ISPRS Journal of Photogrammetry and Remote Sensing 86 (2013) 21–40

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

Automated detection of buildings from single VHR multispectral images using shadow information and graph cuts

Ali Ozgun Ok *

Department of Civil Engineering, Faculty of Engineering, Mersin University, 33343 Mersin, Turkey

ARTICLE INFO

Article history:
Received 21 October 2012
Received in revised form 4 September 2013
Accepted 5 September 2013
Available online 29 September 2013

Keywords:
Building detection
Shadow evidence
Graph cuts
Satellite imagery

ABSTRACT

In this study, we propose a novel methodology for automated detection of buildings from single very-high-resolution (VHR) multispectral images. The methodology uses the principal evidence of buildings: the shadows that they cast. We model the directional spatial relationship between buildings and their shadows using a recently proposed probabilistic landscape approach. An effective shadow post-processing step is developed to focus on landscapes that belong to building regions. The building regions are detected using an original two-level graph theory approach. In the first level, each shadow region is addressed separately, and building regions are identified via iterative graph cuts designed in two-label partitioning. The final building regions are characterised in a second level in which the previously labelled building regions are subjected to a single-step multi-label graph optimisation performed over the entire image domain. Numerical assessments performed on 16 VHR GeoEye-1 images demonstrate that the proposed approach is highly robust and reliable. A distinctive specialty of the proposed approach is its applicability to buildings with diverse characteristics as well as to VHR images with significantly different illumination properties.

1. Introduction

The detection of buildings from very-high-resolution (VHR) images is of great practical interest for a number of applications; including urban monitoring, change detection, estimation of human population, among others. Manual processing of images requires continuous human labour and attention. Hence, a large number of researchers have invested effort in developing approaches that require little or no human intervention. Because of those valuable efforts, an extensive number of research papers have been published whose primary objective is to detect and describe building objects from remotely sensed images. Furthermore, a number of review papers have described the work carried out in the past and further elaborated on the advantages and disadvantages of the developed approaches. In the context of aerial imagery, Mayer (1999) reported and discussed previous work published until the mid-1990s in an exceptional review in which the models and strategies of the developed approaches were deeply investigated and summarised. Baltsavias (2004) focused on different aspects of knowledge that can be used for object extraction, and this survey emphasised key points for the practical use of object extraction. The review section of Ünsalan and Boyer (2005) extended Mayer's work through a comparative analysis of the performances of the approaches developed until late 2003. With a major emphasis on the reconstruction of buildings, Brenner (2005) also presented a review that examined both image- and LIDAR-based reconstruction strategies together with their principal achievements. More recently, Haala and Kada (2010) published a review paper that investigated both the approaches devoted to building reconstruction from airborne and LIDAR elevation data and those addressing terrestrial data that provided facades for buildings.

This study is devoted to the automated detection of buildings from a single optical VHR satellite image. Therefore, in the following section, we primarily consider the previous studies aimed at automatically detecting buildings from monocular optical VHR image datasets and only briefly discuss other approaches that involved (aerial/satellite) image datasets in their proposed framework (Fig. 1).

1.1. Previous work

The pioneering studies for the automated detection of buildings were devoted to monocular aerial images in which low-level features, edge/line segments and/or corners were grouped to form building hypotheses (e.g., Huertas and Nevatia, 1988; Irvin and McKeown, 1989; Mohan and Nevatia, 1989; Liow and Pavlidis,

* Tel.: +90 324 361 0001/7263; fax: +90 324 361 0032.
E-mail addresses: oozgun@mersin.edu.tr, ozguneo@gmail.com

0924-2716/$ - see front matter © 2013 International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS) Published by Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.isprsjprs.2013.09.004

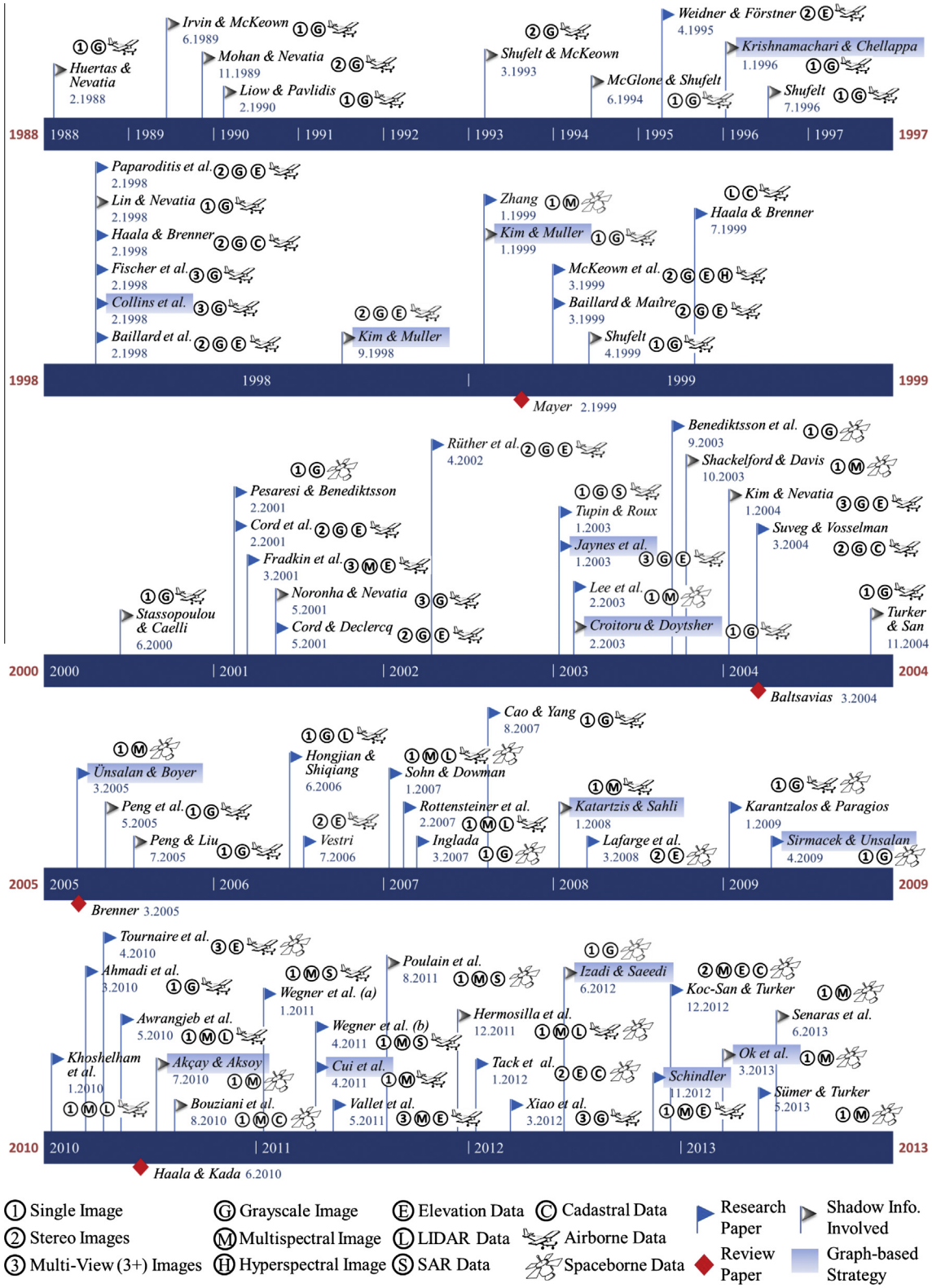

Fig. 1. Summary of the previous work.

1990). A study performed by Shufelt (1999) presented comparative results of different standalone systems (Irvin and McKeown, 1989; McGlone and Shufelt, 1994; Shufelt, 1996) developed for building detection from a single aerial image, and he concluded that none of the developed systems were capable of handling all of the challenges in building detection. Most of those methods rely on low-level edge information extracted from a single panchromatic image, and the shapes of the buildings are assumed to match a specific type of hypothesis to facilitate this detection.

The advances in low-orbit Earth imaging technologies have significantly increased the popularity of space-borne VHR sensors for a wide variety of remote-sensing-related applications. Thus, VHR

satellite images with multispectral information have motivated researchers to develop new approaches for building detection. A number of researchers have favoured classification strategies for the detection of buildings and urban regions. For instance, Zhang (1999) proposed a two-level framework for this classification. In the first level, a fused image was classified in an unsupervised manner using ISODATA clustering. The second level was devoted to filtering based on a modified co-occurrence matrix, and this process eventually improved the classification results of the first level. The study reported performance improvements of approximately 26% compared with the results achieved through normal texture filtering. Nevertheless, the results presented are not sufficiently detailed for investigation of urban buildings. Later, Pesaresi and Benediktsson (2001) and Benediktsson et al. (2003) tested differential morphological profiles (DMPs) for the classification of IRS 1-C and IKONOS-2 image data sets. In those approaches, the features were extracted from the DMPs defined by morphological transforms, and a neural network was used to classify the urban areas. Their approach assumes that the morphological profile of a structure is composed of only one significant derivative maximum; however, that is usually not the case for the structures commonly observed in complex environments. Another classification method was suggested by Lee et al. (2003), who used Extraction and Classification of Homogeneous (ECHO) classifier and ISODATA segmentation to detect buildings from IKONOS-2 images. In their approach, the supervised ECHO classifier detected the approximate locations and shapes of buildings, and ISODATA segmentation followed by Hough transformation were implemented to extract the building boundaries. Unfortunately, the detection step was completely based on the results of the supervised classifier, which might be the main reason for the moderate building detection rates (64.4%) achieved. In the same year, Shackelford and Davis (2003) combined the fuzzy pixel-based and object-based approaches to classify urban land cover from pan-sharpened multispectral IKONOS-2 imagery. This group integrated spectral and spatial information to discriminate urban land cover classes using the fuzzy pixel-based classifier. Subsequently, the shape, spectral, and neighbourhood information was used by the object-based classifier to complete the classification. The approach was tested only on an image patch selected from IKONOS-2 imagery, and the misclassifications between the classes of building and impervious surfaces produced the largest source of error in which 19% of building reference pixels were misclassified as impervious surfaces. Ünsalan and Boyer (2005) later proposed a system to simultaneously detect houses and streets in residential areas using IKONOS-2 images. Their approach introduced a k-means clustering process that takes into account the spatial information and applies a shape-based strategy to focus on the regions that belong to houses and streets. Their method produced a highly successful detection rate (94.8%) for house detection. However, the strategy presented is only practical for the type of house and street formations observed in North America because of the assumptions involved during the detection. With a different approach, Inglada (2007) attempted to recognise man-made objects from SPOT-5 scenes using support vector machine (SVM) classification. This work identified the objects based on their geometric descriptions, i.e., geometric invariants and Fourier–Mellin descriptors, and performed a supervised SVM classification to learn and separate various classes, including the isolated buildings. Nevertheless, this approach is not tailored to detecting building regions in an image but only the patches with certain sizes that are labelled as buildings. In a recent work, Senaras et al. (2013) proposed a decision fusion approach to detect building regions from a single VHR optical satellite image. The approach employed a two-layer hierarchical ensemble learning architecture in which the first level was devoted to individual base-layer classifiers, whereas in the second level, the output decision spaces of each classifier were fused in a hierarchical ensemble-learning algorithm known as Fuzzy Stacked Generalisation (FSG). The approach was assessed on test patches selected from a single QuickBird image, and an overall performance of 84% was reported. Nevertheless, their approach assumes statistical stability of the training and test data, and if this is not the case, the performance values will decrease dramatically. In another recent work, Sumer and Turker (2013) proposed an approach based on an adaptive fuzzy-genetic algorithm for building detection using an IKONOS-2 satellite image. Their approach combined a hybrid system of evolutionary techniques with a traditional classification method (Fisher's linear discriminant) and an adaptive fuzzy logic component. The ten image patches selected yielded detection performances in the range of 50–91%, and the authors argued that the results achieved might not be the optimal solution. In addition, careful selection of parameters is required to achieve the best detection results.

A different method used to extract buildings from single images was based on active contours. Peng et al. (2005) developed a method that combined the regional features of an image with context, and partial snakes were proposed using the direction of the cast shadows. Their approach assessed only the quality of the shapes of the buildings detected, and this work stated that the approach was not designed for complex buildings located in urban environments. The study conducted by Cao and Yang (2007) presented a method based on a modified Chan–Vese model. A three-stage level-set evolution strategy was proposed to minimise the modified model. Unfortunately, the method only extracts the boundaries of the man-made regions from aerial images and not individual buildings. Karantzalos and Paragios (2009) later incorporated the prior shape knowledge of buildings in active contours to extract buildings. The objective function for minimisation was designed to select both the most appropriate prior building model and the transformation used to relate the model to the image. The authors reported overall detection performances of greater than 80%. Nevertheless, prior building shape models are the key input to achieving such performance ratios. In a more recent research, Ahmadi et al. (2010) developed a new model based on level-set formulation to extract buildings using active contours. Their snake model was initialised by selecting sample data from the buildings and background, and tests carried out on a single aerial image indicated a completeness ratio of 80%. However, the number of building and background classes must be precisely known a priori to achieve the best results.

Graph-based approaches were also introduced in several works and were tested with monocular optical images. In Krishnamachari and Chellappa (1996), Markov Random Fields (MRFs) were used to group line segments for delineation of buildings of specific shapes. Lastly, the shapes of the grouped line segments were improved with active contours. The tests were performed on aerial images from rectangular buildings, but no quantitative results were provided. Katartzis and Sahli (2008) also used MRFs in a stochastic framework for the identification of building rooftops. This group used MRFs to define the dependencies between the building hypotheses, and a stochastic optimisation process was performed for verification of the hypotheses. The tests were conducted on a set of aerial image patches covering a few isolated buildings, and the approach yielded good detection and reconstruction performances. However, the approach is valid only for building rooftops with low inclination. Kim and Muller (1999) modelled low-level linear features as the vertices of a graph, and the building hypotheses were extracted by searching closed loops in the line relation graph. Because geometric relationships between line segments are considered during the generation of the hypotheses, the approach is limited to certain building shapes. In another study, Croitoru and Doytsher (2003) proposed Right-Angle Graphs

(RAG) to detect buildings with right-angled shapes. During the generation of the graph, the Hough space and the right-angle corners that may belong to the buildings were considered. This approach is only practical in urban areas with regular building layouts because a priori knowledge of the building models is necessary to initialise the approach. The system developed by Ünsalan and Boyer (2005) also used a graph to detect houses and streets, and the vertices of the graph were generated from binary balloons used to represent house or street segments. As mentioned earlier, their approach is limited to the detection of the house and street formations observed in North America. Sirmacek and Unsalan (2009) used point features for the detection of buildings, and in their work, the vertices of the graph acted as the key points provided by the scale invariant feature transform (SIFT). This work robustly evaluated the approach on a large set of IKONOS-2 images acquired over regularly developed environments, and successful results of building detection (88.4%) were reported. Nonetheless, their approach is designed primarily for environments with isolated buildings, and specific building templates are one of the prerequisites for the detection of buildings. In Cui et al. (2011), the vertices of the graph represented the building corners, and the building edges connecting two corners represented the edges of the graph; the building hypotheses were formed after detecting the cycles in the graph. However, the output of this approach is completely based on the quality of the initial segmentation step performed to identify the building locations. In one of the latest studies, Izadi and Saeedi (2012) developed an approach that employed image primitives, i.e., lines and line intersections, to infer a set of building rooftop hypotheses using a graph-based search. This group validated their approach on 20 QuickBird image patches and reported good detection (95.2%) and reconstruction results. Unfortunately, their approach works only for flat or flat-looking roofs, and the assumption of smoothness of the rooftops might only be valid for certain roof shapes. In a rather recent paper, Schindler (2012) conducted a review of graph cuts for image classification, and the paper clearly emphasised the importance of smoothness priors included in a classification framework. This work found that the assumption of smoothness enforced during the classification could improve the performances up to 33%. However, it should be noted that although the proposed framework is capable of solely handling the images, the presented results take advantage of the height information used during the classification.

A common fact for most of the building detection approaches developed for single image processing is that they employ the evidence of shadows for verification of the generated hypotheses. Because a cast shadow is notably strong evidence of an off-terrain object, its presence can be efficiently used to verify the existence of a building structure. Consequently, strategies for shadow verification are investigated thoroughly in previous work (e.g., Huertas and Nevatia, 1988; Irvin and McKeown, 1989; Mohan and Nevatia, 1989; Liow and Pavlidis, 1990; Shufelt and McKeown, 1993; McGlone and Shufelt, 1994; Lin and Nevatia, 1998; Stassopoulou and Caelli, 2000; Turker and San, 2004; Peng and Liu, 2005; Katartzis and Sahli, 2008; Izadi and Saeedi, 2012). In most of those studies, the shadow areas were not explicitly extracted for verification, and only a dark region was searched in the close neighbourhood of a building hypothesis. However, as explicitly stated in Irvin and McKeown (1989), shadows provide a large amount of information on the structure of man-made objects without the necessity of developing an explicit model of a building's structure. Therefore, if shadow regions can be accurately extracted from images, this information could be highly useful for automated detection of buildings with arbitrary shapes. A number of researchers have also published works with the sole purpose of shadow detection, and the developed approaches frequently exploit multi-band information. For example, Polidorio et al. (2003) proposed an approach based on a colour image (RGB). This group found that the saturation and intensity components of a shadow region follow certain characteristics, and the shadow regions could be detected by applying a thresholding scheme. In Sarabandi et al. (2004), the colour space $C_1C_2C_3$ (Gevers and Smeulders, 1999) generated from RGB images was suggested for shadow detection. This group also argued that the best non-linear transformation for the purpose of shadow detection was the colour space $C_1C_2C_3$. Tsai (2006) compared five different colour spaces and found Hue-Saturation-Intensity (HSI) space to be the optimal choice for automated shadow detection. Chung et al. (2009) further improved on Tsai's work by employing a ratio index and an automated thresholding strategy. In a recent work, a new index for detecting shadow areas was proposed by Teke et al. (2011). Their approach also relies on the HSI space but takes advantage of near-infrared (NIR) information via a darkness image generated from a false colour composite in which the green, red, and NIR bands were used. To the best of our knowledge, only two studies have exploited shadow regions to estimate the possible locations of buildings in early stages of their methodology. The first study was performed by Akçay and Aksoy (2010), who detected candidate building regions using both shadow information and directional spatial constraints. Thereafter, the final buildings were determined after clustering the candidate regions using minimum spanning trees. Their approach fully relies on segments generated by a watershed algorithm, and a major limitation is the user-defined thresholds that might degrade the performance of the approach in different images. The second study was performed by Ok et al. (2013) in which the shadow evidence was considered to focus on building regions. The directional spatial relationship between buildings and their shadows was the basis of the approach. The final building regions were detected using iterative graph cuts in which the information obtained from the shadow regions was effectively applied. The approach was tested for a variety of images, including those with notably challenging conditions. However, the quality of the shadow mask generated at the early stage of processing is the main concern and might seriously affect the final performance of the approach.

Obviously, the approaches developed for building detection are not limited to monocular images. Many researchers have developed strategies that benefit from stereo/multiple images. Suveg and Vosselman (2004) classified those strategies into two distinct groups: (i) systems that are extensions of monocular approaches, and (ii) systems that are originally designed to work with stereo/multiple images. The first group handles overlapping images one by one, similar to the monocular image processing, and uses additional images for verification (e.g., Mohan and Nevatia, 1989; Collins et al., 1998; Noronha and Nevatia, 2001; Xiao et al., 2012). The second group benefits from the stereo/multiple images at the earliest stages of the processing (e.g., Fischer et al., 1998; Cord and Declercq, 2001; Cord et al., 2001; Fradkin et al., 2001; Kim and Nevatia, 2004). Both groups of approaches were also evaluated in a study conducted by Paparoditis et al. (1998). In a different work, Kim and Muller (1998) used the information extracted from stereo/multiple images to assign the heights in a 2-D building hypothesis, whereas Jaynes et al. (2003) used that information only at the reconstruction stage of the detected buildings. In the studies conducted by Baillard et al. (1998) and Baillard and Maitre (1999), a stereo-based Digital Surface Model (DSM) was used to extract buildings. The DSMs are generally used to extract raised structures, and for that purpose, approaches such as thresholding (e.g., Weidner and Förstner, 1995; Rüther et al., 2002), local evidence (Paparoditis et al., 1998), automated segmentation and classification (e.g., Baillard et al., 1998; Baillard and Maitre, 1999; Cord and Declercq, 2001; Cord et al., 2001), marked point processes (e.g., Lafarge et al., 2008; Tournaire et al., 2010) or an existing Digital Terrain

Model (DTM) (e.g., Vestri, 2006; Koc-San and Turker, 2012) were applied.

A different trend in building detection is to fuse different data sources with images. The data sources used along with images cover a wide range. Until now, LIDAR data (e.g., Haala and Brenner, 1998; Haala and Brenner, 1999; Hongjian and Shiqiang, 2006; Rottensteiner et al., 2007; Sohn and Dowman, 2007; Awrangjeb et al., 2010; Khoshelham et al., 2010; Hermosilla et al., 2011), SAR data (e.g., Tupin and Roux, 2003; Poulain et al., 2011; Wegner et al., 2011a), hyper-spectral data (McKeown et al., 1999), or existing building layers in 2-D GIS databases (e.g., Haala and Brenner, 1998; Haala and Brenner, 1999; Suveg and Vosselman, 2004; Bouziani et al., 2010; Vallet et al., 2011; Koc-San and Turker, 2012; Tack et al., 2012) have been successfully treated as supplementary sources to the image data. In addition, libraries of predefined building models (e.g., Fischer et al., 1998; Croitoru and Doytsher, 2003; Jaynes et al., 2003; Suveg and Vosselman, 2004) can also serve as additional data sources.

To summarise, an automated approach must cope with a number of aspects, including the following: (i) different characteristics of the used images (e.g., spatial, spectral, and radiometric quality), (ii) unpredictable environmental and illumination factors during imaging (e.g., haze, snow cover, cloud cover, shadows and shades), (iii) the problems associated with perspective viewing (e.g., occlusions, relief distortions), and (iv) similar characteristics of building regions and their background (e.g., spectral, textural and shape-based). Therefore, if the current level of computational intelligence is taken into account, reliable extraction of building boundaries in an automated manner is still problematic, and development of new approaches is essential.

1.2. Proposed approach

This study presents an original approach for the automated detection of buildings from single very-high-resolution (VHR) multispectral images (Fig. 2). Our method is based on the fact that a 3D building structure should cast a shadow under suitable imaging conditions (Noronha and Nevatia, 2001). Therefore, the methodology begins with the detection and post-processing of shadow areas. To generate a probabilistic landscape for buildings, a recently proposed approach is used that benefits from the shadow areas and the prior knowledge of solar illumination direction. Thereafter, a landscape pruning process ensures that the landscapes generated from the shadow regions are not caused by vegetation cover. To accurately detect the building regions, we propose a graph theory framework consisting of two sequential levels. In the first level, the building regions are detected by iterative binary-label graph cut partitioning performed in automatically selected regions-of-interest (ROIs). After the partitioning, with the aid of the pre-computed shadow and vegetation masks, we divide the image into four distinct classes: *building*, *vegetation*, *shadow*, and *others*. In the second level, these four classes are optimised in single-step multi-label graph partitioning. This optimisation is performed over the entire image domain, and the final building regions are characterised after the second level partitioning. To do this, we extract the regions belonging to the building class after the second-level optimisation and confirm these regions with the previously generated probabilistic landscape. For the regions that could not be confirmed, we exploit the shadow information that may be revealed after the second-level partitioning, and the rejected regions are further tested for new shadow evidence. Lastly, the building mask is generated after applying a post-processing step to all regions validated with the shadow evidence. The assessments of the proposed approach are performed on a set of VHR GeoEye-1 images, and the experiments demonstrate the effectiveness and generality of the proposed framework for the detection of buildings in complex environments.

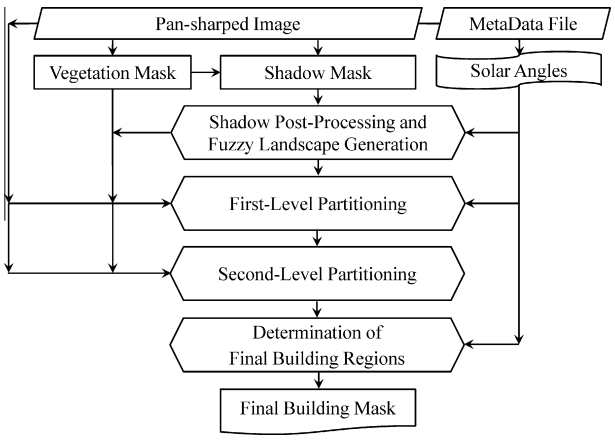

Fig. 2. Proposed methodology for building detection.

1.3. Contributions

Our new approach expands upon the recent work presented in Ok et al. (2013) in which we employed an approach for detection of buildings using shadow information. Numerous aspects of the approach are significantly improved in this study.

The first two contributions of this study are related to the post-processing of the shadow areas. To detect the shadows, we use a shadow detection method developed by Teke et al. (2011). This method is highly suited to detection of shadows in a generic manner with good precision. However, a major disadvantage of this approach is that the detection results are relatively sparse due to histogram-based thresholding. To overcome the problem in this study, the pixels detected as shadow are introduced as seed points into a constrained region-growth process. Additionally, we propose a new approach to enforce a minimum height threshold for buildings. In our previous work, we investigated the length of the shadow objects in the direction of illumination to enforce a height threshold for the buildings. However, this procedure may fail for the cases in which the cast shadow of a building is merged with a neighbouring shadow area due to objects such as garden walls or other non-building objects. In this study, the improved shadow regions allow us to develop an efficient morphological processing step to filter out the shadow areas corresponding to relatively short objects even though the shadows of the short objects are combined with the cast shadows of the buildings.

The key contribution of our work, however, is the development of a two-level graph partitioning framework to detect buildings. Our previous approach depends on a single-level graph partitioning performed locally for each shadow region. Although this configuration performs fairly well and was found to be robust for buildings with arbitrary shapes, sizes and colours, the local processing ignores the global evidence because of the ROI-related strategy, and the results of the building detection fully depend on the results achieved within the ROIs selected. Furthermore, the performance of the approach is limited by the performance of the shadow detector because building regions cannot be recovered if their shadows were overlooked. To mitigate those problems, in this study, we propose a graph partitioning framework that evaluates both local and global evidence in two different levels. In the first level, we form the building regions via iterative binary-label graph partitioning. Thereafter, we initiate the second-level processing step over the entire image domain, which is designed on a multi-label graph partitioning process that makes use of the information extracted from the first level. Global processing of

the information not only allows us to alleviate the ROI-related problem stated above but also provides additional opportunities to recover buildings that were completely missed after the local investigation. In fact, our fourth contribution is related to this proposal. After the second-level partitioning process, new shadow and/or building regions may arise over the entire image domain. Thus, we further validate these newly found regions with the known directional information between the shadows and buildings. This final contribution in our approach eliminates the requirement for a complete preliminary shadow mask to detect buildings.

The remainder of this study is organised as follows. Our new building detection approach is introduced in Section 2, and Section 3 presents the experiments and the results of our approach. Lastly, concluding remarks and suggestions for possible future works are provided in Section 4.

2. Methodology

The proposed approach uses single ortho-rectified multi-spectral images (Fig. 2). By default, the sun angles at acquisition time are supplied as meta-data[1]. First, two masks are extracted, a vegetation mask and a shadow mask. Thereafter, the shadow mask is post-processed with the aid of the solar information in the metadata file, and the fuzzy landscapes are generated. The approach involves two levels of graph partitioning to detect building regions. The first level helps to identify the initial building regions using iterative binary-label graph cuts, whereas the second level improves the regions that belong to the building class with multi-label graph partitioning. The last step confirms the building regions detected using the evidence of shadows. The individual stages of our methodology will be described in the subsequent sections. Certain of these stages have been previously described in detail in Ok et al. (2013), and therefore, these stages are only briefly reviewed to provide a complete overview of the methodology.

2.1. Vegetation extraction and shadow detection

To extract the vegetated areas, we use the well-known Normalised Differential Vegetation Index (NDVI), which allows discrimination of healthy vegetation. Larger values indicate the presence of vegetation. We use automatic histogram thresholding (Otsu, 1975) to compute a binary vegetation mask M_V.

In a recent study, a new index for detecting shadow areas was proposed by Teke et al. (2011), and in this work, we use this approach for two important reasons: (i) the approach benefits from the NIR channel present in most VHR satellite images, and (ii) the approach is fully independent of user- and data-dependent thresholds. Shadow regions in the general exhibit lower radiance values over the entire spectrum, and sensor irradiance from shadow regions decreases from short to long wavelengths due to scattering (Adeline et al., 2013). Therefore, it is easier to distinguish shadow regions from non-shadow areas with NIR images. In Teke et al. (2011), this fact is taken into account with an index that computes a ratio using saturation and intensity components of the HSI space. The HSI space is created from a false colour composite image (NIR, R, G) and shows that the saturation component of the HSI space of shadow regions is much higher than that of other regions because of the definition of the saturation in the HSI space, in which very dark or very light colours are considered as fully saturated. Otsu's method is applied to detect the shadow areas (as also used in the case of vegetation extraction). Because the thresholding scheme detects shadow (dark colour) and vegetation (light colour) regions at the same time, the regions that belong to the vegetation cover are subtracted to obtain a binary shadow mask M_S.

2.2. Post-processing of the shadow mask

In this study, we propose two methods for estimating the shadow mask M_S in the detection of buildings. The first method aims to improve the shape and boundaries of the detected shadow regions. Fig. 3 illustrates two GeoEye-1 images and their corresponding detected shadow masks. As shown in Fig. 3b and e, although the shadow areas are detected with good precision, their shapes and boundaries are imperfect. To avoid this problem, we propose a constrained region-growth process. As a well-known approach, region growth is an iterative region-based segmentation that examines the neighbouring pixels of initial seed points to determine whether the neighbours of the pixels should be included in a region. In this study, each pixel labelled as shadow in M_S is considered as an initial seed point. Thereafter, the regions are grown from the seed points to the adjacent pixels depending on an 8-connected neighbourhood pixel similarity criterion. The similarity is assessed on the normalised intensity band of the Hue-Saturation-Intensity (HSI) space generated from the false colour composite image (NIR, R, G), which is the same band combination used to detect shadow areas. For the criterion, we assessed the difference between the intensity of a pixel and the mean intensity of the region. During the growth process, the pixel with the smallest difference in the neighbourhood is allocated to the region, and the process continues iteratively until the difference between the mean intensity of the region and the intensities of all of the neighbourhood pixels is larger than a threshold (T_I). However, it should be noted that a number of regions erroneously detected by the shadow detector might also grow after this process. Based on our experiments, we realised that comparing the number of pixels involved in the shadow region before and after the growth process with a ratio (T_R) is useful for identifying those cases because almost all of the wrongly labelled shadow regions have a tendency to cover sizeable regions after growth. Thus, the ratio T_R controls the growth process and allows us to identify and eliminate the regions that are exceedingly large after the growth process. Furthermore, because we address relatively dark areas, the pixels corresponding to shadow areas should have stronger saturation levels. Therefore, we constrain the growth process by masking out the pixels with lower saturation values than the normalised intensity component.

The goal of the second post-processing step applied to the shadow mask is to estimate the height difference of the objects compared with the terrain height. In our previous work and for the same purpose, we investigated the length of each shadow component in the direction of illumination to enforce a pre-defined height threshold for buildings. Although that approach was found to be effective and performs well for most cases, the success of the approach may be reduced dramatically for certain illumination conditions (e.g., based on a flat terrain assumption, acute solar elevation angles ($\phi < 30°$) always lead to large cast shadow regions on the Earth's surface). Thus, for these cases, it is more likely to observe merged shadow components cast by multiple separate objects (Fig. 3d). This fact and the fact that a cast shadow of a non-building man-made object merged with the cast shadow of a building might result in a joint deterioration of the pruning process. Fig. 4b shows the shadow mask after applying the pruning process (Ok et al., 2013) to the shadow mask in Fig. 3b. As shown in Fig. 4b, for a given height threshold (T_{height} = 3 m), all of the shadow pixels of non-building objects (e.g., walls, cars) that are independent from the cast shadow of the building are successfully removed.

[1] Solar angles are assumed to be fixed during the period of image acquisition. The azimuth angle (A) in an orthorectified image space is the angle computed from north in a clockwise direction, which also corresponds to the opposite direction of solar illumination in image space. The elevation angle (ϕ) is the angle between the direction of the geometric center of the sun and the horizon.

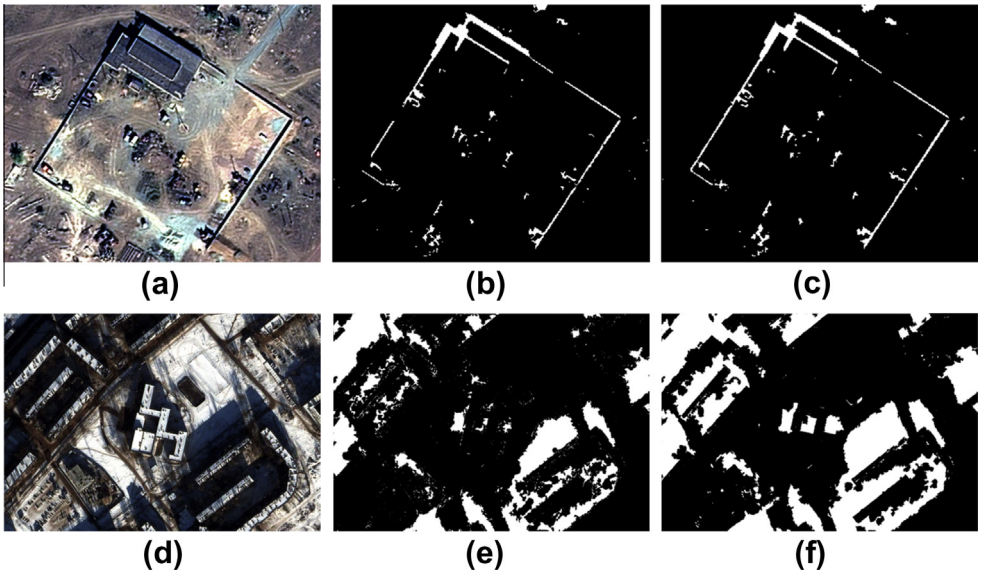

Fig. 3. (a and d) GeoEye-1 pan-sharped images (RGB), (b and e) the shadow masks detected (Teke et al., 2011), and (c and f) the shadow masks after applying the constrained region growing.

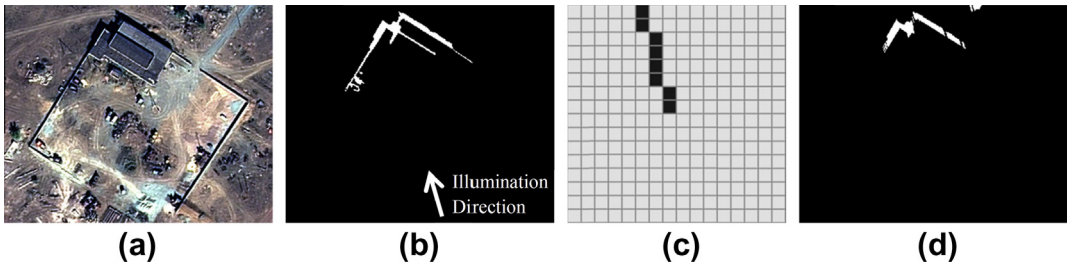

Fig. 4. (a) The GeoEye-1 pan-sharped image (RGB) provided in Fig. 3a, with solar azimuth (A) and elevation (ϕ) angles of 164.87° and 39.01°, respectively. (b) The shadow mask after applying the pruning process (Ok et al., 2013) to the shadow mask in Fig. 3b (with T_{height} = 3 m). (c) The directional flat structuring element generated for the given solar angles, and (d) the shadow mask after carrying out the proposed morphological opening with the structuring element to the shadow mask in Fig. 3c.

However, the cast shadows of the wall structures in the vicinity of the building object have survived because those shadow regions are merged with the cast shadow of the building whose elevation over the terrain surface exceeds the given height threshold. Therefore, it is not possible to eliminate the cast shadows of those objects only by investigating the length of the shadow components in the direction of illumination. In this study, we propose a morphological process to solve this problem. The idea is to apply a morphological opening to each shadow component using a specific directional flat structuring element generated from the known solar information. To do this, with the aid of the Bresenham line discretisation algorithm (Bresenham, 1965), we generate a flat structuring element ($v_{L,\lambda+\pi}$) that maintains the directional information λ ($\lambda = A - \pi/2$) with a minimally connected single edge segment L (Fig. 4c). Based on the assumption that the surface on which shadows fall is flat, the length of the edge segment (l) in the image space for a given illumination direction depends on three parameters: the solar elevation angle (ϕ), the height threshold applied (T_{height}), and the resolution of the image used (c):

$$l = \left\lceil \frac{T_{height}}{\tan\phi . c} \right\rceil \tag{1}$$

where the operator $\lceil . \rceil$ rounds to the next-larger integer. Once the structuring element is generated, we label each shadow component with an 8-neighbourhood connectivity analysis and apply the morphological opening to each shadow component independently. We lastly combine all output components into a single image to achieve the final post-processed shadow mask M_{PS}. Fig. 4d shows the result of the morphological post-processing applied to the mask provided in Fig. 3c. The result achieved indicates that after the post-processing, all cast shadows except those that correspond to the main body of the building are successfully eliminated. However, as visible in Fig. 4d, the proposed post-processing may also negatively affect certain parts of the shadow areas, primarily due to other nearby objects (e.g., trees, vehicles) that occlude the shadow region of the buildings, and as a result, the shadow areas may split into multiple pieces due to occlusion. Because our building detection framework relies on shadow evidence, one might think that this would reduce the quality of the final detection. However, as a result of the proposed second-level partitioning step in Section 2.5, we have the ability to recover the buildings and their components even though their cast shadows are partially or fully missing.

2.3. Generation and pruning of fuzzy landscapes

We model the spatial arrangement between buildings and their shadows with a morphological fuzzy relation approach. Given a shadow object B and a non-flat line-based structuring element $v_{L,\lambda,\sigma,\kappa}$, the landscape $\beta_\lambda(B)$ around the shadow object along the given direction λ can be defined as a fuzzy set of membership values in image space:

$$\beta_\lambda(B) = (B^{per} \oplus v_{L,\lambda,\sigma,\kappa}) \cap B^C. \tag{2}$$

In Eq. (2), B^{per} represents the perimeter pixels of the shadow object B computed in 8-neighbourhood connectivity, B^C is the complement of the shadow object B, and the operators $\oplus$ and $\cap$ denote the morphological dilation and a fuzzy intersection, respectively. The landscape membership values are defined in the range of 0–1, and it is shown that the membership values of the landscapes generated using Eq. (2) decrease when moving away from the shadow object and are bounded in a region defined by the object's extents and the direction defined by angle λ. In Eq. (2), we use a line-based non-flat structuring element $v_{L,\lambda,\sigma,\kappa}$ generated by combining two different structuring elements with a pixel-wise multiplication operator ($*$):

$$v_{L,\lambda,\sigma,\kappa} = v_{L,\kappa,\lambda} {}^{*} v_{\sigma,\kappa}. \tag{3}$$

In Eq. (3), $v_{\sigma,\kappa}$ is a Gaussian non-flat structuring element with kernel size κ, and the rate of decrease of the membership values within the element is controlled by a single parameter σ. However, the flat structuring element $v_{L,\kappa,\lambda}$ is responsible for providing directional information, where L denotes the line segment and λ is the angle where the line is directed.

During the pruning step, we investigate the vegetation evidence within the directional neighbourhood of the shadow regions. At the end of this step, we remove the landscapes that might be generated by the cast shadows of vegetation canopies (Fig. 5). To do this, we define a search region (Fig. 5c) in the immediate vicinity of each shadow object by applying two thresholds (T_{low}, T_{high}) to the membership values of the fuzzy landscapes generated (Fig. 5b). Once the region is defined, we search for vegetation evidence within the defined region using the vegetation mask M_{V} and reject a fuzzy landscape region generated from a cast shadow if there is substantial evidence of vegetation ($\geqslant T_{\text{veg}}$) within the search region. This search is performed with a computed ratio in which the denominator is the total number of pixels defined in the region and the numerator is the total number of pixels labelled as vegetation in the vegetation mask.

2.4. First-level partitioning

For the first-level processing, we considered the building detection task as a two-class partitioning problem in which a given building region must be accurately separated from its background. To solve the partitioning, we used the GrabCut approach (Rother et al., 2004) in which an iterative binary-label graph-cut optimisation is performed.

GrabCut was originally developed as a semi-automated foreground/background partitioning algorithm. Given a group of pixels interactively labelled by the user, GrabCut partitions the pixels in an image using a graph theory approach. Given a set of image pixels $\mathbf{z} = (z_1, z_2, \ldots, z_N)$ in an image space, each pixel is assigned an initial labelling from a tri-map $T = \{T_{\text{B}}, T_{\text{F}}, T_{\text{U}}\}$, where T_{B} and T_{F} represent the background and foreground label information provided by the user, respectively, and T_{U} denotes the unlabelled pixels. In addition, each pixel has an initially assigned value $\underline{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_N)$ corresponding to background or foreground, where $\alpha_n \in \{0, 1\}$ and the underline operator indicates the parameters to be estimated/solved.

At the first stage of the algorithm, two Gaussian Mixture Models (GMMs) with K components for the foreground (K_{F}) and the background classes (K_{B}) are constructed from the pixels manually labelled by the user. We define $\mathbf{k} = \{k_1, k_2, \ldots, k_N\}$ with $k_n \in \{1, \ldots, K\}$ as the vector representing the mixture components for each pixel. Next, the Gibbs energy function for the partitioning can be written as

$$E(\underline{\alpha}, \mathbf{k}, \underline{\theta}, \mathbf{z}) = U(\underline{\alpha}, \mathbf{k}, \underline{\theta}, \mathbf{z}) + V(\underline{\alpha}, \mathbf{z}) \tag{4}$$

where $\underline{\theta}$ denotes the probability density function to be obtained by mixture modelling for each pixel. In Eq. (4), $U(\underline{\alpha}, \mathbf{k}, \underline{\theta}, \mathbf{z})$ denotes the fit of the background/foreground mixture models to the data $\mathbf{z}$ by considering α values and is defined as

$$U(\underline{\alpha}, \mathbf{k}, \underline{\theta}, \mathbf{z}) = -\sum_{n} D(\alpha_n, k_n, \underline{\theta}, z_n) \tag{5}$$

where $D(\alpha_n, k_n, \underline{\theta}, z_n)$ favour the label preferences for each pixel z_n based on the observed pixel values. Additionally, $V(\underline{\alpha}, \mathbf{z})$ is the boundary smoothness and is written as

$$V(\underline{\alpha}, \mathbf{z}) = \gamma_1 \sum_{(m,n)\in C} [\alpha_n \neq \alpha_m] e^{-\beta \|z_m - z_n\|^2} \tag{6}$$

where the term $[\alpha_n \neq \alpha_m]$ is a binary indicator function that takes on value 1 if $\alpha_n \neq \alpha_m$, C is the set of neighbouring pixel pairs computed in the 8-neighbourhood, and β and γ_1 are the constants that determine the degree of smoothness. The smoothness term β is computed automatically after evaluating all of the pixels in an image, and the other smoothness term γ_1 is fixed to a constant value after investigating a set of images. To complete the partitioning and to estimate the final labels of all pixels in the image, a minimum-cut/max-flow algorithm is applied (Rother et al., 2004).

In our previous work, we integrated the GrabCut approach in an automated building detection framework. In that approach, the pixels corresponding to the foreground/building (T_F) and background/non-building (T_B) classes are labelled automatically using the shadow regions and the generated fuzzy landscapes. To do this, we define the T_F region in the vicinity of each shadow object whose extents are outlined after applying a double thresholding (η_1, η_2) to the membership values of the fuzzy landscape generated. To

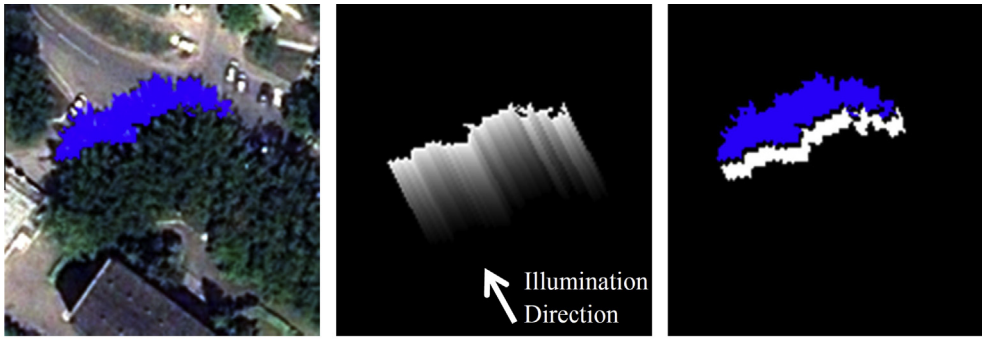

Fig. 5. (a) A detected shadow region (blue) of a vegetation canopy, and (b) generated fuzzy landscape with parameters $\lambda = 62.8°$ and $\kappa = 80$ pixels. (c) Search region (white) generated in the immediate vicinity of shadow region after thresholding the landscape in (b) with values 0.7 and 0.9 for T_{low} and T_{high}, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

acquire a fully reliable T_F region, a refinement procedure that involves a single parameter, i.e., shrinking distance (d), is also developed. For each shadow component, a bounding box whose extent is automatically determined after dilating the shadow region is generated to select the T_B and to define the ROI region in which the GrabCut partitioning is performed. The dilation is performed with a flat line structuring element defined in the opposite direction of illumination, and the extent of the ROI and the covering bounding box is controlled by a single distance parameter, the ROI size (Fig. 6b). Once the bounding box is selected, the pixels corresponding to background information (T_B) within the selected bounding box are automatically determined. To do this, the shadow and vegetation regions as well as the regions outside the ROI region within the bounding box are labelled as T_B.

In this study, we modify the configuration in which the pixels initially labelled as T_U are also involved in the calculation of the foreground GMMs. Fig. 6a clarifies the modification, which provides an automatically selected bounding box covering a building. In our previous work, because each building region is detected locally with the shadow information, the foreground GMMs were estimated from the ROI region (T_F and T_U) (Fig. 6b). This configuration allows selected of complete boundaries for most of the buildings. However, this process eventually causes the area of certain buildings to be overestimated. Therefore, in this study, we modify the configuration used at the initial step, and the foreground GMMs are only estimated using the pixels labelled as T_F (Fig. 6c). As a result, the building regions are detected with relatively reduced completeness but with almost no over-detection. In this way, after performing the local processing for each shadow region, we achieve a set of regions that indicate the most probable locations for buildings in the entire image domain.

2.5. Second-level partitioning

The entire image can be divided into four distinct classes with the aid of the pre-computed vegetation and shadow masks after detection of the building regions from the first-level partitioning (Fig. 7c). To do this, we assign unique labels for the regions that belong to each class: *building*, *vegetation* and *shadow*. Thereafter, the remaining regions that do not correspond to any of these three classes are assigned to a fourth class, *others* (Fig. 7d). However, after the first-level partitioning, miss-classifications of buildings vegetation and shadow pixels might have occurred. Therefore, the goal of the second-level partitioning developed in this study is to identify and revise the incorrectly assigned labels among the four classes by investigating the global evidence over the entire image space.

For the second-level partitioning, we propose a multi-label graph-cut optimisation. Given a set of pixels $\mathbf{z} = (z_1, z_2, \ldots, z_N)$, and a set of class labels $L \in \{1, \ldots l\}$ where $l = 4$, our aim is to find the optimal mapping from the data $\mathbf{z}$ to the class labels L. Each pixel has an initially assigned value $\underline{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_N)$ corresponding to each class label L, where $\alpha_n \in \{1, 2, 3, 4\}$. We use the Gibbs energy function provided in Eq. (4) and initialise a GMM with K components for each of the four classes. Similar to the case in Eq. (5), the terms $D(\alpha_n, k_n, \underline{\theta}, z_n)$ favour the label preferences for each pixel z_n based on the observed pixel values, and the smaller the value of $D(\alpha_n, k_n, \underline{\theta}, z_n)$, the more likely the label l for pixel z_n. We also follow the same expression for the spatial smoothness term $V(\underline{\alpha}, \mathbf{z})$ provided in Eq. (6), which states the smoothness priors in relation to the optimal mapping from data $\mathbf{z}$ to class labels L.

To minimise the energy $E(\underline{\alpha}, \mathbf{k}, \underline{\theta}, \mathbf{z})$ provided in Eq. (4) for multi-label optimisation using graph cuts, a special graph is constructed (Boykov et al., 2001) that depends on the smoothness term $V(\underline{\alpha}, \mathbf{k}, \underline{\theta}, \mathbf{z})$ and the number of labels L. For the optimisation, we used α-expansion move algorithm (Boykov et al., 2001). During the second-level partitioning, we prefer a non-iterative process to limit the computation time because realistic images contain $> 10^6$ pixels. For the parameters, we performed a large number of tests that consider the parameter combinations, and the detailed experiments are described in Section 3.2. Fig. 7e illustrates the results of the second-level partitioning applied to the image shown in Fig. 7d, with parameters $K_1 = K_4 = 8$, and $K_2 = K_3 = 2$ for the number of components and $\gamma_2 = 5$ for the smoothing constant.

2.6. Determination of final building regions

After the second-level partitioning, we extract the regions labelled as buildings from the result of the optimisation. Despite the fact that the second-level partitioning clearly identifies and correctly labels most of the building regions, several false positives occur in the final result due to spectral similarities between building and non-building areas. These false positives must still be rejected.

For the verification, first, we reconsider the shadow information. To do this, we benefit from the landscapes generated (cf. Section 2.3) using the foreground mask T_F constructed prior to the first-level partitioning. Therefore, if a region labelled as a building after the second-level partitioning has an overlap with the building evidence in T_F, the region is verified (Fig. 7f). However, such a verification process uses the post-processed shadow mask M_{PS}

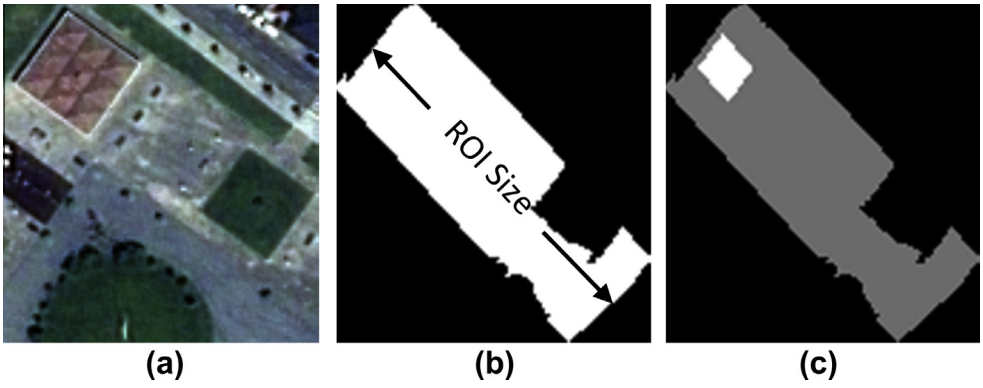

Fig. 6. (a) A bounding box in which first level partitioning is performed. (b) The configuration used for the estimation of the foreground/background GMMs in Ok et al. (2013), colours white and black denote the pixels utilised to estimate foreground and background GMMs, respectively. (c) The configuration proposed in this study. Colours white and black correspond to the pixels utilised to estimate foreground and background GMMs, respectively, and grey colour indicates the pixels that are not involved in the estimation process.

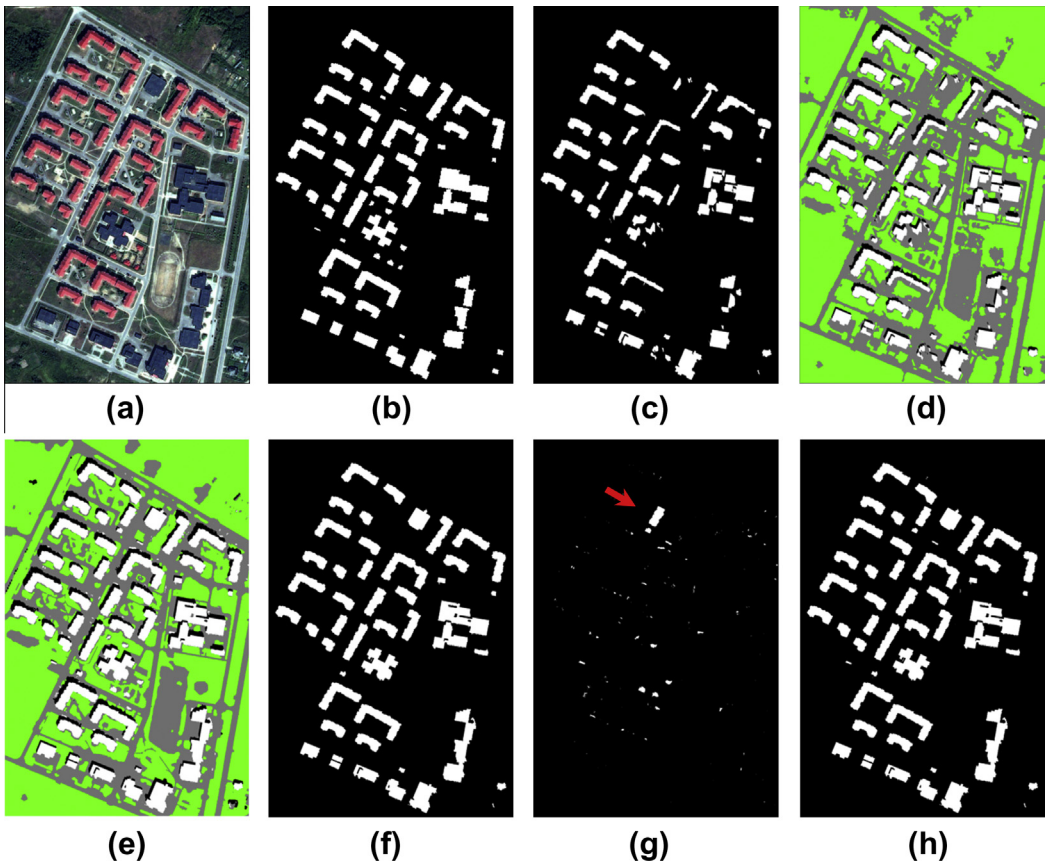

Fig. 7. (a) A GeoEye-1 pan-sharped image (Test patch #12 – RGB) and (b) the corresponding ground truth, (c) building regions detected after first level partitioning (for parameters, see Table 1), and (d) the image with four different classes prepared for the second level partitioning, (e) final result of the second level partitioning (for parameters, see Table 1). In (d) and (e), colours white, green, black, and grey indicate the regions for classes building, vegetation, shadow and other, respectively. Confirmed and rejected regions after the verification with T_F are illustrated in (f) and (g), respectively. (h) building regions after the final shadow verification and post-processing. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

generated via an index (cf. Sections 2.1 and 2.2). Thus, although most of the regions that belong to buildings can be easily confirmed with the shadow evidence in M_{PS}, for the cases in which the index fails to detect shadow areas, the information in M_{PS} is insufficient for the verification. Because of this, certain correctly recovered regions that correspond to buildings after the second-level partitioning might be rejected as well. To mitigate this problem, we propose a novel verification approach. After the second-level partitioning, we apply a mapping from data $\mathbf{z}$ to all class labels $L \in \{1, 2, 3, 4\}$. Thus, because our primary task is to find the correct mapping from data $\mathbf{z}$ to class buildings in the entire image space, the optimisation eventually has an effect on the other three classes, including the class of shadow. Therefore, the second-level partitioning may reveal new shadow evidence in the entire image space that was missing in the shadow mask M_{PS}. Thus, we take advantage of the new shadow evidence to further verify the regions rejected after investigating the foreground mask T_F. To do this, first, we extract the shadow information from the second-level partitioning result and generate a new shadow mask M_S[2]. Next, we follow the same procedure developed to generate fuzzy landscapes from M_S except for the region-growth post-processing because the shadow boundaries from the second level are already correct. We lastly generate a new foreground mask T_F[2] from the shadow mask M_{PS}[2] using a membership threshold (η_F), and the rejected regions in the previous verification are further validated with the new foreground evidence. As a result, the rejected regions that have an overlap with the evidence in T_F[2] are also accepted. The two buildings shown with a red arrow in Fig. 7g are good examples of building regions recovered after such a validation. Note that the cast shadows of those two buildings are not detected in the M_S. Therefore, neither the buildings nor their shadows exist in Fig. 7d. However, the building regions as well as their shadow areas are recovered after the second-level partitioning. Therefore, in the final result, these buildings are accepted as building regions (Fig. 7h). As a final point, after the verification, small-sized artefacts may still exist amongst the accepted building regions. Therefore, to remove artefacts, we employ a simple post-processing step that involves a threshold (T_{area}) to define the minimum area enclosed by a single building region.

3. Experiments

3.1. Input dataset and strategy for accuracy assessment

The experiments are performed on a set of ortho-rectified[2] GeoEye-1 pan-sharpened images with a spatial resolution of 50 cm. All images contained four bands (R, G, B and NIR) with a radiometric resolution of 11 bits per band. The assessments of the proposed approach are performed over 16 test images, which include different types of urban buildings as well as different lighting conditions (Figs. 8 and 9). The solar elevation angles in the tested range are between 15.80° and 66.79°, which shows that a wide range

[2] The images are not true-orthoimages because relief distortions occur as off-terrain objects were not taken into account during ortho-rectification.

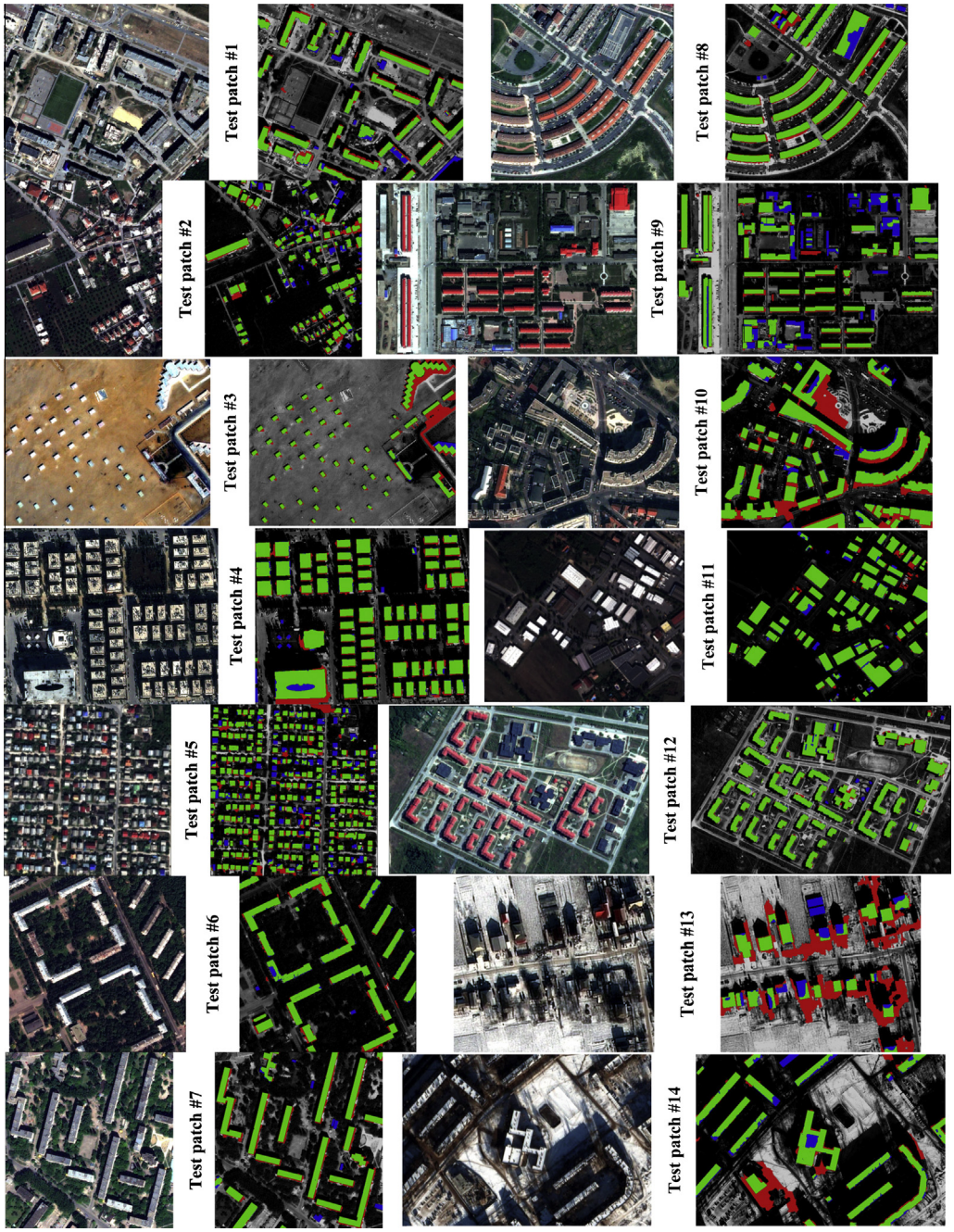

Fig. 8. (first column) Test patches #1–7. (second column) Detected building regions for test patches #1–7. (third column) Test patches #8–14. (fourth column) Detected building regions for test patches #8–14. Green, red and blue colours represent *TP*, *FP* and *FN* pixels, respectively. Note that the test patch #12 shown in Fig. 7a is rotated 90° CCW for a better visualization.

of cast shadow lengths is visible and is considered within the selected image dataset. The mean values of the satellite azimuth and elevation indicate that the images are acquired with unrestricted azimuth angles and with a maximum off-nadir angle of 18 degrees. Among the test images selected, 14 were previously used in the evaluation of our previous approach (Fig. 8). However, the other two test sites (Fig. 9) were intentionally selected to demonstrate the full potential of the proposed approach.

The reference data consisting of building regions were produced manually by a qualified human operator. Two interesting aspects should be pointed out with respect to the reference data. First, buildings that are partially visible at the image boundaries were included in the reference data. Second, three images are used for testing the exhibited difficult environmental conditions: desert (#3) and snow cover (#13 and #14).

The final performance of the proposed approach is assessed by comparing the results of the proposed approach with the reference data. In this study, we use three well-known quality measures (Precision, Recall, and F_1-score) to evaluate the pixel-based performance of the proposed approach as applied in Aksoy et al. (2012) and Ok et al. (2013):

$$Precision = \frac{|TP|}{|TP| + |FP|} \tag{7}$$

$$Recall = \frac{|TP|}{|TP| + |FN|} \tag{8}$$

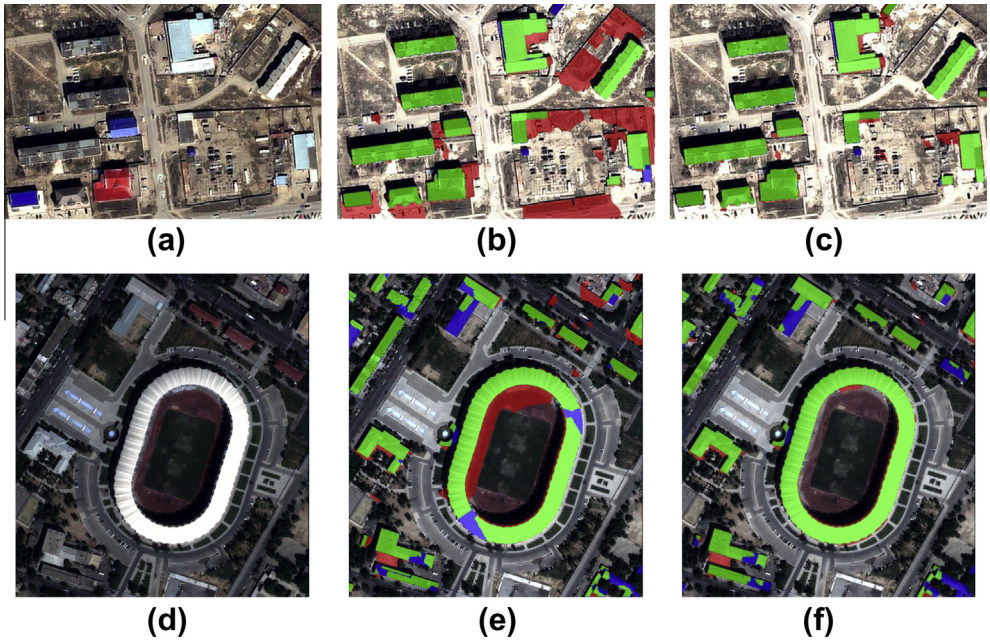

Fig. 9. Two test patches and the results of the building detection. (a and d) The test patches #15 and #16, respectively. (b and e) The results achieved by the approach in Ok et al. (2013). (c and f) The results achieved by the proposed approach. Green, red and blue colours represent *TP*, *FP* and *FN* pixels, respectively.

$$F_1 = \frac{2 * precision * recall}{precision + recall} \tag{9}$$

where *TP* are true positives, *FP* are false positives, and *FN* are false negatives. The operator $|\cdot|$ denotes the number of pixels assigned to each distinct category, and the F_1-score combines Precision and Recall into a single number.

The object-based performance of the proposed approach can also be evaluated with the measures given in Eqs. (7)–(9). For all numerical results provided, we label an output building object as *TP* if it has at least a 60% pixel overlap ratio with a building object in the reference data. However, we label an output object as *FP* if the output object of the proposed approach does not coincide with any of the building objects in the reference data, and we label an output object as *FN* if the output object corresponds to a reference object with a limited amount of overlap (<60%). Thus, it is possible to compute object-based Precision, Recall and F_1-score results for each test image. We also carried out a detailed investigation in which the results of the object-based Precision, Recall and F_1-score for specific overlap ratio intervals are visualised explicitly in a 2D graph. In the graph, the x-axis represents the applied overlap ratio, and the y-axis represents the performance of the proposed approach.

3.2. Selection of parameters

All parameters required to initialise the proposed approach are listed in Table 1. To select the best parameter configuration, we performed a large number of tests on different parameters and investigated the effects of each parameter on the detection performances using the quality measures given in Eqs. (7)–(9). In Figs. 10 and 11, we omit the computed F_1-scores and provide the results of the precision and recall ratios to provide a clear visualisation.

Three parameters (T_I, T_R, and T_{height}) are required for post-processing of the shadow mask M_S (cf. Section 2.2). The effects of different intensity thresholds (T_I) on the computed performances are shown in Fig. 10a. The results indicate improvements as the threshold values increase for both the object- and pixel-based recall ratios, whereas at the same time, performance decreases are observed for the precision ratios. In view of the computed performance measures, we select the intensity threshold as 0.05, which also maximises the F_1-scores computed for both the object- and pixel-based cases. The effects of different ratio thresholds (T_R) on the computed performances are illustrated in Fig. 10b. The ratio threshold also controls the growth process and compares the size of the shadow regions before and after growth. As shown in Fig. 10b, the ratio thresholds between 0.1 and 0.3 provide the best performances. However, larger thresholds significantly reduce the computed recall ratios. Thus, we conclude that ratio thresholds above the value of 0.3 are not suitable for processing of the proposed approach. For the threshold T_R, we select the ratio value of 0.2, which provides the best performance among the thresholds tested. As a final parameter of the post-processing stage, six different values (0–5 m) are used as height threshold T_{height} (Fig. 10c). Although a minimum height of 4 m works best at the pixel level, 2 m is a more suitable threshold at the object level because many small buildings are less than 4 m high. As a compromise, we set T_{height} = 3 m.

Considering the foreground information collected for the first level partitioning (cf. Sections 2.3 and 2.4), five parameters (κ, σ, η_1, η_2, and d) jointly control the number of pixels labelled as building regions. In this study, we fixed the fuzzy landscape parameters κ and σ to 40 m and 100, respectively, which characterise the neighbourhood of a shadow region. The first foreground membership threshold η_1 and the shrinking distance d are also fixed to 0.9 and 2 m, respectively, which ensure that incorrect pixels around the border of the shadow regions are rejected. Thereafter, we investigated different parameter configurations for the second foreground membership threshold (η_2) to assess the impact of the number of pixels labelled as a building region on the performances computed (Fig. 10d). As illustrated in Fig. 10d, if large numbers of pixels are collected as a building region from the directional neighbourhood of the shadow regions ($\eta_2 < 0.4$), performance improvements are observed for the pixel-based recall ratios, whereas performance decreases are apparent for the pixel-based precision ratios. However, if fewer pixels are collected

Table 1
Parameter settings of the proposed approach.

Section	Parameter		Value
Post-processing of the shadow mask (Section 2.2)	Intensity threshold (T_I)		0.05
	Ratio threshold (T_R)		0.2
	Height threshold (T_{height})		3 m
The generation and pruning of fuzzy landscapes (Section 2.3)	Kernel size (κ)		40 m
	Sigma (σ)		100
	Search region		0.7 – 0.9
	Thresholds (T_{low}–T_{high})		
	Vegetation ratio threshold (T_{veg})		0.7
First-level partitioning (Section 2.4)	GMM components for each class	Foreground (K_F)	5
		Background (K_B)	5
	Smoothing constant (γ_1)		50
	Foreground thresholds (η_1–η_2)		0.9–0.4
	Shrinking distance (d)		2 m
	ROI size		50 m
Second-level partitioning (Section 2.5)	GMM components for each class	Building (K_1)	8
		Vegetation (K_2)	2
		Shadow (K_3)	2
		Others (K_4)	8
	Smoothing constant (γ_2)		5
Determination of final building regions (Section 2.6)	Foreground threshold (η_F)		0.8
	Minimum area (T_{area})		30 m^2

($\eta_2 > 0.4$), the results are reversed. Similar trends are also visible and valid for the object-based recall and precision ratios. Thus, we can conclude that the number of pixels assigned as foreground information affects the performance of the proposed approach, and the optimal threshold for η_2 is 0.4 for our dataset. Vegetation detection in the directional neighbourhood requires three parameters (T_{low}, T_{high}, and T_{veg}). The first two parameters T_{low} and T_{high} are fixed to 0.7 and 0.9, respectively, to ensure that the immediate vicinity of the shadow regions is searched for vegetation evidence. To assess the influence of the vegetation ratio threshold (T_{veg}) on the performance of the proposed approach, we investigate the performance for different thresholds (Fig. 10e). We found that except for low thresholds (<0.4), the results computed are rather stable and consistent. Thus, we select the threshold T_{veg} as 0.7, which provides the best performances for both the object and pixel-based measures.

The iterative binary-label graph cut optimisation performed in the first-level partitioning (cf. Section 2.4) is controlled by three parameters (K_F, K_B, and γ_1). For the number of GMM components, we investigate different K values with $K \in \{2,3, \ldots, 10\}$ for each of the foreground (K_F) and background (K_B) classes, and the results of the evaluations are shown in Fig. 10f and g. According to the results achieved, we conclude that different numbers of GMM components affect the performance of both the pixel-based and object-based ratios by less than 1%. For the smoothing constant γ_1, the object- and pixel-based ratios become stable and consistent for smoothing constants larger than 20, for which the best overall performances are observed (Fig. 10h). Thus, in the study, we preferred to fix the threshold values as provided by Rother et al. (2004), which are 5, 5, and 50 for the parameters K_F, K_B, and γ_1, respectively. For the parameter ROI size, we also found that the performances computed are not significantly affected even for small ROI sizes such as 10 m (Fig. 10i). Therefore, we fix the ROI size as 50 m for which marginally better pixel- and object-based performances are observed.

In this study, one major task is to consider the parameters used for the second-level partitioning (cf. Section 2.5), in which the regions of buildings are eventually characterised. As shown in Table 1, the performance of the second-level partitioning is controlled by the number of GMM components used for each class (K_1, K_2, K_3, and K_4) and the smoothing constant γ_2. We ran a large number of tests with different numbers of GMM components, up to a maximum of 20. Furthermore, we repeated the tests for different values of the smoothing constant, $\gamma_2 \in \{1,2, \ldots, 50\}$. In the following, we summarise the outcomes achieved after such detailed investigation. First, we found that adjusting the number of GMM components for the classes of vegetation (K_2) and shadow (K_3) influence the overall performance only slightly (<1%) because these two classes have significantly different spectral characteristics compared with the classes observed in any portion of the test images. However, note that intra-class differences may exists between each of the two classes (e.g., the spectral reflectance of a shadow falling on a dark or bright surface) over the entire image domain. Therefore, in this study, we preferred a mixture model with two components used to characterise each of the vegetation and shadow classes (Table 1). Nevertheless, Fig. 11 proves that the number of GMM components assigned to the classes building (K_1) and others (K_4) have a significant impact on the overall performance. In Fig. 11, the results of the proposed approach are illustrated for mixture models with different numbers of components assigned to the classes building and others in combination with five different values for the smoothing constant γ_2. As clearly shown in the figures, using a limited number of mixture model components ($\leqslant 4$) for each class leads to noticeable performance decreases, especially for the recall values computed. Actually, this result is expected mostly for the classes of building and others because such a limited numbers of mixture model components are not sufficient to properly handle large intra-class variation. However, eight mixture model components for both the classes building and others seem to represent a value for which the performance of the proposed approach becomes stable and well-balanced for the performance measures. For large GMM components (>8), we do not observe any significant differences in the computed pixel- and object-based performances. Therefore, in this study, we describe each of the classes of building and others with an 8-component GMM (Table 1). For the smoothing constant γ_2, first, we found that large smoothing constants ($\gamma_2 > 10$) negatively affect the computed pixel-based precision ratios (Fig. 11d and e). Second, smoothing constants lower than 5 deteriorate the recall ratios computed in an object-based manner (Fig. 11a). For smoothing values $5 \leqslant \gamma_2 \leqslant 10$ the results are rather similar (Fig. 11b and c). To select the appropriate smoothing constant value, we also investigated the results visually. The example given in Fig. 12 shows the effect of the smoothing constant γ_2 on the detection results of the test image #10. As shown in Fig. 12c, when $\gamma_2 = 10$, the neighbouring regions tend to merge into one object (e.g., the buildings with white and red roof colours in the lower left corner of Fig. 12a). However, this result is not commonly observed for the results in which $\gamma_2 = 5$ (Fig. 12b). Therefore, in all of our experiments, we fixed the smoothing constant γ_2 to 5 (Table 1).

During the selection of the final building regions, we employed two thresholds (η_F and T_{area}). For the foreground threshold (η_F), we empirically found that a membership threshold of 0.8 successfully generates the foreground evidence required for the final validation.

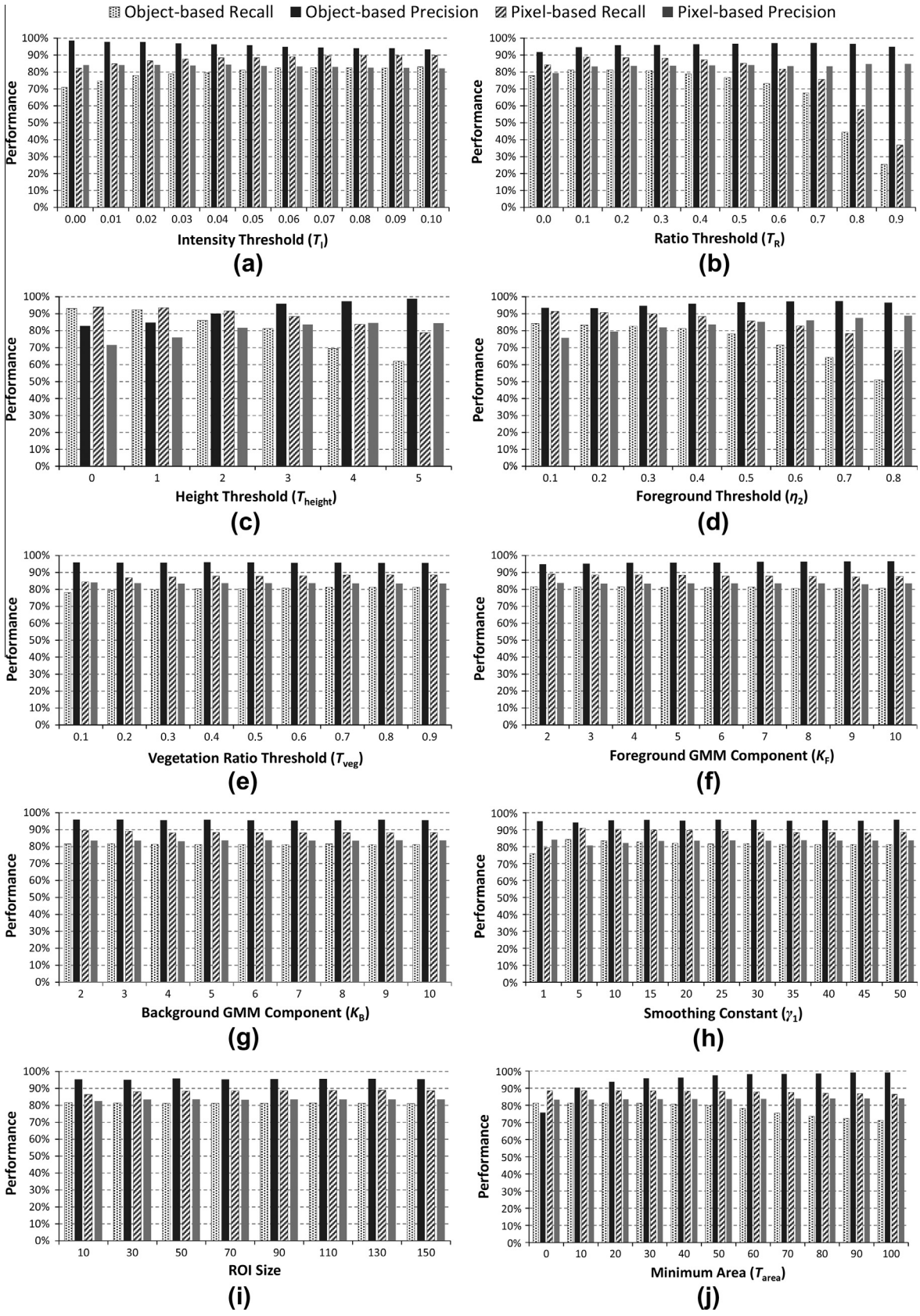

Fig. 10. Effects of different threshold values on the performance results of the proposed approach.

According to Fig. 10j, the minimum area threshold (T_{area}) affects the pixel-based performances by at most 2%. Because the eliminated regions with the threshold T_{area} are relatively small, such stable pixel-based performance results are not surprising.

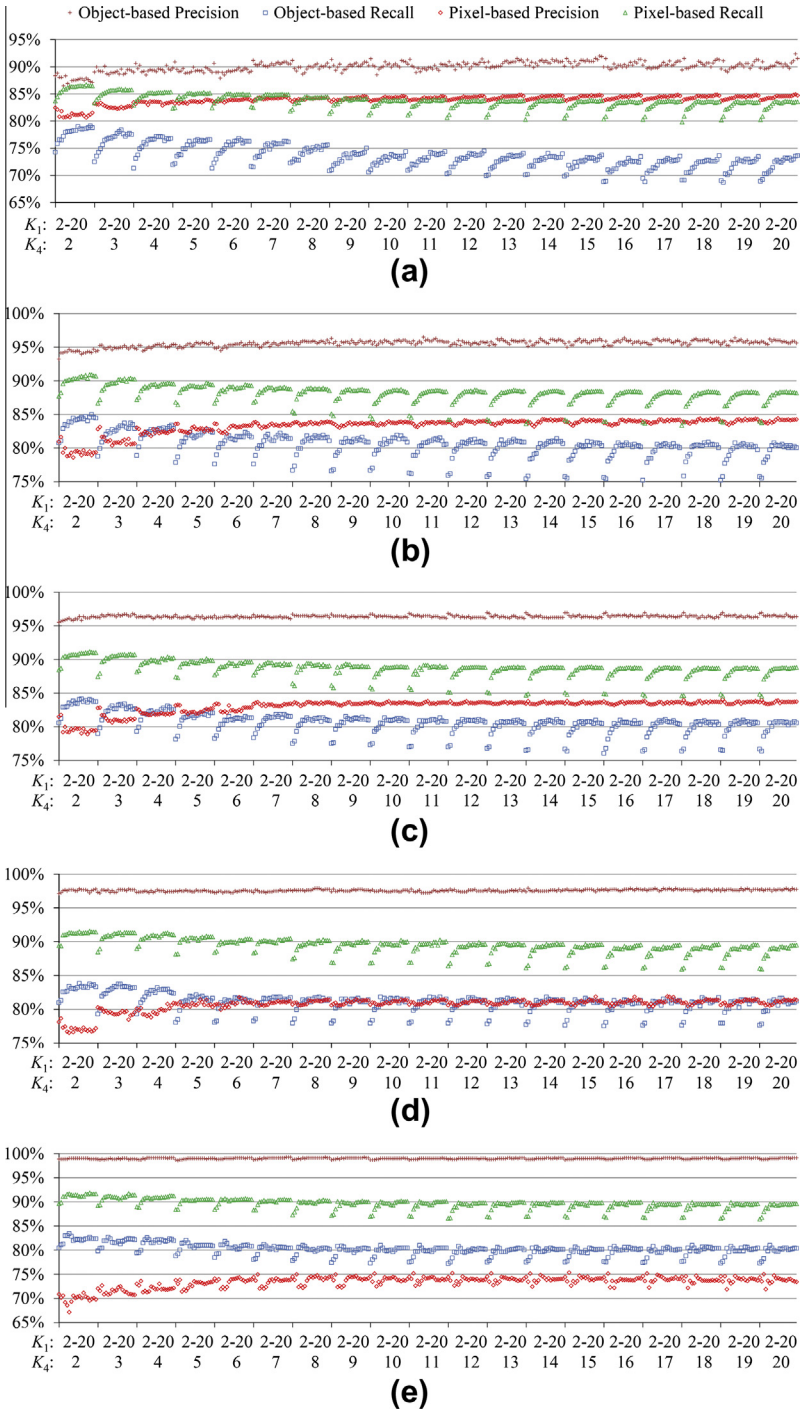

Fig. 11. Overall results for 16 test images utilized with (a) $\gamma_2 = 1$, (b) $\gamma_2 = 5$, (c) $\gamma_2 = 10$, (d) $\gamma_2 = 20$, and (e) $\gamma_2 = 50$. The values in x-axis present the number of GMM components used for the classes Building (K_1) and Others (K_4), respectively. For the experiments, mixture models with two components were used for each of the vegetation and shadow classes.

However, in an object-based perspective, small regions may have strong influence on the computed results. Thus, at the object level, recall decreases and precision increases. We found that the optimum value for our dataset (in terms of F_1-score) was 30 m^2.

In the next section, we present and discuss the results of the proposed approach with the parameter settings given in Table 1. Note that the distance- and area-related parameters given in Table 1 are defined in object space. Therefore, in our approach, these values are converted to pixels according to the resolution of the GeoEye-1 image.

3.3. Results and discussion

We illustrate the detection results of the proposed approach in Figs. 8 and 9. As stated in Section 3.1, the 14 GeoEye-1 test images presented in Fig. 8 are also used during the evaluation of the approach presented in Ok et al. (2013), and the pixel-level results are presented in this study. However, the last two test images (#15 and #16) are original to this study, and the results of the proposed approach as well as the results of our previous approach are presented in Fig. 9. In addition to visual illustration, the numerical

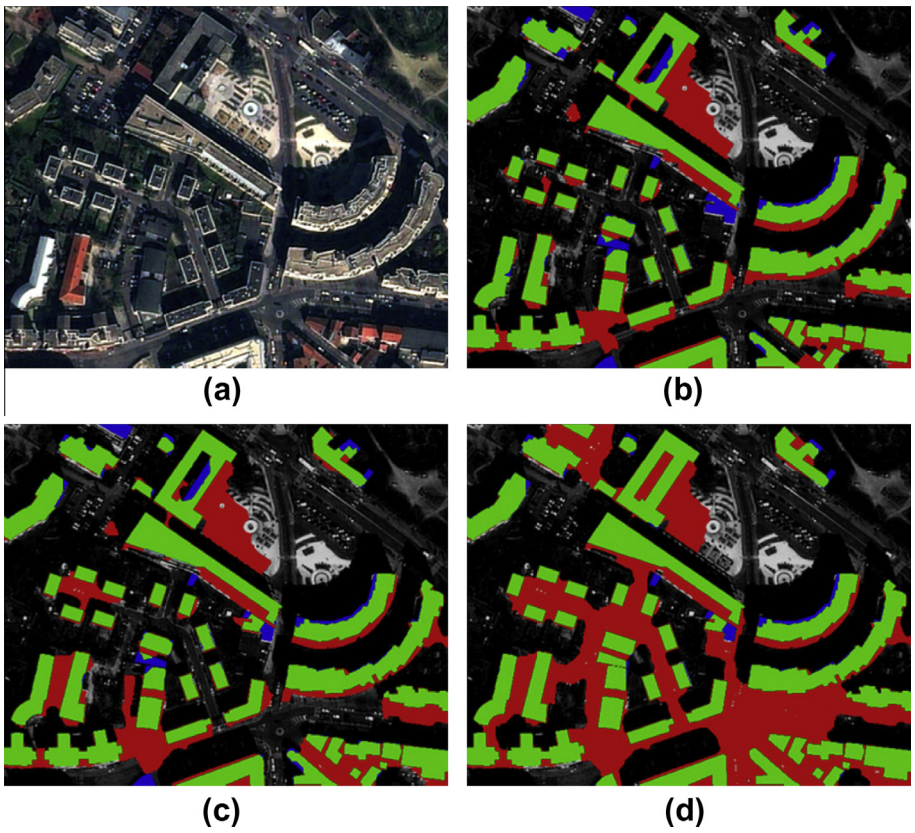

Fig. 12. Effect of smoothing constant γ_2 on the detection results. (a) A GeoEye-1 pan-sharped image (Test patch #10 – RGB). Illustrations of results for (b) $\gamma_2 = 5$, (c) $\gamma_2 = 10$, and (d) $\gamma_2 = 20$. Green, red and blue colours represent *TP*, *FP* and *FN* pixels, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

results of the proposed approach are listed in Table 2. For a numerical assessment for all test images, we also present the results of our previous work, which provide an opportunity to carry out a fair quantitative comparison.

According to Figs. 8 and 9, the results of the proposed approach give the strong impression that the developed approach is highly robust and that the detected regions are quite convincing and representative. As shown, most of the buildings are detected successfully without the strict limitations influenced by the well-known complex characteristics of buildings, e.g., roof colour and texture, shape, size and orientation. It is also evident that the approach distinctively separates building regions from other areas except for a few test images with challenging environmental and illumination conditions. The numerical results in Table 2 also support these findings. Considering the pixel-based assessment, the overall mean ratios of precision and recall are computed as 83.6% and 88.4%, respectively. The computed pixel-based F_1-scores for all test images are approximately 86%, a result that is quite promising for such a diverse and challenging set of test data. It is important to emphasise that our test data involve three highly challenging images (#3, #13, and #14) whose results also contribute to the overall computed performance. Therefore, we would expect better overall pixel-based performance from our approach if the results of these three images were excluded from the evaluation. For the object-based assessment, the overall mean ratios of precision and recall are computed as 95.9% and 81.1%, respectively, and this result corresponds to an overall object-based F_1-score of approximately 88%. As stated earlier, these object-based results are computed based on an overlapping threshold of 60%. However, if we require a strict overlapping threshold of 80%, the precision ratio of our approach drops by approximately 1%, whereas the computed recall decreases to a ratio of approximately 75% (Fig. 13). However, if we accept a building object as detected on the condition that any portion of the building is labelled by the proposed approach (e.g., Lin and Nevatia, 1998; Noronha and Nevatia, 2001; Ünsalan and Boyer, 2005; Sirmacek and Unsalan, 2009), the overall performance reaches a maximum of 96.1%, 89.1%, and 92.5% for the object-based precision, recall and F_1 score measures, respectively (Fig. 13). Considering the complexities in the test images involved, we believe that the performance results achieved by the proposed approach are quite exceptional.

According to the results in Table 2, test image #13 produces the lowest pixel-based performance for all three measures. However, this result was expected because test image #13 is likely the most challenging case out of all the test images involved. The main reason for this challenge is the severe shading effects that occur on the gable roofs. Furthermore, the region growth (cf. Section 2.2) breaks down for several building shadows due to bright snow cover. In contrast, test image #14, which is also acquired in snowy weather conditions, provided quite satisfactory results. In that case, most of the buildings had flat roofs covered with snow. However, due to the region growth step (Fig. 3f) and the second-level partitioning (cf. Section 2.5), almost all of the buildings could be detected, with few false positives.

The second lowest pixel-based performance is achieved for test image #3. This image shows a desert area in which the buildings are located far apart. Two reasons are proposed for the poor pixel-based performance: (i) a strong haze over the entire image, and (ii) the spectral reflectance of the background and the buildings is nearly identical in certain parts (e.g., upper right) of the image. As a result, although the buildings are detected successfully, a dramatic decrease in the pixel-based precision is inevitable due

Table 2
Numerical results of the proposed building detection approach. The values in brackets present the results of the previous approach presented in Ok et al. (2013). Note that the results of three challenging test images (#3, #13, and #14) are also involved in the analysis and contribute to the overall performance computed.

Test image	Number of reference		Pixel-based performance (%)	Object-based performance (%)
	Pixels	Objects	Precision–Recall–F_1-score	Precision–Recall–F_1-score
#1	102,896	39	81.2–88.2–84.6 {82.0–81.1–81.5}	90.3–71.8–80.0 {86.7–66.7–75.4}
#2	59,361	96	76.8–83.3–79.9 {72.3–73.4–72.8}	98.6–75.0–85.2 {98.4–64.6–78.0}
#3	19,923	52	58.3–94.3–72.1 {70.0–95.3–80.7}	100–100 - 100 {100–100 - 100}
#4	104,328	80	85.1–94.3–89.5 {83.1–89.4–86.1}	100–88.8–94.1 {95.8–85.0–90.1}
#5	151,558	290	76.3–80.0–78.1 {74.2–76.2–75.2}	99.6–80.0–88.1 {98.6–75.2–85.3}
#6	82,918	24	89.3–94.7–91.9 {84.6–83.0–83.8}	84.0–87.5–85.7 {70.4–79.2–74.5}
#7	81,186	26	82.1–95.3–88.2 {76.5–90.3–82.9}	100–76.9–87.0 {100–76.9–87.0}
#8	61,155	23	85.2–92.0–88.5 {77.6–90.7–83.6}	95.0–82.6–88.4 {86.4–82.6–84.4}
#9	208,310	70	86.1–75.7–80.6 {78.5–77.6–78.0}	88.2–64.3–74.4 {92.6–71.4–80.6}
#10	56,845	43	70.1–91.5–79.4 {75.9–83.1–79.3}	97.5–90.7–94.0 {94.9–86.1–90.2}
#11	210,108	69	93.4–93.3–93.3 {87.5–88.6–88.0}	89.7–88.4–89.1 {77.0–82.6–79.7}
#12	165,750	59	92.8–95.5–94.1 {88.2–77.6–82.5}	100–88.1–93.7 {97.6–69.5–81.2}
#13	11,352	23	30.1–70.9–42.3 {42.2–64.5–51.0}	94.1–69.6–80.0 {77.8–60.9–68.3}
#14	71,777	21	76.6–86.5–81.3 {79.0–62.1–69.5}	86.4–90.5–88.4 {61.9–61.9–61.9}
#15	32,302	15	88.6–97.4–92.8 {52.0–95.2–67.3}	82.4–93.3–87.5 {72.2–86.7–78.8}
#16	111,837	22	90.3–87.0–88.6 {76.4–83.8–79.9}	87.5–63.6–73.7 {68.2–68.2–68.2}
Overall	1,531,606	952	83.6–88.4–85.9 {78.8–81.8–80.3}	95.9–81.1–87.9 {91.2–76.1–83.0}
min	19,923	15	30.1–70.9–42.3 {42.2–62.1–51.0}	82.4–63.6–73.7 {61.9–60.9–61.9}
max	210,108	290	93.4–97.4–94.1 {87.5–95.3–88.0}	100–100 - 100 {100–100 - 100}

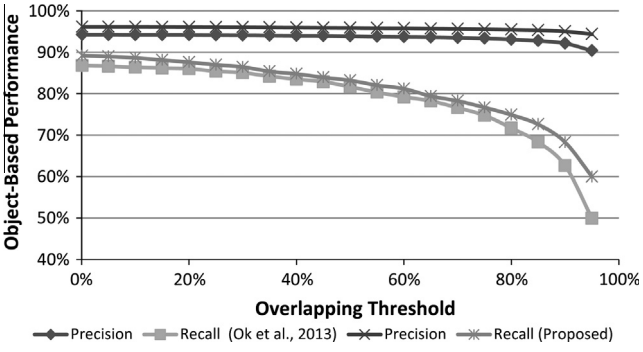

Fig. 13. Object-based performance of the proposed approach for different overlapping thresholds considering all of 16 test patches.

to over-detection. Apart from these test images, the proposed approach provides much better performance results. However, the pixel- and object-based recall ratios computed for test image #9 are not as good as those of the other test sites. This result is primarily due to low buildings with grey roofs in the upper region of the image. These buildings have spectral signatures similar to that of the background, and therefore, the global information exploited in the second-level partitioning does not aid in recovering the missing portions. Beyond that, a reasonable pixel-based precision ratio of 70.1% is computed for the demanding test image #10, in which a large amount of over-detection is observed in the upper centre of the image. In that portion of the image, a component of the gable roof is wrongly detected as a shadow region. As a result, the automatically selected T_F region in the first-level partitioning contains a number of pixels corresponding to the background near the building, thus leading to false positive detections.

The results prove that the approach is generic for different roof colours and types and has the ability to detect arbitrarily shaped buildings in complex environments. Furthermore, the approach is not limited to urban areas but can also be used for building detection in rural environments (#2). Provided that the shadows cast by buildings are not completely occluded, test images #6-7 demonstrate that the approach can recover building regions located in relatively dense vegetation canopies. In addition, if a building is under construction and has a substantial elevation above ground surface, the approach successfully detects it. Several buildings under construction in the test image #1 are good examples of this observation. In addition to these distinctive characteristics, the proposed approach offers further advantages that are clearly highlighted in Fig. 9. As shown in test image #15, if a shadow of a man-made non-building object is attached to the cast shadow of the building, the detection performance of our previous approach deteriorates (Fig. 9b). However, the morphological shadow post-processing (cf. Section 2.2) proposed in this study successfully solves these problems (Fig. 9c). Additionally, the modification step applied in Section 2.4 further improves precision, e.g., the two buildings in the lower left corner in Fig. 9c. As a result, for test patch #15, the

Table 3
Elapsed time of each section of the proposed building detection approach.

	Vegetation extraction and shadow detection (Section 2.1)	Post-processing of the shadow mask (Section 2.2)	Generation and pruning of fuzzy landscapes (Section 2.3)	First-level partitioning (Section 2.4)	Second-level partitioning (Section 2.5)	Determination of final building regions (Section 2.6)	Total
Elapsed time (min: sec)	01:27	05:26	00:08	05:01	36:35	00:25	49:02
Percentage	3%	11%	<1%	10%	75%	<1%	100%

proposed approach provides a notably large improvement of more than 36% for the computed precision ratio (Table 2). Further improvements are also visible in test patch #16, e.g., the upper right corner of Fig. 9e, and clearly demonstrate the difference between the results achieved by the two approaches. However, the major improvement in test image #16 is observed for the ridge of the stadium, and this observation supports the global information integrated into the developed approach. As shown in Fig. 9e, our previous approach partially detects the stadium ridge with a large over-detection. The partial detection is caused by the parameter ROI size (cf. Table 1) used to define the extent of the search region of the local processing. According to our tests, if large ROI sizes (>75 m) are selected, our previous approach handles the partial detection problem that occurred on the stadium ridge. However, for those cases, we observed a worse case of over-detection on the running track within the stadium. In contrast, the proposed approach completely recovered the ridge of the stadium with minimal over-detection (Fig. 9f). It is also noteworthy that there is no need to adjust the ROI size because the situation is addressed by the global second level.

In spite of those improvements and advantages, the proposed approach still contains three major limitations. First, we cannot detect building regions whose shadows are not visible. Note that the detection performance is not limited by the initial shadow mask, and our approach is able to recover the shadow regions missed during the first detection step. Nevertheless, the proposed approach will break down for building regions where the cast shadows are fully (self-) occluded. Second, we still cannot make a distinction between the cast shadows of two specific man-made objects, such as a building and a bridge. Although the proposed morphological shadow post-processing (cf. Section 2.2) better addresses the shadows cast by footbridges used by pedestrians (e.g., the bridge located in the upper right of test image #16), the proposed approach still cannot address large bridges (e.g., used for vehicular traffic) without increasing the height threshold (T_{height}). Third, a group of buildings might be labelled as a single building in the output of the proposed approach due to two specific reasons: (i) the cast shadows of adjoining buildings cannot be separated, and (ii) the smoothing constant in the second level partitioning tends to over-smooth by merging nearby buildings.

Theoretically, the proposed approach is not expected to provide a result for two specific cases. First, there are specific dates/times at which the sun is directly above the building area at noon between the latitudes 23.5°N and 23.5°S. Therefore, a building might not cast a shadow at these times. However, in practice, this condition does not seriously affect the implementation of the proposed approach due to the characteristics of the VHR satellite imaging. Most of the VHR platforms (e.g., Ikonos-2, QuickBird, GeoEye, and WorldView) operate in a sun-synchronous orbit and pass over each area before noon, near 10:30 a.m. local time. Therefore, we can always expect a shadow to be cast by a building object on Earth during the imaging. Second, there is no specific shadow direction at the poles, but shadows are cast to the north or south for the southern and northern poles, respectively. However, because there are no buildings to detect at the poles, this case does not affect the implementation of the proposed approach either.

3.4. Computational time

All experiments were performed on a laptop computer with a CPU Intel Core2 Duo processor with 2.53 GHz and 4 GB RAM. The framework of the approach was developed in a MatLab environment, and the sections related to GrabCut (Ok et al., 2013) and the multi-label graph optimisation (Boykov et al., 2001; Boykov and Kolmogorov, 2004; Kolmogorov and Zabih, 2004) were implemented in C++ code. The number of pixels in each test image is between 128,734 (382 × 337) and 1,251,489 (901 × 1389) pixels, and the elapsed time required for each section in the proposed approach is presented in Table 3. The building detection process for all 16 test images required approximately 49 min, which corresponds to an average processing time of approximately 3 min per image. As shown in the table, 75% of the total time is spent on the second-level partitioning (cf. Section 2.5). However, an interesting point should be highlighted: only 15% of the total time given for that section is consumed by multi-label graph optimization, whereas the rest is spent on the estimation of the GMM parameters of four classes. Because the parameters of a mixture model with an a priori given number of components are also modelled iteratively by Expectation–Maximization (McLachlan and Peel, 2000) routine in MatLab environment, the estimation requires a significant amount of time especially for the classes with relatively large number of components (K_1 and K_4). Nevertheless, we believe that a significant reduction in the amount of time required for that section can be achieved with a better implementation. Among the other sections, the highest computation time is spent on Section 2.2. This is related with the constraint region growth, which also performs an iterative search to grow the shadow pixels.

4. Conclusions

In this study, a novel approach is developed for the detection of buildings from single VHR multispectral images. First, the vegetation and shadow areas are extracted using the multi-spectral information widely accessible in most VHR satellite images. Thereafter, a post-processing step prepares the shadow areas for the detection of buildings by applying prior knowledge of the solar illumination angles during image acquisition. The spatial relationship between the buildings and their cast shadows is modelled using a fuzzy landscape approach, and a pruning process applied to the landscapes eliminates evidence belonging to vegetation canopies. The building regions are detected via an original two-level graph partitioning approach. In the first level, the building regions are labelled by investigating the local evidence. For this purpose, ROIs are generated using the shadow regions, and for each ROI, the building regions are determined by iterative graph cuts. Thereafter, these building regions are supplied to the second level of the process in which multi-label graph optimisation is performed over the entire domain. Lastly, the building regions labelled in the second level are verified with the aid of the shadow evidence. Experiments performed on 16 test images selected from GeoEye-1 VHR images demonstrate that the proposed approach detects the building regions with a high success rate. A distinctive property of the

proposed approach is its applicability to buildings with diverse characteristics and to VHR images with significantly different illumination properties.

The proposed approach can also be successfully applied to other multispectral VHR images (B, G, R, and NIR). In our approach, the distance- and area-related parameters are defined in object space. Therefore, these parameters can be generalised across different image resolutions. In the future, we plan to carry out tests to determine whether thresholds other than those defined in object space can be directly transferred to images from other VHR sensors. More importantly, our future work will focus on overcoming or at least reducing the limitations of the proposed approach. The final validation step in our method uses only evidence from shadows. However, the second-level partitioning step is able to correctly identify buildings whose cast shadows are completely occluded. Therefore, further evidence other than the shadows can be integrated into the verification step to protect building regions that are correctly identified. To separate large bridges from buildings, we plan to integrate an existing road network into the detection framework. In this way, we aim to provide a stable and consistent solution for that particular limitation. A segment-based strategy might be helpful for identifying individual buildings in a set of buildings that are erroneously labelled as a single connected region. Furthermore, an irregular graph generated from the segments also could be integrated into the framework. This adaptation would be beneficial in terms of computation time and could potentially improve the detection performance as well (Wegner et al., 2011b). We lastly plan to improve the boundaries of the detected building regions using edge simplification.

Acknowledgment

The images utilised in this work were provided by HAVELSAN A.Ş. The author would like to thank Baris Yuksel for a useful discussion on the application of the GrabCut approach. The author also acknowledges the accuracy assessment tool written by Caglar Senaras. The author is grateful to two anonymous reviewers and the associate editor for their helpful comments.

References

Adeline, K.R.M., Chen, M., Briottet, X., Pang, S.K., Paparoditis, N., 2013. Shadow detection in very high spatial resolution aerial images: a comparative study. ISPRS Journal of Photogrammetry and Remote Sensing 80, 21–38.

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. International Journal of Applied Earth Observation and Geoinformation 12 (3), 150–157.

Akçay, H.G., Aksoy, S., 2010. Building detection using directional spatial constraints. In: IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 1932–1935.

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 (8), 3117–3131.

Awrangjeb, M., Ravanbakhsh, M., Fraser, C.S., 2010. Automatic detection of residential buildings using LIDAR data and multispectral imagery. ISPRS Journal of Photogrammetry and Remote Sensing 65 (5), 457–467.

Baillard, C., Maitre, H., 1999. 3-D reconstruction of urban scenes from aerial stereo imagery: a focusing strategy. Computer Vision and Image Understanding 76 (3), 244–258.

Baillard, C., Dissard, O., Jamet, O., Maître, H., 1998. Extraction and textural characterization of above-ground areas from aerial stereo pairs: a quality assessment. ISPRS Journal of Photogrammetry and Remote Sensing 53 (2), 130–141.

Baltsavias, E.P., 2004. Object extraction and revision by image analysis using existing geodata and knowledge: current status and steps towards operational systems. ISPRS Journal of Photogrammetry and Remote Sensing 58 (3-4), 129–151.

Benediktsson, J.A., Pesaresi, M., Arnason, K., 2003. Classification and feature extraction for remote sensing images from urban areas based on morphological transformations. IEEE Transactions on Geoscience and Remote Sensing 41 (9), 1940–1949.

Bouziani, M., Goita, K., He, D.-C., 2010. Rule-based classification of a very high resolution image in an urban environment using multispectral segmentation guided by cartographic data. IEEE Transactions on Geoscience and Remote Sensing 48 (8), 3198–3211.

Boykov, Y., Kolmogorov, V., 2004. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (9), 1124–1137.

Boykov, Y., Veksler, O., Zabih, R., 2001. Fast approximate energy minimization via graph cuts. IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (11), 1222–1239.

Brenner, C., 2005. Building reconstruction from images and laser scanning. International Journal of Applied Earth Observation and Geoinformation 6 (3-4), 187–198.

Bresenham, J.E., 1965. Algorithm for computer control of a digital plotter. IBM Systems Journal 4 (1), 25–30.

Cao, G., Yang, X., 2007. Man-made object detection in aerial images using multi-stage level set evolution. International Journal of Remote Sensing 28 (8), 1747–1757.

Chung, K.-L., Lin, Y.-R., Huang, Y.-H., 2009. Efficient shadow detection of colour aerial images based on successive thresholding scheme. IEEE Transactions on Geoscience and Remote Sensing 47 (2), 671–682.

Collins, R.T., Jaynes, C.O., Cheng, Y.-Q., Wang, X., Stolle, F., Riseman, E.M., Hanson, A.R., 1998. The ascender system: automated site modeling from multiple aerial images. Computer Vision and Image Understanding 72 (2), 143–162.

Cord, M., Declercq, D., 2001. Three-dimensional building detection and modeling using a statistical approach. IEEE Transactions on Image Processing 10 (5), 715–723.

Cord, M., Jordan, M., Cocquerez, J.-P., 2001. Accurate building structure recovery from high resolution aerial imagery. Computer Vision and Image Understanding 82 (2), 138–173.

Croitoru, A., Doytsher, Y., 2003. Monocular right-angle building hypothesis generation in regularized urban areas by pose clustering. Photogrammetric Engineering and Remote Sensing 69 (2), 151–169.

Cui, S., Yan, Q., Reinartz, P., 2011. Graph search and its application in building extraction from high resolution remote sensing imagery, Search Algorithms and Applications. Nashat Mansour (Ed.), InTech.

Fischer, A., Kolbe, T.H., Lang, F., Cremers, A.B., Förstner, W., Plümer, L., Steinhage, V., 1998. Extracting buildings from aerial images using hierarchical aggregation in 2D and 3D. Computer Vision and Image Understanding 72 (2), 185–203.

Fradkin, M., Maitre, H., Roux, M., 2001. Building detection from multiple aerial images in dense urban areas. Computer Vision and Image Understanding 82 (3), 181–207.

Gevers, T., Smeulders, A.W.M., 1999. Color-based object recognition. Pattern Recognition 32 (3), 453–464.

Haala, N., Brenner, C., 1998. Interpretation of urban surface models using 2d building information. Computer Vision and Image Understanding 72 (2), 204–214.

Haala, N., Brenner, C., 1999. Virtual city models from laser altimeter and 2d map data. Photogrammetric Engineering and Remote Sensing 65 (7), 787–795.

Haala, N., Kada, M., 2010. An update on automatic 3D building reconstruction. ISPRS Journal of Photogrammetry and Remote Sensing 65 (6), 570–580.

Hermosilla, T., Ruiz, L.A., Recio, J.A., Estornell, J., 2011. Evaluation of automatic building detection approaches combining high resolution images and LIDAR data. Remote Sensing 3 (12), 1188–1210.

Hongjian, Y., Shiqiang, Z., 2006. 3D building reconstruction from aerial CCD image and sparse laser sample data. Optics and Lasers in Engineering 44 (6), 555–566.

Huertas, A., Nevatia, R., 1988. Detecting buildings in aerial images. Computer Vision, Graphics, and Image Processing 41 (2), 131–152.

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 (3), 236–248.

Irvin, R.B., McKeown, D.M., 1989. Methods for exploiting the relationship between buildings and their shadows in aerial imagery. IEEE Transactions on Systems, Man, and Cybernetics 19 (6), 1564–1575.

Izadi, M., Saeedi, P., 2012. Three-dimensional polygonal building model estimation from single satellite images. IEEE Transactions on Geoscience and Remote Sensing 50 (6), 2254–2272.

Jaynes, C., Riseman, E., Hanson, A., 2003. Recognition and reconstruction of buildings from multiple aerial images. Computer Vision and Image Understanding 90 (1), 68–98.

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 (1), 133–144.

Katartzis, A., Sahli, H., 2008. A stochastic framework for the identification of building rooftops using a single remote sensing image. IEEE Transactions on Geoscience and Remote Sensing 46 (1), 259–271.

Khoshelham, K., Nardinocchi, C., Frontoni, E., Mancini, A., Zingaretti, P., 2010. Performance evaluation of automated approaches to building detection in multi-source aerial data. ISPRS Journal of Photogrammetry and Remote Sensing 65 (1), 123–133.

Kim, T., Muller, J.-P., 1998. A technique for 3D building reconstruction. Photogrammetric Engineering and Remote Sensing 64 (9), 923–930.

Kim, T.J., Muller, J.P., 1999. Development of a graph-based approach for building detection. Image and Vision Computing 17 (1), 3–14.

Kim, Z., Nevatia, R., 2004. Automatic description of complex buildings from multiple images. Computer Vision and Image Understanding 96 (1), 60–95.
Koc-San, D., Turker, M., 2012. A model-based approach for automatic building database updating from high-resolution space imagery. International Journal of Remote Sensing 33 (13), 4193–4218.
Kolmogorov, V., Zabih, R., 2004. What energy functions can be minimized via graph cuts? IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (2), 147–159.
Krishnamachari, S., Chellappa, R., 1996. Delineating buildings by grouping lines with MRFs. IEEE Transactions on Image Processing 5 (1), 164–168.
Lafarge, F., Descombes, X., Zerubia, J., Pierrot-Deseilligny, M., 2008. Automatic building extraction from DEMs using an object approach and application to the 3D-city modeling. ISPRS Journal of Photogrammetry and Remote Sensing 63 (3), 365–381.
Lee, D.S., Shan, J., Bethel, J.S., 2003. Class-guided building extraction from Ikonos imagery. Photogrammetric Engineering and Remote Sensing 69 (2), 143–150.
Lin, C., Nevatia, R., 1998. Building detection and description from a single intensity image. Computer Vision and Image Understanding 72 (2), 101–121.
Liow, Y.T., Pavlidis, T., 1990. Use of shadows for extracting buildings in aerial images. Computer Vision, Graphics, and Image Processing 49 (2), 242–277.
Mayer, H., 1999. Automatic object extraction from aerial imagery-a survey focusing on buildings. Computer Vision and Image Understanding 74 (2), 138–149.
McGlone, J.C., Shufelt, J.A., 1994. Projective and object space geometry for monocular building extraction. In: Proc. of Computer Vision and Pattern Recognition, pp. 54–61.
McKeown, D.M., Cochran, S.D., Ford, S.J., McGlone, J.C., Shufelt, J.A., Yocum, D.A., 1999. Fusion of HYDICE hyperspectral data with panchromatic imagery for cartographic feature extraction. IEEE Transactions on Geoscience and Remote Sensing 37 (3), 1261–1277.
McLachlan, G.J., Peel, D., 2000. Finite mixture models. Wiley, New York.
Mohan, R., Nevatia, R., 1989. Using perceptual organization to extract 3D structures. IEEE Transactions on Pattern Analysis and Machine Intelligence 11 (11), 1121–1139.
Noronha, S., Nevatia, R., 2001. Detection and modeling of buildings from multiple aerial images. IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (5), 501–518.
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 (3), 1701–1717.
Otsu, N., 1975. A threshold selection method from gray-level histograms. Automatica 11, 285–296.
Paparoditis, N., Cord, M., Jordan, M., Cocquerez, J.P., 1998. Building detection and reconstruction from mid- and high-resolution aerial imagery. Computer Vision and Image Understanding 72 (2), 122–142.
Peng, J., Liu, Y.C., 2005. Model and context-driven building extraction in dense urban aerial images. International Journal of Remote Sensing 26 (7), 1289–1307.
Peng, J., Zhang, D., Liu, Y.C., 2005. An improved snake model for building detection from urban aerial images. Pattern Recognition Letters 26 (5), 587–595.
Pesaresi, M., Benediktsson, J.A., 2001. A new approach for the morphological segmentation of high-resolution satellite imagery. IEEE Transactions on Geoscience and Remote Sensing 39 (2), 309–320.
Polidorio, A.M., Flores, F.C., Imai, N.N., Tommaselli, A.M.G., Franco, C., 2003. Automatic shadow segmentation in aerial colour images, XVI Brazilian Symposium on Computer Graphics and Image Processing, pp. 270–277.
Poulain, V., Inglada, J., Spigai, M., Tourneret, J.-Y., Marthon, P., 2011. High-resolution optical and sar image fusion for building database updating. IEEE Transactions on Geoscience and Remote Sensing 49 (8), 2900–2910.
Rother, C., Kolmogorov, V., Blake, A., 2004. Grabcut: interactive foreground extraction using iterated graph cuts. ACM Transactions on Graphics 23 (3), 309–314.
Rottensteiner, F., Trinder, J., Clode, S., Kubik, K., 2007. Building detection by fusion of airborne laser scanner data and multi-spectral images: performance evaluation and sensitivity analysis. ISPRS Journal of Photogrammetry and Remote Sensing 62 (2), 135–149.
Rüther, H., Martine, H.M., Mtalo, E.G., 2002. Application of snakes and dynamic programming optimisation technique in modeling of buildings in informal settlement areas. ISPRS Journal of Photogrammetry and Remote Sensing 56 (4), 269–282.
Sarabandi, P., Yamazaki, F., Matsuoka, M., Kiremidjian, A., 2004. Shadow detection and radiometric restoration in satellite high resolution images. In: IEEE International Geoscience and Remote Sensing, Symposium, pp. 3744–3747.
Schindler, K., 2012. An overview and comparison of smooth labeling methods for land-cover classification. IEEE Transactions on Geoscience and Remote Sensing 50 (11), 4534–4545.
Senaras, C., Özay, M., Vural, F.Y., 2013. Building detection with decision fusion. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 6 (3), 1295–1304.
Shackelford, A.K., Davis, C.H., 2003. A combined fuzzy pixel-based and object-based approach for classification of high-resolution multispectral data over urban areas. IEEE Transactions on Geoscience and Remote Sensing 41 (10), 2354–2363.
Shufelt, J.A., 1996. Exploiting photogrammetric methods for building extraction in aerial images. International Archives of Photogrammetry and Remote Sensing 31 (Part B6), pp. 74–79.
Shufelt, J.A., 1999. Performance evaluation and analysis of monocular building extraction from aerial imagery. IEEE Transactions on Pattern Analysis and Machine Intelligence 21 (4), 311–326.
Shufelt, J.A., McKeown, D.M., 1993. Fusion of monocular cues to detect man-made structures in aerial imagery. CVGIP-Image Understanding 57 (3), 307–330.
Sirmacek, B., Unsalan, C., 2009. Urban-area and building detection using SIFT keypoints and graph theory. IEEE Transactions on Geoscience and Remote Sensing 47 (4), 1156–1167.
Sohn, G., Dowman, I., 2007. Data fusion of high-resolution satellite imagery and LiDAR data for automatic building extraction. ISPRS Journal of Photogrammetry and Remote Sensing 62 (1), 43–63.
Stassopoulou, A., Caelli, T., 2000. Building detection using bayesian networks. International Journal of Pattern Recognition and Artificial Intelligence 14 (6), 715–733.
Sümer, E., Turker, M., 2013. An adaptive fuzzy-genetic algorithm approach for building detection using high-resolution satellite images. Computers, Environment and Urban Systems 39, 48–62.
Suveg, I., Vosselman, G., 2004. Reconstruction of 3D building models from aerial images and maps. ISPRS Journal of Photogrammetry and Remote Sensing 58 (3–4), 202–224.
Tack, F., Buyuksalih, G., Goossens, R., 2012. 3D building reconstruction based on given ground plan information and surface models extracted from spaceborne imagery. ISPRS Journal of Photogrammetry and Remote Sensing 67, 52–64.
Teke, M., Başeski, E., Ok, A.Ö., Yüksel, B., Şenaras, Ç., 2011. Multi-spectral false colour shadow detection. In: Stilla, U., Rottensteiner, F., Mayer, H., Jutzi, B., Butenuth, M. (Eds.), Photogrammetric Image Analysis. Springer, Berlin Heidelberg, Berlin, Heidelberg, pp. 109–119.
Tournaire, O., Brédif, M., Boldo, D., Durupt, M., 2010. An efficient stochastic approach for building footprint extraction from digital elevation models. ISPRS Journal of Photogrammetry and Remote Sensing 65 (4), 317–327.
Tsai, V.J.D., 2006. A comparative study on shadow compensation of colour aerial images in invariant colour models. IEEE Transactions on Geoscience and Remote Sensing 44 (6), 1661–1671.
Tupin, F., Roux, M., 2003. Detection of building outlines based on the fusion of SAR and optical features. ISPRS Journal of Photogrammetry and Remote Sensing 58 (1–2), 71–82.
Turker, M., San, B.T., 2004. Detection of collapsed buildings caused by the 1999 Izmit, Turkey earthquake through digital analysis of post-event aerial photographs. International Journal of Remote Sensing 25 (21), 4701–4714.
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 (3), 423–461.
Vallet, B., Pierrot-Deseilligny, M., Boldo, D., Brédif, M., 2011. Building footprint database improvement for 3D reconstruction: a split and merge approach and its evaluation. ISPRS Journal of Photogrammetry and Remote Sensing 66 (5), 732–742.
Vestri, C., 2006. Using range data in automatic modeling of buildings. Image and Vision Computing 24 (7), 709–719.
Wegner, J.D., Hansch, R., Thiele, A., Soergel, U., 2011a. Building detection from one orthophoto and high-resolution InSAR data using conditional random fields. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 4 (1), 83–91.
Wegner, J.D., Soergel, U., Rosenhahn, B., 2011b. Segment-based building detection with conditional random fields, In: Stilla, U., Gamba, P., Juergens, C., Maktav, D. (Eds.). Proceedings of 6th IEEE/GRSS/ISPRS Joint Urban Remote Sensing Event, pp. 205–208.
Weidner, U., Förstner, W., 1995. Towards automatic building extraction from high-resolution digital elevation models. ISPRS Journal of Photogrammetry and Remote Sensing 50 (4), 38–49.
Xiao, J., Gerke, M., Vosselman, G., 2012. Building extraction from oblique airborne imagery based on robust façade detection. ISPRS Journal of Photogrammetry and Remote Sensing 68, 56–68.
Zhang, Y., 1999. Optimisation of building detection in satellite images by combining multispectral classification and texture filtering. ISPRS Journal of Photogrammetry and Remote Sensing 54 (1), 50–60.