Multimed Tools Appl (2018) 77:3339–3351
DOI 10.1007/s11042-017-5093-z

Building detection from orthophotos using binary feature classification

Yan Hu [1,2] · Xiangyun Hu [1] · Penglong Li [2] · Yi Ding [2]

Received: 30 April 2017 / Revised: 24 July 2017 / Accepted: 8 August 2017 /
Published online: 30 August 2017

Abstract Building detection in orthophotos is crucial for various applications, such as urban planning and real-estate management. In order to realize accurate and fast building detection, a non-interactive approach based on binary feature classification is brought forward in this paper. The proposed approach includes two major stages, i.e., building area detection and building contours extraction. In the first stage, a sequence of intersections is obtained by superpixel segmentation in the subsampled orthophoto, and then building area is reserved roughly according to the classification of intersections. In the second stage, the sequence of intersections is updated by superpixel segmentation in the building area from original orthophoto, and then building contours is extracted in accordance with the classification of intersections likewise. The local feature of the intersections is descripted employing our extremely compact binary descriptor, and is classified using binary bag-of-features. Experiments show that benefiting from binary description and making full use of texture details and color channels, the proposed descriptor is not only computationally frugal, but also accurate. Experiments are also conducted on orthophotos with different roof colors, textures, shapes, sizes and orientations, and demonstrate that the proposed approach are capable of achieving desirable results.

Keywords Building detection · Machine learning · Local feature · Descriptor · Classifier

1 Introduction

Building detection is crucial for various applications, including urban planning, real-estate management, and disaster relief [11]. Thus, building detection in remote sensing, specifically in high-resolution aerial orthophotos, has been a popular research topic.

✉ Yan Hu
huyan_dl023@cqu.edu.cn

[1] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China

[2] Chongqing Geomatics Center, Chongqing 401121, China

Most building detection approaches usually rely on image segmentation, such as watershed, statistical region merging (SRM), mean shift and grab-cut. The original grab-cut is a semi-automated foreground/background partitioning algorithm. Given a group of pixels interactively labelled as foreground/background by the user, it partitions the rest of the pixels in an image using a graph-based approach [27]. Most building detection applications based on grab-cut work both slow and costly, for the reason that they need operating by human experts. Some researchers improved grab-cut by iterative optimization algorithms, such as the bio-inspired bacterial foraging optimization (BFO) [25], to realize automatic building detection. However, experiments showed that in several test cases, some road segments and bridges were erroneously detected as buildings currently [14].

Though building detection can be treated as a typical segmentation task, it can also be treated as a recognition task. In recent years, machine learning approaches are evolving exponentially [15]. In multimedia event detection (MED) [8, 9, 33], object recognition [19] and motion detection [10], many machine learning methods have been adopted and guarantee desired performance [6, 20]. Machine learning approaches also have been shown to be successful in addressing building detection with high accuracy and high robustness. Furthermore, these approaches need user interaction in the training stage only, and are fully automatic in the detection stage. There has been a significant amount of past work on classification and segmentation of remote sensing imagery using machine learning. For a recent review, please refer to [4, 12].

Vakalopoulou et al. propose a supervised building detection procedure based on the ImageNet framework, while integrating certain spectral information by employing multi-spectral band combinations into the training procedure [30]. The building detection was addressed through a binary classification procedure based on support vector machine (SVM) classifier. The experimental results indicate the quite promising potentials of this approach. Making use of elevation data such as a digital surface model (DSM), Volpi et al. propose a hybrid network that combines the pre-trained image features with DSM features that are trained from scratch [31]. The hybrid network improves the labelling accuracy on the highest-resolution imagery.

Adhering to the direction of finding an accurate and fast way of building detection, we proposed a non-interactive two-stage approach. In the first stage, a sequence of intersections is obtained by simple linear iterative clustering (SLIC) superpixel segmentation in the subsampled orthophoto [1], and then building area is reserved roughly according to the classification of intersections. In the second stage, the sequence of intersections is updated by SLIC in the building area from original orthophoto, and then building contours is extracted in accordance with the classification of intersections likewise. The key to achieve desirable results is the way to describe and classify the features of the patches around intersections. In this paper, we propose a compact binary descriptor, which is a hybrid bit string. This descriptor contains the texture and similarity comparisons in L channel of LAB color space, and color comparisons in A channel and B channel of LAB color space. In order to cooperate with the proposed descriptor, we also realize a binary bag-of-features (BoF) classifier.

Experiments show that the proposed descriptor is not only computationally frugal, but also accurate. Accordingly, the proposed approach is capable of achieving desirable results.

The remainder of this paper is organized as follows. Section 2 mainly introduces the proposed binary descriptor and binary BoF classifier. Section 3 briefly introduces the proposed two-stage building detection approach. In Section 4, experimental details and results are presented. Lastly in Section 5, conclusions are presented.

2 Compact descriptor and binary classifier

This section is split into three main subsections. The first subsection gives a brief introduction to the related work of local feature descriptors, especially binary descriptors. The second subsection describes the implementation details of the proposed compact binary descriptor. The third subsection briefly introduces the binary BoF classifiers.

2.1 Local feature descriptor

2.1.1 Floating-point descriptors

Hu's moment invariants as a shape feature is a milestone, which has been widely used for feature description due to its scaling and rotation invariance [13]. Scale-invariant feature transform (SIFT) is a benchmark for local feature description, because of its excellent performance, which is invariant to a variety of common image transformations [21, 22]. The SIFT descriptor is a 3D histogram of gradient locations and orientations. Location is quantized into a 4 × 4 location grid, and the gradient angle is quantized into eight orientations, resulting in a 128-dimensional descriptor. Speeded up robust features (SURF) is another commonly used method performing approximately as well as SIFT with lower computational cost [3]. Like SIFT, SURF relies on local gradient histograms but uses integral images to speed up the computation. Different parameter settings are possible but, since the 64-dimensional version already yields good performance, that version has become a de facto standard. We proposed a lightweight approach with the name of region-restricted rapid keypoint registration (R^3KR), which makes use of a 12-dimensional orientation descriptor and a two-stage strategy to reduce the computational cost [17].

Though these local feature algorithms have obtained notable description capability when there are large viewpoint and illumination changes, their additional processing to eliminate the second-order effects brings much computational cost [18].

2.1.2 Binary descriptors

Recently, many efforts have been made to enhance the efficiency of matching by employing binary descriptors instead of floating-point ones.

Locally binary pattern (LBP) builds a Census-like bit string where the neighborhoods are taken to be circles of fixed radius [23, 24]. However, LBP translates the bit strings into its decimal representation and build a histogram of these decimal values. The concatenation of these histogram values has been found to result in stable descriptors. The binary robust invariant scalable keypoints (BRISK) method adopts a circular pattern with 60 sampling points, of which the long-distance pairs are used for computing the orientation and the short-distance ones for building descriptors [16]. In order to compute the feature locations, it uses the AGAST corner detector, which improves FAST by increasing speed while maintaining the same detection performance. For scale invariance, BRISK detects keypoints in a scale-space pyramid, performing non-maxima suppression and interpolation across all scales. Fast retina keypoint (FREAK) is another typical binary descriptors inspired by the Human Visual System, and more precisely the retina [2]. A cascade of binary strings is computed by efficiently comparing pairs of image intensities over a retinal sampling pattern. Interestingly, selecting pairs to reduce the dimensionality of the descriptor yields a highly structured pattern that mimics the saccadic search of the human eyes.

Binary robust independent elementary features (BRIEF) is a representative descriptor which directly computes the descriptor bit-stream quite fast, based on simple intensity difference tests in a smoothed patch [5]. It uses a sampling pattern consisting of 128, 256, or 512 comparisons, with sample points selected randomly from an isotropic Gaussian distribution centered at the feature location. BRIEF descriptor often cooperate with the efficient CenSurE detector. For its simple construction and compact storage, BRIEF has the lowest compute and storage requirements. Rublee et al. further proposed oriented FAST and rotated BRIEF (ORB) on the basis of BRIEF [28]. ORB overcomes the lack of rotation invariance of BRIEF. It computes a local orientation using intensity centroid, which is a weighted averaging of pixel intensities in the local patch assumed not be coincident with the center of the feature [26]. In addition, it also uses a learning method to obtain binary tests with lower correlation, so that the descriptor becomes more discriminative accordingly.

2.2 Proposed binary hybrid descriptor

Aerial orthophotos include not only texture information, but also color information. The buildings appeared in orthophotos is distinctive for their geometric features and color features. Hence, we build the descriptor on the premise of making full use of color information. In order to reduce the computational burden, the proposed BUT descriptor was formed by a set of bit tests.

2.2.1 Orientations

Inspired from ORB, we use a simple but effective way to compute the patch orientation, i.e., the intensity centroid [26]. The intensity centroid assumes that the intensity of a patch is offset from its center, and this vector can be treat as an orientation of the patch. Thus we can define the moments of a patch as follows

$$m_{pq} = \sum_{x,y} x^p y^q I(x,y), \tag{1}$$

and with these moments, we may find the centroid as

$$C = \left(\frac{m_{10}}{m_{00}}, \frac{m_{01}}{m_{00}} \right). \tag{2}$$

We can construct a vector from the center O, to the centroid C. Accordingly, the orientation of the patch is

$$\theta = \mathrm{atan2}(m_{01}, m_{10}), \tag{3}$$

where atan2 is the quadrant-aware version of arctan.

To improve the rotation invariance of this measure, we make sure that moments are computed with x and y remaining within a circular region of radius R. In our experiments, R is empirically set to 15, which is the same with the radius size of standard ORB.

2.2.2 Descriptor

The proposed descriptor is a hybrid bit string description of an image patch. It contains four sets of binary intensity tests. For a smoothed image patch **P**, the binary test τ_1, which describes texture feature, can be defined by

$$\tau_1(\mathbf{P}; a, b) = \begin{cases} 1 & \mathbf{P}_{\mathrm{L}}(a) < \mathbf{P}_{\mathrm{L}}(b) \\ 0 & \mathbf{P}_{\mathrm{L}}(a) \geq \mathbf{P}_{\mathrm{L}}(b) \end{cases}, \tag{4}$$

where $\mathbf{P}_{\mathrm{L}}(a)$ represents the intensity of point a in L channel of patch **P**. Another binary test τ_2, which describes similarity, can be defined by

$$\tau_2(\mathbf{P}; a, b) = \begin{cases} 1 & |\mathbf{P}_{\mathrm{L}}(a) - \mathbf{P}_{\mathrm{L}}(b)| < S \\ 0 & |\mathbf{P}_{\mathrm{L}}(a) - \mathbf{P}_{\mathrm{L}}(b)| \geq S \end{cases}, \tag{5}$$

where S is the similarity threshold. In our experiments, S is set to 5. The binary test τ_3 and τ_4, which describes color information, can be defined by

$$\tau_3(\mathbf{P}; a, b) = \begin{cases} 1 & (\mathbf{P}_{\mathrm{A}}(a) + \mathbf{P}_{\mathrm{A}}(b)) < C_{\mathrm{A}} \\ 0 & (\mathbf{P}_{\mathrm{A}}(a) + \mathbf{P}_{\mathrm{A}}(b)) \geq C_{\mathrm{A}} \end{cases}, \tag{6}$$

$$\tau_4(\mathbf{P}; a, b) = \begin{cases} 1 & (\mathbf{P}_{\mathrm{B}}(a) + \mathbf{P}_{\mathrm{B}}(b)) < C_{\mathrm{B}} \\ 0 & (\mathbf{P}_{\mathrm{B}}(a) + \mathbf{P}_{\mathrm{B}}(b)) \geq C_{\mathrm{B}} \end{cases}, \tag{7}$$

where $\mathbf{P}_{\mathrm{A}}(a)$ and $\mathbf{P}_{\mathrm{B}}(a)$ represent the intensity of point a in A channel and B channel of patch **P** respectively, and C_{A} and C_{B} represent the color threshold of A channel and that of B channel respectively. In our experiments, C_{A} and C_{B} are both set to 255.

Finally, the proposed hybrid feature of patch **P** can be defined as a vector of 4 × n binary tests

$$f_n(\mathbf{p}) = \sum_{0 \leq j < 4} \sum_{0 \leq i < n} 2^{n \times j + i} \tau_j(\mathbf{P}; a_i, b_i). \tag{8}$$

Here, the test pairs (a_i, b_i) are selected in accordance with the Gaussian distribution around the center of the patch, and n = 256.

2.3 Binary BoF classifier

Image classification is a process classifying images based on their features [7]. In the proposed building detection approach, bag-of-features (BoF) model is adopted to handle the classification task of intersections [29].

Owing to its simplicity, robustness and notable performance, BoF model is successfully applied to object and natural scene classification. Originally, the idea is derived from ‘bag of words’ representation for text categorization [32]. By extracting representative words in the training set of numerous sentences, a dictionary is formed and a meaningful sentence is substituted by the frequency of occurrence of the words in the dictionary, which is regarded as a ‘bag’. When a new sentence comes, it can be coded by the dictionary and classified into a specific category by computing its similarity with trained ‘bags’.

In order to predict the categories of the proposed binary descriptor, we realize a binary BoF classifier. The binary BoF model can be divided into four parts, including patches description, codebook and bags formation, distribution of codewords computation and test samples classification. Given a collection of binary features, codebook (or dictionary) is formed by performing k-means clustering algorithm, where k is the size of codebook. While in text retrieval the dictionary size emerges naturally from the training codebook and the subsequent dimensionality reduction techniques, in the visual case, how to choose the best dictionary size is an open research issue. In our work, the dictionary size is chosen empirically, and an explicit effort to keep it small, since large dictionary sizes have an important impact on processing time. Codewords are then defined as the centers of clusters. Thus descriptors in each training patch can be coded by hard assignment to the nearest codeword, yielding a histogram $n(w_i, d_j)$ counting the frequency of occurrence of each codeword, where w_i denotes the ith visual word in the k-size dictionary, and d_j is the jth category. The histogram is treated as a 'bag'. We use Bayesian approaches to model the distribution of codewords. The choice of a linear kernel is motivated by the fact that it normally suffices to keep high accuracy for binary codewords. After representing a test patch as a histogram, the most similar training histogram can be found by calculating Hamming distance, and corresponding category label of the patch is also returned.

3 Two-stage building detection approach

For the reason that the buildings appeared in orthophotos are of a variety of colors and shapes, man-made non-building structures, such as park trails and highways are usually similar to building roofs. In order to reduce the erroneous detection, we propose a coarse-to-fine framework, which contains two major stages. The flow chart of our building detection approach is shown in Fig. 1.

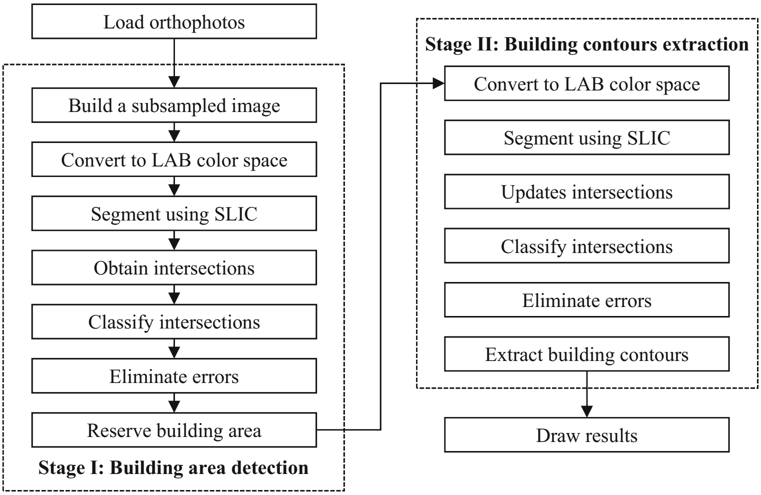
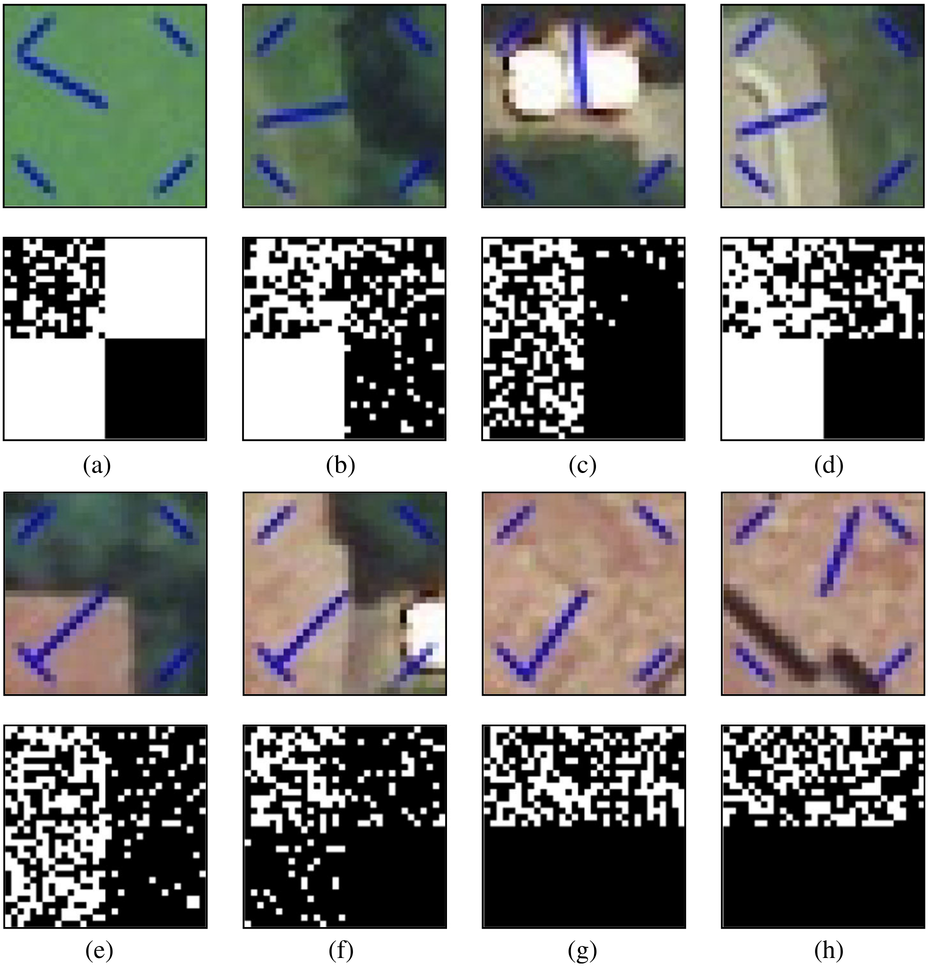


Fig. 1 The flow chart of the proposed building detection approach

3.1 Building area detection

The purpose of the first stage is to find the rough building area from a subsampled orthophoto. The original orthophoto is resized to a subsampled one with the pixel resolution of 1.0 m/pixel. We convert the subsampled orthophoto from RGB color space to LAB color space, in order to effectivelly make use of texture details and color channels. Then, the subsampled orthophoto is super segmented by SLIC, and obtain a sequence of intersections. Thus we can extract the local features of the intersections employing the proposed hybrid descriptor, and classify the intersections into two categories using the above-mentioned binary BoF classifier.

One category is non-building intersections, such as the intersections located in forest and road. The other category is building intersection, including the intersection located in roof and around building edge. Eventually, the approximate area of the building can be obtained according to the categories of the intersections, and the rough building area can be reserved by fitting a rectangle. Most erroneous classification can be discarded due to a constraint of minimum area pixels.


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

(e) (f) (g) (h)

Fig. 2 Original patch and the proposed binary descriptor

3.2 Building contours extraction

The basic modules in this stage is similar to the first stage. We convert the building area detected through stage I from RGB color space to LAB color space in the original orthophoto, and then update the sequence of intersections by SLIC likewise. In this stage, the intersections are classified into three categories. One category is the non-building intersections, another category is the intersections located on the building edge, and the other category is the intersections located in roof. Therefore, we can extract the building contours accurately according to the categories of the intersections.

4 Experiments

4.1 Performance of description

In this subsection, we mainly focus on evaluating the performance of the proposed descriptor. We have built a test library with 10 large-scale aerial orthophotos, in which various types of buildings are included. The spatial resolution of these aerial orthophotos is 0.2 m per pixel.

All orthophotos are segmented using SLIC, in which the density of seeds is 1600 pixels. Therefore, we can obtain thousands upon thousands intersections of superpixels. In each test, we randomly select 500 intersections, and generate patches around the centers of the selected intersections. Each patch is at first described using SURF and the proposed binary descriptor.

Figure 2 shows the original patches and the corresponding binary descriptors. The size of each original patch is 31 pixels × 31 pixels. The orientation of the proposed binary descriptor is computed by intensity centroid method, and is drawn is the patch. For the proposed binary descriptor consists of four 256-bit binary string, the descriptor can be depicted as a 32 pixels × 32 pixels binary image with four 16 pixels × 16 pixels blocks. In Fig. 2, (a) to (d) are obtained from non-building intersections, (e) to (f) are obtained from building edge, and (g) to (h) are obtained from building roof.

Figure 3 shows the Recall- (1-Precision) curve, in which asterisk (*) stands for the proposed method, and plus (+) stands for the SURF + BoF. Namely, SURF descriptor

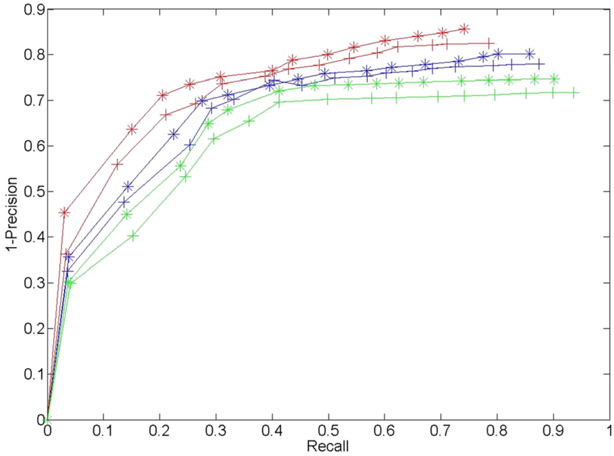


Fig. 3 Performance evaluation through Recall- (1-Precision)

is classified by a classifier which is pre-trained using a standard BoF, and the proposed descriptor is classified by the binary BoF classifier. Red curve shows the accuracy of building roof intersections, blue curve shows the accuracy of non-building intersections, and green curve shows the accuracy of building edge intersections. Experiments show that the classification accuracy of the proposed descriptor outperforms SURF + BoF in most cases.

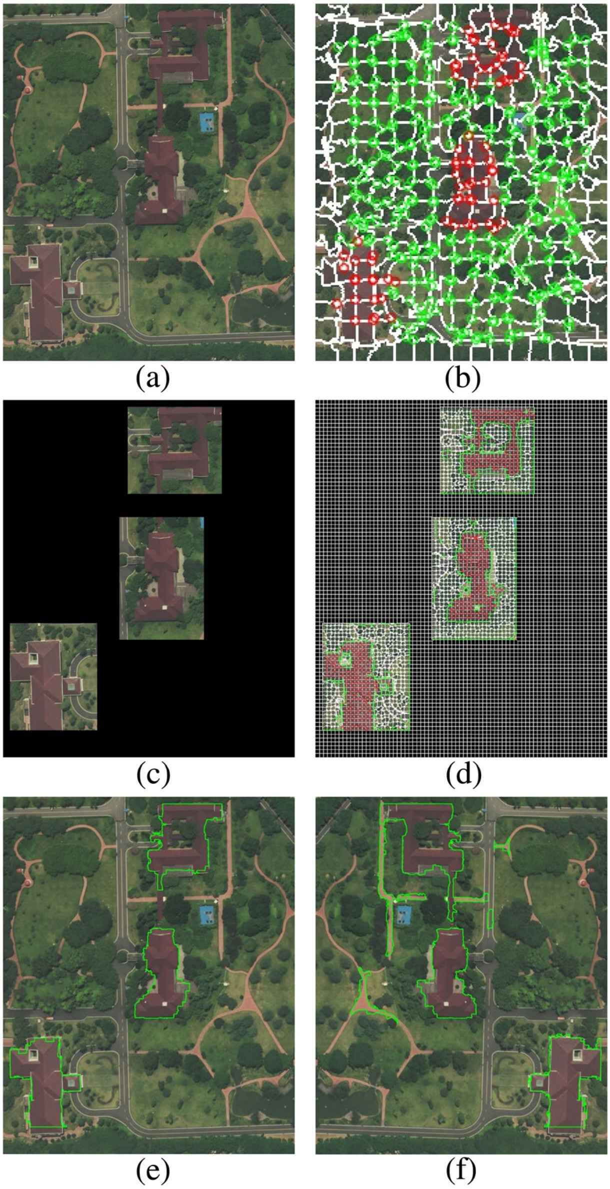


Fig. 4 The major stages of the proposed approach

4.2 Performance of building detection

Figure 4 shows the major stages of the proposed building detection approach and SURF + BoF. In Fig. 4, (a) is the original orthophoto, (b) is the classification results in stage I, (c) is the detected building area in stage I, (d) is the classification results in stage II, (e) is the extracted building contours, and (f) is building contours extracted using standard SURF + BoF. The experiments show that the contours extracted using the proposed approach is more reliable than SURF + BoF.

Figure 5 shows the building contours extracted using the proposed approach on two orthophotos. Figure 6 shows the Intersection over Union (IoU) test results on another three orthophotos. The Area of Overlap is the area of overlap between the extracted building contours and the ground-truth building contours, and the Area of Union is the area encompassed by both the extracted building contours and the ground-truth building contours. In Fig. 6, the Area of Overlap is covered by a green mask, and the Area of Union is covered by a red mask.

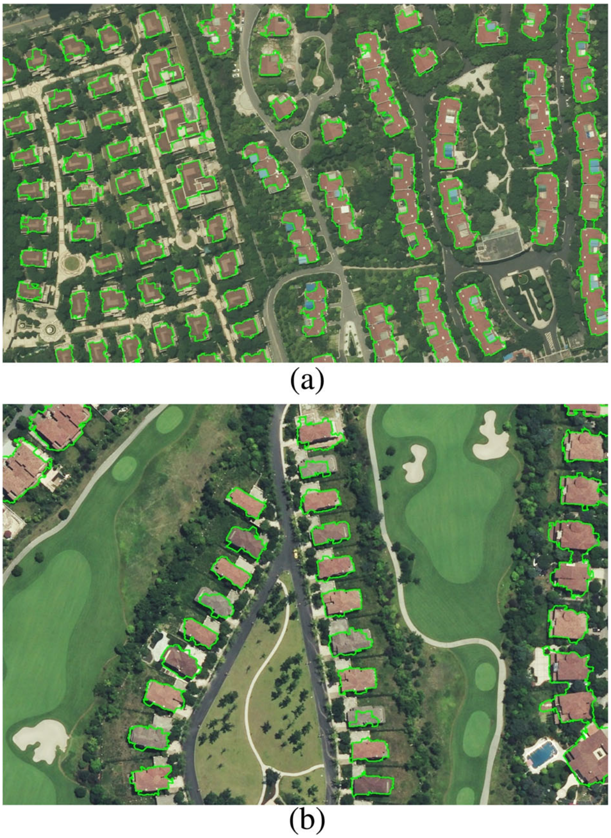


Fig. 5 The building contours extracted using the proposed approach

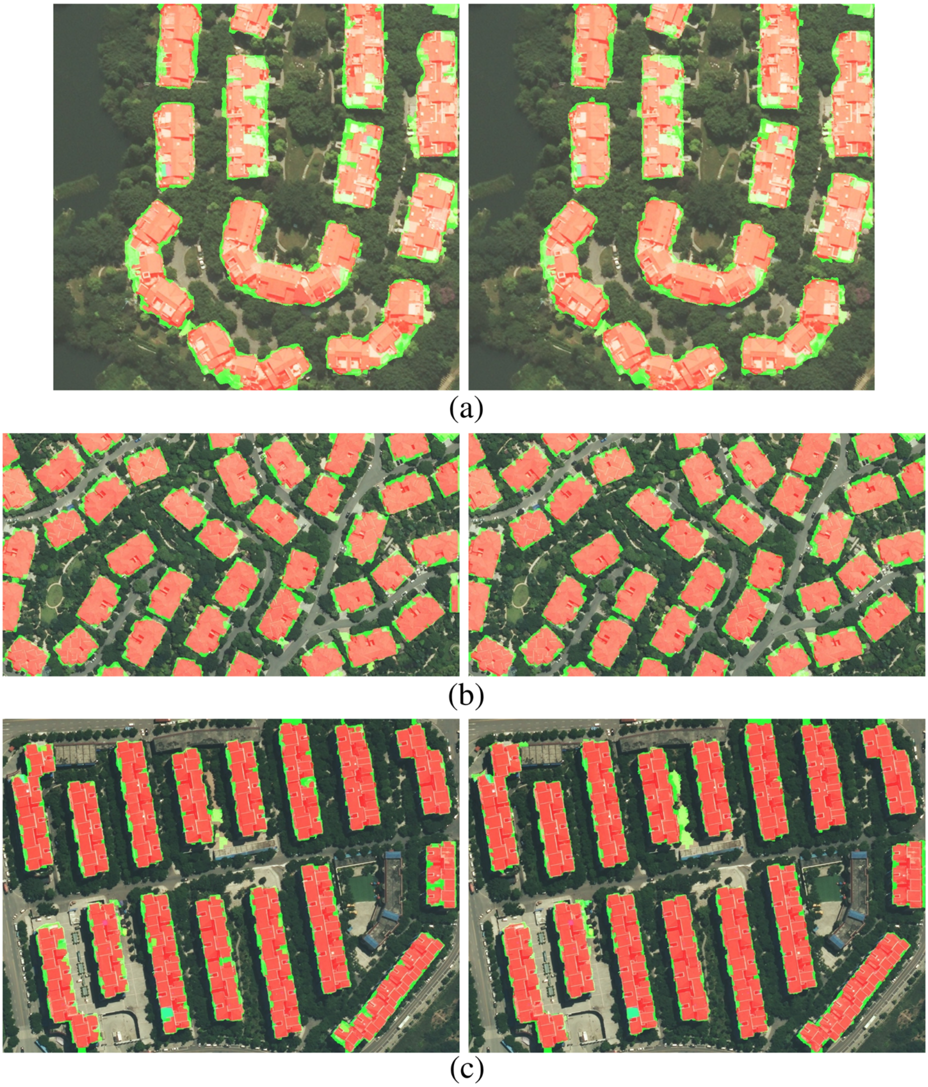


Fig. 6 Comparison of the Intersection over Union test

IoU score can be computed via

$$IoU = \frac{Area\ of\ Overlap}{Area\ of\ Union}. \tag{9}$$

In Table 1, compared the IoU scores and time cost between the proposed approach and SURF + BoF on the original orthophotos of Figs. 4 and 5. The experiments show that the average IoU score of the proposed approach is 2.14% higher than SURF + BoF, however the time cost of the proposed approach takes only 52.6% of SURF + BoF.

Table 1 Comparison of the Intersection over Union score and time cost

Orthophotos No.		1#	2#	3#	4#	5#	Average
IoU (%)	Proposed method	85.72	80.36	77.65	82.24	87.20	82.63
	SURF + BoF	82.85	77.13	76.02	81.89	84.57	80.49
Time cost (s)	Proposed method	1.58	1.27	0.70	1.16	1.29	1.20
	SURF + BoF	3.11	2.35	1.34	2.32	2.28	2.28

5 Conclusions

Benefiting from binary description and making full use of texture details and color channels, the proposed descriptor is not only computationally frugal, but also accurate. Accordingly, the proposed two-stage building detection approach achieves desirable results. In future, deep learning technology will be employed to improve the accuracy of building area detection and to considerably reduce the time cost.

Acknowledgements This research was supported by the Key Laboratory for Earth Observation, National Administration of Surveying, Mapping and Geoinformation of China (K2015009).

References

1. Achanta R, Shaji A, Smith K, Lucchi A, Fua P, Susstrunk S (2012) SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans Pattern Anal Mach Intell 34(11):2274–2282
2. Alahi A, Ortiz R, Vandergheynst P (2012) FREAK: fast retina keypoint. In: IEEE Conference on computer vision and pattern recognition, Providence, USA, 16-21 Jun, pp 510–517
3. Bay H, Ess A, Tuytelaars T, Van Gool L (2008) Speeded-up robust features (SURF). Comput Vis Image Underst 110(3):346–359
4. Bruzzone L, Demir B (2014) A review of modern approaches to classification of remote sensing data. In: Manakos I, Braun M (eds) Land use and land cover mapping in Europe: practices and trends. Springer, Dordrecht, pp 127–143
5. Calonder M, Lepetit V, Ozuysal M, Trzcinski T, Strecha C, Fua P (2012) BRIEF: computing a local binary descriptor very fast. IEEE Trans Pattern Anal Mach Intell 34(7):1281–1298
6. Chang X, Nie F, Yang Y, Huang H (2014) A convex formulation for semi-supervised multi-label feature selection. In: AAAI Conference on artificial intelligence, Quebec City, Canada, 27-31 Jul, pp 1171–1177
7. Chang X, Nie F, Wang S, Yang Y, Zhou X, Zhang C (2016) Compound rank-k projections for bilinear analysis. IEEE Trans Neural Netw Learn Syst 27(7):1502–1513
8. Chang X, Ma Z, Yang Y, Zeng Z, Hauptmann AG (2017) Bi-level semantic representation analysis for multimedia event detection. IEEE Trans Cybern 47(5):1180–1197
9. Chang X, Yu YL, Yang Y, Xing EP (2017) Semantic pooling for complex event analysis in untrimmed videos. IEEE Trans Pattern Anal Mach Intell 39(8):1617–1632
10. Chang X, Ma Z, Lin M, Yang Y, Hauptmann AG (2017) Feature interaction augmented sparse learning for fast Kinect motion detection. IEEE Trans Image Process 26(8):3911–3920
11. Dornaika F, Moujahid A, Merabet M, Ruichek Y (2016) Building detection from orthophotos using a machine learning approach: an empirical study on image segmentation and descriptors. Expert Syst Appl 58:130–142
12. Ghamisi P, Dalla M, Benediktsson JA (2015) A survey on spectral spatial classification techniques based on attribute proles. IEEE Trans Geosci Remote Sens 53(5):2335–2353
13. Hu MK (1962) Visual pattern recognition by moment invariants. IRE Trans Inf Theory 8(2):179–187
14. Khurana M, Wadhwa V (2015) Automatic building detection using modified grab cut algorithm from high resolution satellite image. Int J Adv Res Comput Commun Eng 4(8):158–164
15. Lecun Y, Bengio Y, Hinton GE (2015) Deep learning. Nature 521:436–444

16. Leutenegger S, Chli M, Siegwart RY (2011) BRISK: binary robust invariant scalable keypoints. In: IEEE International conference on computer vision, Barcelona, Spain, 6-13 Nov, pp 2548–2555
17. Li Z, Gong W, Nee AYC, Ong SK (2009) Region-restricted rapid keypoint registration. Opt Express 17(24): 22096–22101
18. Li Z, Gong W, Nee AYC, Ong SK (2009) The effectiveness of detector combinations. Opt Express 17(9): 7407–7418
19. Liu L, Wiliem A, Chen S, Lovell BC (2014) Automatic image attribute selection for zero-shot learning of object categories. In: International conference on pattern recognition, Stockholm, Sweden, 24-28 Aug, pp 2619–2624
20. Liu L, Nie F, Zhang T, Wiliem A, Lovell BC (2016) Unsupervised automatic attribute discovery method via multi-graph clustering. In: International conference on pattern recognition, Cancun, Mexico, 4-8 Dec, pp 1713–1718
21. Liu L, Wiliem A, Chen S, Lovell BC (2017) What is the best way for extracting meaningful attributes from pictures? Pattern Recogn 64:314–326
22. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
23. Ojala T, Pietikainen M, Harwood D (1996) Comparative study of texture measures with classification based on feature distributions. Pattern Recogn 29(1):51–59
24. Ojala T, Pietikainen M, Maenpaa T (2002) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans Pattern Anal Mach Intell 24(7):971–987
25. Passino KM (2002) Biomimicry of bacterial foraging for distributed optimization and control. IEEE Control Syst Mag 22(3):52–67
26. Rosin PL (1999) Measuring corner properties. Comput Vis Image Underst 73(2):291–307
27. Rother C, Kolmogorov V, Blake A (2004) Grabcut: interactive foreground extraction using iterated graph cuts. ACM Trans Graph 23(3):309–314
28. Rublee E, Rabaud V, Konolige K, Bradski G (2011) ORB: an efficient alternative to SIFT or SURF. In: IEEE International conference on computer vision, Barcelona, Spain, 6-13 Nov, pp 2564–2571
29. Sivic J, Zisserman A (2009) Efficient visual search of videos cast as text retrieval. IEEE Trans Pattern Anal Mach Intell 31(4):591–606
30. Vakalopoulou M, Karantzalos K, Komodakis N, Paragios N (2015) Building detection in very high resolution multispectral data with deep learning features. In: IEEE International conference on geoscience and remote sensing symposium, Milan, Italy, 26-31 Jul, pp 1873–1876
31. Volpi M, Tuia D (2017) Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. IEEE Trans Geosci Remote Sens 55(2):881–893
32. Yang F, Lu H, Zhang W, Yang G (2012) Visual tracking via bag of features. IET Image Process 6(2):115–128
33. Zhang T, Liu L, Wiliem A, Lovell B (2016) Is Alice chasing or being chased? : determining subject and object of activities in videos. In: IEEE Winter conference on applications of computer vision, Lake Placid, USA, 7-9 Mar, pp 1–7

Yan Hu is pursuing her PhD from Wuhan University, China. She is also a senior engineer in Chongqing Geomatics Center, China. Her research interest is photogrammetry and remote sensing.