MRF-based Segmentation and Unsupervised Classification for Building and Road Detection in Peri-urban Areas of High-resolution Satellite Images

Ilias Grinias[a,*], Costas Panagiotakis[b], Georgios Tziritas[c]

[a]*Department of Civil Engineering, Surveying and Geomatics, Technological Educational Institute of Central Macedonia, 62124 Serres, Greece*
[b]*Department of Business Administration, Technological Educational Institute of Crete, 72100 Agios Nikolaos, Greece*
[c]*Department of Computer Science , University of Crete, Greece, 70013 Heraklion, Greece*

Abstract

We present in this article a new method on unsupervised semantic parsing and structure recognition in peri-urban areas using satellite images. The automatic "building" and "road" detection is based on regions extracted by an unsupervised segmentation method. We propose a novel segmentation algorithm based on a Markov random field model and we give an extensive data analysis for determining relevant features for the classification problem. The novelty of the segmentation algorithm lies on the class-driven vector data quantization and clustering and the estimation of the likelihoods given the resulting clusters. We have evaluated the reachability of a good classification rate using the *Random Forest* method. We found that, with a limited number of features, among them some new defined in this article, we can obtain good classification performance. Our main contribution lies again on the data analysis and the estimation of likelihoods. Finally, we propose a new method for completing the road network exploiting its connectivity, and the local and global properties of the road network.

Keywords: building/road extraction, satellite images, image segmentation, feature analysis, Random Forest, unsupervised classification

1. Introduction

Automatic detection of buildings and roads in aerial/satellite images is of great importance in a wide range of areas, such as urban planning, urban area monitoring/detection, change detection, construction and update of GIS maps, transportation and telecommunication. Such man-made structures appear with high density and regular patterns in scenes of urban areas, while, by contrary, in rural regions is not unusual

*Corresponding author
Email addresses: elgrinias@gmail.com (Ilias Grinias), cpanag@staff.teicrete.gr (Costas Panagiotakis), tziritas@csd.uoc.gr (Georgios Tziritas)

to find only few buildings spread out over large distances and accessible by a sparse network of, often not paved, roads. Peri-urban areas (Ravetz et al. (2013)) are defined as the transition zones where urban and rural uses mix. As in case of suburban regions, buildings are in neighborhoods and are surrounded by yards in varying densities and directions, while their road network usually follows relatively regular patterns and is often sparser than that of suburban areas.

1.1. Object detection in Remote Sensing Imagery (RSI)

Although aerial images have traditionally been used to extract buildings/roads for mapping applications (Mayer (1999); Ahmadi et al. (2010); Hu et al. (2007); Mnih and Hinton (2010)), the successive launching of high spatial resolution commercial satellites IKONOS, QuickBird, WorldView (1,2 and 3) and Geoeye-1, has led to high-resolution, cost-effective satellite imagery. One of the main difficulties of image processing tasks when moving from (either aerial or satellite) images with low (coarser than 10m) and medium (of a few meters) resolutions to high (metric or sub-metric) resolution ones, is to be able to deal with the high complexity of the image content. This high complexity is mainly due to the fact that the elements or objects of interest are not any more only individual pixels or surfaces, but complex, structured groups of pixels (Inglada (2007); Blaschke (2010)).

Objects under detection or localization may be man-made ones, such as vehicles, ships, buildings and roads, that have sharp boundaries and are not part of the background, as well as landscape objects, such as trees and land-use/land-cover parcels that often are not characterized by clear boundaries and hence, may be considered natural parts of the background environment. Even though with the advances of remote sensing technology a greater range of man-made objects become separable from their background, the explosion in the availability of high-resolution Remote Sensing Imagery (RSI) underscores the need for automated satellite image interpretation methods. Such imagery has greatly increased the number of possible applications, but at the cost of an increase in the amount of required manual processing. Recent applications of large-scale machine learning to such high-resolution imagery (Inglada (2007); Kluckner and Bischof (2009); Paisitkriangkrai et al. (2015); Volpi and Ferrari (2015); Vakalopoulou et al. (2015)) have produced object detectors characterized by high levels of accuracy, reinforcing the belief that automated aerial/satellite image interpretation systems are within reach.

In RSI applications, aerial/satellite image interpretation is usually formulated as a pixel labeling task. Given an image, the goal is to produce either a complete semantic segmentation of the image into classes of objects such as "building", "road", "tree", "grass", and "water" or a binary classification of the image for a single object class. A very recent and complete review of object detection and localization methods in RSI, is found in (Cheng and Han (2016)). According to this review, the very large number of object detection methods can generally be divided into four, not necessarily independent, main categories: template matching-based methods, knowledge-based methods, Object Based Image Analysis (OBIA)-based methods, and machine

learning-based methods. Among them, OBIA-based or Geographic OBIA (GEOBIA)-based methods (Blaschke (2010); Blaschke et al. (2014)) have become a very promising alternative for detecting objects in high-resolution (sub-meter) RSI. Those methods consist of two main parts, namely, image segmentation in "homogeneous" pixel regions or (hopefully) "meaningful" objects of interest, followed by feature classification of the resulting objects based on various extracted features of objects such as spectral information, texture, shape, size, geometry and semantic features (Blaschke et al. (2014)).

1.2. Previous works on building and road network extraction

Several methods of the four categories have been proposed in literature for extracting the man-made objects and in particular road network and buildings, from satellite imagery of low, medium or high spatial resolution, using spectral or/and shape and structural (topology) properties, as it is described in detail in the following paragraphs. Several methods rely on supervised, ground truth based classification (Paisitkriangkrai et al. (2015); Vakalopoulou et al. (2015)) and in some of them previously stored information, such as road centerlines in vector data format, is also assumed and used (Yuan and Cheriyadat (2013)).

Road and building detection are applied either on their own or simultaneously (Ünsalan and Boyer (2005)), since, depending on the methodology followed, the joint detection approach may improve the detection of both. An interesting, template-matching method has been proposed by Karantzalos and Argialas (2009), capable of detecting either road network or buildings by tuning the parameters of a level set (Sethian (1999)) algorithm.

Aytekin et al. (2012) detect buildings using spectral properties in conjunction with spatial properties, both of which provide complementary information to each other. Natural and man-made regions are classified and segmented using Normalized Difference Vegetation Index (NDVI)(Rouse et al. (1974); Myneni et al. (1995)). Shadow regions are detected, and the rest of the image, consisting of man-made areas only, is partitioned by mean shift segmentation (Cheng (1995); Comaniciu and Meer (2002)). Resulting segments, whose shape is irrelevant to that of buildings are eliminated using morphological operations. Karantzalos and Paragios (2009) introduced competing shape priors, and building extraction is addressed through a segmentation approach that involves the use of a data-driven term constrained from the prior models.

A method based on local feature point extraction using Gabor filters (Jain et al. (1997)) is described in (Sirmaçek and Ünsalan (2010)). Local feature points vote for the candidate urban areas and final urban area is detected using an optimal decision-making approach on the vote distribution. In (Sirmaçek and Ünsalan (2011)) building detection is achieved using probability density functions of four, locally extracted, feature values. Finally, Benedek et al. (2012) introduced a global, probabilistic optimization process to find the optimal configuration of buildings, considering the observed data, prior knowledge, and interactions between the neighboring building parts. Since the method integrates building extraction with change detection in aerial and satellite

imagery, the authors, apart from the results for change detection, provide quantitative performance results for building detection in a benchmark dataset that they created and consists mainly by the images of their freely available SZTAKI-INRIA building detection dataset[1].

Recently, novel semantic labeling/segmentation techniques of high accuracy have been proposed, using Convolutional Neural Network (CNN) features (Jin and Davis (2007); Wang et al. (2015)) and Conditional Random Fields (CRFs) (Lafferty et al. (2001); Kumar and Hebert (2003)) to smooth region labeling, while respecting the edges of the image (Paisitkriangkrai et al. (2015)). Volpi and Ferrari (2015) model the segmentation problem by a CRF as well, employing Structured Support Vector Machines (SSVM) (Tsochantaridis et al. (2005); Finley and Joachims (2008)) to learn both the weights of a set of visual descriptors and local class interactions. In (Vakalopoulou et al. (2015)) an automated building detection framework is proposed based on deep convolutional networks. The core of the developed method is based on a supervised classification procedure employing a very large training dataset. Using a Markov Random Field (MRF) model (Li (2009)) the classification result is improved. Experimental results are given on the data set used also in our work. Their quantitative validation indicates that this approach is quite promising.

A comprehensive review of automatic road network extraction techniques for GIS update is found in (Mena (2003)). In (Laptev et al. (2000)) roads are automatically extracted using the multi-scale detection of roads in combination with geometry-constrained edge extraction using ribbon snakes (Mille et al. (2008)). Shadows and partially occluded areas are detected, as the bridges between partially (dis-)connected road segments and the road network is constructed after extracting crossings with varying shape and topology. In (Huang and Zhang (2009)), spectral and structural features are extracted in a number of scales and classified using Support Vector Machines (Vapnik (1995)). A majority voting approach is then used to integrate the multi-scale road information at the decision level in order to extract road centerlines and roads map.

In (Das et al. (2011)) a multistage framework for road network extraction is proposed by fusing region and boundary information to segment the image and then applying morphological operations to reject false positives. In (Ünsalan and Sirmaçek (2012)) probabilistic and graph theoretic methods are used to extract centerline and shape of roads.

Valero et al. (2010) build a granulometry chain using Path Openings and Path Closings (Talbot and Appleton (2007)) to construct Morphological Profiles. For each pixel, the Morphological Profile constitutes the feature vector on which the road extraction is based. In (Hu et al. (2007)), local homogeneous regions are enclosed by polygons, called footprints of pixels upon which road detection is achieved using tree expansion and pruning techniques. In (Silva and Centeno (2010)), the centerline is modeled as being a chain of short line segments. Radon transform (Herman (2009)) is used to detect

[1] http://web.eee.sztaki.hu/remotesensing/building_benchmark.html

seed line segments and for the automatic extraction of the roads centerlines starting from the seeds. Mnih and Hinton (2010) follow an approach based on neural networks to face the problems introduced to road extraction by shadows and occlusions.

1.3. The proposed method

In the proposed, automatic and large-scale OBIA-based work, roads and buildings are automatically and jointly extracted from high-resolution satellite images, using object based features, computed from visual and shape cues. In the segmentation stage, rural parts of peri-urban areas except rural roads, are detected and removed and an initial classification of the remaining urbanized areas, followed by MRF-based segmentation, lead to the extraction of regions or objects contained in only those areas. The unsupervised classification stage that follows, classifies those regions in classes "building", "road" and "other", which are further subdivided in sub-classes for obtaining (whenever possible) unimodal probability distributions. As the process is fully unsupervised, an extensive, automatic learning stage provides both the likelihood functions of sub-classes on relevant data, consisting of both appearance and shape features and then, the final classification result by the application of the maximum likelihood principle. After the OBIA-based modules, a global post-processing method for road extraction is applied in order to further improve the detection of both buildings and roads.

To get the upper limits of the classification efficiency that can be reached by segmentation on a given dataset, a fully supervised procedure using *Random Forests* (Breiman (2001)) has been developed and independently performed. For the needs of training, the object level matching between extracted regions and ground truth objects of classes is conducted by object based criteria (Jiang et al. (2006); Everingham et al. (2010); Özdemir et al. (2010); Ok et al. (2013); Ok (2013); Cheng et al. (2013); Han et al. (2014); Wiedemann et al. (1998)) that either have been used elsewhere before or are newly defined in this work. In addition, a novel object level performance metric has been implemented in order to measure the efficiency of building detection, motivated by the well studied application-dependence of object level performance metrics (Cheng and Han (2016); Rutzinger et al. (2009)) and the subjectivity of their thresholds (Shufelt (1999); Rutzinger et al. (2009)).

1.4. Article outline

The article is organized as follows. In Section 2 we present visual appearance and shape features used in our work and a correlation analysis of these features. We give also analysis of relevant data in order to automatically obtain parameters for segmentation and classification tasks. In Section 3 we present our segmentation algorithm using vector quantization, automatic probability distribution estimation of clusters features and MRF model optimization. In Section 4 we give results on region-based supervised classification using *Random Forests* in order to both assess the segmentation outcome and determine limits of reachable performance in building and road detection. The various object matching criteria used in this work are also described in detail. In

Section 5 we present the proposed unsupervised classification method and results on the whole data set. In Section 6 we present a global method for road extraction based on the initial classification and the properties of road network. In Section 7 we give final global results and, especially in the case of buildings, we provide a detailed performance evaluation using object level performance metrics that have been used before, as well as using the proposed object level performance metric. Furthermore, the pixel level performance of our building detection system on the SZTAKI-INRIA benchmark dataset is presented. Finally, Section 8 concludes our work.

2. Data and feature description and analysis

In the proposed work, roads and buildings are automatically and jointly extracted from high-resolution satellite images, using region features, computed from visual and shape descriptors.

The data set contains multi-spectral images for remote sensing a region in Attica. The acquisitions of years 2006, 2007 and 2009 were given by Quickbird satellite with 4 spectral bands (B, G, R and NIR), while those of years 2010 and 2011 were given by a WorldView-2 satellite with 8 spectral bands. For the most recent data we limited the number of bands to the four primary (bands: 2, 3, 5 and 7), because these spectral bands are related to the four bands of Quickbird satellite and contain sufficient information for detecting man-made objects. The image size is 6794*7884 pixels, corresponding to approximately 1400 ha.

Concerning visual features, we have transformed the data for using a component related to the luminance and three normalized difference components. If $X(c), c = 1, 2, 3, 4$ are the four channels (B, G, R and NIR) in the data, we use the normalized root squared luminance $Y = \sqrt{0.114R + 0.587G + 0.299B}$ $(0 \leq Y \leq 1)$, and the following normalized inter-band differences

$$X_d(c) = \frac{X(c+1) - X(c)}{X(c+1) + X(c)}, \quad c = 1, 2, 3. \tag{1}$$

As the fourth channel is the NIR, $X_d(3)$ component is the well-known Normalized Difference Vegetation Index (NDVI), with very good properties for discriminating vegetation areas from urbanized and bare ground regions. The three normalized difference components are illustrated in Fig. 1 for a large image block. In addition to the discriminative power of $X_d(3)$, it appears that $X_d(2)$ could be useful for extracting brick rooftops. The luminance is gamma corrected ($\gamma = 0.5$), because the images in the whole dataset are mostly dark. In addition, this correction decreases the correlation coefficient, as measured on the data set, between luminance component and inter-band difference components. Therefore, a better class discrimination could be expected. Dealing with regions, the appearance features are given by the median value over the region of the corrected luminance and the three normalized difference components.

Because of the discriminant properties of components $X_d(2)$ and $X_d(3)$, we considered them as the most useful visual components for detecting buildings and roads. For

(a) (b) (c) (d)

Figure 1: Data components for a subimage from the 2010 dataset: (a) RGB, (b) $X_d(1)$, (c) $X_d(2)$ and (d) $X_d(3)$.

this reason we give a detailed data analysis of these two, jointly considered, components. The analysis is performed for each year separately, as, even for data captured by the same satellite, the appearance could significantly change.

In Fig. 2 we give the median value of $X_d(2)$ given $X_d(3)$, as well as two vigintiles at (0.05, 0.95). More exactly, for an interval of $X_d(3)$ values, the value of $X_d(2)$ where the cumulative probability is 0.05, 0.5 and 0.95 are determined and plotted in Fig. 2. The contours of the 2-dimensional conditional distribution are depicted for the five years. Strong similarities between these conditional distributions appear, while at the same time it is shown that the analysis should be adapted for each year. In addition, in Fig. 3 we give the estimated joint probability density function of $X_d(2)$ and $X_d(3)$, confirming the need for adaptive data analysis.

One of the first stages in our algorithmic approach consists in discriminating vegetation regions from non-vegetation, as we are dealing with data from peri-urban areas. The joint or the conditional distribution of $(X_d(2), X_d(3))$ can be assumed as the mixture of two distributions corresponding to these two land covers. We observe that for years 2006 and 2007 the joint distribution is confusing. However the two classes exist for these datasets too, indicating that the intra-class variation is large in comparison with the inter-class distance. Therefore it would be possible, although difficult, to obtain discriminating boundaries for all years with suitable adaptations.

We describe hereafter our adaptive method for finding a decision boundary in the 2-dimensional space of $(X_d(2), X_d(3))$, for separating the vegetation areas using data-adaptive discriminant functions, even with ostensibly unimodal distributed data. For simplicity and robustness, we tried to determine piece-wise linear discriminant functions based on the estimated distributions illustrated on Figure 2. The first linear discriminant function is found as the best separator on the principal axis of the $(X_d(2), X_d(3))$ data set. The direction of the principal axis is determined by Principal Component Analysis (PCA) with a mean absolute deviation criterion. Plots of the estimated probability distribution of the projected on the principal axis data, are given in Figure 4. As expected for years 2006 and 2007 this probability distribution is ostensibly unimodal. The threshold V_1 results from the distribution of $X_d(3) - \lambda X_d(2)$, where λ is the slope of the principal axis, as it follows straightforwardly for bimodal distributions (years 2009,

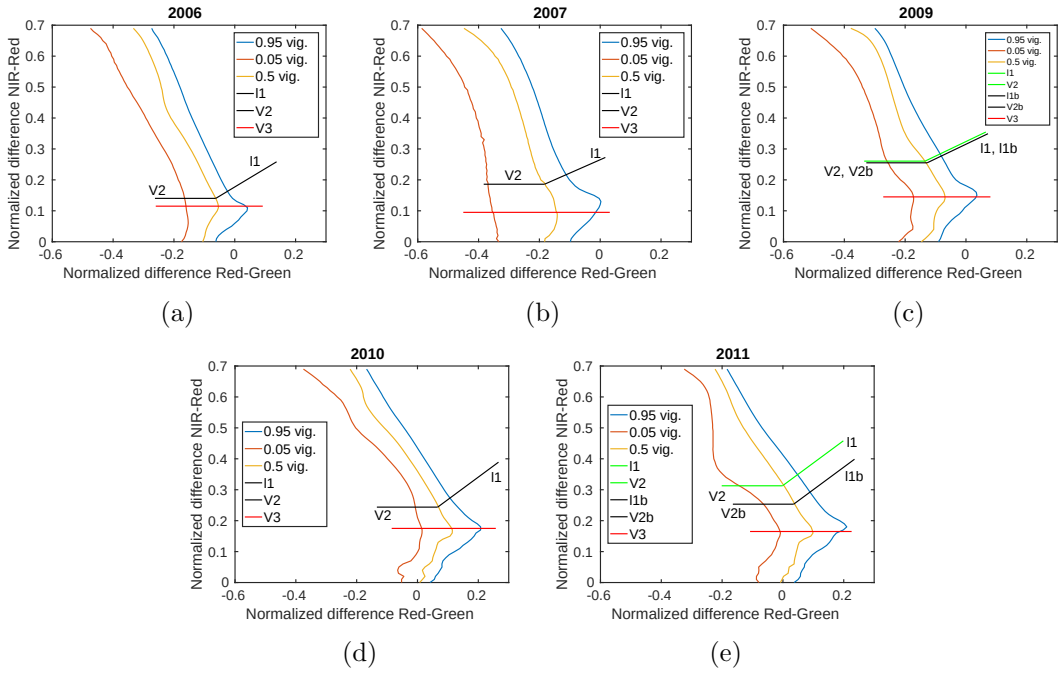


Figure 2: Median and two vigintiles (0.05, 0.95) of $X_d(2)$ given $X_d(3)$ for (a) 2006, (b) 2007, (c) 2009, (d) 2010 and (e) 2011.

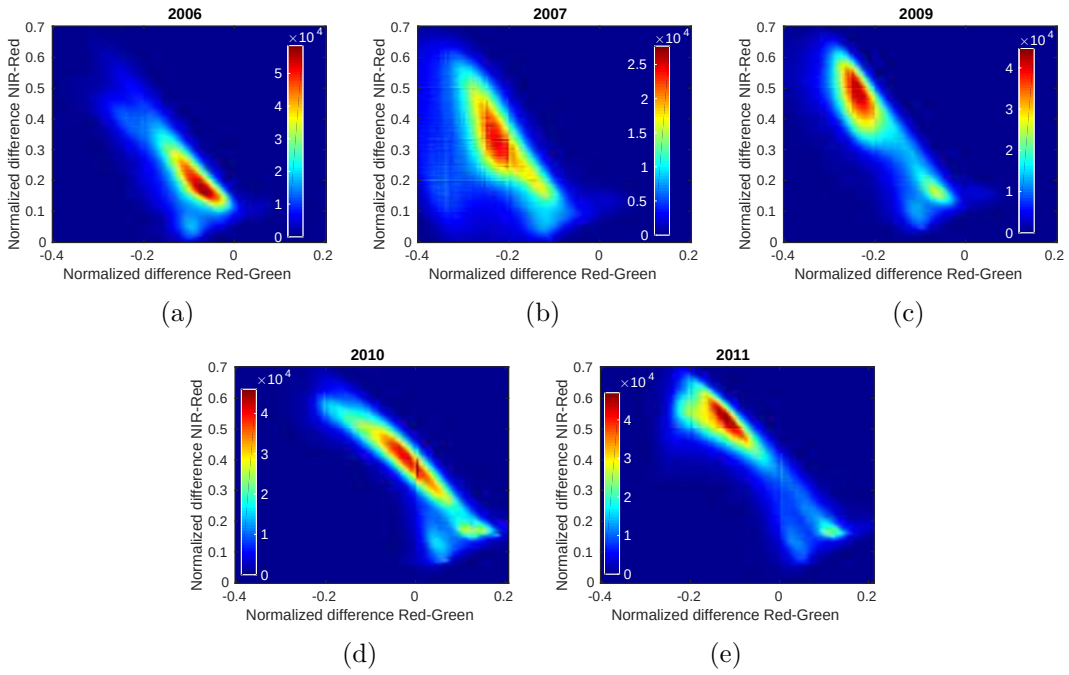


Figure 3: Joint probability density of $(X_d(2), X_d(3))$ for (a) 2006, (b) 2007, (c) 2009, (d) 2010 and (e) 2011.

2010 and 2011), while for unimodal distributions V_1 is estimated from a point where the first derivative of the probability density function is low. From the intersection point of the line $X_d(3) - \lambda X_d(2) = V_1$ with the median of $X_d(2)$ given $X_d(3)$ is defined a threshold on $X_d(3)$ providing another horizontal line, determined by a threshold V_2 on $X_d(3)$. The two discriminant line segments are illustrated with black color (years 2006, 2007 and 2010) or with green color (2009 and 2011) in the same figure (Fig. 2). For the data of these two years (2009 and 2011) two different discriminating lines could be determined using two equally acceptable criteria, an assertion that is corroborated by the density functions given in Fig. 3. This case occurs when the probability density function corresponding to the first mode (years 2009 to 2011) presents positive skewness. The lower threshold corresponds to an hypothesized symmetric distribution. In Fig. 2, a second set of discriminant functions for these years is given in black. From our experience we consider that the 'black' threshold could be retained for the "building" class, while the 'green' upper threshold could be applicable for the "road" class. If only one discriminant line is determined, it holds for both classes.

In addition, a threshold on $X_d(3)$ is needed for almost sure discrimination of bare soil regions, which is also estimated (V_3). Assuming that the largest median $X_d(2)$ is mainly due to bare soil regions, V_3 is estimated by the position of the peak in the median of $X_d(2)$ given $X_d(3)$ (see Fig. 2). All the parameters estimated on the whole ($X_d(2)$, $X_d(3)$) data set are given in Table 1.

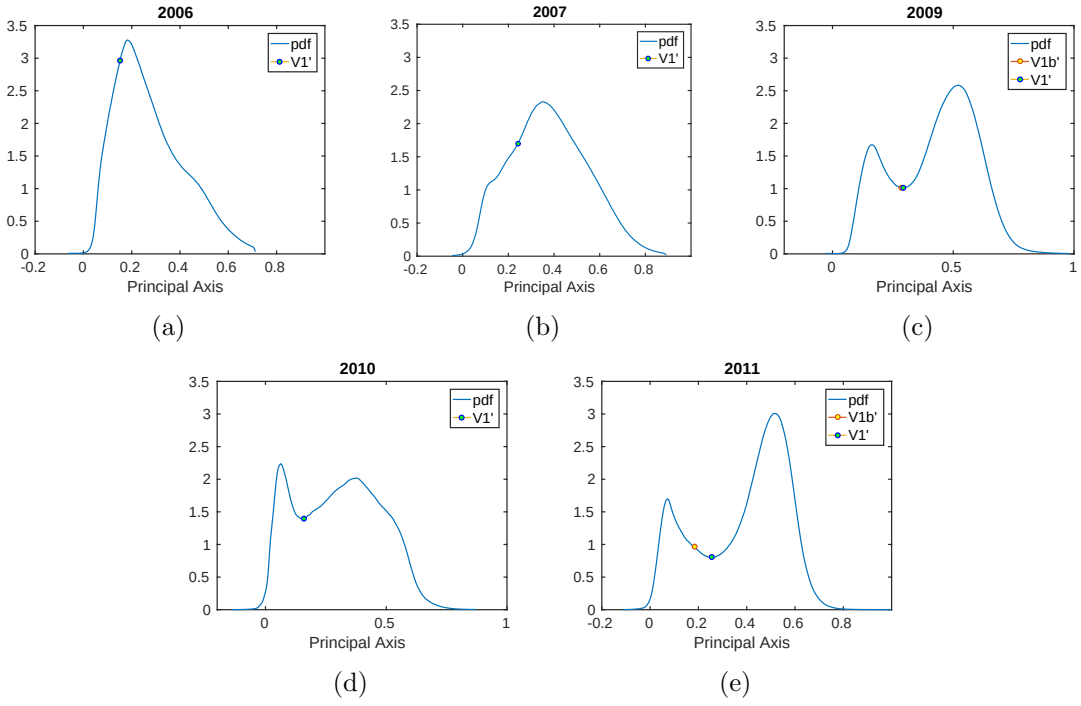


Figure 4: Probability density function of $X_d(3) - \lambda X_d(2)$ for (a) 2006, (b) 2007, (c) 2009, (d) 2010 and (e) 2011.

Table 1: Linear discrimination parameters.

Year	λ	V_1	V_2	V_{1b}	V_{2b}	V_3
2006	0.5914	0.1786	0.1450	0.1786	0.1450	0.1150
2007	0.4706	0.2705	0.1950	0.2705	0.1950	0.0950
2009	0.4706	0.3276	0.2650	0.3146	0.2550	0.1450
2010	0.7265	0.1923	0.2450	0.1923	0.2450	0.1750
2011	0.6796	0.3285	0.3350	0.2287	0.2550	0.1650

We can also estimate discriminant values for 'brick rooftop color'. From the distributions it appears that only a threshold c_2 on $X_d(2)$ is sufficient for a good detection of 'brick rooftop color', as the higher values of $X_d(2)$ correspond to this color. We propose to estimate this threshold from the distribution of $X_d(2)$ given that $X_d(3)$ values are limited to soil or urbanized areas. The c_2 estimation results show that the *'brick'* threshold is mainly determined by the spectral properties of the satellite bands. For the Quickbird satellite the estimated value is $c_2 = 0$, while for WorldView-2 we found $c_2 = 0.18$. In any case, our estimation method gives a different result for each year dataset.

Since the proposed classification method is region-based, shape features are also measured for all the regions extracted by the image segmentation algorithm. We have introduced two new shape features, presented below in Eqs. (2) and (4), with good discriminative power between buildings and roads. The shape features used in our work are as follows:

- the area A of the region,
- the normalized mean distance from the boundary over all pixels q,

$$\overline{D} = \frac{\overline{d(q,B)}}{\sqrt{A}}, \tag{2}$$

  B being the region boundary. In continuous space, for a circle the measure is maximum and equal to $\frac{1}{3\sqrt{\pi}}$, while for a rectangle with side ratio β $(0 < \beta \leq 1)$, the measure is

$$\overline{D}(\beta) = \frac{\sqrt{\beta}(3-\beta)}{12}, \tag{3}$$

  an increasing function of β.

- the normalized mean distance of boundary from centroid C_R,

$$D_B = \frac{\overline{d(C_R,B)}}{\sqrt{A}}. \tag{4}$$

  For a square $D_B \approx 0.57$, and it increases for a rectangle as β decreases.

We have also investigated the existing correlations between the previously defined features, both shape and visual. It might be useful to measure correlations on all the segments that could be extracted from the satellite images, excepting vegetation regions. However, the correlation is possibly class-dependent, mainly in the case of shape features. For this reason, for measuring correlation indexes we make use of the ground truth on segments extracted with our unsupervised segmentation algorithm. As "building" are considered the segments having Jaccard similarity coefficient (Eq. 7) with a true building greater than 0.45. As "road" are considered the segments for which the length of the ground truth road center-lines passing through them, divided by the length of their skeleton pixels, is above 0.35. The choice of the two thresholds guarantee that selected regions belong to the corresponding class of objects with high confidence given the ground-truth, as it is evident from the description of matching criteria and the meaning of their thresholds in Section 4.

In Table 2, the Spearman correlation index is given for "building" and "road" regions for the year 2006. Similar results have been obtained for the other years.

Table 2: Spearman rank correlation index for the visual and shape features (2800 "Building" regions and 1796 "Road" regions).

	"Building"	"Road"
$(X_d(2), X_d(3))$	0.6651	0.4453
$(X_d(2), Y)$	−0.5125	0.1998
$(X_d(3), Y)$	−0.8365	−0.3832
$(\sqrt{A}, \overline{D})$	−0.4927	−0.9140
$(\sqrt{A}, D_B)$	0.2855	0.8041
$(\overline{D}, D_B)$	−0.3707	−0.9270

As expected, the correlation between shape and visual features is low. Furthermore, as it is evident from Table 2:

- visual features are more correlated in case of buildings in comparison with the case of roads,
- $\sqrt{A}$ is highly correlated with both normalized mean distances, $\overline{D}$ and D_B, for "road" class and
- a strong correlation exists between $\overline{D}$ and D_B for the "road" class, and a much weaker one for the "building" class.

3. Unsupervised segmentation

In peri-urban areas, urban and rural uses mix. Having as objective the detection of buildings and roads, we propose as first stage the localization of areas with buildings, and at the same time the detection of possible road segments traversing soil or vegetation regions. It is equivalent to detect transitional and special areas as soil regions and

fields with vegetation. For detecting vegetation we use the separation criteria resulting from the analysis presented in the previous Section. On the other hand, it is assumed that bare soil regions are large with almost homogeneous visual appearance.

The remaining regions are then independently segmented using a Markovian approach for pixel labeling, where the classes correspond to homogeneous visual appearance, as our final objective is to achieve region-based classification. As the classes are unknown, as well as their number, we have to cluster data and estimate the class probability densities, or the likelihoods. The probability density function estimation is locally adaptive and is based on data vector quantization for identifying classes of distinctive appearance.

Finally, the successive stages for unsupervised segmentation are:

1. detection of rural and transitional regions,
2. vector quantization and initial classification of the data in urbanized areas,
3. Markov random field optimization for the final segmentation.

These stages are summarized in more detail in the flow chart of Fig. 5 and described hereafter.

3.1. Rural and transitional regions detection

For detecting large uniform regions, the gradient magnitude on the four spectral channels of the initial data is computed and thresholded. A low threshold is used for discriminating almost surely the homogeneous regions. It is estimated by fixing the percentage of homogeneity at 40%. The minimum size is fixed at 2500 pixels, which is about 600 squared meters. For all uniform regions the median and the variance of the appearance vector are estimated separately for its components, as well as the three shape features presented in Section 2. Initially, regions that are not dominated by vegetation and have a relatively large width, are considered. Taking into account that the the average road width is about $\overline{W}_r = 20$ pixels (10 meters) and that the width of a rectangle could be approximated as $4\sqrt{A}\,\overline{D}$, according to the definition in Eq. (2), the initial map of soil areas would satisfy the following conditions:

$$0 \leq X_d(3) \leq V_2 \;\text{ and }\; \sqrt{A}\,\overline{D} > 0.25\overline{W}_r.$$

From the resulting initial group of connected components, which may not exclusively contain soil regions, the median X_s is estimated, which is then considered that represents the dominant visual appearance of soil regions. The mean absolute deviation, σ_s, is also estimated. Finally, as almost surely non urbanized regions are considered those satisfying

$$X_d(3) \geq 0 \text{ and } (X_d(3) \geq V_3 \text{ or } |X - X_s| < 1.5\sigma_s) \text{ and } (D_B < 5\overline{D}{+}0.5 \text{ or } \sqrt{A}\,\overline{D} > 0.3\overline{W}_r),$$

where $X = (Y, X_d(1), X_d(2), X_d(3))$ is the representative visual data vector of the region and the dissimilarity is measured by the first order Minkowski distance (Xu and Wunsch (2009)). The relationship between the two distances $(D_B, \overline{D})$ characterizes

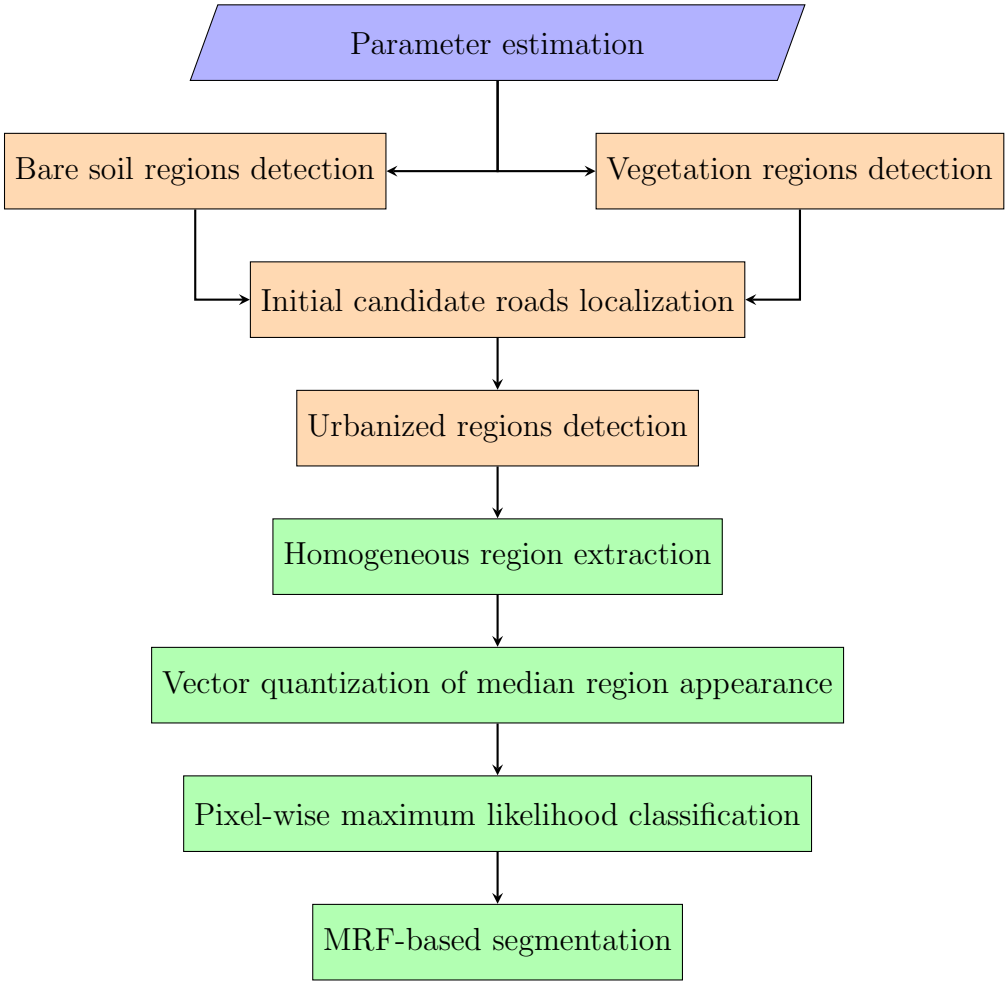


Figure 5: Flow chart of segmentation module: the 'green' processes are executed independently for all the connected components of the urbanized area.

non-elongated regions and is evident by experimental results, as those depicted in Fig. 6. As the threshold on gradient magnitude is low, a growing procedure is applied on the detected regions, in order to agglomerate neighboring pixels with similar visual appearance.

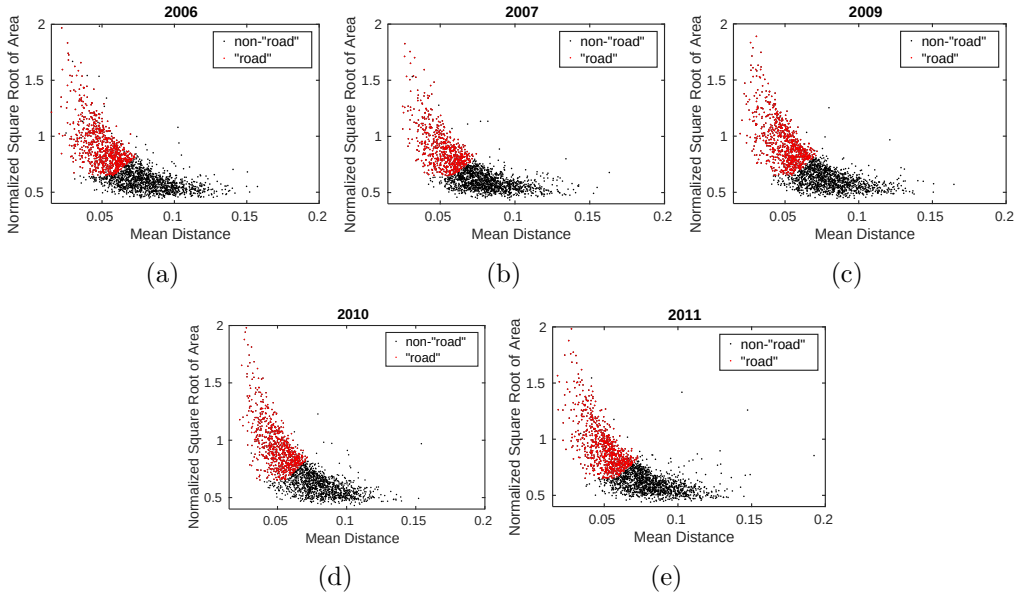


Figure 6: Shape feature data classification for candidate "road" segments for (a) 2006, (b) 2007, (c) 2009, (d) 2010 and (e) 2011.

All the remaining areas are considered as urbanized, except vegetation fields, which are excluded by the following additional conditions on luminance and *NDVI*,

$$X_d(3) \leq \max(V_1 + \lambda X_d(2), V_2) \quad \text{and} \quad Y \geq 0.25$$

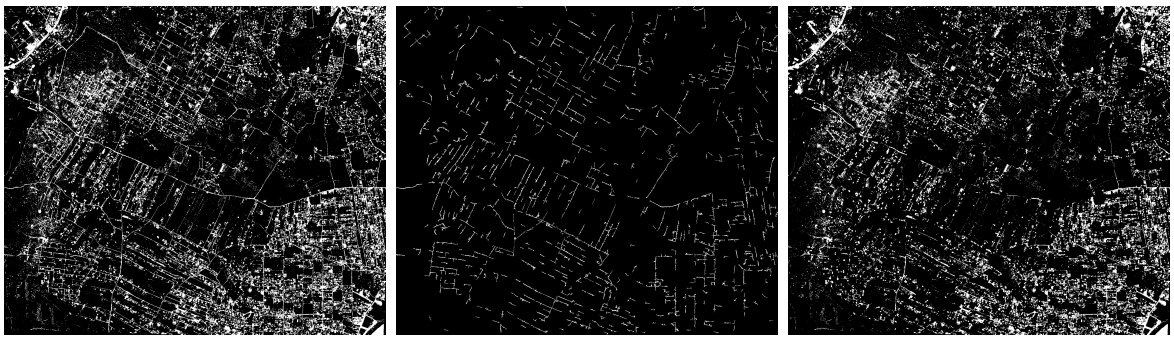
Roads are also included in the urbanized areas, but segments of roads might traverse rural areas as well, which could be detected as elongated segments after subtracting vegetation and bare soil regions. Morphological operations allow us to determine candidate road segments. The shape features of the extracted segments are computed. Automatic clustering using *k-means* algorithm with *cosine* distance is performed on $(D_B, \overline{D})$ features for separating a group of candidate road segments. In addition, the road width could be bounded as follows:

$$D_B > 0.65 \quad \text{and} \quad \sqrt{A}\,\overline{D} < -2.5 \log_{10}(\overline{D}).$$

The constant thresholds above are justified both by the fact that $D_B \approx 0.57$ for a square and from experimental results for the whole data set, shown in Fig. 6. In this figure, the clustering result is illustrated, with the data corresponding to the road segments depicted in red. Knowing that an important number of rooftops are much

more luminous than "road" regions, we obtain another threshold for detecting such a 'rooftop color'. This threshold, Y_1, is given by the median of "road" luminance, increased by 3 times the standard deviation of the "road" luminance measured by Y.

The localization result of urbanized regions for year 2006 is given in Fig. 7. At first, vegetation and bare soil regions are excluded and then, candidate roads and regions containing buildings (and probably roads) are localized.



(a) (b) (c)

Figure 7: The result of (a) initial urbanized areas, (b) candidate roads and (c) final urbanized areas extracted for 2006.

After extracting the "road" candidate segments, the remaining areas are segmented using the visual data. These areas are all the connected components of the image resulting from the localization and subtraction of vegetation areas, soil regions and elongated segments separating these areas. These components are mainly located in urbanized areas and their shape and size as well as their visual appearance may be quite different. Therefore, the segmentation of the resulting areas should be adapted to the local data.

3.2. Vector quantization and initial classification of urbanized areas

The estimation of the data distribution given an appearance class is based on data clustering. For efficiency reasons and for using globally reliable features, homogeneous regions are extracted, having again as criterion the low value of the gradient magnitude. We then compute appearance features for all the homogeneous regions extracted. The appearance is summarized by the mean intensity of the luminance Y and the normalized chromatic differences $X_d(c), c = 1, 2, 3$.

In order to obtain a content-guided, visual data vector quantization, we perform a partial initial classification of the extracted regions. The objective pursued is to have, if possible, clusters that contain data from the two classes ("building" and "road") to be detected. However, reliable criteria could be assumed for only two possible groups of buildings : 'brick rooftop color' ($X_d(2) \geq c_2$) and 'luminous rooftop color' ($X_d(2) < c_2, Y \geq Y_1$). For these sub-classes only one representative data vector is extracted. It is difficult to characterize the other regions by their visual appearance, because they

could be either "building", or "road", or even "other". In this last confusing case, we estimate an initial maximum number of classes related to the deviation of the visual appearance vector and the number of homogeneous regions belonging to this class. The *k-means* algorithm with a 'cityblock' distance, is then initialized using the method given in Kauffman and Rousseeuw (1990). In addition, as the number of classes is unknown, a stopping criterion is introduced based on the intra-cluster deviation. If the rate of improvement in the sum of intra-class distances is less than 5%, no more classes are added.

A pixel-wise classification is obtained based on the distance of the data vector to the cluster representative vectors. Therefore the data are vector quantized and the pixels are labeled according to the nearest representative vector. First order Minkowski distance is used for labeling, the number of labels being equal to the number of clusters. The labels determine data clusters, which will serve to obtain probability density functions of the data in the sub-classes defined by the clusters. For simplifying the computations, as the regions considered are almost homogeneous, a Gaussian assumption is adopted. Therefore, we have only to estimate the mean value and the covariance matrix.

3.3. Markov random field segmentation

We propose to optimize a discrete Markov random field (MRF) in order to obtain a regularized label field. In this manner, we aim at capturing the local interactions between pixels, which will help us to refine and correct the class labels of the minimum distance classification. The problem can be formulated as follows: we seek to assign a class label $l(q)$ (from a discrete set of class labels $\mathcal{L}$) to each node (pixel) of a graph $q \in \mathcal{V}$, so that the following cost is minimized:

$$\sum\nolimits_{q\in\mathcal{V}} \mathfrak{c}_l(q) + \sum\nolimits_{(q,p)\in\mathcal{E}} w(l(q), l(p)), \tag{5}$$

where $\mathcal{E}$ is the set of the graph edges. The graph is composed by the pixels of the connected components of urbanized areas. The optimization is implemented independently on all the connected components extracted, according to the number of labels and the priors estimated on each of them.

The singleton potentials, or priors, are based on the pixel-wise computed Gaussian probability density functions. Then the dissimilarities of pixels to the clusters l are given by

$$\mathfrak{c}_l(q) = -\ln p_l(X(q)), \tag{6}$$

where $X(q)$ is the vector data, consisting of the normalized luminance and inter-band differences, at pixel q. The pairwise potentials are set according to the Potts function, with all weights w set equal to a constant w_0, in case of different classification, and zero potential when the two neighboring points are assigned the same class. The regularization constant w_0 is data adapted, having as relevant statistics the mean value on the minimum dissimilarities at each point.

For minimizing the MRF energy in Eq. (5), we make use of the *primal-dual* method Komodakis and Tziritas (2007), which casts the MRF optimization problem as an integer programming problem and then makes use of the duality theory of linear programming in order to derive solutions that have been proved to be almost optimal. The optimization code can be found in `http://www.csd.uoc.gr/~komod/FastPD/`.

The result for the largest connected component detected for year 2006 is illustrated in Fig. 8. In the same area other connected components exist and they are segmented separately. An MRF model is estimated and used for each connected component of the area detected as urbanized.

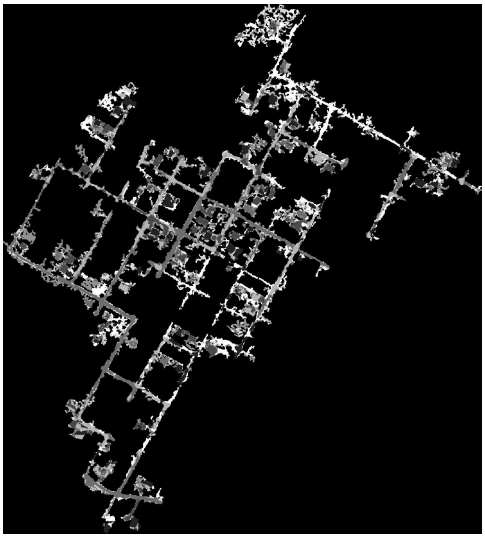


Figure 8: The segmentation result of the largest urbanized component for 2006.

4. Random Forest based segment analysis

Even if our objective is to perform unsupervised structure recognition, we developed a Random Forest (RF) (Breiman (2001)) based classifier, for evaluating the role of features and for comparing our results to the optimum, considering that RF discovers the best discrimination functions for a given set of features. Indeed, RF constitute a mechanism to estimate *a posteriori* probabilities $\Pr(c|\mathbf{x})$ of class c given the (even high dimensional) feature vector $\mathbf{x}$, without direct modeling or any assumption made about the conditional probabilities $p(\mathbf{x}|c)$.

Supervised Random Forests analysis is applied on a number of combinations of the features described in the previous paragraphs, derived from the segments extracted by segmentation. The discriminative power (or *importance*) of features can be determined as one of the results of RF training. Optimal performance of features is measured in RF testing phase, using a new evaluation criterion for class "building" and a matching criterion that has been used before (Wiedemann et al. (1998); Mnih and Hinton (2010)) for the class "road", respectively.

Features are classified in three classes denoted as "other", "building" and "road" respectively. For the needs of training, a region R is considered to belong to class "building", if the ratio

$$I_R = \frac{\sum_{i=1}^{K} |R \cap B_i|}{|R|}$$

is above 0.5, where $|R \cap B_i|$ is the area of its intersection with the K ground truth (real) buildings B_i and $|R|$ is the area of the region. Although I_R does not measure how well the shape of region R fits to the shape of real buildings, it is assumed that, because of segmentation, even in the case that region R spans more than one real buildings, those buildings undergo the same appearance characteristics.

A region R is considered to belong to class "road", if the length of the ground truth (real) road center-lines passing through R, divided by the length of the skeleton pixels of R is above 0.1. The low value of that criterion is justified by the often large number of skeleton points of extracted regions due to their noisy boundaries and the fact that the non systematic, wrong assignments in the input data of training do not affect the prediction ability of RFs.

The remaining regions are assigned to class "other". The feature sets that were tested for their efficiency in classification are given in the corresponding rows of Table 3. In each row, the identifier as well as the features that are included in the corresponding set are reported.

Table 3: Feature sets used for RF training.

ID	Features
FS_0	$\{X_d(1), X_d(2), X_d(3), Y, \sqrt{A}, \overline{D}, D_B\}$
FS_1	$\{X_d(2), X_d(3), Y, \sqrt{A}, \overline{D}, D_B\}$
FS_2	$\{X_d(2), X_d(3), Y, \sqrt{A}, \overline{D}\}$

1000 trees were used for training in any case (feature set and year), considering equal prior probabilities of classes. The outcome of training is the *a posteriori* probability $\Pr(c|\mathbf{x})$ of class $c \in \{$"other", "building", "road"$\}$ given the feature vector $\mathbf{x}$. Feature importance of feature set FS_0 for each year is depicted in Fig. 9. In the first three plots, the color difference $X_d(1)$ is the least important feature in discriminating the three classes, although, that feature plays an important role in year 2010 and is important in year 2011.

For each region R, with feature vector $\mathbf{x}$, the *a posteriori* probability $\Pr(c|\mathbf{x})$ of class c given $\mathbf{x}$ is computed using the trained RF, for classes "building" and "road". If this probability is above a threshold T_c, the region is classified in class c. Optimal thresholds are determined by the Precision-Recall Curve (PRC) of each class as described below.

In order to measure the accuracy of detecting ground truth (real) buildings using the trained RF, a new criterion that establishes a one-to-one relationship between regions and ground truth buildings and, at the same time, scales the contribution of detected regions to the detection of real buildings, has been defined. First, the best fit region

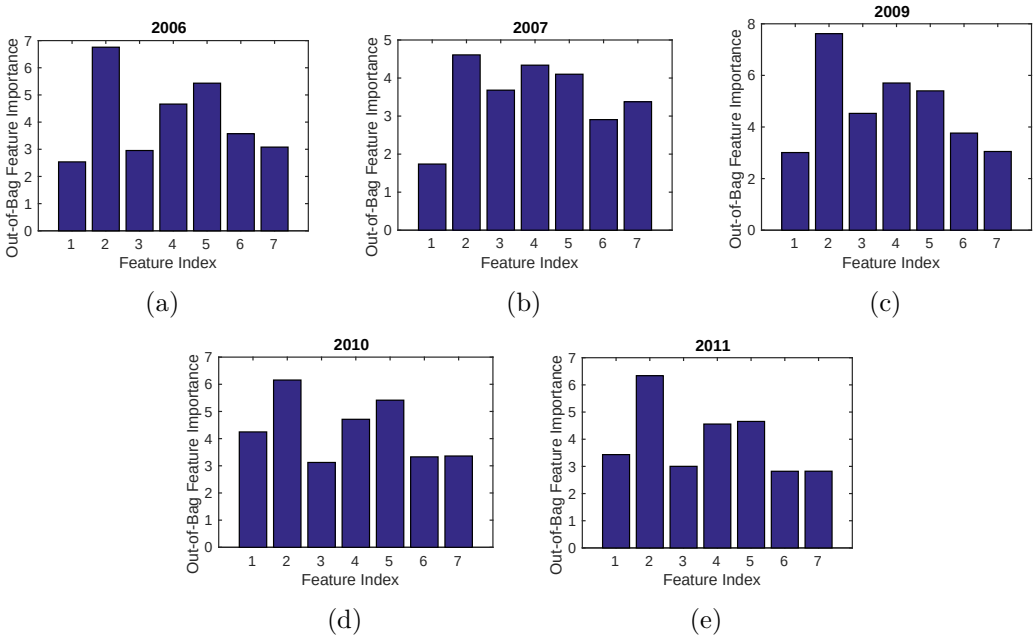


Figure 9: Feature importance for feature set FS_0 and years (a) 2006, (b) 2007, (c) 2009, (d) 2010 and (e) 2011. "Feature Index" refers to the order of features of FS_0 in Table 3.

R_{B_i} for each building B_i is determined as:

$$R_{B_i} = \arg\max_{R_j}\{J(R_j; B_i)\},$$

where

$$J(R; B_i) = |R \cap B_i|/|R \cup B_i| \tag{7}$$

is the Jaccard similarity coefficient and R_j are the regions assigned by RF in class "building". Detectability of B_i is then measured using the formula:

$$TP_{B_i} = \min\{1, J(R_{B_i}; B_i)/T_B\} \tag{8}$$

where T_B is a threshold. The overall true positiveness, TP, of building detection is the sum of TP_{B_i} for all real buildings B_i and Recall and Precision are computed by ratios TP/K and TP/L respectively, where K is the number of real buildings and L the number of "building" regions. Finally, the optimal threshold T_c is determined for class "building" by the break-even point where Precision equals Recall. The break-even point obtained by each feature set is depicted in Table 4, for $T_B = 0.25$.

Setting T_B to 0.25, accounts for cases of segmentation regions corresponding to groups of real buildings that are placed very close to each other and undergo the same appearance. Such groups cannot be further separated by the segmentation process due to the low contrast and weak edges of objects that characterize the satellite images of our dataset. A reference building B_i that is contained in such a detected region R, will

be considered fully (i.e with $TP_{B_i} = 1$ in Eq.(8)) TP, if it overlaps with R in more than 25% of its area. Roughly speaking, such a region will contribute to the true positiveness of groups of at most four buildings. On the opposite direction, over-segmented large buildings will be considered fully detected if at least one of the regions included in them covers more than 25% of their area. This technique constitutes an alternative to "invasive" approaches such as the *topological clarification* that is described in Rutzinger et al. (2009), which splits regions of small buildings and merges sub-regions of the same reference building before evaluation.

Table 4: Break-even points of "building" detection for $T_B = 0.25$.

FeatSet	2006	2007	2009	2010	2011
FS_0	0.8097	0.7607	0.8267	0.8059	0.8235
FS_1	0.8091	0.7600	0.8267	0.8053	0.8234
FS_2	0.8084	0.7564	0.8267	0.8053	0.8184

In case of roads, efficiency of detection is measured using the ground truth (real or reference) centerlines of roads against the skeleton points of regions detected to belong to class "road". Similar to Wiedemann et al. (1998) and Mnih and Hinton (2010), for each such region, its skeleton points that are placed in a distance less than d (pixels) from real road centerlines are considered as detected or True Positives (TP). Otherwise, they are considered False Positives (FP). The same way, centerline points that are placed further than a distance d from all skeleton points of detected regions, are considered False Negatives (FN). Precision and Recall are computed as in the case of buildings and the optimal threshold T_c is determined. The break-even point obtained by each feature set and year is depicted in Table 5, for $d = 10$, a relaxation distance threshold that is justified by the fact that the average width of roads for the five years is about $\overline{W}_r = 20$ pixels (10m) and the reference centerlines are not always placed exactly at the center of roads. The overall performance results indicate that the three feature sets are statistically equivalent.

Table 5: Break-even points of "road" detection for $d = 10$.

FeatSet	2006	2007	2009	2010	2011
FS_0	0.8094	0.7846	0.8457	0.8495	0.8301
FS_1	0.8093	0.7846	0.8456	0.8496	0.8302
FS_2	0.8090	0.7849	0.8451	0.8478	0.8299

By far the worst performance in building/road detection is that of year 2007, a conclusion that holds using either RF or the new unsupervised learning method of Section 5. In order to detect the source of this behavior, *a posteriori* probabilities in the feature space of the two appearance features $(X_d(2), X_d(3))$ and the shape features $(\overline{D}, 0.1\sqrt{A})$ have been separately computed using 100×100 2D histograms, for all years. *A posteriori* probabilities of appearance and shape features are depicted in Fig. 10 and Fig. 11 respectively. In all plots, green color corresponds to class "other", red

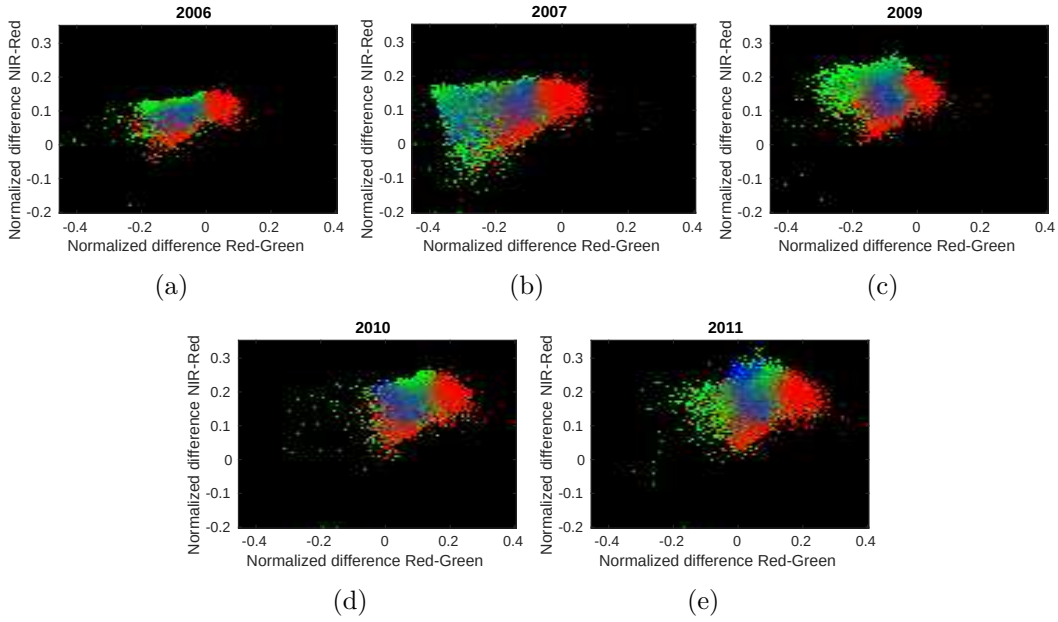


Figure 10: A posteriori probabilities of $(X_d(2), X_d(3))$ for the three classes using 2D histograms for years 2006 (a), 2007 (b), 2009 (c), 2010 (d) and 2011 (e).

to "building" and blue to "road" respectively, while the color of each bin comes of the weighted mixing of these colors using the *a posteriori* probabilities as weights. Compared to the other years, the distributions of appearance features for year 2007 are quite different and specially for class "road", which is separated in two sub-classes. On the contrary, the distribution of shape features in year 2007, follows that of the other years, for all classes.

As in the case of feature importance analysis, the fact that the images of the first three years have been captured by satellites with different settings compared to those of 2010 and 2011, is reflected in the analysis of appearance features. A closer look at the plots of Fig. 10, shows that building/road distributions in (d) and (e), are translated in the feature space, compared to the corresponding distributions of years 2006, 2007 and 2009.

5. Unsupervised learning and classification

We present in this Section our unsupervised, object-based classification method. We consider three main classes: "building", "road" and "other", which are subdivided in sub-classes for obtaining, if possible, unimodal probability distributions. As the process is fully unsupervised, an extensive learning stage provides estimates of distributions of relevant data. Relevant data are: appearance features, consisting of the luminance Y and normalized differences $(X_d(2), X_d(3))$ and shape features $(\sqrt{A}, \overline{D}, D_B)$. It is assumed that the appearance features are independent from those of shape for all

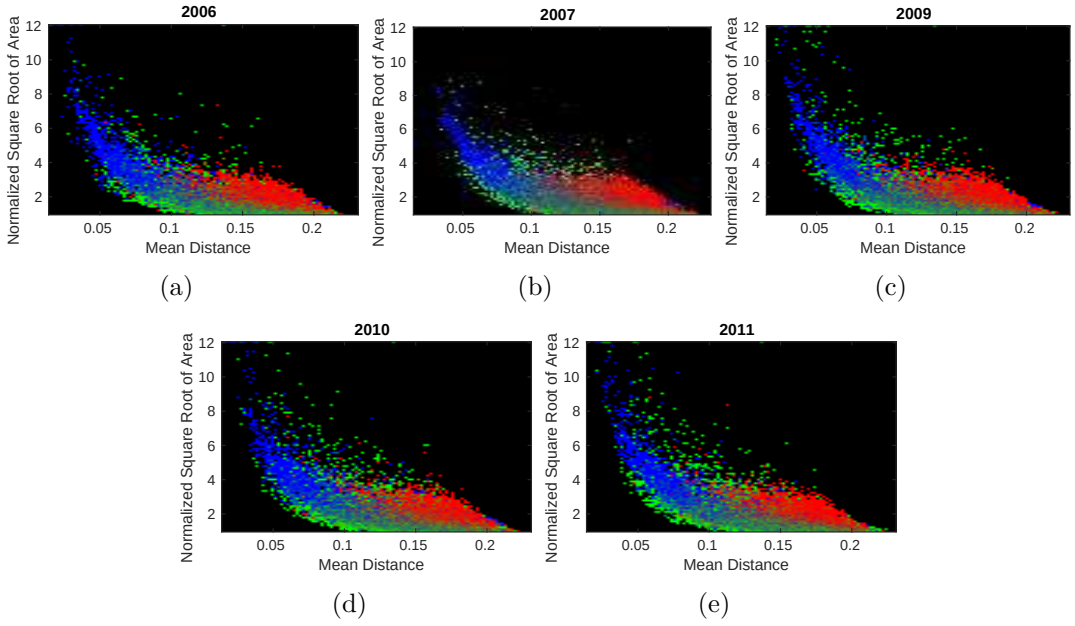


Figure 11: A posteriori probabilities of $(\overline{D}, 0.1\sqrt{A})$ for the three classes using 2D histograms for years 2006 (a), 2007 (b), 2009 (c), 2010 (d) and 2011 (e).

classes. All appearance distributions as well as the distribution of $\overline{D}$ are assumed to be Gaussians. Even if the Gaussian assumption might not be the best in some cases, it is retained for robustness and regularization reasons.

The whole procedure being completely automatic, the parameters needed for the initial discrimination are estimated from the segments. The first parameter estimated is a threshold on $X_d(2)$, noted as c_{2m}, which rejects outliers, that is, segments which could be considered as not belonging to man-made classes. The following conditions provide the *a priori* classification of segments as "building"

$$\sqrt{A} > A_b \text{ and } X_d(2) > c_{2m} \text{ and } \overline{D} > F_b \text{ and } S < 1.1 \text{ and } D_B < G_b,$$

where S is the ratio of the area covered by the object with holes filled to the region area. The threshold A_b on the region size is related to the real size of the buildings knowing the image resolution. It is set to 12 for our data, which corresponds to approximately 35 square meters. F_b is a discriminative threshold on $\overline{D}$ fixed at 0.1, a value that corresponds to a rectangle with sides ratio about 5. In a similar way, G_b is set to 0.75, knowing that for a square $D_B = 0.57$. Three sub-classes are determined for the "building" class using the *k-means* algorithm on visual data $(X_d(2), X_d(3), Y)$.

For the *a priori* classification of segments as "road", a clustering of all the segments is performed on shape features $(\overline{D}, D_B)$. "Road" segments are characterized by low values in $\overline{D}$ and high values in D_B. This property provides an initial group, which is refined with the following criteria

$$X_d(2) > c_{2m} \text{ and } X_d(3) > 0 \text{ and } -0.75 \log_{10} \overline{D} < \sqrt{A}\,\overline{D} < -3 \log_{10} \overline{D} \text{ and}$$

$$\overline{D} \leq F_b \text{ and } D_B > 0.55 \text{ and } S < 1.01$$

The bounds on $\sqrt{A}\ \overline{D}$ can be inferred from the knowledge of the average width of roads. Indeed, the product $\sqrt{A}\ \overline{D}$ is strongly correlated with the road width, which could be approximately considered as *a priori* known. The other thresholding parameters result from empirical experiments and mainly the cluster analysis of the shape features $(\overline{D}, D_B)$ shown in Fig. 6. Then, two sub-classes are determined for the "road" class using the *k-means* algorithm on visual data $(X_d(2), X_d(3), Y)$.

The remaining segments are initially classified as "other", and then subdivided in two sub-classes using again the *k-means* algorithm on visual data $(X_d(2), X_d(3), Y)$ with the 'cityblock' distance.

We present now the estimation of probability density functions for the shape features obtained using the kernel method given in Botev et al. (2010). The estimated probability density functions of shape feature $\overline{D}$ for the different sub-classes resulting from the above criteria and cluster analysis are given in Fig. 12 for the whole data set. Their stability over different years is illustrated in these experimental results. Even if they are not fitting well the Gaussian distribution, we consider that the Gaussian assumption ensures the robustness of the estimator, as far as it is unimodal, taking into account the unavoidable weakness of the above *a priori* classification based on strong, but not sure assertions. On the other hand, the shape of the density function depends on the bandwidth of the kernel which controls its regularity, the more regular being finally the Gaussian density function.

The size feature $\sqrt{A}$, could be assumed not depended on the normalized mean distance $\overline{D}$ for the "building" class, while for the other two classes a strong correlation exists. The probability density function of $\sqrt{A}$ given the class and, if it depends on, given the normalized mean distance $\overline{D}$ is assumed to be Gaussian for the "building" and "road" classes, while it is assumed to be exponential for the "other" class. Noting the variable corresponding to feature $\overline{D}$ by U and that corresponding to $\sqrt{A}$ by V, we can write

$$p_{U,V}(u, v) = p_{V|U}(v|u)p_U(u). \tag{9}$$

The estimated mean and standard deviation of $\sqrt{A}$ given $\overline{D}$ for the "road" class and the estimated $\lambda(u)$ for the class "other", are depicted in Fig. 13 (data set 2011). The mean value of the size feature for the "building" class is also given in black.

More exactly, for the class "road" we assume a Gaussian conditional distribution, with mean and standard deviation given empirically as

$$\mu(u) = \frac{-2\log_{10} u}{u}, \sigma(u) = 0.25\mu(u).$$

These empirical relationships are supported by experimental results similar to those of Fig. 6. For the class "other" we assume an exponential distribution, as follows:

$$p_{V|U}(v|u) = \frac{1}{\lambda(u)} e^{-\frac{v - v_m(u)}{\lambda(u)}}, v \geq v_m(u) \tag{10}$$

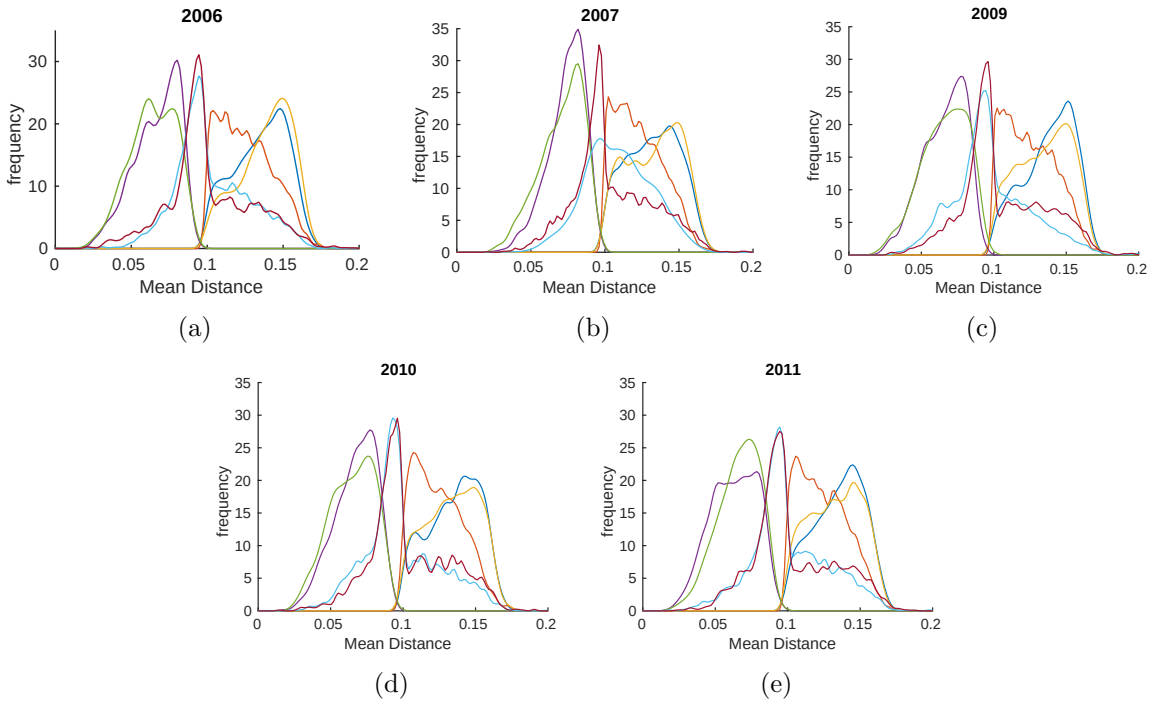
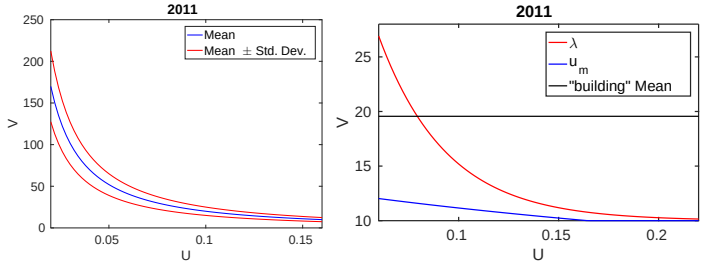


Figure 12: The probability density function of $\overline{D}$ for all the sub-classes and for all years: (a) 2006, (b) 2007, (c) 2009, (d) 2010 and (e) 2011.

where

$$\lambda(u) = e^{\alpha_1 u + \alpha_0} \quad \text{and} \quad v_m(u) = \max\left\{10, \frac{1}{\beta_1 u + \beta_0}\right\}.$$

Parameters $\alpha_1, \alpha_0, \beta_1, \beta_0$ are estimated from the data using linear regression.



(a) (b)

Figure 13: For the 2011 data set: (a) the conditional mean and mean ± standard deviation functions of V given U for "road" class and (b) parameters $(\lambda(u), v_m(u))$ for "other" class and mean value (in black) for the "building" class.

The normalized mean distance D_B of boundary from the centroid is assumed to be independent from the other shape features and distributed according to the gamma distribution for the classes "building" and "other". The corresponding sets of parameters are estimated for these classes. For the "road" class a gamma distribution is also assumed, but the two distances, D_B and $\overline{D}$, are strongly related, and linked by the elongation parameter β. The distance D_B could be assumed linearly depending on β, while $\overline{D}$ could be considered proportional to $\sqrt{\beta}$. Finally, the shape parameter of the gamma distribution is given the value 3, an integer approximation of the estimated value and the scale parameter θ of the D_B distribution relates to the average length of the road segments, which in turn relates both to a constant denoted as α_1 and to an additional term which is, according to the above explication, inversely proportional to $\overline{D}^2$:

$$\theta = \alpha_0 + \frac{\alpha_1}{1 + 10000\overline{D}^2}.$$

The numerical values of parameters α are estimated from experiments on real data.

From the likelihood functions we obtain the *a posteriori* probabilities for all segments and all classes, $\Pr(c|\mathbf{x})$, $\mathbf{x}$ being all the data, summarized by the appearance and shape features described above. A Bayesian approach is adopted, where equal costs are assumed for false classifications, as well as equal *a priori* probabilities for the seven sub-classes. The decision inferred by the *maximum a prosteriori* probability criterion on the sub-classes determines the final decision for any segment, "building", "road" or "other".

Results on Recall, Precision and F-measure are given in Table 6 for "building" and "road" classes and for all years using parameters of Table 1. The methods for measuring the accuracy have been given in Section 4 (testing phase of RF analysis).

Table 6: Measures on classification results using parameters of Table 1.

	Buildings			*Roads*		
year	Recall	Precision	F-measure	Recall	Precision	F-measure
2006	0.7407	0.7326	0.7366	0.7057	0.7106	0.7081
2007	0.7373	0.6701	0.7021	0.6228	0.6325	0.6276
2009	0.8064	0.7077	0.7539	0.7835	0.6685	0.7214
2010	0.7610	0.7027	0.7307	0.7571	0.6874	0.7206
2011	0.7704	0.7170	0.7427	0.7224	0.6573	0.6883

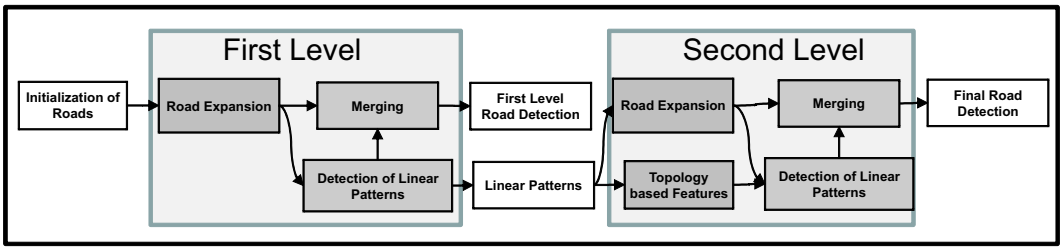


Figure 14: The schema of the road detection method.

6. Global road detection

The methodology applied for the global detection of the roads consists of two levels. In the first level, shape and appearance based features are combined with a linear pattern detection method. Firstly, we exploit the road connectivity by expanding the regions having high *a posteriori* probability to be "road" segments. In the second step, linear patterns of roads are detected. Then, the results of the first two steps are merged. The second level of road detection takes into account topology based features in a global sense, computed from the detected linear patterns of the first level. Figure 14 presents an overview of the proposed global road detection method.

6.1. Initialization and expansion of "road" segments

Initially, the regions are divided into three sets $\{HR, LR, NR\}$ that consist of regions having high, low and almost zero *a posteriori* probability to be "road", respectively. Firstly, a binary image B_{NDVI} is computed by applying a threshold (V_2 of Table 1) on NDVI.

The set HR consists of regions that are classified as "road". The set NR consists of regions R that satisfy at least one of the following criteria: the probability to be "road" is lower than 0.01 or are classified as "building". The set LR consists of regions having low probability to be roads, so the expansion of HR set may include these regions. This set is given by the regions that do not belong in HR and NR sets. Figure 15(b) shows an example of this classification applied on image of Fig. 15(a), where the sets HR, LR and NR are depicted with white, gray and black colors, respectively.

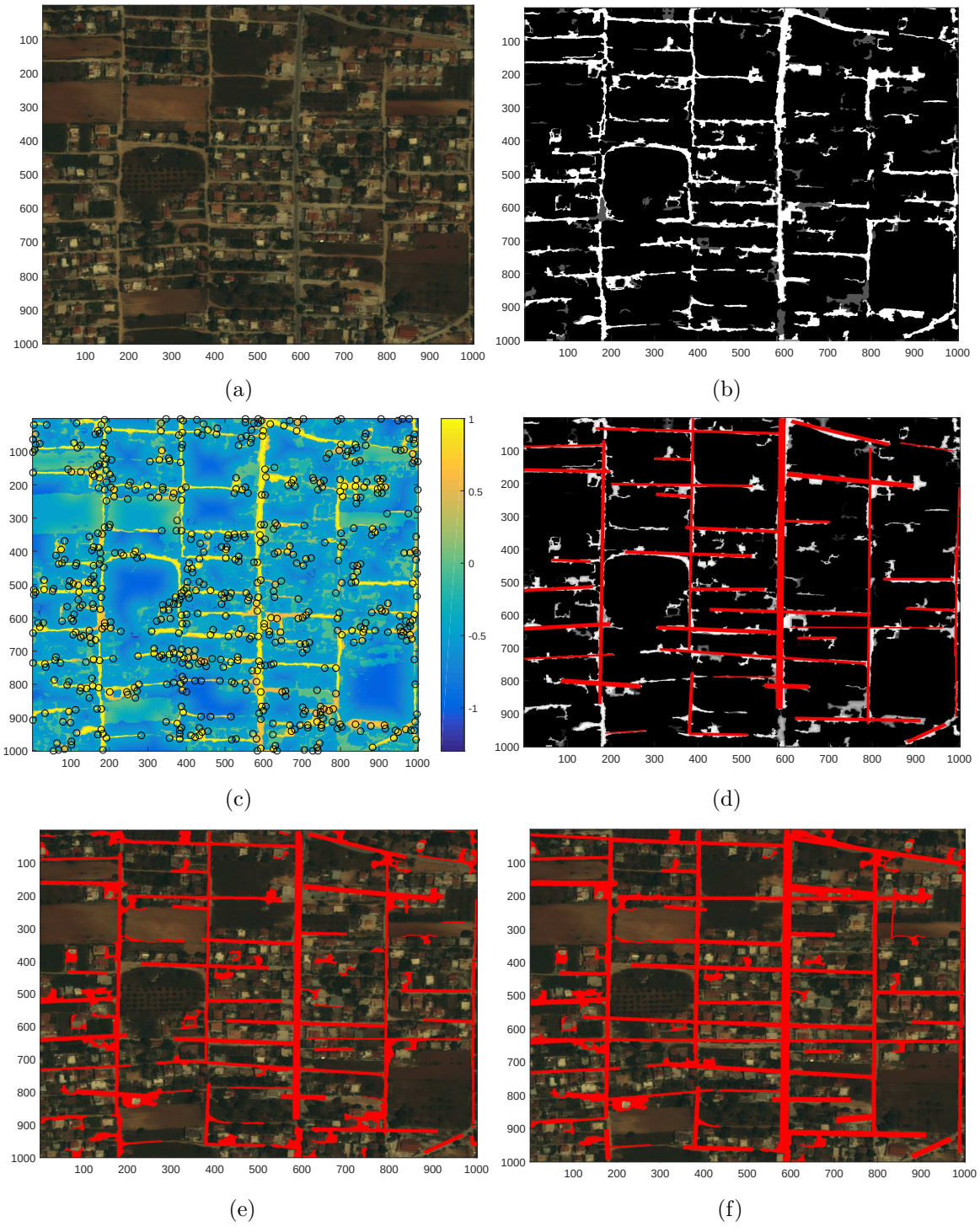


Figure 15: **(a)** Original image. **(b)** Initial Classification. **(c)** The PEP set projected on I_m. **(d)** The detected line segments using LPSA algorithm. **(e)** The final result of the first level of road detection method. **(f)** The final result of the second level of road detection method.

Next, the road expansion set RE is computed from the sets $\{HR, LR, NR\}$. According to the Road Expansion method, a region that belongs to HR or to LR and it is adjacent to a region of HR should belong to the road expansion set RE, resulting at an initial road map.

6.2. Linear pattern detection

In this step, we propose a method to fill some road gaps and to correct segmentation errors using linear patterns. This method is inspired by the work in Panagiotakis and Kokinou (2015), where linear patterns of geological faults are detected. The problem of linear pattern detection is reduced to an optimization problem that takes into account the road shape and topology features, such as the distance and angle between the road line segments structures. Initially, a sampling set $PEP = \{p_k, k \in \{1, ..., N\}\}$ of the most probable end points of line segments is computed. Then a sampling process is executed on the set of line segments that correspond to each pair of points of PEP with distance less than a threshold (e.g. 500 pixels), getting the linear patterns of roads. The distance threshold is used in order to decrease the computation cost.The size of PEP should be low in order to get a computationally effective method Panagiotakis and Kokinou (2015). In this work, PEP is given by the end points and the joints of skeleton Lam et al. (1992) of RE, reducing the size of PEP without loosing in detection accuracy. In the second level of road detection, the intersections between the extensions of the detected line segments are also added in PEP.

Next, we compute the road enhancement image I_m, where the road pixels are enhanced taking into account the shape characteristics of RE, the *a posteriori* probability of a region to be road Pr_{road}, the *a posteriori* probability of a region to be building Pr_{build} and the B_{NDVI} image as defined in the initialization step of this section. I_m is positive on pixels that probably belong to "road" class and negative on pixels that probably do not belong to this class. Let DF_{RE} and $DF_{\overline{RE}}$ be the distance transform of the binary images that correspond on RE and its complementary $\overline{RE}$ set, respectively. The pixels of RE that are close to its skeleton are enhanced (local maxima of DF_{RE}), while the pixels of non-road expansion set $\overline{SE}$ that are close to the middle of wide regions get negative values according to the following equation

$$SC_{RE}(p) = \begin{cases} e^{-(\frac{DF_{RE}(p) - DF_{RE}(lm(p))}{DF_{RE}(lm(p))})^2}, & p \in RE \\ -1 + e^{-(\frac{DF_{\overline{RE}}(p)}{W_r})^2}, & - \end{cases} \quad (11)$$

where W_r denotes a user defined parameter that corresponds on the maximum width of roads e.g. 40 pixels and $lm(p)$ corresponds on the maximum of DF_{RE} in a small neighborhood of p (e.g. 5×5 pixels) ($DF_{RE}(lm(p)) \geq DF_{RE}(p)$). It holds that when p is classified as "road", the higher value of $Pr_{road}(p) + SC_{RE}(p)$ the higher value of probability of p to belong to "road" class. When p is not classified as "road",the lower value of $SC_{RE}(p) - Pr_{build}(p) - B_{NDVI}(p)$, the lower value of probability of p to not belong to "road" class. Thus, the road enhancement image I_m is given by the following

equation

$$Im(p) = \frac{1}{2}(Pr_{road}(p) + SC_{RE}(p) - Pr_{build}(p) - B_{NDVI}(p)) \quad (12)$$

It holds that $I_m \in [-1.5, 1]$. However, usually it holds that $Pr_{build}(p) + B_{NDVI}(p) \leq 1$, which means that $I_m \in [-1, 1]$.

Figure 15(c) depicts an example of PEP set of Fig. 15(a), projected on I_m. Thanks to the combination of shape and content based characteristics of the previous stages, it holds that most of the pixels that belong on the road network have been correctly emphasized, with low number of false alarms. However, due to segmentation errors there exist several road gaps(discontinuities) and small errors on borders. The goal of the Linear Pattern Detection step is to correct them.

Then, the problem formulation for the detection of linear patterns of roads is described. Let $G = (S, W)$ be the complete graph that represents the possible end points and the connections between them. Let the line segment $\overline{p_i p_j}$, where p_i and $p_j \in S$ and its two parallel equal length lines segments $\overline{p'_i p'_j}$ and $\overline{p''_i p''_j}$ that are equidistant from $\overline{p_i p_j}$, with W_r distance. Due to W_r, it holds that if the line segment $\overline{p_i p_j}$ lies on a road line segment, then the line segments $\overline{p'_i p'_j}$ and $\overline{p''_i p''_j}$ probably lie out of road.

The edge weight between two points $p_i, p_j \in S$ is related to the potency of the corresponding linear road pattern (line segment $\overline{p_i p_j}$). This means that the values $I_m(x)$, where x belongs on $\overline{p_i p_j}$, should be high. At the same time, the values $I_m(x)$ on points x, that belong on line segments $\overline{p'_i p'_j}$ and $\overline{p''_i p''_j}$, should be low (negative). Therefore, in this work the positive edge weight $W(p_i, p_j)$ is given by:

$$\widehat{W_s}(p_i, p_j) = \sum_{x \in \overline{p_i p_j}} I_m(x) \quad (13)$$

$$\widehat{W_p}(p'_i, p'_j) = \sum_{x \in \overline{p'_i p'_j} \wedge I_m(x) > 0} I_m(x) \quad (14)$$

$$\widehat{W}(p_i, p_j) = \widehat{W_s}(p_i, p_j) - \frac{1}{4}(\widehat{W_p}(p'_i, p'_j) + \widehat{W_p}(p''_i, p''_j)) \quad (15)$$

$$W(p_i, p_j) = max(0, \widehat{W}(p_i, p_j)) \cdot W_d(p_i, p_j) \quad (16)$$

where $W_d(p_i, p_j)$ is only used on the second level of road detection taking into account parallelism and perpendicularity of roads in a global sense. The edge weight of Eq. (16) has the advantage that it is higher on large and strong linear patterns, since it holds that if $\overline{ab} \supset \overline{cd}$ then $\widehat{W_s}(a, b) \geq \widehat{W_s}(c, d)$, where $\overline{ab}$ is a strong linear pattern. Let $L = \{\overline{p_i p_j}, p_i, p_j \in S\}$ be a set of representative linear patterns (line segments) of the roads. If we select as the optimal representative linear patterns L_{opt}, the edges that maximize the sum of the corresponding edge weights then we will oversample high "energy" roads.

Let us denote as $S_T(\overline{ab}, \overline{cd})$ the topology similarity between the line segment $\overline{ab}$, $\overline{cd}$ taking into account the angle between the two line segments θ, their shortest Euclidean distance $d(\overline{ab}, \overline{cd})$, and the Euclidean distances between the points of line segment $\overline{ab}$

and line segment $\overline{cd}$, $d(y, \overline{cd})$, $y \in \overline{ab}$:

$$S_T(\overline{ab}, \overline{cd}) = \begin{cases} 0 & , d(\overline{ab}, \overline{cd}) \geq \frac{W_r}{2} \\ \frac{1}{|\overline{ab}|_2} \sum_{y \in \overline{ab}} e^{-(\frac{d(y,\overline{cd})}{W_r})^4} \cdot cos(\theta) & , - \end{cases} \quad (17)$$

$S_T(\overline{ab}, \overline{cd}) \in [0, 1]$ is close to one when the line segments are quite "similar" (e.g. they are very close and parallel) meaning that $\overline{ab}$ is also represented by $\overline{cd}$. $S_T(\overline{ab}, \overline{cd})$ is zero when the line segments are not "similar" e.g. they are far away or vertical.

Then, L_{opt} is given by the set of edges that maximize the following function $ME(L)$ (see Eq. (19)), in order to take into account the similarity between the linear patterns, topology constraints and the representativeness attribute of L, emphasizing the fact that L should equally describe all the road parts.

$$\widehat{S}_T(\overline{p_i p_j}) = \max_{\overline{p_m p_n} \in L - \overline{p_i p_j}} S_T(\overline{p_i p_j}, \overline{p_m p_n}) \quad (18)$$

$$ME(L) = \sum_{\overline{p_i p_j} \in L} W(p_i, p_j) \cdot (1 - \widehat{S}_T(\overline{p_i p_j})) \quad (19)$$

The maximization of $ME(.)$ is sub-optimally solved by the LPSA (linear patterns selection algorithm) Panagiotakis and Kokinou (2015) by sequentially selecting the most representative linear patterns until $W(a, b) < W_r$. A robust estimation of linear pattern width is given getting the median value of $DF_{RE}(lm(p))$, $p \in \overline{p_i p_j}$. Figure 15(d) depicts the result of this method projected on I_m for the image of Fig. 15(a), where most of the roads are well represented.

6.3. Merging of linear patterns and road expansion image

The detection of linear patterns gives high performance results on straight parts of roads, but fails on turnings that are well captured by the Road Expansion algorithm. So, in this step we merge their results. Let LPI and REI be the resulting images of the LPSA and Road Expansion algorithms respectively. The "Road Detection" map is initialized by the LPI in order to keep the straight line borders. Then, the dilation of the LPI is computed using a circle structuring element of $\frac{W_r}{2}$ radius. Let MD be the image of points belonging to the "Road Expansion", but not belonging to the above dilated image, consisting mainly of turning parts and some noisy regions. For eliminating the noisy segments we retain the connected components of MD of an area of at least 100 pixels and we take the union of these components with the LPI. Figure 15(e) depicts with red color a result of the proposed merging technique algorithm.

6.4. Second Level of road detection

In the second level of road detection method, the three steps of the first level are repeated using as input the detected linear patterns of the first level (see Fig. 14), taking into account topology based features like parallelism and perpendicularity of roads in a global sense. Firstly, the probability density function of orientation of linear

patterns (PDF_d) is computed. All the angles are transformed in $[0, \pi)$. In order to take into account the perpendicularity between roads, for each orientation we have also added its vertical in the estimation of PDF_d. Then, in the second level of road detection method the $W_d(p_i, p_j)$ of Eq. 16 is defined by $W_d(p_i, p_j) = \frac{1}{2} + \frac{1}{2} \cdot PDF_d(\measuredangle\overline{p_i p_j})$, where $\measuredangle\overline{p_i p_j} \in [0, \pi)$ denotes the orientation of the line segment $\overline{p_i p_j}$. Figure 15(f) depicts the final result of the second level of road detection method using red color. It holds that some parts of detected roads (e.g. close to points (550, 550), (600,400), (650,250)) have been expanded improving the detection of the first level (see Fig. 15(e)).

Results of the first and second level of road detection on Recall, Precision and F-measure are given in Table 7. Similarly with Section 4, the computation of Precision and Recall is done using $d = 10$. According to our experiments the execution of the second level increases the Recall and F-measure of the first level at about 2.5% and 0.5%, respectively. The average F-measure is 75% improving the result of unsupervised classification of Table 6 at about 5.6%. In addition, there exists a significant improvement on road boundary detection that is not captured by the F-measure. This improvement can be visually seen if we compare the classification result of Fig. 15(b) with the global road detection result of Fig. 15(f). Even if the density, the color and the size of roads vary, the proposed method gives high performance results.

Table 7: Results of the first and second level of road detection on Recall, Precision and F-measure.

year	*First Level*			*Second Level*		
	Recall	Precision	F-measure	Recall	Precision	F-measure
2006	0.7309	0.7898	0.7592	0.7518	0.7786	0.7650
2007	0.6513	0.7003	0.6749	0.6765	0.7004	0.6882
2009	0.8122	0.7361	0.7723	0.8346	0.7230	0.7749
2010	0.7900	0.7541	0.7717	0.8144	0.7414	0.7762
2011	0.7640	0.7198	0.7412	0.7977	0.6974	0.7442

7. Experimental Results

The goal of this section is to present qualitative and numerical results of our method and, especially in the case of buildings, to provide a detailed performance evaluation using object level performance metrics that have been used before, as well as using the proposed object level performance metric described in Section 4 (testing phase of RF analysis). Furthermore, the pixel level performance of our building detection system on the SZTAKI-INRIA benchmark dataset is presented. By contrary, we do not give pixel level performance evaluation results for our dataset, since, as it is analyzed in the excellent work of Rutzinger et al. (2009), such an evaluation makes no sense in cases where datasets consist of a large number of small buildings (and a few large ones) and errors in the delineation of ground truth buildings are, almost surely, present.

After road network extraction, a final step is performed on regions that have been detected as buildings, in order to exclude by them subparts that overlap with the

detected roads. The resulting, altered or (possibly) newly created, regions are excluded by the segmentation map, if their size is smaller than a predefined threshold (set to 100 pixels in our case). In Fig. 16 are illustrated our final results of building detection (red color) juxtaposed to the ground truth (green color boundary), and simultaneously those of road detection (blue color) and the ground truth centerlines (white color), for a 1500×1500 tile of years 2007 and 2010. The results for the overall dataset are found in the tab "WP5" of page `http://erato.survey.ntua.gr/thalis/`.

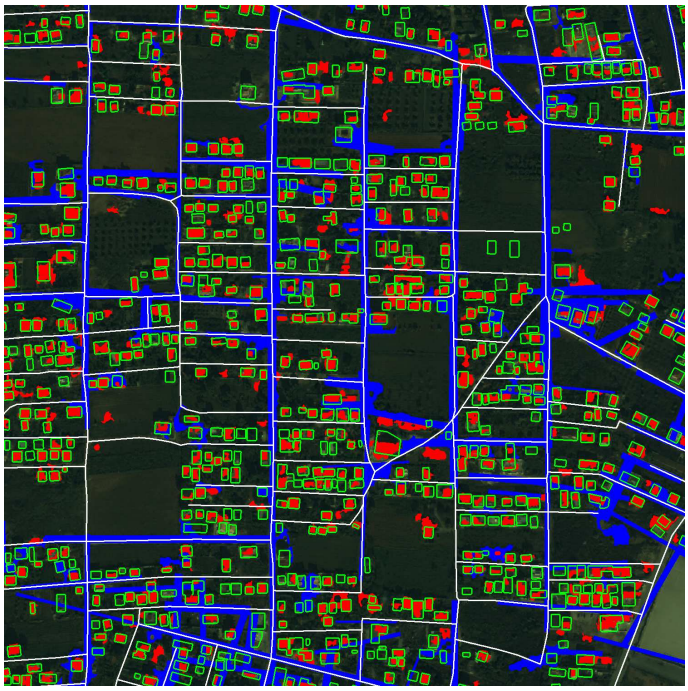
7.1. Object level evaluation of building detection

We compare in Table 8 the initial and the global classification result for "building" class to the best expected. We see that our unsupervised method performs quite well as the F-measure is near to the best expected, where the *Random Forest* classifier is trained using the whole ground truth. In any case, the performance is limited by the segmentation result, as our method is region-based. It is worth noting that the accuracy is measured by one-to-one correspondence of classified regions and the ground truth using the Jaccard similarity coefficient.

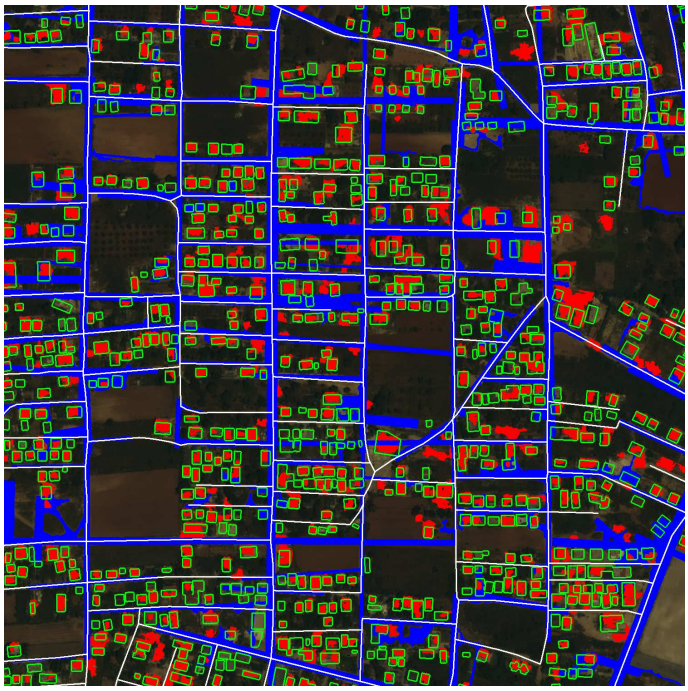
Table 8: Comparison of F-measure obtained for "building" to the best one for the same parameter set.

Year	2006	2007	2009	2010	2011
Random Forest (FS_1)	0.8091	0.7600	0.8267	0.8053	0.8234
Unsupervised (I)	0.7366	0.7021	0.7539	0.7307	0.7427
Unsupervised (G)	0.7418	0.7065	0.7531	0.7365	0.7482

Object level performance is quantified in a number of previous works (Ok et al. (2013); Ok (2013); Cheng et al. (2013); Han et al. (2014)) using the Jaccard similarity coefficient (Eq. 7) or other intersection-based overlap ratios, to compute a matching value between detected and ground truth objects. Jaccard similarity coefficient of the bounding boxes of detected and reference objects has been also used in Pascal challenge (Everingham et al. (2010)) of detecting various classes of objects. In that case, true positives are defined as the detected objects whose bounding boxes with those of reference buildings give Jaccard coefficient value greater than, or equal to a threshold T_{ov}. If the coefficient of bounding boxes is less than T_{ov}, the detected object is considered as FP. Furthermore, if several output bounding boxes overlap with a single ground truth bounding box in more than T_{ov}, only one is considered as TP and the others are considered as FP, to avoid multiple overlaps. Ignoring the technically substantial difference of using bounding boxes of objects instead of their areas, this approach is close to the proposed performance metric, since a one-to-one relationship between bounding boxes of detected objects and buildings is established, although it lucks the scaling of contribution of detected regions in the definitions of Recall and Precision and it tests the true positiveness of regions instead of buildings. Consequently, threshold T_{ov} controls mainly the accuracy of matching between bounding boxes and by no means relates to the counting of small buildings that may be contained in a

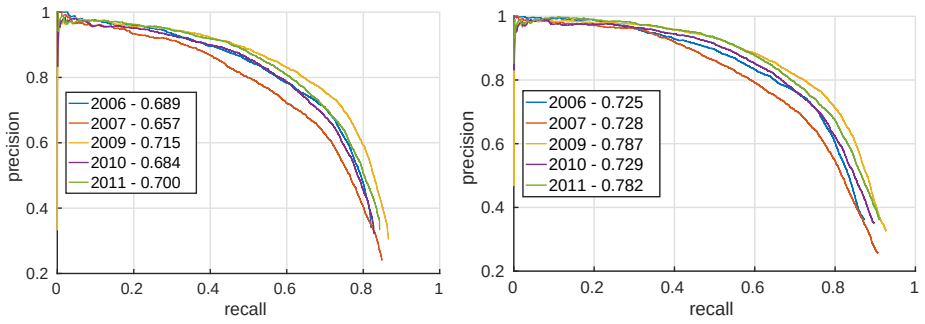


(a)



(b)

Figure 16: Final results of unsupervised detection for years 2007 (a) and 2010 (b). Detected buildings/roads are illustrated in red/blue color, respectively; green boxes and white lines correspond to the building ground truth and the centerlines of road ground truth, respectively.



(a) (b)

Figure 17: Precision Recall Curves (PRC) and Average Precision (AP) performance value for all years, using the performance evaluation of (a) the Pascal challenge method with $T_{ov} = 0.25$ and (b) the proposed method with $T_B = 0.25$.

single region, making the performance evaluation of Pascal challenge more application-independent (or application-insensitive).

Similar to other object based detection approaches (Cheng et al. (2013); Han et al. (2014)), given those definitions of TP and FP, we also adopt the Average Precision (AP) of Pascal challenge (Everingham et al. (2010)), to evaluate our unsupervised "building" detection approach, using $T_{ov} = 0.25$, to quantify the differences with the proposed performance evaluation method. The PRC as well as the AP for each year are depicted in Fig. 17(a). The confidence for each detected object, required by the computation of AP, is provided by the *a posteriori* probability of the class "building" given the feature vector of the output object. In Fig. 17(b), the PRC and AP values, that are obtained for all years using the proposed performance evaluation method, are shown. As expected by the discussion above, AP is lower in the plot of Fig. 17(a) for all years, while the better quality of the proposed, unsupervised building detection method for years 2009 and 2011 is highlighted by the AP of the proposed performance evaluation metric for those years, in Fig. 17(b).

7.2. Evaluation of the overlap threshold

The role of the overlap threshold in the new performance evaluation metrics has been also investigated, given the segmentation objects for each year and the fact that our dataset consists of many small buildings and consequently, errors in ground truth buildings are, almost certainly, present. The AP obtained for each year and for threshold values

$$\{1/20, 1/10, 1/7, 1/5, 1/4, 1/3, 1/2\},$$

roughly corresponding to groups of 20, 10, 7, 5, 4, 3 and 2 buildings respectively, is graphically depicted in the plot of Fig. 18. In that plot, the measured AP drops steeply for thresholds above 0.33, while at exactly 0.25, the AP of the five years quantifies their segmentation efficiency in two groups, the first one consisting of years 2006, 2007 and 2010 and the second including years 2009 and 2011, for whom the better segmentation

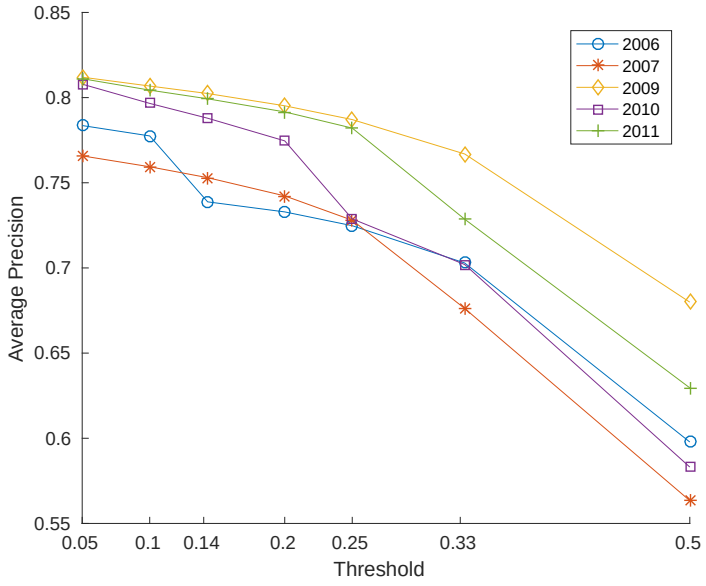


Figure 18: Average Precision (AP) values, for all years and thresholds $\{1/20, 1/10, 1/7, 1/5, 1/4, 1/3, 1/2\}$. .

quality is achieved. However, at 0.33 the order of years corresponds to the perceived quality of segmentation, upon which building detection is performed.

Another interesting outcome of the plot is that the order of classification efficiency of years, changes for thresholds below 0.25, indicating that evaluation below this threshold value should be avoided, given the dataset. Concluding, the threshold value which achieves reliable and acceptable performance evaluation results, strongly depends on the size of buildings and most likely on other parameters, such as the quality of RSI under examination.

7.3. Comparison to other building detection methods

In what follows, we give comparison results for the SZTAKI-INRIA building detection dataset (Benedek et al. (2012)) [2], in order to test the robustness and the (as much as possible) effortless re-usability of our building detection approach to aerial-borne and satellite-borne input images with quite different capturing conditions and visual content. A description of the dataset is also included in (Cheng and Han (2016)). It consists of 2 aerial (with code-names BUDAPEST and SZADA) and 4 satellite (code-named CÔTE D' AZUR, BODENSEE, NORMANDY and MANCHESTER) high-resolution study cases, where only the appearance channels Red, Green and Blue are available and no resolution information is provided. Benedek et al. (2012) used this dataset in order to compare their Marked Point Process (MPP) (Descombes and

[2] http://web.eee.sztaki.hu/remotesensing/building_benchmark.html

Zerubia (2002)) building detection algorithm, to a number of other building detection methods (Sirmaçek and Ünsalan (2009); Sirmaçek and Ünsalan (2011); Sirmaçek and Ünsalan (2008); Müller and Zaum (2005)), concluding that their model surpasses the other methods, using either pixel or object level evaluation metrics.

Although the normalized difference NIR-Red was absent, our automatic OBIA-based detection system has been performed on this dataset almost "as is", with the additional effort of tuning a number of parameters, a fact that highlights its re-usability. The segmentation and classification results of our method on the sub-image of BUDAPEST, selected by Benedek et al. (2012) to demonstrate their results, are shown in images (a) and (b) of Fig. 19, respectively. The ground truth of this sub-image is depicted in Fig. 19(c). As it is evident by those images, our method over-segments rooftops to their parts when they are characterized by obviously perceived, large intensity (luminance) differences. However, those parts are classified to class "building" by our unsupervised classifier and could be merged back to whole buildings, although we consider that such a post-processing modification to our detection system is beyond the scope of this work.

(a) (b) (c)

Figure 19: Segmentation (a), unsupervised classification (b) and ground truth (c) for the sub-image of BUDAPEST selected by Benedek et al. (2012) to demonstrate their results. Segmentation regions that are classified to class "building" by the proposed unsupervised classifier are shown with their natural color in image (b).

The pixel level performance criteria used in (Benedek et al. (2012)) were the same as in other works (Aksoy et al. (2012); Ok et al. (2013); Ok (2013)). Detected (or output) and ground truth (reference) buildings are used to define the four categories of pixel areas, namely, True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). TP represents those pixel areas containing both detected and reference buildings. TN represents areas without reference or detected buildings. FP represents areas containing detected buildings but without reference buildings. FN represents undetected building areas.

Following the same approach, the performance evaluation results of our work for the images in this dataset are given in Table 9 (columns "Prop."). Precision (Pr) and Recall

(Rc) for the images contained in the dataset are reproduced by (Benedek et al. (2012)) for methods Edge Verification (EV) (Sirmaçek and Ünsalan (2008)) and Segment-Merge (SM) (Müller and Zaum (2005)) as well. In that table as well as in the freely available dataset, images ABIDJAN and BEIDJING of (Benedek et al. (2012)) are not included. Furthermore, our results refer to the overall set of the reference buildings that are actually included in the dataset, which differs from the set of reference buildings reported by Benedek et al. (2012), since a small portion of them (about 5% or less per case study) is used for training only. Taking also into account that Benedek et al. (2012) incorporate as prior knowledge to their method the assumption that buildings are rectangles, our building detection method achieves performance near that of MPP, even if the differences in ground truth data that were used by our two approaches, do not permit the accurate comparison between them using pixel level performance metrics and make object level performance comparison not applicable, due to the subjectivity that they introduce. In order to give a more fair comparison in pixel level, Precision and Recall values of the methods reported in (Benedek et al. (2012)) were used to compute TP and FP pixel numbers on the overall set of reference buildings for each case study of the dataset and the overall F-measure for each method, shown in the corresponding column of Table 9, was recomputed.

Table 9: Performance evaluation results of the proposed method (Prop.) for the SZTAKI-INRIA building detection dataset, according to the pixel level metrics used by Benedek et al. (2012) to evaluate their MPP model. Precision (Pr) and Recall (Rc) for the images contained in the dataset are reproduced by (Benedek et al. (2012)) for methods Edge Verification (EV) (Sirmaçek and Ünsalan (2008)) and Segment-Merge (SM) (Müller and Zaum (2005)) as well.

Data Set	EV		SM		MPP		Prop.	
	Pr	Rc	Pr	Rc	Pr	Rc	Pr	Rc
BUDAPEST	0.73	0.46	0.84	0.61	0.82	0.71	0.79	0.61
SZADA	0.61	0.62	0.79	0.71	0.93	0.75	0.74	0.67
CÔTE D' AZUR	0.73	0.51	0.75	0.61	0.83	0.69	0.76	0.67
BODENSEE	0.56	0.30	0.59	0.41	0.73	0.51	0.71	0.68
NORMANDY	0.60	0.32	0.62	0.55	0.78	0.60	0.72	0.58
MANCHESTER	0.64	0.38	0.60	0.56	0.86	0.63	0.79	0.72
Overall F-measure	0.502		0.614		0.721		0.699	

The proposed method has been implemented in Matlab, except of the MRF minimization of the segmentation module, which has been developed as a C++ executable and is called by the segmentation script, using file reading/writing in order to pass to it input and get back the segmentation result. We tested our implementation on a 64-bit, Dell Inspiron laptop machine running under Ubuntu 14.04 LTS operating system, equipped with an Intel Core i7-4500U CPU @1.80GHz×4 and 8GB SDRAM. We provide execution times for segmentation, classification and in total, per case study in Table 10. Case sizes in KiloPixels are also reported since they are different from those of Benedek et al. (2012). Most of the segmentation times reported in the corresponding

column of Table 10 is consumed by the MRF minimization method. Our implementation is fast, consuming on average 31.7 sec. per MegaPixel of visual data, compared to the corresponding results of Benedek et al. (2012), where the best reported, average computational time, achieved by the Gabor method (Sirmaçek and Ünsalan (2011)), is 51 sec. per MegaPixel. It is worth noting that execution times of our system could be further reduced, since our code is not computationally optimized.

Table 10: Execution times of the proposed algorithm.

Data Set	Size (kPix)	Execution Time (seconds)		
		Segmentation	Classification	Total
BUDAPEST	294	8.8	1.6	10.4
SZADA	1544	39.2	3.5	42.7
CÖTE D' AZUR	743	20	3	23
BODENSEE	563	23	3.4	26.4
NORMANDY	1170	30	6.2	36.2
MANCHESTER	1126	28.6	5.1	33.7
Average	907	25	3.8	28.8

7.4. Road detection performance evaluation

Concerning the "road" class we should compare the limit provided by the *Random Forest* learning algorithm with our results on independent and on global decisions. These results are given in Table 11. The gain of the global approach is about 5.6% in

Table 11: Comparison of F-measure obtained for "road" to the best one for the same parameter set.

Year	2006	2007	2009	2010	2011
Random Forest (FS_1)	0.8093	0.7846	0.8456	0.8496	0.8302
Unsupervised (I)	0.7081	0.6276	0.7214	0.7206	0.6883
Unsupervised (G)	0.7650	0.6882	0.7749	0.7762	0.7442

F-measure. Both the Recall and the Precision rate are improved by the global road detection method. Finally, 80% of the road network has been correctly extracted with a relatively low number of false positives. In addition, thanks to linear pattern fitting method the road boundaries are also well localized in high accuracy.

8. Conclusions

Building recognition and road extraction in peri-urban environment have been considered. Our approach was based on objects (regions) extracted by an automatic segmentation algorithm. The unsupervised classification that follows the segmentation stage, is based on a Bayesian approach using a small number of visual and shape features. The initial, independently classified "road" segments are then used for extracting

a road network as complete as possible. A new, global algorithm, that exploits the connectivity, rectilinearity, parallelism and perpendicularity properties of road network, is proposed, giving a substantial improvement in the quality of road detection.

We present now our main conclusions from our experience concerning the problem of building localization and road detection. Visual and shape features have been evaluated based on extensive data analysis. The main conclusion on visual features is that the two normalized spectral band differences (near infrared − red and red − green) and the luminance, are relevant to the detection of man-made objects in satellite images and preferable to the original captured data. Concerning the shape features we have tried to reduce their number, and for the main unsupervised classification module we found powerful both the mean distance of region pixels to region boundary and the mean distance of boundary pixels from the region centroid. The object size is also relevant, as its joint distribution with mean distances is different for the different classes.

Our approach is object-based, therefore an accurate segmentation method was needed. Hence, in a first stage the vegetation and soil areas have been localized. After the extraction of the urbanized areas, we found that a set of likely "road" segments could be identified. Finally, areas of various shapes and sizes have been segmented by a graph-based unsupervised segmentation algorithm. We adopted a Markovian model, with a number of sub-classes greater than that of the indented classes, in order to correctly model the mixture of the classes to be detected, in the feature space. We performed the important step of obtaining the distributions for the sub-classes, after pixel-wise vector quantization of the visual appearance data. After extraction of objects, we have proposed an unsupervised classification method for detecting "building" and "road" segments. Both visual appearance and region shape features were employed. We found preferable to identify sub-classes with almost identical and, if possible, unimodal distribution, for each class. We obtained good enough classification results in comparison with the reachability threshold set by a *Random Forest* classifier trained on the whole ground truth.

Conclusively, as it is evident by the results of the previous Section, we consider that the proposed framework for simultaneous building and road detection in peri-urban areas is solid, robust and, the most important, unsupervised, with promising quantitative and qualitative results, comparable to those obtained by supervised methods. Future work will include the application of our system in purely urban scenes where the distinction between man-made classes appears to be more challenging. Furthermore, an extension of the proposed methodology may include the detection of other classes of objects such as trees, grass, rivers and lakes/sea.

Acknowledgments

The authors would like to thank the associate editor and all anonymous reviewers for their constructive comments. This work has been partially co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference

Framework (NSRF) - Research Funding Program: THALIS-NTUA-UrbanMonitor.

References

Ahmadi, S., Zoej, M.J.V., Ebadi, H., Moghaddam, H.A., Mohammadzadeh, A., 2010. Automatic urban building boundary extraction from high resolution aerial images using an innovative model of active contours. Int. J. Applied Earth Observation and Geoinformation 12, 150–157. URL: http://dx.doi.org/10.1016/j.jag.2010.02.001, doi:10.1016/j.jag.2010.02.001.

Aksoy, S., Yalniz, I.Z., Tasdemir, K., 2012. Automatic detection and segmentation of orchards using very high resolution imagery. IEEE Transactions on Geoscience and Remote Sensing 50, 3117–3131. doi:10.1109/TGRS.2011.2180912.

Aytekin, O., Erener, A., Ulusoy, I., Duzgun, S., 2012. Unsupervised building detection in complex urban environments from multispectral satellite imagery. International Journal of Remote Sensing 33, 2152–2177. URL: http://dx.doi.org/10.1080/01431161.2011.606852, doi:10.1080/01431161.2011.606852.

Benedek, C., Descombes, X., Zerubia, J., 2012. Building development monitoring in multitemporal remotely sensed image pairs with stochastic birth-death dynamics. IEEE Transactions on Pattern Analysis and Machine Intelligence 34, 33–50. doi:10.1109/TPAMI.2011.94.

Blaschke, T., 2010. Object based image analysis for remote sensing. ISPRS Journal of Photogrammetry and Remote Sensing 65, 2 – 16. URL: http://www.sciencedirect.com/science/article/pii/S0924271609000884, doi:http://dx.doi.org/10.1016/j.isprsjprs.2009.06.004.

Blaschke, T., Hay, G.J., Kelly, M., Lang, S., Hofmann, P., Addink, E., Feitosa, R.Q., van der Meer, F., van der Werff, H., van Coillie, F., Tiede, D., 2014. Geographic Object-Based Image Analysis Towards a new paradigm. ISPRS Journal of Photogrammetry and Remote Sensing 87, 180 – 191. URL: http://www.sciencedirect.com/science/article/pii/S0924271613002220, doi:http://dx.doi.org/10.1016/j.isprsjprs.2013.09.014.

Botev, Z., Grotowski, J., Kroese, D.P., 2010. Kernel density estimation via diffusion. Annals of Statistics 38, 2916–2957.

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.

Cheng, G., Han, J., 2016. A survey on object detection in optical remote sensing images. ISPRS Journal of Photogrammetry and Remote Sensing 117, 11 – 28. URL: http://www.sciencedirect.com/science/article/pii/S0924271616300144, doi:http://dx.doi.org/10.1016/j.isprsjprs.2016.03.014.

Cheng, G., Han, J., Guo, L., Qian, X., Zhou, P., Yao, X., Hu, X., 2013. Object detection in remote sensing imagery using a discriminatively trained mixture model. ISPRS Journal of Photogrammetry and Remote Sensing 85, 32 – 43. URL: http://www.sciencedirect.com/science/article/pii/S0924271613001809, doi:http://dx.doi.org/10.1016/j.isprsjprs.2013.08.001.

Cheng, Y., 1995. Mean shift, mode seeking, and clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence 17, 790–799.

Comaniciu, D., Meer, P., 2002. Mean shift: a robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 24, 603–619. doi:10.1109/34.1000236.

Das, S., Mirnalinee, T.T., Varghese, K., 2011. Use of salient features for the design of a multistage framework to extract roads from high-resolution multispectral satellite images. IEEE Transactions on Geoscience and Remote Sensing 49, 3906–3931. URL: http://dx.doi.org/10.1109/TGRS.2011.2136381, doi:10.1109/TGRS.2011.2136381.

Descombes, X., Zerubia, J., 2002. Marked point process in image analysis. IEEE Signal Processing Magazine 19, 77–84. doi:10.1109/MSP.2002.1028354.

Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A., 2010. The Pascal Visual Object Classes (VOC) Challenge. International Journal of Computer Vision 88, 303–338. URL: http://dx.doi.org/10.1007/s11263-009-0275-4, doi:10.1007/s11263-009-0275-4.

Finley, T., Joachims, T., 2008. Training Structural SVMs when Exact Inference is Intractable, in: Proceedings of the 25th International Conference on Machine Learning, ACM, New York, NY, USA. pp. 304–311. URL: http://doi.acm.org/10.1145/1390156.1390195, doi:10.1145/1390156.1390195.

Han, J., Zhou, P., Zhang, D., Cheng, G., Guo, L., Liu, Z., Bu, S., Wu, J., 2014. Efficient, simultaneous detection of multi-class geospatial targets based on visual saliency modeling and discriminative learning of sparse coding. ISPRS Journal of Photogrammetry and Remote Sensing 89, 37 – 48. URL: http://www.sciencedirect.com/science/article/pii/S0924271614000033, doi:http://dx.doi.org/10.1016/j.isprsjprs.2013.12.011.

Herman, G.T., 2009. Fundamentals of Computerized Tomography: Image Reconstruction from Projections. 2nd ed., Springer Publishing Company, Incorporated.

Hu, J., Razdan, A., Femiani, J., Cui, M., Wonka, P., 2007. Road network extraction and intersection detection from aerial images by tracking road footprints. IEEE Transactions on Geoscience and Remote Sensing 45, 4144–4157. URL: http://dx.doi.org/10.1109/TGRS.2007.906107, doi:10.1109/TGRS.2007.906107.

Huang, X., Zhang, L., 2009. Road centreline extraction from highresolution imagery based on multiscale structural features and support vector machines. International Journal of Remote Sensing 30, 1977–1987. URL: http://dx.doi.org/10.1080/01431160802546837, doi:10.1080/01431160802546837.

Inglada, J., 2007. Automatic recognition of man-made objects in high resolution optical remote sensing images by SVM classification of geometric image features. ISPRS Journal of Photogrammetry and Remote Sensing 62, 236 – 248. URL: http://www.sciencedirect.com/science/article/pii/S092427160700055X, doi:http://dx.doi.org/10.1016/j.isprsjprs.2007.05.011.

Jain, A.K., Ratha, N.K., Lakshmanan, S., 1997. Object detection using gabor filters. Pattern Recognition 30, 295 – 309. URL: http://www.sciencedirect.com/science/article/pii/S0031320396000684, doi:http://dx.doi.org/10.1016/S0031-3203(96)00068-4.

Jiang, X., Marti, C., Irniger, C., Bunke, H., 2006. Distance measures for image segmentation evaluation. EURASIP J. Appl. Signal Process. 2006, 209–209. URL: http://dx.doi.org/10.1155/ASP/2006/35909, doi:10.1155/ASP/2006/35909.

Jin, X., Davis, C.H., 2007. Vehicle detection from high-resolution satellite imagery using morphological shared-weight neural networks. Image and Vision Computing 25, 1422 – 1431. URL: http://www.sciencedirect.com/science/article/pii/S0262885606003544, doi:http://dx.doi.org/10.1016/j.imavis.2006.12.011.

Karantzalos, K., Argialas, D., 2009. A region-based level set segmentation for automatic detection of man-made objects from aerial and satellite images. Photogrammetric Engineering & Remote Sensing 75, 667–677.

Karantzalos, K., Paragios, N., 2009. Recognition-driven two-dimensional competing priors toward automatic and accurate building detection. IEEE Transactions on Geoscience and Remote Sensing 47, 133–144. URL: http://dx.doi.org/10.1109/TGRS.2008.2002027, doi:10.1109/TGRS.2008.2002027.

Kauffman, L., Rousseeuw, P., 1990. Finding Groups in Data. An Introduction to Cluster Analysis. J. Wiley and sons.

Kluckner, S., Bischof, H., 2009. Semantic classification by covariance descriptors within a randomized forest, in: Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on, pp. 665–672. doi:10.1109/ICCVW.2009.5457638.

Komodakis, N., Tziritas, G., 2007. Approximate labeling via graph cuts based on linear programming. IEEE Transactions on Pattern Analysis and Machine Intelligence 29, 1436–1453. URL: http://doi.ieeecomputersociety.org/10.1109/TPAMI.2007.1061, doi:10.1109/TPAMI.2007.1061.

Kumar, S., Hebert, M., 2003. Discriminative random fields: A discriminative framework for contextual interaction in classification, pp. 1150–1157. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-0344120654&partnerID=40&md5=5d4cd5a9f9774313672e05d445f103d3. cited By 186.

Lafferty, J.D., McCallum, A., Pereira, F.C.N., 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, in: Proceedings of the Eighteenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. pp. 282–289. URL: http://dl.acm.org/citation.cfm?id=645530.655813.

Lam, L., Lee, S.W., Suen, C.Y., 1992. Thinning methodologies-a comprehensive survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 14, 869–885.

Laptev, I., Mayer, H., Lindeberg, T., Eckstein, W., Steger, C., Baumgartner, A., 2000. Automatic extraction of roads from aerial images based on scale space and snakes. Mach. Vis. Appl. 12, 23–31. URL: http://dx.doi.org/10.1007/s001380050121, doi:10.1007/s001380050121.

Li, S.Z., 2009. Markov Random Field Modeling in Image Analysis. 3rd ed., Springer Publishing Company, Incorporated.

Mayer, H., 1999. Automatic Object Extraction from Aerial ImageryA Survey Focusing on Buildings. Computer Vision and Image Understanding 74, 138 – 149. URL: http://www.sciencedirect.com/science/article/pii/S1077314299907506, doi:http://dx.doi.org/10.1006/cviu.1999.0750.

Mena, J.B., 2003. State of the art on automatic road extraction for GIS update: a novel classification. Pattern Recognition Letters 24, 3037–3058. URL: http://dx.doi.org/10.1016/S0167-8655(03)00164-8, doi:10.1016/S0167-8655(03)00164-8.

Mille, J., Bon, R., Cohen, L., 2008. Region-based 2D deformable generalized cylinder for narrow structures segmentation, in: European Conference on Computer Vision (ECCV), Springer. pp. 392–404. URL: http://liris.cnrs.fr/publis/?id=4478.

Mnih, V., Hinton, G.E., 2010. Learning to detect roads in high-resolution aerial images, in: Daniilidis, K., Maragos, P., Paragios, N. (Eds.), Proceedings, Part VI ECCV 2010 - 11th European Conference on Computer Vision, Springer. pp. 210–223. URL: http://dx.doi.org/10.1007/978-3-642-15567-3, doi:10.1007/978-3-642-15567-3.

Müller, S., Zaum, D.W., 2005. Robust building detection in aerial images, in: Proc. Joint Workshop of ISPRS and DAGM: CMRT05, Vienna, Austria, pp. 143–148. URL: ftp://ftp.tnt.uni-hannover.de/pub/papers/2005/CMRT05-SMDWZ.pdf.

Myneni, R.B., Hall, F.G., Sellers, P., Marshak, A., 1995. The interpretation of spectral vegetation indexes. IEEE Transactions on Geoscience and Remote Sensing 33, 481–486. doi:10.1109/36.377948.

Ok, A.O., 2013. Automated detection of buildings from single VHR multispectral images using shadow information and graph cuts. ISPRS Journal of Photogrammetry and Remote Sensing 86, 21 – 40. URL: http://www.sciencedirect.com/science/article/pii/S0924271613002050, doi:http://dx.doi.org/10.1016/j.isprsjprs.2013.09.004.

Ok, A.O., Senaras, C., Yuksel, B., 2013. Automated Detection of Arbitrarily Shaped Buildings in Complex Environments From Monocular VHR Optical Satellite Imagery. IEEE Transactions on Geoscience and Remote Sensing 51, 1701–1717. doi:10.1109/TGRS.2012.2207123.

Özdemir, B., Aksoy, S., Eckert, S., Pesaresi, M., Ehrlich, D., 2010. Performance measures for object detection evaluation. Pattern Recognition Letters 31, 1128 – 1137. URL: http://www.sciencedirect.com/science/article/pii/S0167865509002918, doi:http://dx.doi.org/10.1016/j.patrec.2009.10.016. Pattern Recognition in Remote Sensing, Fifth IAPR Workshop on Pattern Recognition in Remote Sensing (PRRS 2008).

Paisitkriangkrai, S., Sherrah, J., Janney, P., Hengel, A.V.D., 2015. Effective semantic pixel labelling with convolutional networks and conditional random fields, in: EARTHVISION 2015 Workshop, Computer Vision and Pattern Recognition Conference.

Panagiotakis, C., Kokinou, E., 2015. Linear pattern detection of geological faults via a topology and shape optimization method. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 8, 3–11.

Ravetz, J., Fertner, C., Nielsen, T., 2013. The dynamics of peri-urbanisation. Springer. pp. 13–44. doi:10.1007/978-3-642-30529-0_2.

Rouse, J.W., Haas, R.H., Schell, J.A., Deering, D.W., 1974. Monitoring vegetation systems in the Great Plains with ERTS, in: NASA Goddard Space Flight Center 3d ERTS-1 Symposium, pp. 309–317.

Rutzinger, M., Rottensteiner, F., Pfeifer, N., 2009. A Comparison of Evaluation Techniques for Building Extraction From Airborne Laser Scanning. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 2, 11–20. doi:10.1109/JSTARS.2009.2012488.

Sethian, J.A., 1999. Level set methods and fast marching methods : evolving interfaces in computational geometry, fluid mechanics, computer vision, and materials science. Cambridge monographs on applied and computational mathematics, Cambridge University Press, Cambridge, GB. URL: http://opac.inria.fr/record=b1095605.

Shufelt, J.A., 1999. Performance evaluation and analysis of monocular building extraction from aerial imagery. IEEE Transactions on Pattern Analysis and Machine Intelligence 21, 311–326. doi:10.1109/34.761262.

Silva, C., Centeno, J.A.S., 2010. Automatic extraction of main roads using aerial images. Pattern Recognition and Image Analysis 20, 225–233.

Sirmaçek, B., Ünsalan, C., 2008. Building detection from aerial images using invariant color features and shadow information, in: Computer and Information Sciences, 2008. ISCIS '08. 23rd International Symposium on, pp. 1–5. doi:10.1109/ISCIS.2008.4717854.

Sirmaçek, B., Ünsalan, C., 2009. Urban-area and building detection using sift keypoints and graph theory. IEEE Transactions on Geoscience and Remote Sensing 47, 1156–1167. doi:10.1109/TGRS.2008.2008440.

Sirmaçek, B., Ünsalan, C., 2010. Urban area detection using local feature points and spatial voting. IEEE Geoscience and Remote Sensing Letters 7, 146–150.

Sirmaçek, B., Ünsalan, C., 2011. A probabilistic framework to detect buildings in aerial and satellite images. IEEE Transactions on Geoscience and Remote Sensing 49, 211–221. URL: http://dx.doi.org/10.1109/TGRS.2010.2053713, doi:10.1109/TGRS.2010.2053713.

Talbot, H., Appleton, B., 2007. Efficient complete and incomplete path openings and closings. Image and Vision Computing 25, 416 – 425. URL: http://www.sciencedirect.com/science/article/pii/S0262885606002575, doi:http://dx.doi.org/10.1016/j.imavis.2006.07.021. international Symposium on Mathematical Morphology 2005.

Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y., 2005. Large Margin Methods for Structured and Interdependent Output Variables. J. Mach. Learn. Res. 6, 1453–1484. URL: http://dl.acm.org/citation.cfm?id=1046920.1088722.

Ünsalan, C., Boyer, K.L., 2005. A system to detect houses and residential street networks in multispectral satellite images. Computer Vision and Image Understanding 98, 423–461. URL: http://dx.doi.org/10.1016/j.cviu.2004.10.006, doi:10.1016/j.cviu.2004.10.006.

Ünsalan, C., Sirmaçek, B., 2012. Road network detection using probabilistic and graph theoretical methods. IEEE Transactions on Geoscience and Remote Sensing 50, 4441–4453. URL: http://dx.doi.org/10.1109/TGRS.2012.2190078, doi:10.1109/TGRS.2012.2190078.

Vakalopoulou, M., Karantzalos, K., Komodakis, N., Paragios, N., 2015. Building detection in very high resolution multispectral data with deep learning features, in: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium.

Valero, S., Chanussot, J., Benediktsson, J.A., Talbot, H., Waske, B., 2010. Advanced directional mathematical morphology for the detection of the road network in very high resolution remote sensing images. Pattern Recognition Letters 31, 1120–1127. URL: http://dx.doi.org/10.1016/j.patrec.2009.12.018, doi:10.1016/j.patrec.2009.12.018.

Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer-Verlag New York, Inc., New York, NY, USA.

Volpi, M., Ferrari, V., 2015. Semantic segmentation of urban scenes by learning local class interactions, in: EARTHVISION 2015 Workshop, Computer Vision and Pattern Recognition Conference.

Wang, J., Song, J., Chen, M., Yang, Z., 2015. Road network extraction: a neural-dynamic framework based on deep learning and a finite state machine. International Journal of Remote Sensing 36, 3144–3169. URL: http://dx.doi.org/10.1080/01431161.2015.1054049, doi:10.1080/01431161.2015.1054049.

Wiedemann, C., Heipke, C., Mayer, H., Jamet, O., 1998. Empirical evaluation of automatically extracted road axes, in: Empirical Evaluation Techniques in Computer Vision, pp. 172–187.

Xu, R., Wunsch, D., 2009. Clustering. Wiley-IEEE Press.

Yuan, J., Cheriyadat, A., 2013. Road segmentation in aerial images by exploiting road vector data, in: Fourth International Conference on Computing for Geospatial Research and Application, COM.Geo '13, San Jose, CA, USA, July 22-24, 2013, IEEE. pp. 16–23. URL: http://dx.doi.org/10.1109/COMGEO.2013.4, doi:10.1109/COMGEO.2013.4.