Detecting Man-Made Structures and Changes in Satellite Imagery with a Content-Based Information Retrieval System Built on Self-Organizing Maps

Matthieu Molinier, Jorma Laaksonen, Senior Member, IEEE, Tuomas Häme

Abstract—The increasing amount and resolution of satellite sensors demand new techniques for browsing remote sensing image archives. Content-based querying allows efficient retrieval of images based on the information they contain, rather than their acquisition date or geographical extent. Self-Organizing Maps (SOMs) have been successfully applied in the PicSOM system to content-based image retrieval (CBIR) in databases of conventional images. In this paper, we investigate and extend the potential of PicSOM for the analysis of remote sensing data. We propose methods for detecting man-made structures, as well as supervised and unsupervised change detection, based on the same framework. In our study, a database was artificially created by splitting each satellite image to be analyzed into small images. After training PicSOM on this imagelet database, both interactive and off-line queries were made to detect man-made structures, as well as changes between two very high-resolution images from different years. Experimental results were both evaluated quantitatively and discussed qualitatively, and suggest this new approach is suitable for analyzing very high-resolution optical satellite imagery. Possible applications of this work include interactive detection of man-made structures, or supervised monitoring of sensitive sites.

Index Terms—Content-based information retrieval, Self-Organizing Maps, man-made structure detection, change detection, high-resolution optical satellite images.

I. INTRODUCTION

Earth Observation data volumes are growing rapidly, with an increase in both the number of satellite sensors and in their resolutions. Yet, it is estimated that only 5% of all Earth Observation data collected up to now has been used. Therefore, traditional remote sensing archiving systems – with queries made typically on sensor type, geographical extents or acquisition date – could become obsolete as the amount of data to be stored, accessed and processed, explodes. Using image content would allow a more efficient use of the potential of PicSOM for the analysis of remote sensing data, and land use identified by SVM classification) and spatial information (Gabor wavelet coefficients as a textural feature). Performance evaluation was run on a database of 128 × 128 pixels extracts of Landsat TM images, with promising results. Other work [2], [5], [7] seemed to focus more on managing large databases of full remote sensing scenes.

Very few works describe the utilization of content-based image retrieval techniques for the purpose of a single satellite scene interpretation, let alone change detection. Techniques inspired from knowledge discovery in database (KDD) were however used in [13], for segmenting images of natural scenes. A method for multisource remote sensing data fusion with multiple Self-Organizing Maps (SOMs) [14] has been introduced in [15]. Spectral, geographical, temporal and spatial features were merged in a data vector. But the SOMs were used in [15] for classification, not for information retrieval from remote sensing image archives nor detection of changes.

Change detection in remote sensing imagery is involved in an ever-growing number of applications, allowing to monitor the impact of natural disasters (e.g. floods, or earthquakes [16]), forest changes [17] or even compliance to nuclear non-proliferation treaties [18]. Classical change detection techniques rely e.g. on image difference or ratio, post-classification comparison, classification of multitemporal datasets or change vector analysis [19]. Pixel-based methods require subpixel registration accuracy, as mis-registrations greater than one pixel produce numerous errors when comparing images [19].
We present an original utilization and improvement of a CBIR system, PicSOM, for the analysis of remote sensing images. In the PicSOM image database browsing system [20], several thousands of images are mapped on Self-Organizing Maps, through the extraction of image descriptors including textural and color features. After the SOMs are trained, the user can visually query the database and the system automatically finds images similar to those selected. This approach has been successfully applied to databases of conventional images [21], [22]. We aim at using the potential of this neurally-motivated CBIR system for detecting man-made structures, or changes between two or more satellite images, with a special interest in changes involving human activity (e.g. buildings, structures). The approach presented in this paper could be applied to various types of sensors, even to synthetic aperture radar (SAR) with appropriate feature descriptors. In this study, we focused on very high-resolution optical images, as change detection is challenging in such imagery – due to the high level of details in the images, and sensitivity to illumination conditions or mis-registration effects.

The key idea of our study is to artificially create an “image database” from each satellite image to be analyzed, by clipping it into thousands of small images, or imagelets. PicSOM can be trained on that virtual database, then visually queried for finding objects of interest like man-made structures. Previous work [23] showed promising results on very high-resolution satellite images, even with features that had been developed for generic images, and did not take into account the spectral or spatial properties of Earth observation (EO) data.

In the present study, we extended PicSOM system so that it handles multispectral imagery and includes features adapted to the detection of man-made structures in high-resolution optical satellite images. Fusion of panchromatic and multispectral information was done naturally within PicSOM, in which several SOMs are trained in parallel (one SOM per feature). Qualitative and quantitative evaluation of the methods were provided for man-made structure detection and change detection, using partially labelled datasets. Possible applications of this work are high-resolution satellite image annotation, or monitoring of sensitive areas for undeclared human activity, both in an interactive way.

This paper is organized as follows. The next section explains how we built a database from one or two satellite images. The Self-Organizing Map and PicSOM are introduced in Section III, and the features extracted for the study in Section IV. Section V explains the use of PicSOM for retrieving man-made structures and detecting changes. Experimental results are presented in Section VI and discussed in Section VII. Conclusions and future work are given in Section VIII.

II. DATABASE PREPARATION

Content-based image retrieval is a technique for finding relevant images in a database. We here present how we created a database of images from one or two satellite images.

A. Satellite imagery

Two QuickBird scenes were acquired in the beginning of September 2002 and in mid June 2005, covering a same coastal area in Finland. They include 4 multispectral channels with a 2.4 m ground resolution – blue (450 – 520 nm), green (520 – 600 nm), red (630 – 690 nm) and near-infrared NIR (760 – 900 nm), and a panchromatic channel (450 – 900 nm) with ground resolution of 0.6 m. Both scenes were remarkably cloud-free, and were co-registered one to another. Slight mis-registration effects remained due to lack of a Digital Elevation Model of the monitored area.

Panchromatic and multispectral images were kept as separate data sources for later feature extraction. For the sole purpose of displaying images in PicSOM, fusion of panchromatic and multispectral images was done by pansharpening – no feature was calculated on the pansharpened images. Pansharpening was also used to check the correct registration between multispectral and panchromatic channels. A same study area of size about 4 × 4 km, shown in Fig. 1, was extracted from both scenes.

Fig. 1. True-color pansharpened 2005 QuickBird study scene.

B. Database generation

An image retrieval system, such as PicSOM, typically requires several thousands of images in a database in order to produce relevant indexing. We artificially created databases from the two study scenes to be analyzed. Each scene was cut into 71 × 71 = 5041 non-overlapping small images or imagelets, of size 100 × 100 pixels for panchromatic data and 25 × 25 pixels for multispectral data.

By this operation, we expected to reduce the amount of contents in each image. The study scene on Fig. 1 is a rich, detailed image with lots of target classes (water, forest, arable land, built-up areas, etc.). On the contrary, imagelets, like shown on Fig. 2, usually contain only two or three distinct target classes. The pansharpened images were cut in the same manner (into 100 × 100 pixels imagelets), to be included in PicSOM for visualization.

In this study, imagelets were named in such a way that it tells their location within the study area – row and column of the original image – and year of acquisition. A triplet (row, column, acquisition date) designated unambiguously an imagelet, regardless of the data type it had been extracted from. We denote by database the reunion of imagelets coming from panchromatic, multispectral and pansharpened data sources, for a given geographical extent – here the study scene. Each acquisition date generates a sub-database of the parent...
In this section we describe the principle of the PicSOM content-based retrieval and browsing system based on Self-Organizing Maps (SOMs). PicSOM has been developed as a framework for studies on content-based information retrieval, and it has been applied in divergent domains of both visual and non-visual data. A central design principle of the system has been its good scalability to large database sizes by efficient off-line indexing of the objects. A demonstration version of PicSOM is displayed at [20].

### A. Self-Organizing Maps

The Self-Organizing Map [14] is a neurally-motivated unsupervised learning technique which has been used in many data-analysis tasks. A genuine feature of the Self-Organizing Map is its ability to form a nonlinear mapping of a high-dimensional input space to a typically two-dimensional grid of artificial neural units. During the training phase of a SOM, the model vectors in its neurons get values which form a topographic or topology-preserving mapping. Through this mapping, vectors that reside near each other in the input space are mapped into nearby map units in the SOM grid. Patterns that are mutually similar in respect to the given feature extraction scheme are thus located near each other on the SOM surface.

The training of a Self-Organizing Map starts from the situation where the model vectors \( \mathbf{m}_i \) of each map unit \( i \) are initialized with random values. For each input feature vector or sample \( \mathbf{x}(t) \), the “winner” or best-matching map unit (BMU) \( c(\mathbf{x}) \) is identified on the map by the condition

\[
\forall i : \| \mathbf{x}(t) - \mathbf{m}_{c(\mathbf{x})}(t) \| \leq \| \mathbf{x}(t) - \mathbf{m}_i(t) \|, \tag{1}
\]

where \( \| \cdot \| \) is commonly the Euclidean norm. After finding the BMU, a subset of the model vectors constituting a neighborhood centered around node \( c(\mathbf{x}) \) are updated as

\[
\mathbf{m}_i(t + 1) = \mathbf{m}_i(t) + h(t; c(\mathbf{x}), i)(\mathbf{x}(t) - \mathbf{m}_i(t)). \tag{2}
\]

Here \( h(t; c(\mathbf{x}), i) \) is the “neighborhood function”, a decreasing function of the distance between the \( i \)-th and \( c(\mathbf{x}) \)-th nodes on the map grid. The training is reiterated over the available samples, and the value of \( h(t; c(\mathbf{x}), i) \) is allowed to decrease in time to guarantee the convergence of the prototype vectors \( \mathbf{m}_i \). Large values of the neighborhood function \( h(t; c(\mathbf{x}), i) \) in the beginning of the training initialize the network, and small values on later iterations are needed in fine-tuning. The outcome of training a SOM is essentially the topographic ordering of the training data samples on the surface of the map grid.

### B. PicSOM system

The PicSOM system [20] used in this study has originally been developed for content-based image retrieval (CBIR) research [21], [22], [25]. It implements two essential CBIR techniques, query by examples (QBE) and relevance feedback. These methods can be used for iterative retrieval of any type of visual or non-visual content. In iterative QBE, the system...
presents in a visual interface some images or other objects to the user, who then marks a subset of them as relevant to the present query. This relevance information is fed back to the system, which then tries to find more similar objects and returns them in the next query round. In Section V we show how the same technique can also be applied in the semi-automated, interactive analysis of satellite images – for e.g. detecting man-made structures.

In PicSOM, multiple SOMs are used in parallel, each created with different low-level visual features. Each imagelet is therefore used multiple times as input to different feature extraction techniques and the resulting feature vectors are used in training different SOMs. When an image SOM has been trained, all the feature vectors \( x \) are once more mapped to it, each one in its best-matching unit. Every SOM unit and its model vector \( m \) is then assigned a visual label from the imagelet whose feature vector was the nearest to the model vector. Figure 3 demonstrates how the feature vectors, BMUs and visual labels can situate in an imagined two-dimensional feature space. The tessellation regions determine which imagelets will be represented by each SOM unit in the index defined by that specific feature extraction scheme.

![Fig. 3. A visualization of the distributions of the feature vectors of images (crosses) and the SOM model vectors (circles) in a two-dimensional space.](image)

The different feature extraction schemes and the SOMs that result from the training impose different similarity functions on the images. As a consequence, two imagelets whose contents are similar to each other with respect to some feature will be mapped to nearby units on that SOM’s surface, whereas they will be located far apart on another feature’s SOM if that feature regards them as dissimilar. Every interactive image query is unique and each user of a CBIR system has his or her own transient view of images’ similarity and relevance to the specific query. Therefore, a system structure capable of holding many simultaneous similarity representations can adapt to different kinds of retrieval tasks. Also in studies made by other researchers, it has been shown that joint multiple SOMs have better discrimination capability than a single SOM [15].

Relevance feedback has been implemented in PicSOM by using the parallel Self-Organizing Maps. Each image presented in the interactive on-line use of the system is graded by the user as either relevant or non-relevant. All these relevance grades are then projected to the best-matching SOM units of the graded images on all the different SOM surfaces. Maps where there are many relevant images mapped in same or nearby SOM units agree well with the user’s conception on the relevance and semantic similarity of the images. When we assume that similar images are located near each other on the SOM surfaces, we are motivated to spread the relevance information placed in the SOM units also to the neighboring units.

This relevance spreading is implemented in PicSOM as follows. All images marked as relevant are first given an equal positive value inversely proportional to their number. Likewise, non-relevant images receive negative values that are inversely proportional to the number of non-relevant images. The overall sum of these relevance values is thus zero. The values are then summed in the BMUs of the images, and the resulting sparse value fields are low-pass filtered to produce the final score or qualification value for each SOM unit.

As the final step, each image is given a total qualification value obtained as a sum of the qualification values from its best-matching or index units from the SOM surfaces of the different features. Those yet unseen images which have the highest qualification values will then be shown to the user on the next query round. In PicSOM, features that fail to coincide with the user’s conceptions always produce lower qualification values than those descriptors that match the user’s expectations. As a consequence, the different features do not need to be explicitly weighted as the system automatically takes care of weighting their opinions.

Qualification values can similarly be calculated off-line for different semantic content classes, if such have been defined for the database. Suppose that all images in a specific ground truth class are marked as relevant, with a positive value in their respective BMUs on the different SOMs. After low-pass filtering, we then obtain class distributions on the SOM surfaces as can be seen in Fig.6. The better a feature is in mapping the semantic class in a dense region, the sharper the resulting distribution on the SOM surface will be.

### IV. Features

Feature extraction and selection is a central task in any content-based image retrieval system. The features are usually low-level, and describe general properties of the images, like color distribution, texture or shapes. In [11], weighted color histograms, Hough transform as a shape descriptor, and textural information based on the Fourier transform were used as features. In [12], features like Haralick descriptors, Gabor features, Gauss-Markov Random Fields parameters were included, as well as geometric features resulting from local statistical analysis of the edges contained in images.

In our study, we used six low-level features automatically extracted from the imagelets. Some were part of PicSOM, three were added in this study for the specific purpose of detecting man-made structures or changes in satellite imagery.
A. Features included in the standard PicSOM

The PicSOM system was originally developed for databases of conventional images (those found e.g. in common web image databases). Generic and low-level images features are integrated within the PicSOM system. We here use the term imagelet to equivalently describe any data source (panchromatic or multispectral) covering the geographical extent of that imagelet. The features were:

- **average rgb** This 3-dimensional feature was calculated as the average values of the red, green and blue (RGB) channels of the pixels in an imagelet.

- **color moments** The RGB color coordinates of the pixels were first transformed to the hue–saturation–value (HSV) coordinate system. Then the three first moments (mean, variance and skewness) of the HSV values were calculated and stored in a 9-dimensional feature.

- **texture** This feature was formed by studying the 8-neighbors of each imagelet pixel. For each 8-neighbor of a given pixel, a counter was incremented when the illumination in that neighbor was larger than in the center pixel. When all imagelet pixels had been scanned, the final counts were divided by the total number of pixels in the imagelet. The resulting 8-dimensional feature vector describes local illumination differences, and thus the small-scale texture of the imagelet.

In a preliminary work [23], the above three low-level image features were extracted from pansharpened RGB imagery. Pansharpening was used in an attempt to keep spatial and spectral information in a same data source. In addition, the radiometry had been reduced from 11 bits to 8 bits, so that images could be displayed in the PicSOM interface. Despite the extraction of features from pansharpened and radiometrically reduced images, results were encouraging.

In the present study, the full 11 bit radiometry was conserved, and data sources were kept separated. Pansharpening was still used to display imagelets in PicSOM, but strictly not for any feature extraction. Those generic features were extracted on the most relevant input data source – see table I.

B. Additional features for satellite imagery

Other features can as well be calculated in another environment then loaded into PicSOM as plain text files. In order to help retrieving man-made structures and detecting changes in satellite imagery, three additional features were included:

- **xy-coordinates** This feature used the spatial location of the imagelet in the original scene as a 2-dimensional feature (row and column indexes). As the images from the two years were co-registered, the value of this feature was always the same for any imagelet position, regardless of the year, image content or data source.

- **NDVI histogram** The Normalized Difference Vegetation Index, or

\[
\text{NDVI} = \frac{\text{NIR} - \text{Red}}{\text{NIR} + \text{Red}},
\]

where NIR means the near-infrared spectral wavelength and Red the visible red light, is a widely used feature in remote sensing to characterize the amount of green vegetation. The bare soil gives low NDVI values whereas the values of green vegetation are high. In land surfaces the index value is usually positive. The NDVI has also a good correlation with the green biomass at low biomass levels, but it saturates when the bare soil fraction is close to 0%.

In this study the NDVI was utilized to separate vegetated surfaces from non-vegetated, such as buildings and roads. Imagelets are likely to have often heterogeneous content, thus the mean or standard deviation of NDVI on the imagelets would not effectively characterize them. Therefore a NDVI histogram was computed on imagelets, with 100 bins ranging from minimum to maximum NDVI values of the uncut study scene. Histograms of different imagelets could easily be compared since the bin positions were independent of imagelets. Negative values of NDVI, usually associated to water or deep shadows, were left unchanged because the study scene contains a significant proportion of water areas, that were part of the zone to monitor. The allure of NDVI histograms was hypothesized to help discriminating imagelets containing e.g. buildings from the completely vegetated imagelets. The counts of the NDVI histogram constituted a 100-dimensional feature vector.

- **edges histogram** It was forecasted in [26] that, in addition to spectral features, structural features (like edges or shapes) would be suited to content based image retrieval in high-resolution satellite images. Pattern directionality [27] has been used in automatic video indexing [28], and recently in systems to manage Earth observation image archives [29]. It is based on the histogram of local gradient magnitude against directional angle. We used a slightly modified version of that feature. Let \( I \) be an imagelet, \( S_x \) and \( S_y \) the \( 3 \times 3 \) Sobel operators extracting vertical and horizontal edges. Noting \( * \) the convolution operation, the imagelet gradient magnitude \( \| G \| \) and angle \( \theta \) are defined at each pixel by:

\[
\| G \| = \sqrt{G_x^2 + G_y^2}, \quad \text{with } G_x = I * S_x \text{ and } G_y = I * S_y
\]

\[
\theta = \tan^{-1} \left( \frac{G_y}{G_x} \right) + \frac{\pi}{2}, \quad \theta \in [0, \pi]
\]

The angle \( \theta \) is counted counter-clockwise and equals 0 when the gradient \( G \) is in the horizontal direction. In [28], a histogram is obtained quantizing \( \theta \) and counting how many pixels have a gradient magnitude \( \| G \| \) higher than a non-critical threshold. Because shadows or details can generate in very high-resolution satellite images strong edges that do not correspond to structures, we avoided choosing a threshold in our study. Instead, for each angle, the histogram count was weighted by the mean of the two extremal gradient magnitudes (minimum and maximum values). Typically, simple buildings are rectangular, with corresponding edges appearing in directions \( \theta_0 \) (direction of the edge with the strongest magnitude) and \( \theta_0 + \frac{\pi}{2} \) mod \( \pi \).
The histogram was then centered around $\theta_0$, by circular shifting. After conversion to degrees, the result was a histogram of gradient magnitude between angles $-90^\circ$ and $+89^\circ$, normalized in angular position. If an imagelet contains a rectangular building, the histogram should show a primary peak in the central position and secondary peaks at both ends. Imagelets containing only forest should show a flat, close to uniform histogram – no direction of the gradient would be clearly dominating.

Figure 4 shows sample imagelets from the 2002 image, with their NDVI and edges histograms. A single building (first imagelet) generates enough gradient magnitude to leave a signature at the center of the edges histogram, and two secondary lobes in the outermost parts. The NDVI histogram shows high values for vegetation, and values closer to zero for the building. The NDVI histogram of the second imagelet, containing forest only, does not include as low NDVI values as the first imagelet with the building. Also, the NDVI histogram is symmetrical around the modal value, whereas the edges histogram is rather uniform and reveals no man-made structure in the imagelet. The third imagelet has a clear bimodal NDVI histogram because it contains forest and a ploughed field with bare soil, in approximately equal proportions. The edges histogram does not show any clear structure here either.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Data source</th>
<th>Dimensionality</th>
</tr>
</thead>
<tbody>
<tr>
<td>average RGB</td>
<td>Multispectral (RGB only)</td>
<td>3</td>
</tr>
<tr>
<td>color moments</td>
<td>Multispectral (RGB only)</td>
<td>9</td>
</tr>
<tr>
<td>texture</td>
<td>Panchromatic</td>
<td>8</td>
</tr>
<tr>
<td>$xy$-coordinates</td>
<td>Index of the imagelet</td>
<td>2</td>
</tr>
<tr>
<td>NDVI histogram</td>
<td>Multispectral (Red and NIR)</td>
<td>100</td>
</tr>
<tr>
<td>edges histogram</td>
<td>Panchromatic</td>
<td>180</td>
</tr>
</tbody>
</table>

TABLE I
Features extracted from the imagelets

The distributions of the ground truth classes on the six different feature SOMs are shown on Fig. 6. The more localized the distributions of feature vectors are on the map, the better it is for detection. In that sense color moment and edges features seem to be good candidates for building detection.

V. USING PICSOM FOR DETECTING MAN-MADE STRUCTURES AND CHANGES

The PicSOM system was originally developed for browsing databases of images. We here explain how we used and extended PicSOM for detecting man-made structures, or changes – both in supervised and unsupervised manners.
A. \textit{Interactive detection of man-made structures with CBIR}

Content-based image retrieval techniques have already been used for remote sensing image archive management. In [11], the user is asked to label several pictures as relevant or non-relevant at each query round. A study reported in [29] also includes query by pictorial examples.

In our study, we have used the PicSOM CBIR system to interactively find imagelets containing man-made objects such as buildings or roads. The system first displays a random selection of imagelets in a web browser. The user then selects all imagelets containing man-made objects – or anything else but water and forest – and sends this information back to the system by pressing the ‘Continue query’ button. In the forthcoming query rounds, the user can then focus the query with their selections more precisely to more specific semantic targets, such as buildings, roads or clearcuts. Fig. 7 shows the user interface of the system in the middle of an interactive query session. The user has selected some man-made objects shown in the middle of the browser window. In the top part, the distribution of those imagelets are shown on the six different SOMs. In the bottom of the interface, some of the new imagelets returned by the system are shown to the user.

B. \textit{Automated detection of man-made structures}

As described in the end of Section III-B, models for semantic imagelet classes, such as buildings or other man-
made structures, can be created with different features from the ground-truth-labeled training data. Imagelets from different areas or different acquisition dates can then be classified by the system to detect the presence of similar structures. In order to achieve this, the qualification value assigned to each imagelet by the PicSOM system is simply interpreted as a discrimination value, which indicates the likeliness that the specific imagelet belongs to that semantic class.

After calculating the discrimination value, PicSOM will also order the test data by their qualification values in the decreasing order of similarity to the class model. If one, instead of the ordered list, desires the system to give a decisive output for each imagelet (either belonging or not to the studied semantic class), a threshold value has to be selected. Standard pattern classification techniques for selecting the threshold and evaluating the performance of the classification system can then be used. Such techniques include e.g. receiver operating characteristic (ROC) curves [30] and recall–precision graphs.

One may note that the above-described detection procedure is analogous to the interactive CBIR operation, in the sense that the labeled training data has now replaced the imagelets graded by the user as relevant. Similarly, the test data is in the role of the yet unseen database images.

C. Supervised content-based change detection

We outlined above how the PicSOM system can be used to create a multi-featural discrimination model for a semantic class of imagelets. If the model is applied to a pair of imagelets, both depicting the same geographical location, but acquired at different dates, one can perform change detection. Assuming that the system produces larger discrimination values for imagelets that portray man-made structures than those that do not, then a temporal increase in this value indicates that a new man-made structure has probably appeared. Vice-versa, if the discrimination value decreases between the two acquisition dates, then it is likely that a building or other non-natural structure has disappeared.

Again all imagelet pairs can be sorted by decreasing order of change in this discrimination value. In that ordering, the first imagelets are those that most probably depict areas where a new building or other structure has appeared. By selecting a proper threshold value or by taking a preset number of the top-ranked imagelets, one can then check if new structures really have appeared.

It must be noted that this kind of change detection is supervised in the sense that one needs to have labeled training data in order to create the class models. The models can be used for change detection in data temporally and geographically independent from the training data, as long as the training samples from that semantic class still match the analyzed imagelets accurately enough. Therefore, models created for detecting buildings in Northern Europe might not be applicable e.g. to Mediterranean regions.

D. Unsupervised content-based change detection

We also devised an unsupervised method for finding pairs of imagelets – one from year 2002 and the other from 2005 – which differed the most in the sense of some of the extracted features. The dissimilarity between imagelets was again defined solely on the Self-Organizing Maps. This means that we did not calculate any pixel-wise differences between the imagelets, but defined the changes relatively to a particular feature extraction scheme. In this manner, change detection is less dependent on small variations in the absolute image coordinates due to inaccurate co-registration.

Some variations in the imagelets are due to various forms of noise (e.g. varying illumination conditions, overlapping shadows), and do not correspond to true changes in the land cover. We assumed that the differences caused by noise lead to situations where the best-matching unit for the calculated feature vector remained the same, or was moved to a neighboring SOM unit. Only true changes in the imagelet’s content should then give rise to such a striking change in the feature vector’s value, that its projection on the SOM surface would be moved to a significantly different location. The substantiality of the change can therefore be measured as the distance on a SOM between the best-matching units of the imagelet’s feature vectors from different years.

The proposed unsupervised change detection technique was as follows. For each imagelet pair from years 2002 and 2005 – excluding the water areas – we solved the two BMUs on one particular feature’s SOM. The Euclidean distances between the BMUs were then calculated, and the imagelet pairs ordered by descending pair-wise BMU distance. A
number of imagelet pairs were then regarded as the locations where the most substantial changes had taken place. Fig. 8 presents the unsupervised change detection approach based on Self-Organizing Maps. The basic framework – decomposition of images into imagelets, feature extraction and mapping on the SOM – is common to content-based supervised and unsupervised detection of man-made structures or changes.

The same procedure can be repeated for all the five visual features defined in Section IV. A feature selection stage can also determine the combination of features that performs best for a given change detection task. In case more than one feature was involved in the change detection framework, the between-BMU distances on different feature SOMs had to be combined. The distances were so first normalized on each feature map separately, into the range [0, 1], by histogram equalization. That process aimed to compensate for cases when the average distances were larger on one map than on the others. These equalized distances are then summed to produce the final value indicating the amount of change.

One must note that this unsupervised detection scheme treats all kinds of changes equally. As a result, one cannot expect good detection performance for any specific type or semantic class of changes. For such needs, provided that labeled training data is available, the supervised detection technique described in the previous section will be more appropriate.

VI. RESULTS

The proposed methods were evaluated quantitatively and qualitatively for different tasks, including interactive and off-line detection of built-up areas, and supervised or unsupervised change detection. The proposed change detection methods were also quantitatively compared to other change detection methods.

A. PicSOM for detection of buildings

The performance of the proposed methods for detecting buildings in very high-resolution images was evaluated quantitatively by ROC curve analysis, and both quantitatively and qualitatively in an interactive experiment on a validation set.

1) Using the 2002 training set to detect buildings in the 2005 set: Considering imagelets from the study scene and their associated labelling in 2002 and 2005, we searched for a good combination of features to detect buildings. A training set was made of the 5041 imagelets of 2002 from study scene, while the testing set contained the imagelets of the 2005 study scene. Those sets included 244 and 266 buildings, respectively. Two-fold cross-validation on the training set was used with sequential forward selection (SFS) [31] to select those features which maximize the area under the ROC curve [30], denoted here as auc. NDVI, edges and color moment features were found to form the optimal feature set, whose combination gave the highest area under curve auc = 0.94. Some ROC curves for the testing set are shown in Fig. 9. The color moment feature alone performed well too, but it did not have as good a sensitivity in the beginning (left part) of the ROC curve as the combination of those three features.

2) Interactive validation: In this experiment, the feature SOMs were trained on the imagelets of both years from the study scene. A validation set was taken from the same original QuickBird images, but in different geographical extents. This set was cut into 60 × 90 = 5400 imagelets for each year, in the same manner as the training set in Section II. PicSOM interface was then used to retrieve, from the 2005 validation set, imagelets that contain buildings.

Two sub-experiments were conducted, repeated by four persons, two of them being external to the present work. In experiment A, the plain content-based image retrieval approach with query by pictorial example and relevance feedback was evaluated. In the first query round, PicSOM presented to the user 20 imagelets picked randomly from the validation set. The user would then select those imagelets containing man-made structures other than roads (e.g. buildings or boats). These imagelets marked as relevant were fed back to PicSOM, that returned in the next query round 20 other imagelets similar to those already selected. The experiment was stopped after 10 query rounds.

In experiment B, feedback was brought to PicSOM both by the user (relevance feedback as in A) and by the class model (imagelets containing buildings in 2002 training set), in equal weight. The initial selection of 20 imagelets presented to a same user was identical in both experiments, but different between the four experimenters.

The number of relevant images returned versus query round is shown in Fig. 10. The lines are the average results over the four experimenters, with standard deviation ±σ marked by ‘*’ and ‘o’. On the average, only 1.5 imagelets from the initial random set of 20 imagelets contained a building. There were probably only a few imagelets containing houses in the validation set, as there were in the training set (only about 200). In both experiments, PicSOM returned a comfortable majority of relevant images already after three query rounds. The system seems to perform slightly better when the class model is integrated into the feedback mechanism. However, the result of experiment A suggests that the plain CBIR system is also performing well in interactive building detection.

Similar results were achieved when visually selecting clearcuts or arable land as a target. This already shows a use-case of the PicSOM system with remote sensing data, as a supervised, general-purpose and interactive tool for satellite image annotation, by visual and intuitive querying. A proposed
Fig. 8. Overview of the SOM-based unsupervised change detection framework. Features are extracted from appropriate data sources - panchromatic or multispectral - in the earlier (2002) and later (2005) imagelet. The imagelets are then mapped on the SOM grid, where the distance between the best matching units (BMU) reflects the change magnitude between the two dates.

Fig. 10. Average number of relevant imagelets versus query round, without (A) and with (B) the class model.

visual output consists of “lighting up”, in the studied satellite image, only the imagelets in which objects of interest have been detected – e.g. buildings in Fig. 11.

B. Supervised change detection

Our proposed supervised change detection method was evaluated in its ability to find buildings that have appeared between years 2002 and 2005. The labelled database contained 40 imagelet pairs where this change really took place. As an independent testing set was not available, we selected the optimal set of features for change detection by optimizing ROC auc value for the training set of 2002–2005 imagelet pairs with the SFS algorithm. A baseline change detection method was also used in quantitative comparison.

1) Baseline: As a baseline for supervised change detection method, we used post-classification comparison [19]. Panchromatic and multispectral data were fused and classified by standard maximum likelihood. Prior probabilities of classes were set to 0.3 for water and forest, and 0.08 for the other 5 classes. Training areas for each class were manually delineated on the 2002 and 2005 full scenes by visual inspection. For each year, exactly half of the buildings from the ground truth built in Section II-C were used in training.

In order to allow comparison with our method, the two classification images were cut into imagelets. For each imagelet pair, the number of pixels classified as buildings in 2005 and that were not buildings in 2002, was counted. Focusing on the detection of appeared buildings, imagelet pairs were sorted by this decreasing value and compared to the ground truth, to build the ROC curve in Fig. 12.

2) Proposed method: It turned out that the best detection accuracy was obtained by using the edges feature alone. Its ROC curve with auc = 0.87 is shown in Fig. 12. Other features and feature combinations performed rather poorly in this task, with an auc at best close to 0.6. Our content-based supervised change detection method is more accurate than the baseline method at the beginning of the ROC curve, with less false positives. Later the baseline outperforms the proposed method. However in the baseline method half of the buildings in the 2005 were manually delineated during training, whereas our method didn’t use any knowledge about buildings in the 2005 image – only in 2002.

Setting the detection threshold in our method at 71 imagelet pairs, 12 true positives were returned (effective changes in built-up areas) and 59 false positives. A higher threshold value returned 11 true positives, and 19 false positives. These 30 locations where changes were detected are shown with circles and red color over the test area in Fig. 13.

C. Unsupervised change detection

We evaluated our unsupervised change detection method with the ground-truth available for buildings in 2002 and 2005. Imagelets containing only water were discarded – they would otherwise had contributed too much in the set of the
“most changed” imagelets. The best feature combination was searched with SFS, maximizing area under ROC curve.

1) Baseline: We considered some variations of image difference [19] as baseline methods for unsupervised change detection. Difference images were computed for panchromatic and each multispectral channel. The square root of the sum of square differences between each multispectral channel was also computed, resulting in one change magnitude image. No threshold was applied to the change magnitude images. They were cut into imagelets, that were then sorted by decreasing order of the sum of the magnitude change over all imagelet pixels. By comparison to the ground truth, ROC curves were produced — Fig. 14.

2) Proposed method: The best result was obtained with the combination of average rgb and edges features, which gave \( \text{auc} = 0.63 \) as shown in Fig. 14. The result is quite reasonable as the procedure detects all changes, not only changes in buildings. Therefore, false positives appear quite often, but not so often as with random picking. However, if we would like to really demonstrate this method’s potential, we would need a ground truth that includes all changes, not only in buildings.

Image differencing on panchromatic channel performed better than our method. Similar results were obtained with the NIR channel only, although slightly less accurate. The three optical channels considered independently in image differencing performed quite poorly, with a ROC curve below the diagonal, similar to that obtained with the combination of all multispectral channels — Fig. 14.

The choice of the threshold (how many imagelet pairs are considered to exhibit striking changes) is problem-dependent both in the supervised and unsupervised methods. For instance, when monitoring a strategic site, missing an appearing target is worse than returning false positives. In earlier work [23], we heuristically proposed to set the threshold at the square root of the number of imagelet pairs in the database. We empirically observed that areas where color moment feature had substantial differences matched well with clearcuts — i.e. forested areas that had been cut in between 2002 and 2005.

VII. DISCUSSION

A. Detection of man-made structures

The set of features that performed best for detecting buildings in 2005 after training PicSOM on the 2002 database (Section VI-A) was formed by NDVI, color moment and edges features. In those cases the surrounding or context of a building is a useful hint, because buildings usually appear in groups and in mutually similar environments. Yet, this feature combination did not work well for supervised change detection. This may be because color moments and NDVI very seldom change when a new building is built, as they tend to appear in the vicinity of existing buildings. Therefore, feature combinations that are good for detecting buildings are not necessarily good for detecting changes in buildings. In other words, context may be important for detecting a building, but it is often non-changing when such a building appears or disappears in a short time frame, like three years in our study.

The targets of interest were small-sized buildings, which made their detection even more challenging. Small buildings can easily remain undetected by feature extraction if masked by high trees or shadows. Detection may also fail when the building is located on an imagelet border, thus only a part of it is present in the imagelet. Moreover, the features should account for different target sizes, and adapt to various spatial resolutions, depending on the sensor. The importance of scale for satellite image description has been emphasized in [26].
B. Change detection

Pixel-based change detection in very high-resolution imagery is a challenging task, limited by the requirement of pixel or sub-pixel accuracy registration. In the context of change detection, a clear advantage of the decomposition in imagelets is that it relaxes this constraint. Slight misregistrations remained between the 2002 and 2005 images, but it did not influence the efficiency of our methods.

A building appearing between 2002 and 2005 was a rare event – only 40 imagelet pairs in the study database. In addition, if the earlier imagelet already contains a building, the appearance of a new building in the later image could remain undetected. In such a case, the increase of the discriminative value associated to buildings might not be striking enough that it would suggest a new building has appeared. Furthermore, our current ground-truth labeling does not take into account the number of objects of a given class in an imagelet.

The proposed change detection method performed slightly better than the baseline method in the supervised case, while needing less user interaction than supervised post-classification change detection. However our unsupervised change detection method didn’t perform as well as simple image differencing on panchromatic or NIR channel. This may be due to the fact the features were not tuned for unsupervised detection.

Some changes were attributed to locations containing in fact the same land cover (often forest) in 2002 and 2005. Those cases were mainly vegetation changes between beginning and end of summer. While this could be interesting for season monitoring applications, it dragged PicSOM away from the goal of detecting appearing or disappearing man-made structures. A clear definition within the system of what changes are of interest is needed. How this should be done remains open. This could be partly circumvented by radiometrically calibrating images before loading them into PicSOM.

A way to refine our change detection method would be to provide two content targets to PicSOM: a content from which the change occurs (earlier target), and a content to which the change occurs (later target). This would allow an intuitive and interactive definition of interesting changes by the user – e.g. by selecting imagelets containing forest as the earlier target, and buildings as the later target, the system would detect newly constructed buildings in forested areas.

The approach developed here could handle more than two satellite images for change detection applications. One could train PicSOM on the imagelets extracted from all images available up to a given time, then query the database for imagelets representative of the objects of interest. The system would then return imagelets where significant changes have been detected in the later scene, according to the distance on a feature SOM between the BMUs of earlier and later imagelets.

C. Size of imagelets and objects of interest

The influence of the imagelet size has to be studied more carefully. If smaller than the objects of interest, it is expected that an imagelet would not restitute all spatial or spectral properties of the target. On the other hand, larger imagelets would contain a too great proportion of perturbing, non-interesting content compared to the spatial extensions of the objects of interest – typically in our study scene, a small isolated building surrounded by forest. In that case the imagelets would be clustered in the SOM according to their dominant content, which would not necessarily be the content of interest.

Therefore the size of imagelets has to be adjusted so that the information they contain is representative of the target. Similarly, the imagelet size should adapt to the scale of interesting changes. In an operative system, the expected dimensions of the interesting targets or scale of changes could be specified by the user or pre-set, depending on the application, then used to determine an appropriate imagelet size.

The 100 \times 100 pixel imagelets, extracted from QuickBird images, seemed to provide a trade-off between the two undesirable situations. Luckily (and surprisingly), not too many buildings in the study scene were split into two or more imagelets. In order to reduce the consequences of “cutting” an object of interest into several non-overlapping imagelets (namely, generating “artificial” objects on the borders of imagelets), overlapping imagelets could be used.

The methods should also make use of spatial information with respect to imagelets: the neighboring imagelets of a given imagelet may bring useful information for the analysis of satellite images. For example, built-up areas are often concentrated, and the proposed methods currently do not specifically handle objects bigger than the imagelets. The use of the coordinate feature could help detecting those bigger structures, by considering the neighboring imagelets. The coordinate feature introduced in this paper (but not really used in the described experiments) is a step in that direction, as the relevance spreading in this case is directly related to favoring neighbors of relevant imagelets in retrieval.

VIII. CONCLUSIONS AND FUTURE WORK

We have presented how a content-based image retrieval system, PicSOM, can be used with remote sensing images for tasks like detection of man-made structures, as well as for supervised or unsupervised change detection. The approaches rely on the decomposition of a satellite image into several thousands small images or imagelets, to generate a database. After extracting features from the imagelets and training Self-Organizing Maps, the imagelet database can be queried. Relevant images can be retrieved either automatically, or interactively by a user in a visual and intuitive manner. The same framework allows detection of man-made structures as well as changes between two images, or both tasks simultaneously.

Quantitative and qualitative evaluation of the proposed methods were performed. The results were encouraging, considering we have presented a new approach to the challenging problem of change detection in very high-resolution images. Future work might include better shape descriptors for the purpose of building detection. The same kind of approach could be applied to radar images, in which case the importance of adapted feature extractors would be critical. Preliminary results on a database of imagelets constructed from an ASAR scene were encouraging.
The versatility of PicSOM will allow several applications to be embedded in a same system, only to be differentiated by the type of query. Further research will aim at a fully operative and interactive system built around PicSOM. Possible applications of the proposed methods include fast interactive analysis of satellite images. After the SOMs have been trained, a new image can be presented to the system in order to be automatically analyzed, or visually queried for structures of interest. The approach could also be used in long-term monitoring of strategic sites. In such a setting, each new image would first be analyzed, then integrated in the training set for improving the knowledge of the system about the area under surveillance.

ACKNOWLEDGMENTS
We are grateful to Ville Viitaniemi and Mats Sjöberg for taking part in the interactive retrieval experiment. We thank the anonymous reviewers for their fruitful comments.

REFERENCES


