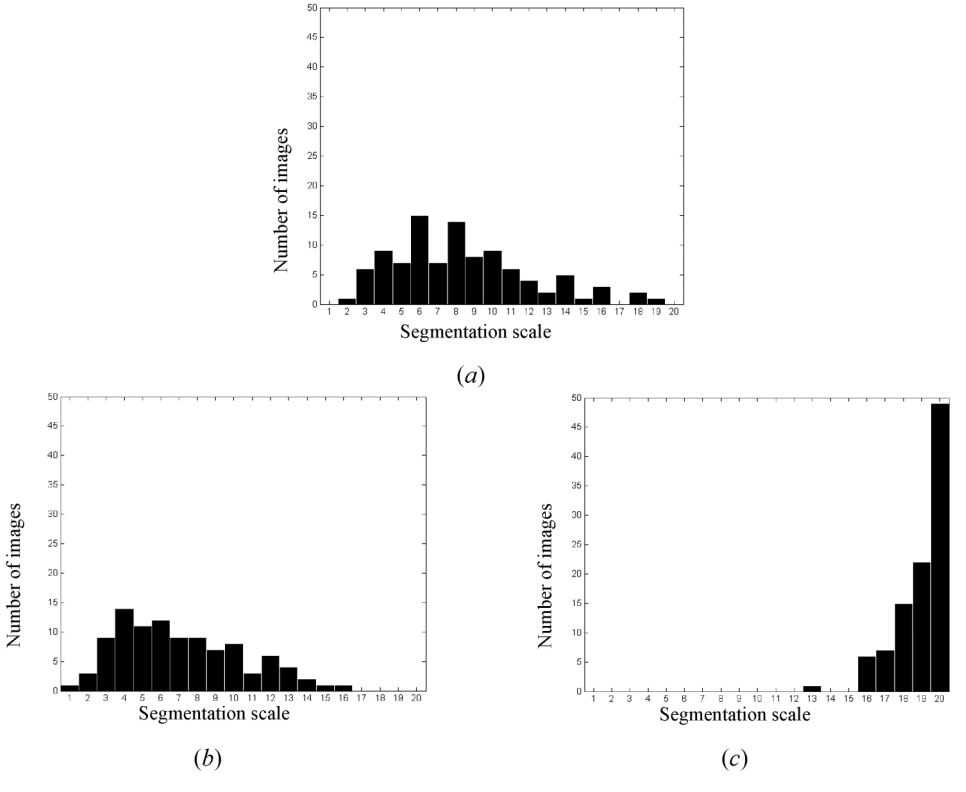


Figure 8. Histograms showing segmentation scale on the x-axis versus on the y-axis a count of the number of images which maximize (*a*) the Rand index, (*b*) the SU metric and (*c*) the PBM index. The segmentation scale increases from left to right.

results at the scales corresponding to the modes of the three individual metric histograms can be seen by referring back to figures 5 and 7. Visual inspection verifies that the segmentations corresponding to the modes of the Rand index (scale 6) and SU metric (scale 4) tend to be close to the relative best. This again shows the similarity of behaviour between the Rand index and SU metric. On the other hand, the segmentation corresponding to the mode of the PBM metric (scale 20) is significantly under-segmented compared the optimal result. Using the histogram distance defined in equation (12), the distance between the cumulative histograms of the Rand index (figure 8(*a*)) and the SU metric (figure 8(*b*)) was calculated to be 121. On the other hand, the distance between the cumulative histograms of the Rand index (figure 8(*a*)) and PBM metric (figure 8(*c*)) was calculated to be 1064. The SU metric histogram is significantly closer to the accurate Rand index histogram than the PBM metric histogram and this again signifies it is a more accurate metric of segmentation quality.

4.2 *Results on remotely sensed data*

To evaluate the accuracy of the SU and PBM metrics across 50 remotely sensed images, segmentation was run on each at 20 scales, ranging from over- to under-segmented. Figure 9 displays an example remotely sensed image. A selection of six corresponding segmentation scales, corresponding to scales 2, 3, 4, 5, 6 and 7, are displayed in figure 10, while segmentation at the scale 20 is displayed in figure 11. From these results we see that for this image, segmentation at or within a single scale of scale 4 is relatively the best and closely resembles the desired primitive-object segmentation, although some over-segmentation is still evident. For example, most trees and roof surfaces are segmented correctly but some over-segmentation is still present within the rear gardens of the terraced houses. The other results contain significantly more over- or under-segmentation. For example the grass area in the right of the image is over-segmented

Figure 9. An image from the remotely sensed dataset. Ordnance Survey Crown Copyright. All rights reserved.

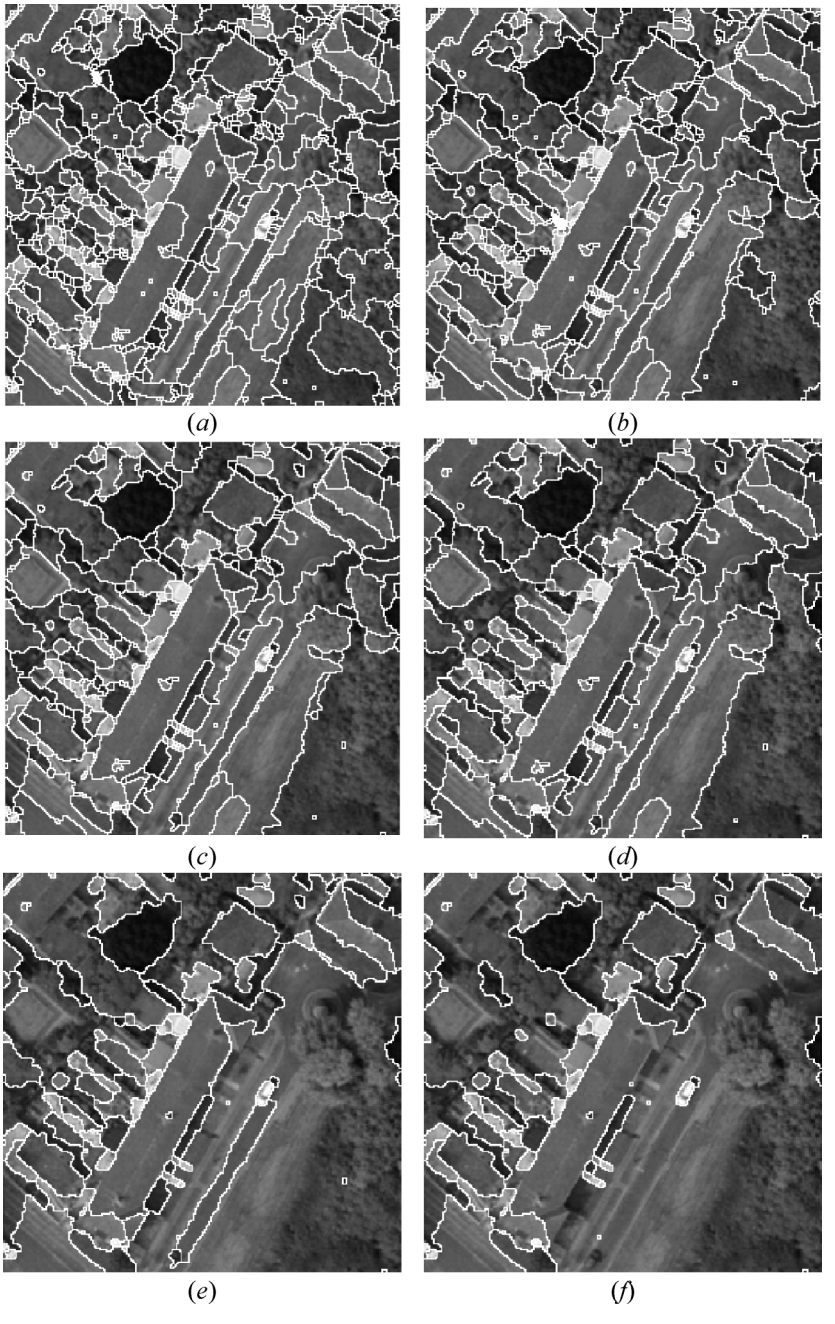


Figure 10. Segmentation results of figure 9 at the scales (*a*) 2, (*b*) 3, (*c*) 4, (*d*) 5, (*e*) 6 and (*f*) 7. Segmentation boundaries are represented by white. Segmentation object boundaries are represented by the colour white.

Figure 11. Segmentation result for figure 9 at segmentation scale 20. Segmentation object boundaries are represented by the colour white. Ordnance Survey Crown Copyright. All rights reserved.

at scales 2 and 3, while under-segmentation is evident within the rear gardens of the terraced houses at scales 6 and 7. The segmentation at scale 20 is extremely under-segmented. It is important to recall that the goal of the work presented in this paper was not to produce an accurate segmentation result. Instead it was to provide a metric that, given a set of segmentation results, assigns the highest relative score to the best segmentation. For this set of segmentations the SU metric produced a best-to-worst scale ordering of 3, 4, 2, 5, 6, 7, 20. Meanwhile on the same set the PBM metric produced a best-to-worst scale ordering of 20, 6, 5, 4, 3, 2. Based on our previous discussion it is clear for this particular image that the SU metric produces a more accurate ordering of segmentation quality relative to the PBM metric. The ordering also represents closely the perceived relative quality of segmentation results.

Figure 12 displays another image in the dataset with corresponding segmentations at scales 2, 3, 4, 5 and 6, while segmentation at scale 20 is displayed in figure 13. In a similar fashion to the results in the previous example the best segmentation results were achieved at or close to scale 4. The segmentations results at scales 2 and 3 exhibit more over-segmentation, for example the building roof surfaces, whereas the segmentations at scales 5 and 6 exhibit more under-segmentation, for example the road surfaces. Again the result at scale 20 is extremely under-segmented. For this set of segmentations the SU metric produced a best-to-worst scale ordering of 2, 4, 3, 5, 6, 20. While on the same set the PBM metric produced a best-to-worst scale ordering of 20, 6, 5, 4, 2, 3. Referring to the previous discussion these orderings show that for this particular image the SU metric targets a slightly over-segmented result. Despite this fact, it still significantly outperforms the BPM metric which targets an extremely over-segmented result. Based on the above facts we can therefore say that, for the two previous examples discussed, the SU metric more closely captures the perceived relative quality of segmentations compared to the BPM metric.

Due to the reasons discussed in §2.1 we do not have accurate primitive-object ground truth for our remotely sensed data. Therefore, unlike the previous section,

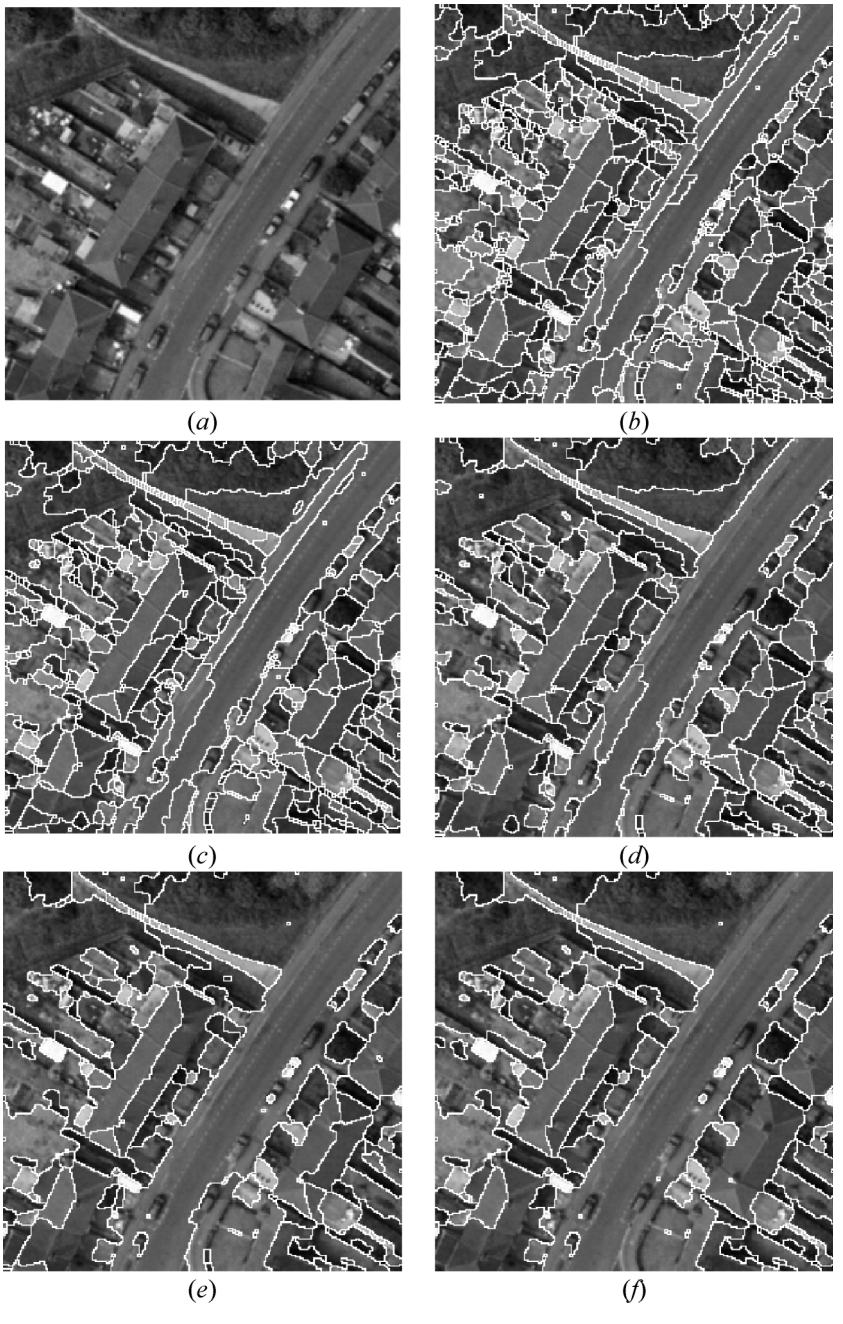


Figure 12. Sample image in (*a*) with corresponding segmentation results at scales (*b*) 2, (*c*) 3, (*d*) 4, (*e*) 5 and (*f*) 6. Segmentation object boundaries are represented by the colour white.

Figure 13. Segmentation result for figure 11(*a*) at segmentation scale 20. Segmentation object boundaries are represented by the colour white. Ordnance Survey Crown Copyright. All rights reserved.

an accurate supervised metric may not be used to measure performance against. To overcome this issue we performed segmentation at multiple scales and used visual inspection to determine at which scale the segmentation algorithm achieved the best results across the entire dataset. Next we constructed a histogram to represent these optimal results and used the distance between this histogram and those histograms produce by the proposed metrics as a quantitative measure of accuracy. The property that segmentations at or close to scale 4 achieved visually the best results while lesser scales were over-segmented and greater scales were under-segmented was uniform across the entire dataset. This is supported by the results achieved on the two previous examples and another set of segmentations displayed in figure 14 which exhibits these same properties. Building on this observation, if we say that scale 4 is uniformly the best result for the dataset in question we can infer that the optimal histogram corresponds to a histogram with all counts of images located at scale 4. Stated explicitly this is a histogram with 50 counts located at scale 4 where 50 is the number of images in the dataset. We refer to this as the optimal histogram derived by visual inspection and it is displayed in figure 15(*a*). Although the actual optimal histogram may not have exactly this shape, its shape would be very close to this. For both the SU and PBM metrics we calculated their corresponding histograms and these are displayed in figures 15(*b*) and 15(*c*) respectively. The spread of both histograms is small, with both having a single significant mode. The mode of the SU metric histogram is located at scale 2, and referring to our previous results we see that this represents a slightly over-segmented result but is still relatively close to the best results located at or close to scale 4. The mode of the BPM metric histogram is located at scale 20 and again referring to previous results we see that this represents an extremely under-segmented result. As discussed earlier, we believe this is due to the failure of BPM to incorporate spatial information. Using the histogram distance defined in equation (12), the distance between the cumulative

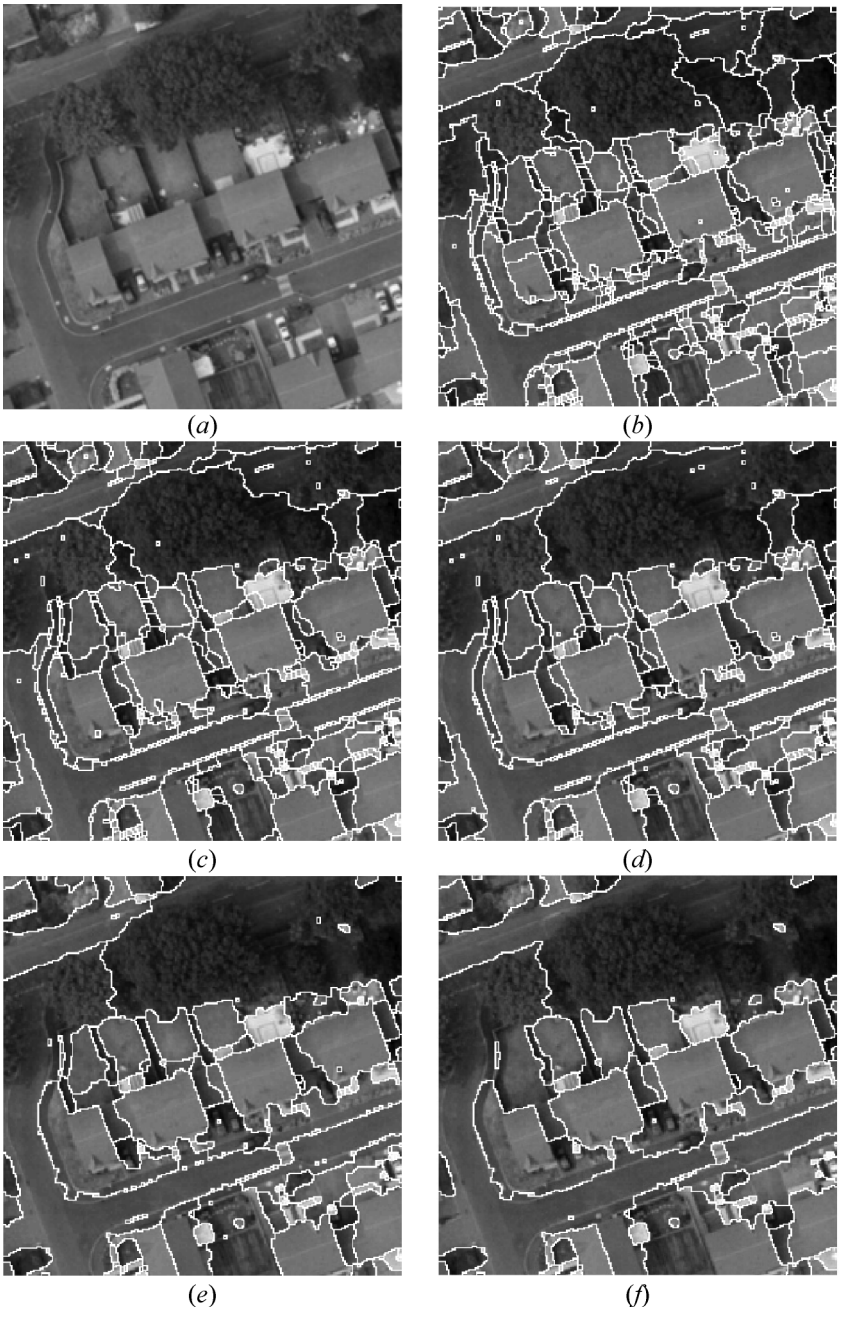


Figure 14. Sample image in (*a*) with corresponding segmentation results at scales (*b*) 2, (*c*) 3, (*d*) 4, (*e*) 5 and (*f*) 6. Segmentation object boundaries are represented by the colour white.

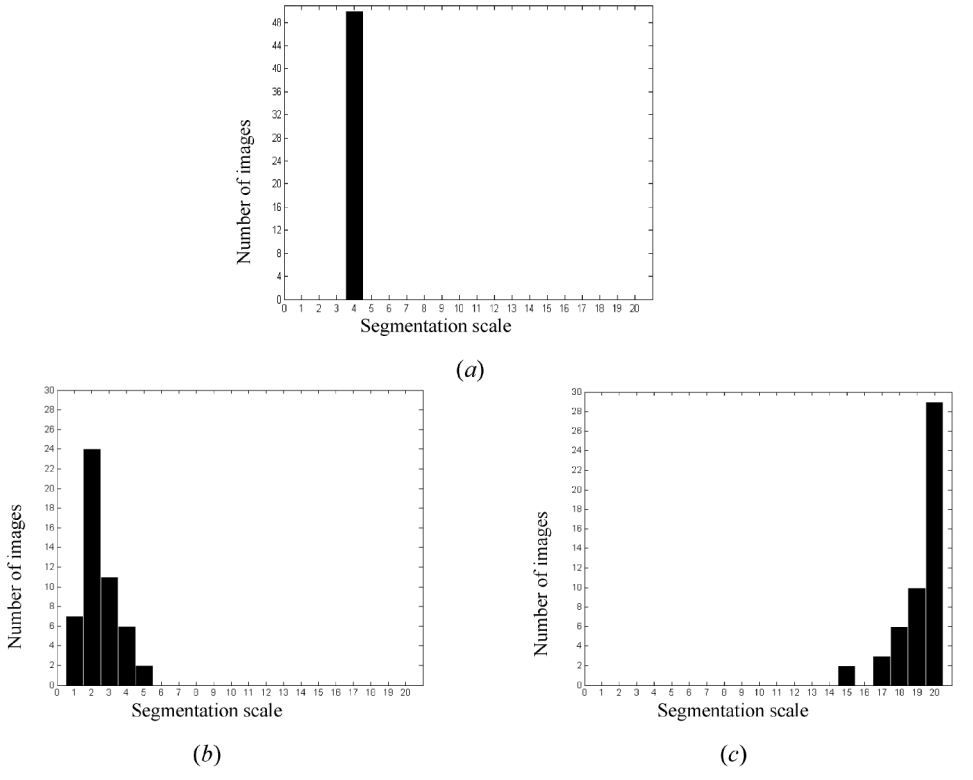


Figure 15. Histograms showing segmentation scale on the x-axis versus on the y-axis a count of the number of images which achieve a maximum at that scale for the optimal result derived by (*a*) visual inspection, (*b*) the SU metric and (*c*) the PBM index. The segmentation scale increases from left to right.

histograms of the optimal histogram defined by visual inspection (figure 15(*a*)) and SU metric (figure 15(*b*)) was calculated to be 82. On the other hand, the distance between the cumulative histograms of the optimal histogram defined by visual inspection (figure 15(*a*)) and PBM metric (figure 15(*c*)) was calculated to be 759. This quantitative result verifies the superior performance of the SU metric compared to the PBM metric and also agrees with our qualitative results.

The fact that the SU metric does not always assign a maximum value to the best result is not a totally unexpected outcome given the metric has a correlation coefficient of 0.72 with an accurate supervised metric. Although this correlation coefficient is high relative to other results published in previous literature, it is still significantly less than the perfect value of 1. Figure 6 illustrates this point. For the synthetic image in this figure the Rand index and SU metric have a very high correlation with a correlation coefficient of 0.87. Despite this the Rand index achieves a maximum at the sixth scale while the SU metric achieves a maximum at the fourth scale. Therefore for an unsupervised metric to be very accurate it must have an extremely high correlation with an accurate supervised metric. This indicates how difficult it is for an unsupervised metric to match the accuracy of a supervised metric.

The segmentation algorithm used to generate the segmentation results in this work operates on the same complementary feature set as that used by the SU and PBM

metrics. It is evident from segmentation results obtained that if this segmentation algorithm is parameterized correctly then it will generate accurate primitive-object segmentation. This indicates that the proposed complementary feature set is an accurate model of the low-level features used in the HVS to define this primitive-object segmentation and satisfies the first of the two requirements for an unsupervised metric stated in §3. The fact that the SU and PBM metrics generally assign a maximum score to a slightly over-segmented and significantly under-segmented result respectively must therefore be an issue with the metric which operates on this feature set and not the feature set itself.

In summary, using two large varied datasets, an in-depth qualitative and quantitative evaluation of both proposed segmentation evaluation metrics was performed. In all results the PBM metric performed poorly and in general targeted results which were significantly under-segmented. In contrast, results achieved by the SU metric were very positive. They indicate that the scale at which the SU metric achieves a maximum is significantly closer to the desired best segmentation. The superior performance of the SU metric relative to the PBM metric demonstrates the need to incorporate the spatial domain when performing unsupervised segmentation evaluation. Quantitative results on a benchmark dataset using a common performance evaluation strategy show the SU metric to outperform existing state-of-the-art metrics for images which contain textured regions.

5. Conclusions

The object-based approach to remote sensing aims to overcome the inaccuracies of pixel-based approaches by incorporating the spatial domain. Modelling the human visual process of primitive-object segmentation is a challenging task. If an accurate implementation is ever to be realized a procedure must be in place to evaluate potential solutions and guide efforts.

An analysis of existing strategies for supervised segmentation evaluation using ground truth shows them to be flawed when applied to primitive-object segmentation. This is due to the inherent inaccuracies of ground truth. Given this difficulty we investigated the possibility of evaluating segmentation in an unsupervised manner without ground truth. Two requirements for any metric which attempts to perform this evaluation were proposed. The first requirement states that any such metric must operate on a feature set which is an accurate model of the low-level features used in the HVS to define primitive objects. The second requirement states that any such metric must operate in the spatial domain. To highlight the importance of the second requirement we proposed two metrics. The first metric, known as the PBM metric, is a function of a complementary feature set and operates in the feature domain. It therefore fails to meet the second requirement. The second metric, known as the SU metric, is also a function of a complementary feature set but operates in the spatial domain and consequently meets both requirements. The PBM metric performs poorly on both a benchmark synthetic dataset and a remotely sensed dataset. In contrast, on the same datasets, the SU metric performs significantly better. An in-depth quantitative and qualitative evaluation indicates that the scale at which the SU metric generally achieves a maximum is very close to the optimal result and significantly closer than that of the BPM metric. The superior performance of the SU metric relative to the BPM metric signifies the need to perform segmentation evaluation in the spatial domain. The accuracy of the SU metric relative to existing metrics is demonstrated by

the fact that on the benchmark synthetic dataset the SU metric outperforms an existing state-of-the-art metric.

As described in this paper the SU metric does not have a perfect positive correlation coefficient of 1 with an accurate supervised metric. Consequently it generally assigns a maximum value to a slightly over-segmented result. Addressing this concern will be the focus of our future work.

Acknowledgements

This work was partly funded by the Irish Research Council for Science, Engineering and Technology (IRCSET) under the Embark Scholarship Scheme and also by a Strategic Research Cluster grant (07/SRC/I 1168) from Science Foundation Ireland under the National Development Plan. The authors would also like to acknowledge the insightful and extremely helpful reviews and comments provide by the editor Prof Costas Varotsos and the anonymous reviewers.

References

Acharyya, M. and Kundu, M.K., 2001, Wavelet-based texture segmentation of remotely sensed images. In *Proceedings of the 11th International Conference on Image Analysis and Processing*, September 2001, Palermo, Italy (Palermo, Italy: IEEE Press), pp. 69–74.

Aplin, P., Atkinson, P.M. and Curran, P.J., 1997, Fine spatial resolution satellite sensors for the next decade. *International Journal of Remote Sensing*, **18**, pp. 3873–3881.

Blaschke, T., 2003, Object-based contextual image classification built on image segmentation. In *IEEE Workshop on Advances in Techniques for Analysis of Remotely Sensed Data*, October 2003, Washington DC (Washington DC: IEEE Press).

Blaschke, T., Lang, S., Lorup, E., Strobl, J. and Zeil, P., 2000, Object-oriented image processing in an integrated GIS/remote sensing environment and perspectives for environmental applications. In *Environment Information for Planning, Politics and the Public*, A. Cremers and K. Greve (Eds), pp. 555–570 (Marburg: Metropolis Verlag).

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

Chabrier, S., Emile, B., Rosenberger, C. and Laurent, H., 2006, Unsupervised performance evaluation of image segmentation. Special issue on performance evaluation in image processing. *EURASIP Journal on Applied Signal Processing*, **2006**, pp. 1–12.

Chen, C.-F. and Lee, J.-M., 2001, The validity measurement of fuzzy c-means classifier for remotely sensed data. In *Proceedings of the 22nd Asian Conference on Remote Sensing*, November 2001, Singapore, pp. 208–211.

Chen, Z., Zhoa, Z., Gong, P. and Zeng, B., 2006, A new process for the segmentation of high resolution remote sensing imagery. *International Journal of Remote Sensing*, **27**, pp. 4991–5001.

Comaniciu, D. and Meer, P., 2002, Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**, pp. 603–619.

Congalton, R.G. and Green, K., 1998, *Assessing the Accuracy of Remotely Sensed Data: principles and practice* (Florida, USA: Lewis Publishing).

Corcoran, P. and Winstanley, A., 2007, Using texture to tackle the problem of scale in landcover classification. In *Object-Based Image Analysis – spatial concepts for knowledge-driven remote sensing applications*, T. Blaschke, S. Lang and G. Hay (Eds), pp. 113–132 (Berlin, Heidelberg: Springer).

Foody, G.M., 2002, Status of land cover classification accuracy assessment. *Remote Sensing of Environment*, **80**, pp. 185–201.

Forsyth, D.A. and Ponce, J., 2002. *Computer Vision: A Modern Approach* (New Jersey, USA: Prentice Hall).

Freixenet, J., Munoz, X., Raba, D., Marti, J. and Cuff, X., 2002, Yet another survey on image segmentation: region and boundary information integration. In *Proceedings of the European Conference on Computer Vision*, , May 2002, Copenhagen, Denmark (Copenhagen, Denmark: IEEE Press), pp. 408–422.

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

Huang, Q. and Dom, B., 1995, Quantitative methods of evaluating image segmentation. In *Proceedings of the IEEE International Conference on Image Processing*, October 1995, Washington, DC (Genoa, Italy: IEEE Press), USA, pp. 53–56.

Jiang, X., Marti, C., Irniger, C. and Bunke, H., 2006, Distance measures for image segmentation evaluation. Special issue on performance evaluation in image processing. *EURASIP Journal on Applied Signal Processing*, **2006**, 1–10.

Kühnert, C., Helbing, D. and West, G.B., 2006, Scaling laws in urban supply networks. *Physica A*, **363**, pp. 96–103.

Martin, D., Fowlkes, C., Tal, D. and Malik, J., 2001, A database of human segmented natural images and its applications to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the International Conference on Computer Vision*, January 2001, Vancouver, Canada (Vancouver, Canada: IEEE Press), pp. 416–423.

Martin, D.R., Fowlkes, C.C. and Malik, J., 2004, Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **26**, pp. 530–549.

McCane, B., 1997, On the evaluation of image segmentation algorithms. In *Proceedings of Digital Image Computing: Techniques and Applications*, December 1997, Auckland, New Zealand (Auckland: APRS), pp. 455–459.

Mitra, P., Shankar, B.U. and Pal, S.K., 2004, Segmentation of multispectral remote sensing images using active support vector machines. *Pattern Recognition Letters*, **25**, pp. 1067–1074.

Neubert, M., Herold, H. and Meinel, G., 2007, Assessing image segmentation quality – concepts, methods and application. In *Object-Based Image Analysis – spatial concepts for knowledge-driven remote sensing applications*, T. Blaschke, S. Lang and G. Hay (Eds), pp. 769–784 (Berlin, Heidelberg: Springer).

Pakhira, M.K., Bandyopadhyay, S. and Maulik, U., 2004, Validity index for crisp and fuzzy clusters. *Pattern Recognition*, **37**, pp. 487–501.

Pal, S.K. and Mitra, P., 2002, Multispectral image segmentation using the rough-set-initialized EM algorithm. *IEEE Transactions on Geoscience and Remote Sensing*, **40**, pp. 2495–2501.

Pal, S.K., Ghosh, A. and Shankar, B.U., 2000, Segmentation of remotely sensed images with fuzzy thresholding, and quantitative evaluation. *International Journal of Remote Sensing*, **21**, pp. 2269–2300.

Rosenberger, C., Chabrier, S., Laurent, H. and Emile, B., 2006, Unsupervised and supervised image segmentation evaluation. In *Advances in Image and Video Segmentation*, Y.-J. Zhang (Eds), pp. 365–393 (Pennsylvania, USA: IRM Press).

Rubner, Y., Tomasi, C. and Guibas, L.J., 1998, A metric for distributions with applications to image databases. In *Proceedings of the IEEE International Conference on Computer Vision*, January 1998, Bombay, India (Bombay, India: IEEE Press), pp. 59–66.

Tan, P.-N., Steinbach, M. and Kumar, V., 2006, *Introduction to Data Mining* (Reading, MA: Addison Wesley).

Theodoridis, S. and Koutroumbas, K., 2006, *Pattern Recognition*, 3rd edn (San Diego, California: Academic Press).

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.

Usamentiaga, R., Garcia, D.F., Lopez, C. and Gonzalez, D., 2006. A method for assessment of segmentation success considering uncertainty in the edge positions. Special issue on performance evaluation in image processing. *EURASIP Journal on Applied Signal Processing*, **2006**, pp. 1–12.

Varotsos, C., 2005, Modern computational techniques for environmental data. Application to the global ozone layer. *Computational Science – ICCS 2005*, PT **3516**, pp. 504–510.

Varotsos, C.A. and Cracknell, A.P., 2004, New features observed in the 11-year solar cycle. *International Journal of Remote Sensing*, **25**, pp. 2141–2157.

Varotsos, C.A., Ondov, J.M., Cracknell, A.P., Efstathiou. M.N. and Assimalopoulos, M.N., 2006, Long-range persistence in global Aerosol Index dynamics. *International Journal of Remote Sensing*, **27**, pp. 3593–3603.

Varotsos, C., Assimakopoulos, M.N., and Efstathiou, M., 2007, Technical note: Long-term memory effect in the atmospheric CO2 concentration at Mauna Loa. *Atmospheric Chemistry and Physics*, **7**, pp. 629–634.

Vecera, S.P. and Farah, M.J., 1997, Is visual image segmentation a bottom-up or an interactive process? *Perception & Psychophysics*, **59**, pp. 1280–1296.

Wirth, M., Fraschini, M., Masek, M. and Bruynooghe, M., 2006, Performance evaluation in image processing. Special issue on performance evaluation in image processing. *EURASIP Journal on Applied Signal Processing*, **2006**, pp. 1–3.

Yang, L., Albregtsen, F., Lonnestad, T. and Grottum, P., 1995, A supervised approach to the evaluation of image segmentation methods. In *Proceedings of the 6th International Conference on Computer Analysis of Images and Patterns*, September 1995, Prague, Czech Republic (Berlin: Springer), pp. 759–765.

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

Zhang, Y.-J., 2006a, An overview of image and video segmentation in the last 40 years. In *Advances in Image and Video Segmentation*, Y.-J. Zhang (Ed.), pp. 1–15 (Pennsylvania, USA: IRM Press).

Zhang, Y.J., 2006b, A summary of recent progress for segmentation evaluation. In *Advances in Image and Video Segmentation*, Y.-J. Zhang (Ed.), pp. 423–440 (Pennsylvania, USA: IRM Press).

Zhu, Z., Yang, L., Stehman, S.V. and Czaplewski, R.L., 2000, Accuracy assessment for the U.S. geological survey regional land-cover mapping program: New York and New Jersey region. *Photogrammetric Engineering and Remote Sensing*, **66**, pp. 1425–1435.