A Hierarchical Building Detection Method for Very High Resolution Remotely Sensed Images Combined with DSM Using Graph Cut Optimization

Rongjun Qin and Wei Fang

Abstract

Detecting buildings in remotely sensed data plays an important role for urban analysis and geographical information systems. This study proposes a hierarchical approach for extracting buildings from very high resolution (9 cm GSD (Ground Sampling Distance)), multi-spectral aerial images and matched DSMs (Digital Surface Models). There are three steps in the proposed method: first, shadows are detected with a morphological index, and corrected for NDVI (Normalized Difference Vegetation Index) computation; second, the NDVI is incorporated using a top-hat reconstruction of the DSM to obtain the initial building mask; finally, a graph cut optimization based on modified superpixel segmentation is carried out to consolidate building segments with high probability and thus eliminates segments that have low probability to be buildings. Experiments were performed over the whole Vaihingen dataset, covering 3.4 km2 with around 3000 buildings. The proposed algorithm effectively extracted 94 percent of the buildings with 87 percent correctness. This demonstrates that the proposed method achieved satisfactory results over a large dataset and has the potential for many practical applications.

Introduction

The identification and localization of buildings in an urban area is very important for planning, building analysis, automatic 3D reconstruction of building models and change detection (Qin and Gruen, 2014). The development of very high resolution (VHR) remote sensing images (Qin *et al.*, 2013) creates a possible avenue to sense individual buildings in an urban scenario, e.g., Ikonos with 1-meter resolution, or Worldview with 0.5-meter resolution. Sensors with even higher resolution are in the planning stages (e.g., Geoeye-2 and Worldview-3 with 0.3-meter resolution). However, this increasing level of detail does not necessarily facilitate building detection with an improved accuracy (Huang and Zhang, 2011). Indeed, more detailed image contents actually increase spectral ambiguities in remotely sensed images, such as symbol patterns on the road, and big vehicles. Therefore, researchers have devoted a lot of effort toward using multi-source data and designing better detection strategies to increase the building detection rate.

Multispectral images provide shadow information as primitives for building locations. Furthermore, shadow information are especially effective in single image based methods (Huang and Zhang, 2012; Ok, 2013; Ok *et al.*, 2013). Meanwhile, NDVI data extracted from a multispectral image can be used as vegetation indicators to eliminate trees. Vector features such as parallel lines and corner junctions reveal the characteristics of rectangular buildings, which have been investigated and used to develop single-image based methods for building detection (Lin and Nevatia, 1998; Sirmacek and Unsalan, 2011; Sirmacek and Unsalan, 2010; Sirmaçek and Unsalan, 2009).

Lidar (Light Detection and Ranging) point clouds provide height information for a ground scene and are used for building detection. By subtracting the DTM (Digital Terrain Model) from the DSM (Digital Surface Model), a nDSM (normalized DSM) can be computed to obtain off-terrain points for building detection (Weidner and Förstner, 1995). In addition, the multi-return characteristics of lidar provide useful information to eliminate the vegetation for point clouds based methods (Ekhtari *et al.*, 2008; Meng *et al.*, 2009), to increase the accuracy of building detection.

Both multispectral image and lidar point clouds have their advantages and deficiencies. Complex algorithms based on a single image usually have assumptions concerning building distribution and sometimes are only able to detect certain types of buildings. For example, methods based on feature point extraction from a single image are only able to detect isolated buildings with regular patterns, and methods relying on parallel lines are not able to detect dome roofs. As compared to multi-spectral images, lidar point clouds provide accurate height information, but less accurate boundaries. There are also null values for lidar point clouds due to occlusion and specular reflection from water surfaces on the roofs. Therefore, integration of both sources is a possible direction for improving building detection accuracy as well as robustness.

There has been a spate of integrated methods proposed in the literature. Rottensteiner *et al.* (2007) and Rottensteiner *et al.* (2005) proposed a supervised classification-based building

Rongjun Qin is with the Singapore ETH Center, Future Cities Laboratory, ETH, Zurich. 1 CREATE Way, #06-01 CREATE Tower, Singapore 138602 (rqin@student.ethz.ch).

Wei Fang is with the Singapore ETH Center, Future Cities Laboratory, ETH, Zurich. 1 CREATE Way, #06-01 CREATE Tower, Singapore 138602, and the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, China. #129, Luoyu Road, Wuchang District, LIESMARS, Wuhan University, Wuhan, P. R. China, 430079.

Photogrammetric Engineering & Remote Sensing
Vol. 80, No. 8, September 2014, pp. 000–000.
0099-1112/14/8009–000

doi: 10.14358/PERS.80.9.000

extraction framework: it fused features extracted from the nDSM and multispectral images such as strength and directness with the Dempster-Shafer algorithm (Shafer, 1976), with a final morphology operation performed to eliminate small segments. However, parameter tuning in the method is dependent on an estimation of wooded areas, and is not able to detect buildings smaller than 30 m^2. Adopting this same supervised approach, Lu *et al.* (2006) fused a matched DSM, a NDVI of the multispectral image and modeled the edges as features and inserted this data into the Dempster-Shafer algorithm, and then the building class was taken as the final output. Nevertheless, this algorithm is strongly dependent on the features extracted in early processing stages, and missed buildings will not be recovered in the classification step.

Instead of extracting information from all over the image, Turlapaty *et al.* (2012) first obtained an initial test dataset by truncating a DSM by a threshold, and then fused image block-based features such as mean, variance, skewness, etc. into a SVM (Support Vector Machine) classifier to separate the building pixels from non-building pixels in the initial test data. This method can effectively reduce false positives, but it cannot handle ground scenes in high relief.

Meng *et al.*, (2012) generated building candidates using a multi-directional ground filter on lidar data to obtain bare ground points, and then NDVI was employed to remove trees. Finally a supervised C4.5 decision tree analysis (Quinlan, 1993) was performed to separate building pixels from non-building pixels. However, about 2.55 percent of tree pixels were identified as buildings, which might have been due to the low NDVI values of vegetation under the shadow.

The aforementioned methods are mostly based on supervised classification. They aim to extract the primitives of each class in a fuzzy way, fusing different types of source data. However, inappropriate feature selection, underestimation of urban classes, and insufficient training samples affect the building detection results (Durrieu *et al.*, 2007). Therefore, some researchers have developed hierarchical approaches, which aim to exclude non-building area/pixels in a step-wise fashion.

Chen *et al.*, (2012) proposed a hierarchical approach for building detection with a combination of nDSM and multispectral images: initial building segments were generated by truncating both nDSM and NDVI sequentially, and final building masks were determined by a set of rules considering the region size and the spatial relation between trees and buildings. However, their method relied on the quality of the nDSM, which preferably uses an accurate DTM, unfortunately not generally available for many areas. Based on nDSM and NDVI, Grigillo *et al.* (2011) obtained the initial building mask in the same manner as (Chen *et al.*, 2012), they eliminated vegetation under the shadows with low NDVI values by truncating areas with low homogeneity. This method works well when trees are isolated, but building boundaries are affected when surrounded by trees. Awrangjeb *et al.* (2010) proposed an iterative way to extend high probability to low probability building masks based on line primitives and the height information. Nevertheless, this method estimated a DTM as the approximate DSM, which failed to detect buildings with high variation in terrain relief.

Semantic information such as shadows can be included in hierarchical approaches. In single image based approaches, shadows provide an active clue for possible locations of the off-terrain objects (Ok, 2013; Ok *et al.*, 2013). When height information is available, shadows hide the spectral signature of some trees, which results in false positives (identifies trees as buildings) (Grigillo *et al.*, 2011). However, there are only a few algorithms that consider this problem.

In most of the hierarchical building detection methods, the initial building masks were obtained by step-wise exclusion of non-building pixels using NDVI and nDSM. However, there were still non-building pixels identified as buildings due to inappropriate thresholds or truncation rules, and the initial building mask need further refinement (Chen *et al.*, 2012; Rottensteiner *et al.*, 2005). Most of the algorithms attributed the refinement to a morphology operator or region-size filter on the binary mask (Meng *et al.*, 2012; Meng *et al.*, 2009; Rottensteiner *et al.*, 2005). Such methods are effective ways to clean the building mask, but can erroneously remove small buildings. Therefore, the refinement of the building mask must consider the original spectral and height information.

The graph cuts algorithm (Vicente *et al.*, 2008) is a powerful optimization tool for solving binary or multiple labeling problems in a connected graph. In this context, the refinement of building masks can be transformed to a binary problem, where building and non-building pixels/segments are regarded as nodes in connected graphs. It is possible to deploy height and spectral information as weighting constraints in the graph, thus achieving better performance.

Therefore, we propose a hierarchical method aiming to integrate the spectral and height information through a graph cut optimization framework for building detection. The proposed method will first generate an initial building mask hierarchically by considering the shadow problem and off-terrain object extraction, and thus refine the initial mask with graph cut optimization. The rest of the sections are organized as follows: the next section presents the general methodological consideration of how the proposed algorithm is going to tackle the aforementioned limits. The subsequent section presents the proposed building detection algorithm in detail. In the next subsequent section, two experiments with the Vaihingen dataset (Cramer, 2010) are presented, along with detailed analysis and discussion. The final section concludes the paper by discussing the pros and cons of the new method and further improvements that can be made.

Overall Methodological Considerations

Due to some of the limitations discussed in the first section, the proposed algorithm will do three things: (a) extract buildings under shadows while eliminating the ambiguities of vegetation found under the shadow area when possible, (b) extract building candidates without the use of a DTM, and (c) integrate the original height and spectral information for building mask refinement, so as to prevent small buildings from being excluded. Therefore, we accordingly proposed this algorithmic strategy as follows:

- For shadows in high resolution images, some still contain weak spectral responses. Thus, we can still recover part of the information by stretching the histograms in the corresponding areas. Since the main purpose is to recover the NDVI under the shadow, we stretch the histogram of both the "Saturation" and "Lightness" channel of the HSL (Hue, Saturation, Lightness) color space: stretching the "Lightness" recovers the illuminations and stretching the "Saturation" ensures that interrelation among each color bands.
- 2D grey level based morphology top-hat reconstruction is commonly used to identify the peak area of a 2D function. This is considered to be a typical feature of off-terrain objects in the DSM. By selecting an appropriate radius in accordance with the DSM resolution, these top hats can be efficiently detected. To provide building-orientated, top-hat reconstruction, the NDVI can be embedded for truncating non-building top-hats.
- Small buildings normally have different color when compared to the surrounding roads/bare soils/grasses, thus spectral connectivity between the segments and the background (non-building pixels) can be used to

strengthen the potential building pixels/segments, and filter out possible non-building pixels/segments which have similar spectral response and height in relation to the surrounding environment.

Hence, our proposed building detection algorithm has three steps: (a) shadow detection and auto-color equalization of the shadow areas, (b) Top-hat reconstruction of the DSM combining NDVI for initial building mask generation, and (c) Graph cut optimization based on superpixel segmentation for initial building mask refinement, with height filtering as post-processing. The proposed workflow is shown in Figure 1; each step will be introduced in detail in the following section.

The Proposed Building Detection Algorithm

Shadow Detection and Auto-color Correction

Shadow Detection

The shadow is regarded as a common feature class in remote sensing images of an urban scenario. On one hand, it is a solid evidence of the off-terrain objects, but on the other hand, objects in the shadow area (Huang *et al.*, 2013), such as buildings and vegetation, can be occluded and missed. Therefore, in the methods combining multispectral and height information, the problem of missed vegetation in the shadows occurs frequently: trees in the shadow area are identified as off-terrain candidates, and their NDVI values are not strong enough for them to be identified and eliminated. Therefore, our major goal is to recover information from under the shadow, in order to compute the NDVI more comprehensively.

The most prominent indicator of a shadow is low luminance. Based on this fact, many shadow detection approaches (Chen *et al.*, 2007; Dare, 2005; Huang and Zhang, 2012; Liu *et al.*, 2011; Tsai, 2006) have been proposed. In this study, we adopt an effective morphology shadow index (MSI) similar to that proposed by Huang and Zhang (Huang and Zhang, 2012), but with modifications for computational efficiency and effectiveness. It adopts morphology top-hat reconstruction based on the brightness image (computed by taking the maximal spectral response of each multispectral channel), to detect dark blobs as the shadow areas. It is particularly effective in urban areas, where the scale of the shadows is attributed to the buildings and trees, which can be assumed to be within a certain range. Different from MSI proposed by Huang and Zhang (2012), which adopts multi-scale structural elements for morphology reconstruction, we use a fixed radius of the multi-directional lines and combine them in a single structural element, to achieve both computational efficiency and accurate results. Moreover, instead of performing shadow detection on the brightness image, we perform shadow detection on the first channel of a PCA (Principal Component Analysis) transformed image (P1 image), since it not only has high intra-pixel variance, but also high inter-pixel variance (Jolliffe, 2005).

To determine the shadow mask on the multispectral image, it is critical to set the threshold for the shadow index, and this varies with the image. In urban scenes with many off-terrain objects, shadow and non-shadow areas are the major domain of luminance. Therefore, we adopt unsupervised K-mean clustering to separate the P1 image into two clusters, and take the mean of the centroid values of the two clusters as the threshold.

Auto-Color Correction

As described in the previous section, the main aim of shadow correction is to compute the NDVI more completely. Due to the local micro-environment in an urban scene, the illumination varies at different locations. Therefore, the correction needs to be done in each local shadow to maximize the intra-variance of the color. The per-pixel shadow mask is segmented using a fast segmentation with 4-neighborhood connectivity (Davies, 2004), and the auto-color correction is performed for each segment.

To maintain the relative relations of the multispectral bands, we treat the near infrared band, red band, and green band as the normal R,G,B band and transform it in to HSL color space (Joblove and Greenberg, 1978), and then a histogram stretching (Awrangjeb *et al.*) is performed on the lightness and saturation for each shadow object:

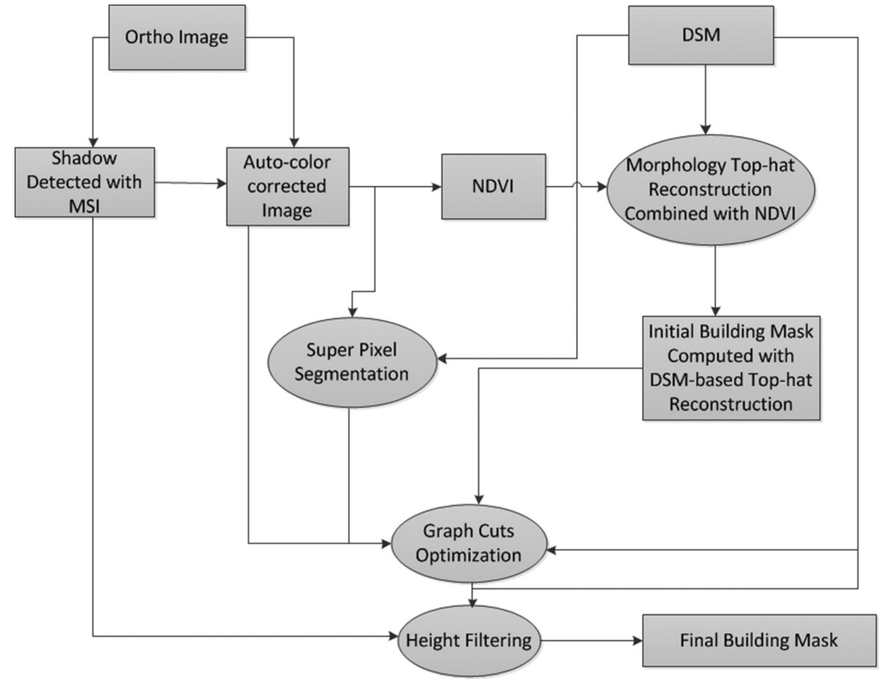
Ortho Image
DSM
Shadow Detected with MSI
Auto-color corrected Image
NDVI
Morphology Top-hat Reconstruction Combined with NDVI
Super Pixel Segmentation
Initial Building Mask Computed with DSM-based Top-hat Reconstruction
Graph Cuts Optimization
Height Filtering
Final Building Mask

Figure 1. The work flow of the proposed algorithm.

$$S' = \frac{S - \min(S)}{\max(S) - \min(S)}, L' \frac{L - \min(L)}{\max(L) - \min(L)} \quad \text{for } S, L \in \textit{shadow object} \qquad (1)$$

where S and L stand for "Saturation" and "Lightness," respectively. After HS (histogram stretching), the image is transformed back to RGB space from the HSL space. The advantage of HS on saturation and lightness is that it enhances the luminance of the shadow area without changing the color information of the original band. Figure 2 shows the detected shadow and the shadow-corrected images. In Figure 2d and 2e, it could be seen clearly that shadow area on the left side of the building is recovered, with trees and grasses more visible in this area. The NDVI of the shadow corrected area in Figure 2e shows that the vegetation under this area reveals much higher response than that in Figure 2d, leading to more complete identification of the vegetation. It should be noted that not only shadow areas covering vegetation can be recovered, but that the shadow on the buildings are corrected as well, which provides more information for the following superpixel segmentation. Thus, the NDVI can be computed based on the shadow-corrected multispectral image.

Top-hat Reconstruction on DSM Combining NDVI

Mathematical Morphology is recognized as a powerful tool for digital image processing, especially for binary/grey level shape analysis (Soille, 2007). Opening and closing operations infer the spatial relations in the image space and provides useful information about the local area. Binary morphology operators have been extended to grey level images (Vincent, 1993), and one of the most useful components is morphology reconstruction, where a mask image J can be reconstructed as B_{JI} from the marker image I by finding the peaks of I which are marked by J. By subtracting B_{JI} from the image mask J, the peaks of J overlaying on I can be extracted, namely top-hat extraction. The marker image I is usually generated by grey level erosion through a pre-defined structural element e; therefore a top-hat reconstruction T_J^e of a grey level image J is computed as follows:

$$T_j^e = J - B_{J\varepsilon(j,e)} \qquad (2)$$

where $\varepsilon(j,e)$ is the grey level morphology erosion, which is defined as follows:

$$\varepsilon(j,e)(i,j) = \min\{J(a,b) \mid , e(a - i, b - j) = 1\} \qquad (3)$$

Having assumed that the buildings dominate the bright region of remote sensing images, Huang and Zhang (2011) adopted top-hat reconstruction for the brightness of images for building detection. However, this method cannot handle dark roofs and bright trees. Therefore, we adopt the same concept of top-hat reconstruction on the DSM. The structural element is formed by combining multi-direction lines with a fixed radius, which is more efficient than "disk-shaped" structural element used in the previous method. The radius of the structural element is estimated as the radius of the circumcircle of the largest buildings in the scene, which is dependent on the resolution of the data source. An increasing radius beyond the circumcircle of the largest buildings will not bring more correct detection, but more computational time.

Since the extracted top-hats from the DSM contain the off-terrain segments, a common strategy is to impose NDVI constraints on the extracted results to eliminate the vegetation. However, one deficiency of top-hat reconstruction is that it

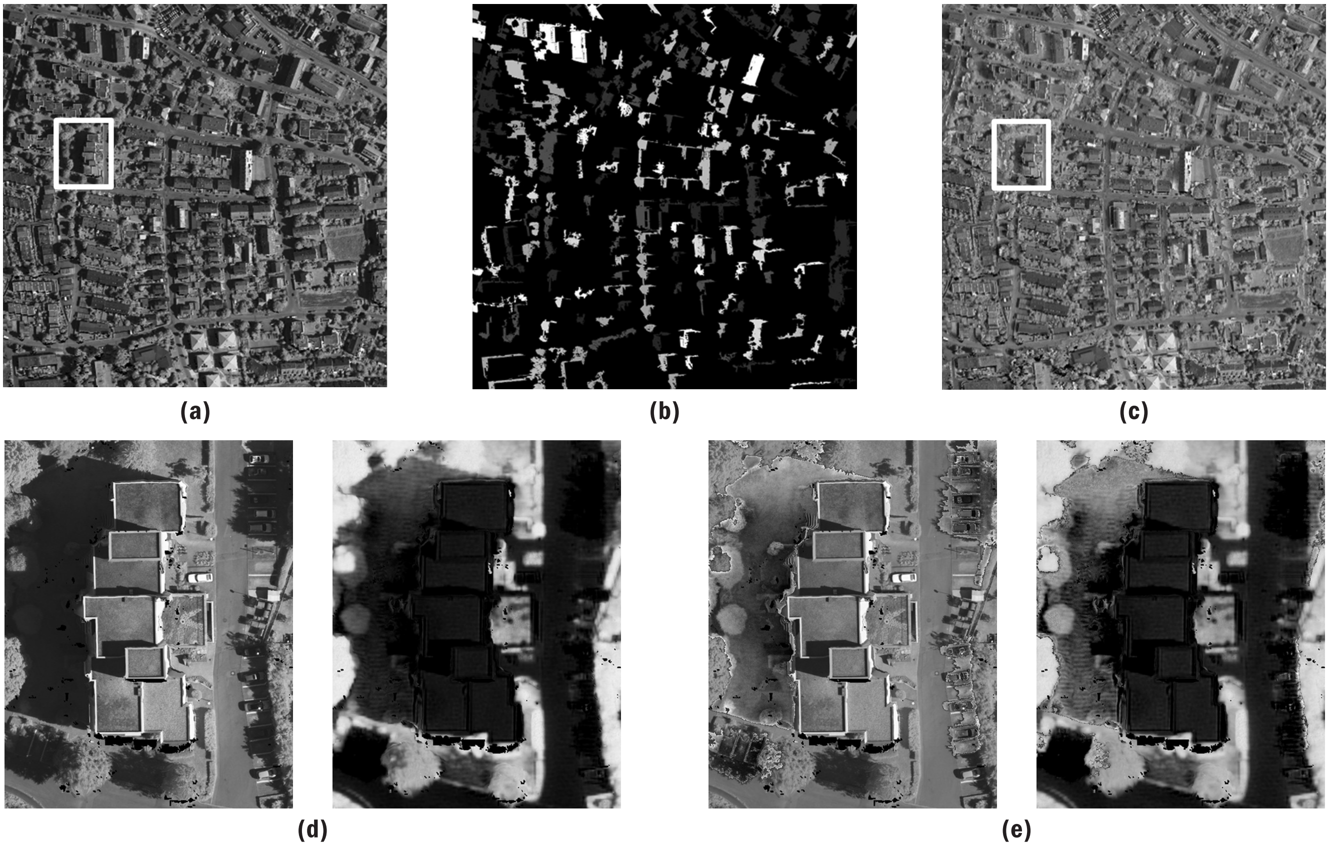

Figure 2. Shadow detection and object-wise auto-color correction: (a) original image, (b) detected shadow objects with grey color coding, (c) color-corrected image, (d) enlarged view of the a shadow area in original image in the white box and the corresponding NDVI, and (e) enlarged view of the color-corrected image in the white box and the corresponding NDVI.

cannot detect buildings that have at least one side connected to a slope or to trees with similar height. Therefore, to avoid such cases, we propose that the top-hat reconstruction should be combined with NDVI truncation during the calculation, instead of a post truncation on the top-hats. Thus we implement the NDVI truncation into the erosion process as follows:

$$\tilde{\varepsilon}(J,e)(i,j)=\begin{cases}\min\left(\left\{J(a,b)\right|,e(a-i,b-j)=1\right), & \text{if NDVI}(i,j)<T_h\\ J(i,j) & \text{otherwise.}\end{cases}\tag{4}$$

where T_h is the NDVI threshold (in our case it is 0.2), and then the top-hat reconstruction combined with NDVI can be computed by replacing with:

$$'T_J^e = J - B_{J\tilde{\varepsilon}(j,e)}\tag{5}$$

Figure 3 shows a comparison between a post truncation on the extracted top-hat and the NDVI truncation during the extraction process. The urban scene shown in Figure 3 is situated in a sloped area, with dense vegetation between buildings and terrain. It can be seen in the white box marked in Figure 3, a post-truncation misses buildings due to the dense canopy, while the NDVI truncation during the reconstruction process extracts more top-hats from buildings. It can also been seen that Figure3b shows more non-building structures, which are caused by isolated grounds surrounded by truncated trees. These non-buildings structures are highly anisometric and can be easily removed by subsequent building mask refinement. Figure 3c shows the used multi-direction structural element, and the advantage of this is that it requires less computation time than normal disk-shaped element and it maintains highly similar results as those of the normal disk-shaped structural elements.

Graph Cut Optimization Based on the Superpixel Segmentation

The top-hat reconstruction of DSM combined with the NDVI provides the coarse height of each pixel relative to the ground, and with a given threshold (we use 1-meter in our experiments), the initial binary building mask can be obtained.

There are still many non-building pixels in the initial mask, e.g., low terrain reliefs (such as low-rise parking lots) on the ground, isolated tree pixels that failed to be truncated by the NDVI, and remaining tree segments hidden under the shadows which were not corrected. To eliminate these non-building segments, the most popular methods are morphology opening/closing operations and a region-size filter. Morphology opening and closing operations aim to eliminate isolated segments, and in the meantime fill up missing segments inside a region. A region-size filter aims to delete isolated pixels that are smaller than a given region size. However, both of the two methods only perform the operation on the binary mask of the building candidates, and can erroneously eliminate small buildings.

It should be noted that the urban objects are strongly related to their surrounding environment, and color and height difference between these objects are good indicators to separate them. For example, low terrain reliefs share the same spectral response with the surrounding ground, as well as similar heights; the remaining tree pixels have similar colors and heights as the surrounding tree pixels which have been eliminated in previous steps. Small isolated buildings usually have distinguishable colors and heights separating them from the surrounding grasses and ground. This context forms strong primitives for the final optimization of the building mask.

Graph Cut Optimization

Graph cut optimization (Vicente *et al.*, 2008) can be adopted to elaborate the connectivity contexts: pixels that have similar color and height must belong to the same categories (in this case, it is either building pixels or non-building pixels). The classic graph cuts algorithm tries to minimize a cost function in the following form:

$$C(X) = \textstyle\sum_{p\in v} D_p(x_p) + \alpha\sum_{(pq)\in E} U_{pq}(x_p, x_q)\tag{6}$$

where (V, E) is a general graph; V is the vertex set, and E is the edge set. $D_p(x_p)$ is the node cost, and $x_p = 0$ in the binary case. $U_{pq}(x_p, x_q)$ is the smooth term that defines the neighborhood relationship for each edge. A large value for $U_{pq}(x_p, x_q)$ means a high penalty for the value of x_p and x_q, and vice versa; α controls the weight of the smooth term in the overall cost $C(X)$. The minimization of the cost function can be then transformed to a max-flow/min-cut problem (Boykov and Kolmogorov, 2004).

In our context, we define $x_p = 1$ as being a building pixel/segment, and $x_p = 0$ as a non-building pixel/segment. Graph cut optimization solves for each node to minimize the total cost $C(X)$. $D(x_p)$ is regarded as the cost of a building pixel/non-building pixel, and $U_{pq}(x_p, x_q)$ can be defined according to the color and height difference:

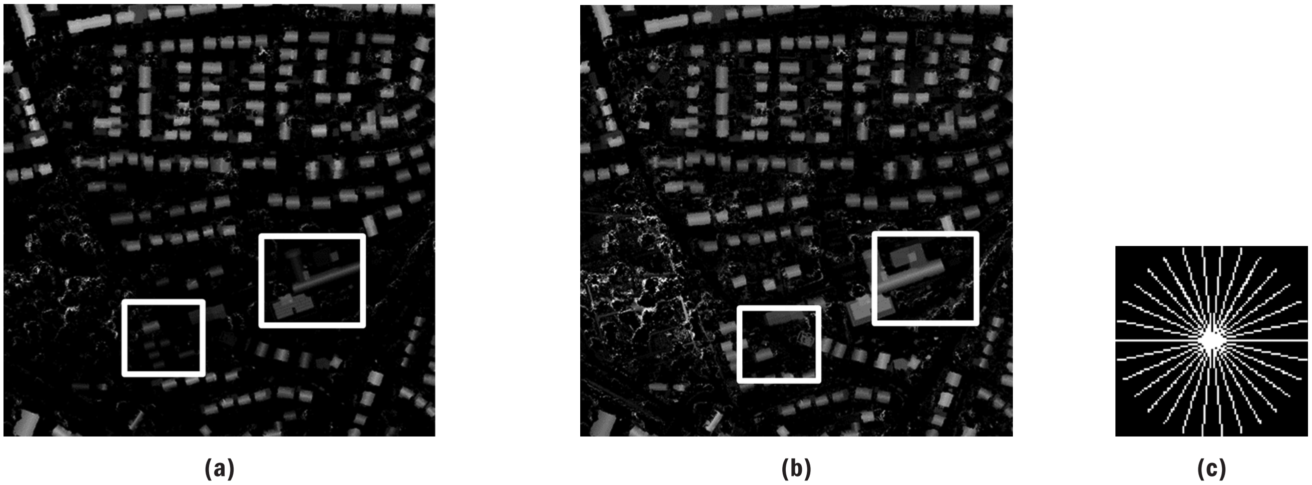

Figure 3. (a) Post-NDVI truncation on the top-hats and the used structural element, (b) top-hats reconstruction NDVI truncation during the top-hat reconstruction, and (c) multi-direction linear structural element with 20 directions.

$$U_{pq}\left(x_p, x_q\right) = \begin{cases} 1-(1-\beta)\left\|I_p - I_q\right\| - \beta\left\|h_p - h_q\right\|, if x_p = x_q \\ 0, \qquad otherwise \end{cases} \tag{7}$$

where I and h indicates the color and the height for the node, and both of them are normalized to [0,1] (the color is normalized for each individual channel). β is a weight parameter controlling the contribution of the height and spectral term, and $\beta \in [0,1]$. In our experiment, we set $\beta = 0.5$ to share the contribution of spectral and height term equally. Graph cuts can be performed in a per-pixel fashion, where each pixel represents a node, and the 4-connectivity/8-connectivity relationships can be used to build up the edges of the graph. However, problems arise for the per-pixel graph cuts:

- It is difficult to find a good indicator to describe the probability of a pixel being a building pixel with unsupervised approaches, since there is no *a priori* information of the building colors. If using height as the indicator, the short buildings/low buildings will be eliminated due to low probability. The most reliable information is the initial building mask, thus the probability is either 0 or 1.
- An image/DSM grid can easily go up to tens of thousands of megapixels, and therefore the computational complexity will be tremendous for the graph-based algorithms.
- A per-pixel operator can be easily affected by noise, which may result in irregular building boundaries.

Graph Cuts on Modified Superpixel Segmentation

To address the problems stemming from pixel-based graph cuts, we perform graph cut optimization on the superpixel segments (Achanta *et al.*, 2012). This has several advantages: The initial cost can be computed by analyzing the statistics from the initial building mask inside the superpixels, to produce a more robust probability measurement; thus segments are more robust to noise, and follow the image boundary well; it is more efficient to solve the graph cuts problem in segments than in a per-pixel fashion.

Superpixel segmentation is an image over-segmentation technique that groups connected pixels with similar color. It adopts a K-mean clustering method over local regions of initially well-distributed seed points. It iteratively groups pixels with the Euclidean distance of the 5D vector constructed by R, G, B components and 2D geometric positioning in the image space (pixel position):

$$G_S = \left\|I_k - I_i\right\| + \tfrac{M}{S}\left\|x_k - x_i\right\| \tag{8}$$

where I_k and I_i are the color (in our case, we use the *CIELAB* color space) of the k_{th} seeds and pixels in its local region, and x_k and x_i represent the respective image coordinates. M and S controls the compactness of the resulting superpixels. A small M/S value results in less compact but more boundary-aware superpixels, and vice versa.

To obtain more meaningful segments, we incorporate the height information for superpixel segmentation as an additional channel:

$$G_S = \left\|I_k - I_i\right\| + \tfrac{M}{S}\left(\left\|x_k - x_i\right\| + \left\|h_k - h_i\right\|\right) \tag{9}$$

where h indicates the corresponding height of the pixel, and it is scaled to the same order of magnitude as the color channels. One essential parameter is the size S of the superpixels, used to determine the number of initial seeds N. The number of superpixels can be estimated as ($ImageWidth \times ImageHeight/S$), which needs to be adjusted according to the image resolution. Thus, S should be small enough to cover small buildings and large enough for efficient computation. After obtaining the superpixels, we compute their connectivity based on neighborhood relations. For each superpixel, we compute its color and height by averaging all the pixels inside it.

For the per-pixel graph cuts, it is difficult to define the probability of being building pixels. However with the superpixel segments, we can compute the probability of a superpixel d $P(a)$ directly from the initial building mask :

$$P(d) = \mathrm{Mean}_{t\in d}\,(T(t)) \tag{10}$$

where $T(t)$ is either 0 or 1, and $P(d)$ is computed by calculating the percentage of the building pixels in a superpixel d. Therefore, we can compute the initial cost $D(x_p)$ of each node p in Equation 6 as:

$$\begin{cases} D\left(x_p = 1\right) = 1 - P\left(p\right) \\ D\left(x_p = 0\right) = P\left(p\right) \end{cases} \tag{11}$$

and the smooth term $U_{pq}(x_p,x_q)$ can be computed according to Equation 7. Figure 4 shows the graph cuts results based on the superpixel segments.

Graph cut optimization eliminates most of the non-building segments, including the small tree segments, low terrain reliefs and small segments on the ground. However, some large shadows attached to buildings might be wrongly identified as part of the buildings, because these shadows failed to be corrected due to very weak spectral response, and these areas are either low terrain reliefs or trees.

Therefore, to eliminate these wrongly identified areas without affecting shadow areas covering the roofs, we adopt a height filter on these shadows: we use a fast region-growing method to segment the resulting building masks, and compare it with the detected shadow area in the first step. The average heights of the shadow areas are then compared to their adjacent building segments: if the difference is larger than a given threshold (e.g., 10 meters), the shadow areas in the building segment will be eliminated.

Unlike morphology/region-size filtering methods, which perform a pure operation on the binary mask of the building candidates graph cut optimization explores the connectivity between the building candidates and non-building pixels with their spectral and height similarity. It can be seen in Figure 4f that very small buildings are kept, while larger non-building candidates are erased.

Experiment and Discussion

An experiment was conducted on the Vaihingen test dataset, which was captured over Vaihingen in Germany (Cramer, 2010) with aerial camera. This dataset contains different scenarios: inner city, high rises, and residential areas. The DSM and orthophoto were generated by INPHO 5.3 software. In this section, we first evaluate the proposed algorithm with the three test sites by comparing the extracted buildings with the reference data, and then compare the whole dataset with the manually sketched building masks on the orthophoto. The performance of the proposed algorithm is measured with the following three metrics (Rottensteiner *et al.*, 2012) for the three test datasets, which were computed as follows for (1) Completeness, (2) Correctness, and (3) Kappa coefficient (KC):

1. $Completeness = TP\ /\ (TP + FN)$
2. $Correctness = TP\ /\ (TP + FP)$
3. $KC = ,\ M = (PD \times PG + ND \times NG)/(Npix \times Npix)$ (12)

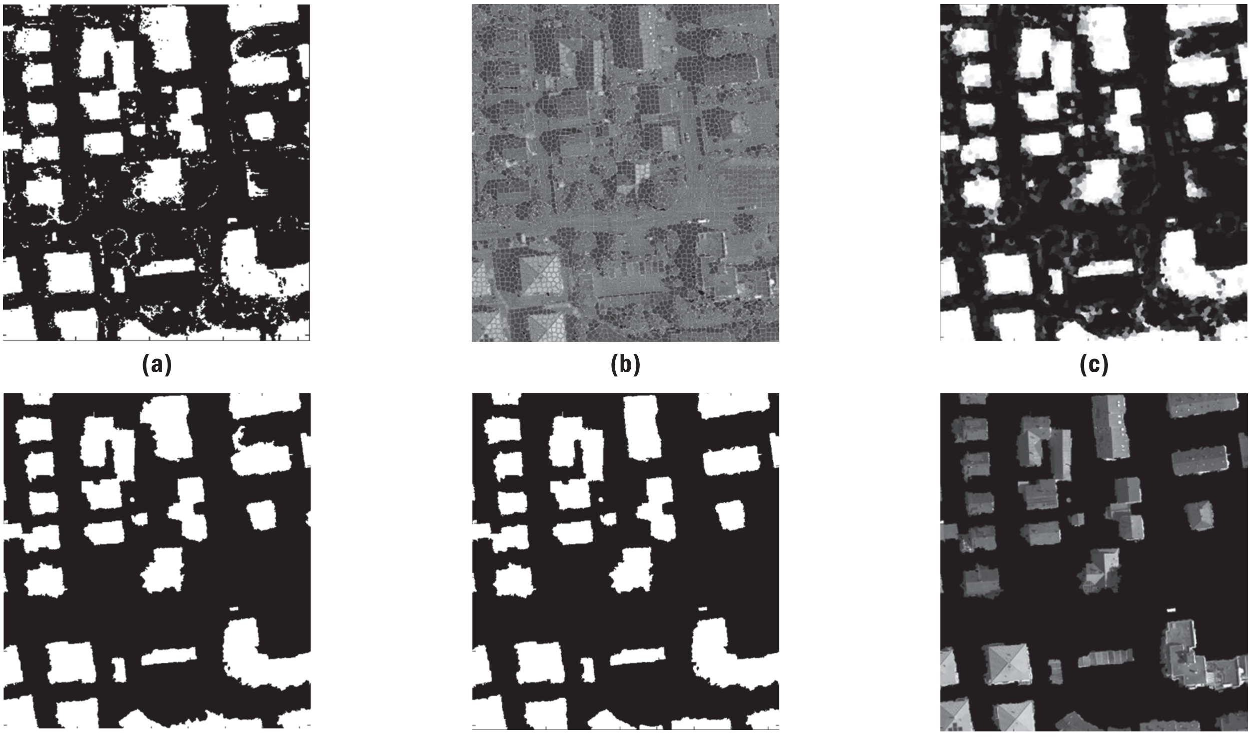

Figure 4. Graph cut optimization for the building mask refinement: (a) initial building mask, (b) superpixel over-segmentation, (c) initial cost, (d) Graph cut optimization, (e) height filter, and (f) detected buildings.

where *TP*, *FP*, *FN* are true positive, false positive and false negative, respectively. *Npix* denotes the number of pixels in the area of interest, and *PD*, *PG*, *ND*, *NG* are the number of detected positives, number of positives for the reference data, number of negatives detected, and number of negatives for the reference data, respectively.

Experiment 1-Vaihingen Test Sites

The three test areas (Figure 5a, 5b, and 5c) consist of different urban scenes. Since the proposed algorithm is a pixel-based approach, therefore the evaluation is conducted in the pixel domain. The building detection results of the proposed algorithm are shown in Figure5d, 5e, and 5f).

Table 1 shows the statistical results for the three datasets. According to (Mayer *et al.*, 2006), a building detection system should have completeness value larger than 70 percent, and a correctness value larger than 85 percent. Our results on this dataset fulfilled this requirement, 90.4 percent and 91.67 percent, respectively. It can be seen that there are still some omission errors on results of the test dataset, which are mainly caused by vegetation on the roofs, as well as the matching errors. We compared the results of the proposed approach with the results published on the ISPRS website (ISPRS, 2013), and the proposed method yields state-of-the-art results.

Experiment 2 - The Whole Vaihingen Dataset

An evaluation of the test dataset is a good way to validate algorithms, but appraisal of the robustness and practicability of a method relies on its performance on large dataset. For example, there are no dome shaped roofs in the test areas, and the terrains of the three test areas are relatively flat. Therefore, we tested the proposed algorithm on the whole dataset (Figure 6a), which covers approximately 3.4 km^2 with 20,250 × 21,300 pixels in the orthophoto, containing over 3,000 buildings. The computation was made by dividing the whole dataset into small tiles, with the same set of parameters adopted for each tile. It can be seen in Figure 6b that the whole area varies in height, there are some buildings sitting on a slope, and irregular shaped buildings and round buildings can also be seen in the whole dataset. Figure 6c illustrates the extraction results of the whole dataset, demonstrating that our algorithm can effectively detect these irregularly shaped buildings, as well as buildings located on slopes. By a comparison with the manually sketched reference, the proposed algorithm achieved 94.2 percent completeness and 87.5 percent correctness over the whole scene, and this demonstrates its robustness and practical potential. Most of the errors come from deep slopes, where top-hat reconstruction fails to extract the building patches, and some of the errors happen on building roofs under vegetation, which are eliminated by NDVI truncation.

Table 1. Building Detection Result of the Three Test Areas

Test dataset	Completeness	Correctness	KC
1	0.9027	0.9126	0.8442
2	0.9407	0.9033	0.9044
3	0.8846	0.9277	0.8663
Overall	0.9039	0.9166	0.8746

Parameter Analysis

There are several tunable parameters in the proposed algorithm: the truncation height T_{hei} for generating the initial mask; the radius r of the morphology top-hat operator; cell size S (pixels) of the superpixel segmentation. Normally, the connectivity weight α of the graph cuts should be adjusted for different scenarios, due to the scale of different initial cost. Since the initial cost is computed from the initial binary mask, there is less variation in the initial values than normal cases. So α is fixed as 0.5 in our algorithm for all the experiments.

T_{hei} should be small enough to keep short buildings, but large enough to get rid of possible disturbances. In our experiments, we set T_{hei} as 1-meter for the truncating threshold. For

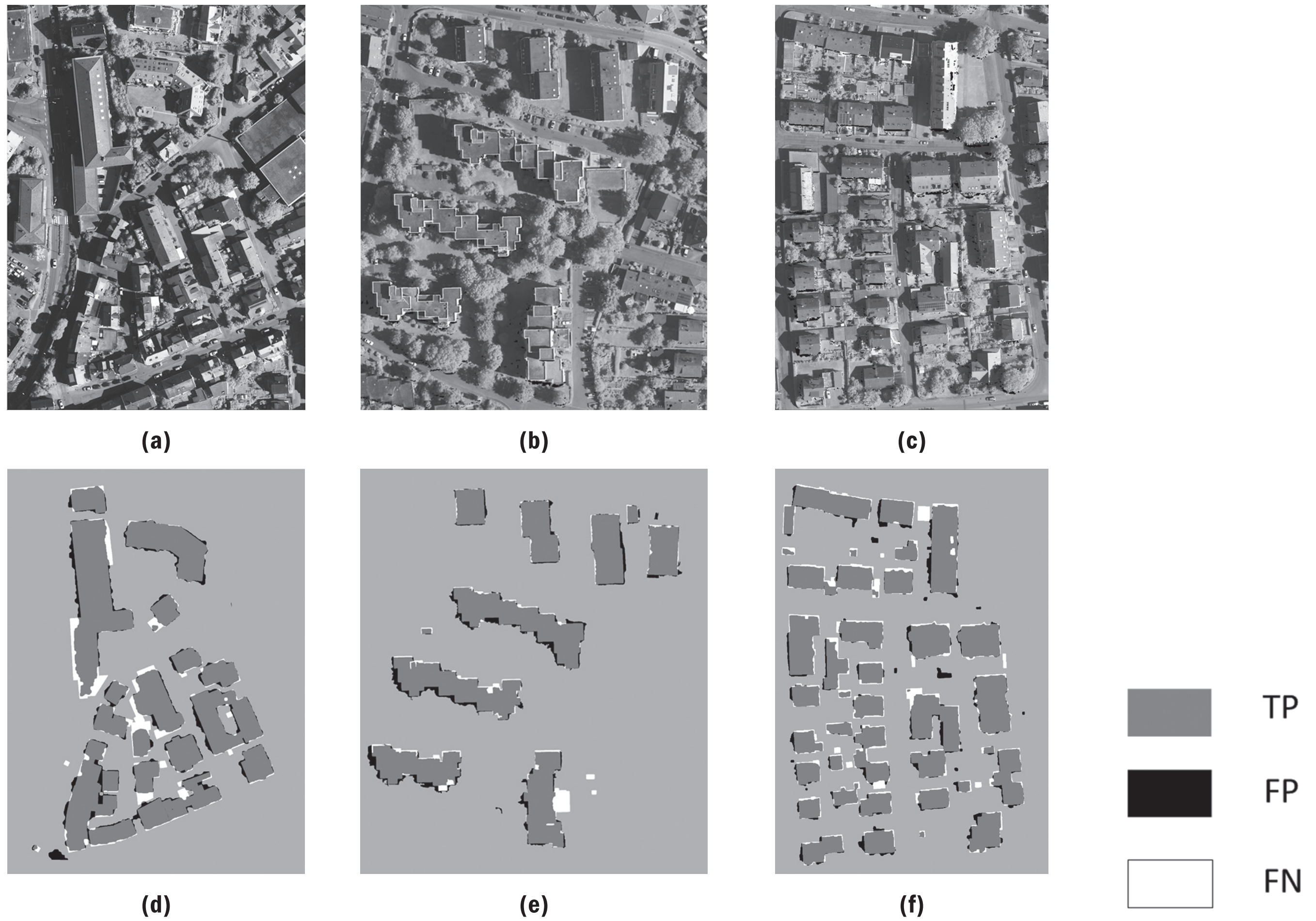

Figure 5. The building detection results of the proposed algorithm: (a, b, c) three test areas in the ISPRS benchmark, and (d, e, f) the detection results. TP: True positive; FP: False Positive; FN: False Negative. The reader may refer to (Rottensteiner *et al.*, 2013) for the state-of-the-art results.

steep areas, it should be small to capture small top-hats (e.g., 0.5-meter), and relatively large for the flat areas (e.g., 2-meters).

As described in the previous section, r should be set as the approximate value of the largest building radius, which is directly dependent on the image resolution. When r becomes larger, the computational load will increase. Figure 7a shows the influence of r on completeness, correctness, and kappa coefficient. It can be seen that the resulting KC is very low when r = 50, and it becomes stable from r = 100. This is mainly because 50 is a smaller than the building scale, which fails to detect top-hats from large buildings. There are slight drops of KC when r changes towards 500, but these drops are not significant. Therefore, r is relatively robust for the proposed algorithm.

Similar to r, the cell size of superpixel segmentation is also dependent on the resolution of the image. S represents the granularity of the superpixel. Figure 7b shows the relationship between S and the final results. We can see that plays an important role in the final result. As S increases from 300, the KC gradually decreases, but correctness remains relatively stable. This means that increasing granularity will not cause more false positives, but rather reduce the true positives.

Figure 7c shows the impact of the spectral and height terms; we fix the other parameters with β varying from 0 to 1. When β = 0, the connectivity of superpixels considers spectral information only, while it considers the height information with β = 1. It can be seen from the Figure 7c that the peak of KC is obtained when β = 0.7, while in general, the KC is relatively stable with the change of β.

Computational Complexity

The computation process of the proposed method is comprised of three major components: (a) morphology top-hat reconstruction, (b) superpixel segmentation, and (c) graph-cut optimization. For other computations such as NDVI computation and height-filtering, they are very fast (less than a second for a 5,000 × 5,000 pixel images) since each pixel only needs to be operated for one time.

The most time-consuming part is the morphology top-hat reconstruction, since it requires a iterative solution. With a fast hybrid algorithm as described in (Vincent, 1993), one could achieve the fastest performance. In our experiment, the top-hat reconstruction of a 5,000 × 5,000 pixel image usually takes about 30 seconds with an Intel Xeron® Processor with 3.10 GHz. The top-hat reconstruction should be computed for two times (one is for the initial building mask generation, and the other one is for the MSI computation).

As described in Achanta *et al.* (2012), the complexity of the superpixel segmentation is practically with respect to the number of pixels, and in our experiment, it usually takes about 40 seconds to segment an 5,000 × 5,000 pixel color image with 20,000 superpixels.

The graph cut algorithm has a low-order polynomial complexity, and the computation time will increase dramatically with increasing number nodes. In our experiments, the nodes are referred to the superpixels, the number of which is much less than the number of pixels. It only takes around four seconds for a graph containing 20,000 nodes.

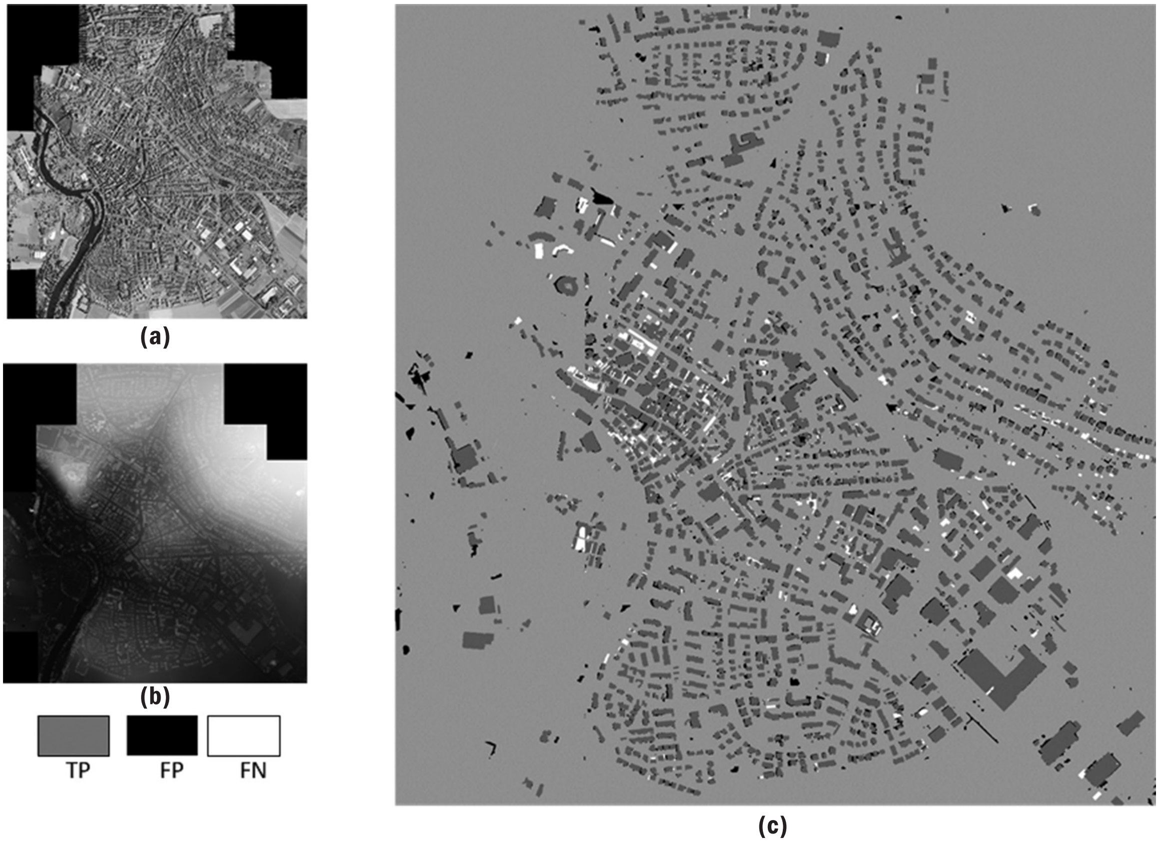

Figure 6. Building detection on the whole Vaihingen dataset: (a) The orthophoto, (b) the matched DSM, and (c) the detection result.

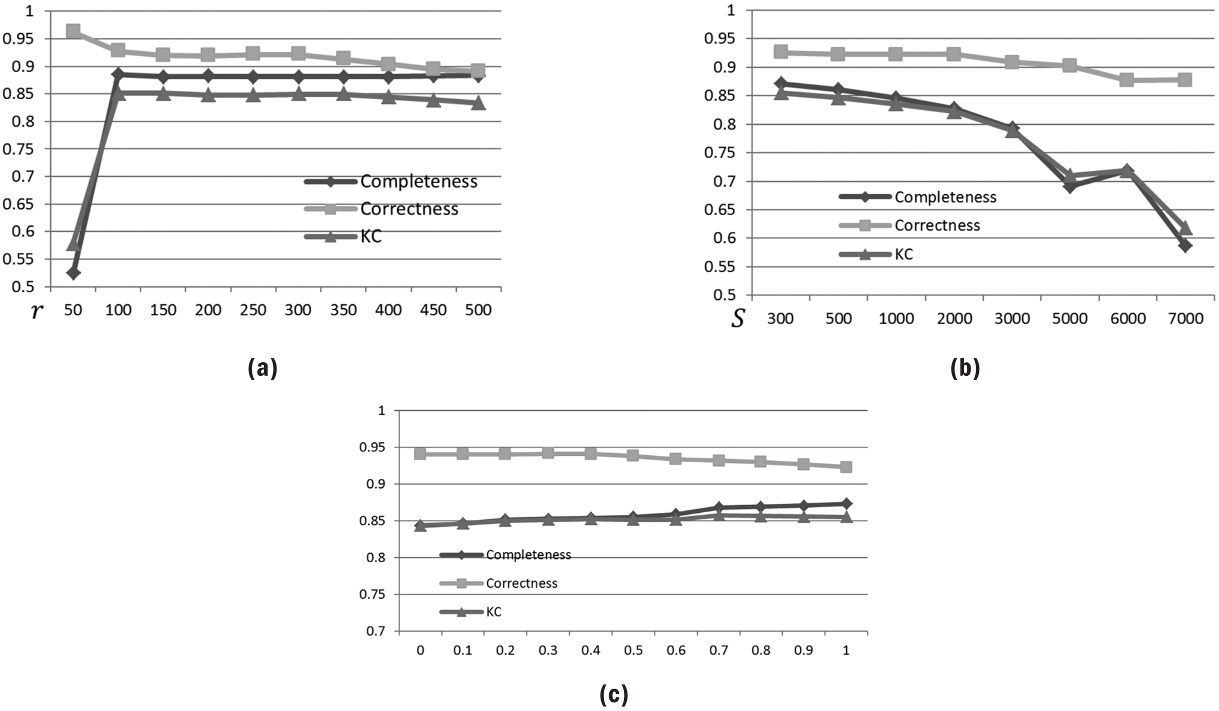

Figure 7. The effect of the parameters of the proposed method: (a) the influence of *r*, (b) the influence of S and (c) the influence of β.

Therefore, the computation time of the proposed methods takes around two minutes for a 5,000 × 5,000 pixel image, which could meet the requirements of most of the applications and is easy to be applied with large datasets.

Conclusions

The current trend in building detection combines the multispectral and height information to achieve more accurate results. However, most of these algorithms do not make the full use of the information concerning shadow areas. Especially in the final separation of building and non-building pixels in multispectral images, few algorithms make use of the connectivity between the building candidates and the non-buildings pixels. This study proposed a hierarchical building detection method that aims to make full use of the information about the color, height, and the connectivity of urban segments in the whole image. The contribution of this paper lies in the following aspects:

1. An object-based shadow correction scheme was proposed to recover information from the shadows to compute the NDVI more comprehensively.
2. A morphology top-hat reconstruction of a DSM combined with a NDVI scheme was adopted to extract the initial building mask, instead of requiring external DTM for nDSM computation.
3. A graph cut algorithm based on modified superpixel segmentation was adopted to fuse the height and multispectral information from DSM and multispectral images. This approach takes into account the connectivity of superpixel segments found all over the image to effectively eliminate non-building segments and retains small buildings that may be erased by traditional morphology/region-size filtering methods.

A comparative study was performed on bench mark data. The experimental results from the proposed method achieved a state-of-the-art level of performance for accuracy and precision. An experiment was performed on the whole Vaihingen dataset which contains over 3,000 buildings varying in shape and size. The proposed algorithm achieved 94.2 percent completeness, 87 percent correctness, with a KC of 0.898, revealing the practical potential of the proposed algorithm.

This study has provided an automatic workflow for building extraction. It can effectively extract 94 percent of the buildings found in a typical urban environment, but there are also several points that need to be improved in the future:

1. Since the proposed method relies purely on the DSM and the generated orthophoto, extracted building boundaries are dependent on the quality of the orthophoto and the DSM. Therefore, better matching algorithms with higher quality around the edges are planned. In addition, better lidar sensors that produce point clouds with clearer edges are expected and will be included for better results.
2. The extraction of the initial mask of a building is based on the top-hat morphology reconstruction. One deficiency in top-hat reconstructions is that it cannot handle buildings that are connected to slopes. Therefore, top-hat reconstruction considering buildings in sloped areas is also planned as a future improvement in the proposed algorithm. Moreover, more automated parameter tuning for truncating thresholds will be investigated to improve the robustness of the proposed method for more challenging datasets.

Acknowledgments

The work was established at the Singapore-ETH Centre for Global Environment Sustainability (SEC), co-funded by the Singapore National Research Foundation (NRF) and ETH Zurich. The Vaihingen data set was provided by the German Society for Photogrammetry, Remote Sensing and Geoinformation (DGPF) (Cramer, 2010), URL: *http://www.ifp.uni-stuttgart.de/dgpf/DKEP-Allg.html*. The authors would like to thank anonymous reviewers for their precious comments and suggestions.

References

Achanta, R., A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, 2012. SLIC superpixels compared to state-of-the-art superpixel methods, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34 (11):2274–2282.

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

Boykov, Y., and V. Kolmogorov, 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.

Chen, L., S. Zhao, W. Han, and Y. Li, 2012. Building detection in an urban area using lidar data and QuickBird imagery, *International Journal of Remote Sensing*, 33(16):5135–5148.

Chen, Y., D. Wen, L. Jing, and P. Shi, 2007. Shadow information recovery in urban areas from very high resolution satellite imagery, *International Journal of Remote Sensing*, 28(15):3249–3254.

Cramer, M., 2010. The DGPF-test on digital airborne camera evaluation overview and test design, *Photogrammetrie-Fernerkundung-Geoinformation*, 2010 (2):73–82.

Dare, P.M., 2005. Shadow analysis in high-resolution satellite imagery of urban areas, *Photogrammetric Engineering & Remote Sensing* 71(2):169–177.

Davies, E.R., 2004.*Machine Vision: Theory, Algorithms, Practicalities*, Fourth edition, Morgan Kaufmann, Burlington, Massachusetts, 934 p.

Durrieu, S., T. Tormos, P. Kosuth, and C. Golden, 2007. Influence of training sampling protocol and of feature space optimization methods on supervised classification results, *Proceedings of IEEE International Geoscience and Remote Sensing Symposium, 2007*, 23-27 July, Barcelona, Spain, pp. 2030–2033.

Ekhtari, N., M. Sahebi, M.V. Zoej, and A. Mohammadzadeh, 2008. Automatic building detection from Lidar point cloud data, *Proceedings of the 21st ISPRS Congress, Commission IV, WG IV/3*, Beijing, China, pp. 473–478

Grigillo, D., M. Kosmatin Fras, and D. Petrovič, 2011. Automatic extraction and building change detection from digital surface model and multispectral orthophoto, *Geodetski Vestnik*, 55(1):28–45.

Huang, X., and L. Zhang, 2012. Morphological building/shadow index for building extraction from high-resolution imagery over urban areas, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(1):161–172.

Huang, X., and L. Zhang, 2011. A multidirectional and multiscale morphological index for automatic building extraction from multispectral GeoEye-1 imagery, *Photogrammetric Engineering & Remote Sensing*, 77(7):721–732.

Huang, X., L. Zhang, and T. Zhu, 2013. Building change detection from multitemporal high- resolution remotely sensed images based on a morphological building index, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, (99):1–11.

ISPRS, 2013. ISPRS Test Project on Urban Classification and 3D Building Reconstruction: Results, URL: *http://www2.isprs.org/commissions/comm3/wg4/tests.html* (last date accessed: 01 July 2014).

Joblove, G.H., and D. Greenberg, 1978. Color spaces for computer graphics, *Proceedings of ACM SIGGRAPH 1978*, 23-25 August, Atlanta, Georgia, pp. 20–25.

Jolliffe, I., 2005.*Principal Component Analysis*, Wiley Online Library.

Lin, C., and R. Nevatia, 1998. Building detection and description from a single intensity image, *Computer Vision and Image Understanding*, 72(2):101–121.

Liu, J., T. Fang, and D. Li, 2011. Shadow detection in remotely sensed images based on self- adaptive feature selection, *IEEE Transactions on Geoscience and Remote Sensing Letters*, 49(12):5092–5103.

Lu, Y.H., J.C. Trinder, and K. Kubik, 2006. Automatic building detection using the Dempster- Shafer algorithm, *Photogrammetric Engineering & Remote Sensing*, 72(4):395–403.

Mayer, H., S. Hinz, U. Bacher, and E. Baltsavias, 2006. A test of automatic road extraction approaches, *International Archives of Photogrammetry, Remote Sensing, and Spatial Information Sciences* 36(3):209–214.

Meng, X., N. Currit, W. Le, and X. Yang, 2012. Detect residential buildings from lidar and aerial photographs through object-oriented land-use classification, *Photogrammetric Engineering and Remote Sensing*, 78(1):35-44.

Meng, X., L. Wang, and N. Currit, 2009. Morphology-based building detection from airborne lidar data, *Photogrammetric Engineering & Remote Sensing*, 75(4):437–442.

Ok, A.O., 2013. Automated detection of buildings from single VHR multispectral images using shadow information and graph cuts, *ISPRS Journal of Photogrammetry and Remote Sensing*, 86(2013):21–40.

Ok, A.O., C. Senaras, and B. Yuksel, 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.

Qin, R., J. Gong, H. Li, and X. Huang, 2013. A coarse elevation map-based registration method for super-resolution of three-line scanner images. *Photogrammetric Engineering & Remote Sensing*, 79(8):717–730.

Qin, R., and A. Gruen, 2014. 3D change detection at street level using mobile laser scanning point clouds and terrestrial images, *ISPRS Journal of Photogrammetry and Remote Sensing*, 90:23–35.

Quinlan, J.R., 1993.*C4. 5: Programs for Machine Learning*, Morgan Kaufmann Series in Machine Learning, Morgan Kaufmann, 302 p.

Rottensteiner, F., G. Sohn, M. Gerke, J.D. Wegner, U. Breitkopf, and J. Jung, 2013. Results of the ISPRS benchmark on urban object detection and 3D building reconstruction, *ISPRS Journal of Photogrammetry and Remote Sensing*, 93:256-271.

Rottensteiner, F., G. Sohn, J. Jung, M. Gerke, C. Baillard, S. Benitez, and U. Breitkopf, 2012. The ISPRS benchmark on urban object classification and 3D building reconstruction, *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information* Sciences, pp. 293–298.

Rottensteiner, F., J. Trinder, S. Clode, and K. Kubik, 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.

Rottensteiner, F., J. Trinder, S. Clode, and K. Kubik, 2005. Using the Dempster–Shafer method for the fusion of LiDAR data and multi-spectral images for building detection, *Information Fusion*, 6(4):283–300.

Shafer, G., 1976. *A Mathematical Theory of Evidence*, Princeton University Press, Princeton, New Jersey.

Sirmacek, B., and C. Unsalan, 2011. A probabilistic framework to detect buildings in aerial and satellite images, *IEEE Transactions on Geoscience and Remote Sensing*, 49(1):211–221.

Sirmacek, B., and C. Unsalan, 2010. Urban area detection using local feature points and spatial voting, *IEEE Geoscience and Remote Sensing Letters*, 7(1):146–150.

Sirmaçek, B., and C. Unsalan, 2009. Urban-area and building detection using SIFT keypoints and graph theory, *IEEE Transactions on Geoscience and Remote Sensing*, 47(4):1156– 1167.

Soille, P., 2007.*Morphological Image Analysis: Principles and Applications*, Second edition, Springer-Verlag, Inc., New York, 391 p.

Tsai, V.J., 2006. A comparative study on shadow compensation of color aerial images in invariant color models, *IEEE Transactions on Geoscience and Remote Sensing*, 44(6):1661–1671.

Turlapaty, A., B. Gokaraju, Q. Du, N.H. Younan, and J.V. Aanstoos, 2012. A hybrid approach for building extraction from space-borne multi-angular optical imagery, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(1):89–100.

Vicente, S., V. Kolmogorov, and C. Rother, 2008. Graph cut based image segmentation with connectivity priors, *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 24-26 June, Anchorage, Alaska, pp. 1–8.

Vincent, L., 1993. Morphological grayscale reconstruction in image analysis: Applications and efficient algorithms, *IEEE Transactions on Image Processing*, 2(2):176–201.

Weidner, U., and W. Förstner, 1995. Towards automatic building extraction from high-resolution digital elevation models, *ISPRS Journal of Photogrammetry and Remote Sensing*, 50(4):38–49.

(Received 23 November 2013; accepted 12 April 2014; final version 14 May 2014)