COMPACT AND HYBRID FEATURE DESCRIPTION FOR BUILDING EXTRACTION

Z. Li [a, b, *], Y. Liu [a], Y. Hu [c], P. Li [c], Y. Ding [c]

[a] Key Laboratory for Optoelectronic Technology and Systems of Ministry of Education, College of Optoelectronic Engineering, Chongqing University, 400044 Chongqing, China - (lizhenghao, 20140813038t)@cqu.edu.cn
[b] Chongqing Academy of Science and Technology, 401123 Chongqing, China - lizhenghao@cqu.edu.cn
[c] Chongqing Geomatics Center, 401121 Chongqing, China - (huyan_dl023, 20110802080, 20160801008)@cqu.edu.cn

Commission II, WG II/6

KEY WORDS: Building Extraction, Machine Learning, Local Feature, Descriptor, Binary Uniformity Tests, Binary Random Trees, Superpixel segmentation

ABSTRACT:

Building extraction in aerial orthophotos is crucial for various applications. Currently, deep learning has been shown to be successful in addressing building extraction with high accuracy and high robustness. However, quite a large number of samples is required in training a classifier when using deep learning model. In order to realize accurate and semi-interactive labelling, the performance of feature description is crucial, as it has significant effect on the accuracy of classification. In this paper, we bring forward a compact and hybrid feature description method, in order to guarantees desirable classification accuracy of the corners on the building roof contours. The proposed descriptor is a hybrid description of an image patch constructed from 4 sets of binary intensity tests. Experiments show that benefiting from binary description and making full use of color channels, this descriptor is not only computationally frugal, but also accurate than SURF for building extraction.

1. INTRODUCTION

Building extraction in aerial orthophotos is crucial for various applications, including urban planning, real-estate management, and disaster relief (Dornaika et al., 2016). Thus, building extraction in remote sensing, specifically in high-resolution aerial orthophotos, has been a popular research topic.

Most building extraction methods usually rely on image segmentation, such as 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 (Rother et al., 2004). For the reason that the building detection applications using grab-cut need operating by human experts, these applications work both slow and costly. Some researchers improved grab-cut by iterative optimization algorithms, such as the bio inspired bacterial foraging optimization (BFO) (Passino, 2002), to realize automatic building detection. However, experiments showed that in several test cases, some road segments and bridges were mistakenly detected as buildings currently (Khurana et al., 2015).

Though building extraction can be treated as a typical segmentation task, it can also be treated as a recognition task. In recent years, machine learning approaches, especially deep learning (also known as deep structured learning or hierarchical learning), are evolving exponentially (Lecun et al., 2015). There has been a significant amount of past work on classifcation and segmentation of remote sensing imagery using machine learning technology. For a recent review, please refer to see (Bruzzone et al., 2014) and (Ghamisi et al., 2015).

Deep learning is a special type of machine learning that involves a deeper level of automation. One of the great challenges of machine learning is feature extraction where the user needs to tell the algorithm what kinds of things it should be looking for, in order to make a decision and just feeding the algorithm with raw data is rarely effective. The algorithm's effectiveness relies heavily on the skill of the user. Deep learning models address this problem as they are capable of learning to focus on the right features by themselves and requires little guidance from the user, making the analysis better than what humans can do.

Currently, deep learning has been shown to be successful in addressing building extraction with high accuracy and high robustness. (Vakalopoulou et al., 2015) propose a supervised building detection procedure based on the ImageNet framework, while integrating certain spectral information by employing multispectral band combinations into the training procedure. The building detection was addressed through a binary classification procedure based on support vector machine (SVM) classifier. The experimental results and the performed quantitative validation indicate the quite promising potentials of this approaches. Making use of elevation data such as a digital surface model (DSM), (Volpi et al., 2017) propose a hybrid network that combines the pre-trained image features with DSM features that are trained from scratch. The hybrid network improves the labelling accuracy on the highest-resolution imagery.

It should be noted that quite a large number of samples is required in training a classifier when using deep learning model. Therefore, many researchers attempt find an effective way to build the training library automatically or semi-interactively. The core to realize accurate and semi-interactive labeling is to detect the candidate corners on the building roof contours automatically.

In this paper, we bring forward a compact and hybrid feature description method, in order to guarantees desirable classification accuracy of the corners on the building roof contours. The proposed descriptor is designated as BUT for the acronym of 'Binary Uniformity Tests'. Each 'Test' includes texture and similarity comparison in L channel of LAB color space, and color comparisons in A channel and B channel of LAB color space. Benefiting from binary description and making full use of color channels, the BUT descriptor is not only computationally frugal, but also accurate.

The remainder of this paper is organized as follows. Section 2 gives a brief introduction to local feature descriptors, especially binary descriptors. Section 3 introduces the implementation details of the proposed BUT descriptor. In section 4, experimental results are presented. Lastly in section 5, conclusions are presented.

2. LOCAL FEATURE DESCRIPTORS

2.1 Floating-Point Descriptors

Scale-invariant feature transform (SIFT) is a *de facto* standard for local feature description, because of its excellent performance, which is invariant to a variety of common image transformations (Lowe, 2004). Speeded up robust features (SURF) is another commonly used method performing approximately as well as SIFT with lower computational cost (Bay et al., 2008). 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 further reduce the computational cost (Li et al., 2009a). 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 (Li et al., 2009b).

2.2 Binary Descriptors

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

Some researchers try to increase the robustness of classification by improving the sampling pattern for 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 (Leutenegger et al., 2011). Fast retina keypoint (FREAK) is another typical one leveraging a novel retina sampling pattern inspired by the human visual system (Alahi et al., 2012).

Binary robust independent elementary features (BRIEF) is a representative example which directly computes the descriptor bit-stream quite fast, based on simple intensity difference tests in a smoothed patch (Calonder et al., 2012). (Rublee et al., 2011) further proposed oriented FAST and rotated BRIEF (ORB) on the basis of BRIEF. The approach computes the orientation of the keypoints utilizing intensity centroid (Rosin, 1999); thus, the descriptor is rotation-invariant and scale-invariant. In addition, it also uses a learning method to obtain binary tests with lower correlation, so that the descriptor becomes more discriminative accordingly.

3. PROPOSED METHOD

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 built 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.

3.1 Orientations

Inspired from ORB, our method uses a simple but effective way to compute the patch orientation, i.e., the intensity centroid (Rosin, 1999). The intensity centroid assumes that the intensity of a patch is offset from its center, and this vector can be treat as an orientation. 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 patch's center O, to the centroid C. 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 T. In our experiments, T is empirically set to 15, which is the same with the radius size of standard ORB.

3.2 BUT Descriptors

The BUT descriptor is a hybrid bit string description of an image patch constructed from 4 sets of binary intensity tests. Consider a smoothed image patch **P**.

The binary test τ_1, which describes texture feature, can be defined by

$$\tau_1(\mathbf{P}; x, y) = \begin{cases} 1 & \mathbf{P}_L(x) < \mathbf{P}_L(y) \\ 0 & \mathbf{P}_L(x) \geq \mathbf{P}_L(y) \end{cases}, \tag{4}$$

where $P_L(x)$ is the intensity of point x in L channel of patch **P**. So, the binary test τ_2, which describes similarity, can be defined by

$$\tau_2(\mathbf{P}; x, y) = \begin{cases} 1 & |\mathbf{P}_L(x) - \mathbf{P}_L(y)| < s \\ 0 & |\mathbf{P}_L(x) - \mathbf{P}_L(y)| \geq s \end{cases}, \tag{5}$$

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

$$\tau_3(\mathbf{P}; x, y) = \begin{cases} 1 & (\mathbf{P}_A(x) + \mathbf{P}_A(y)) < c_A \\ 0 & (\mathbf{P}_A(x) + \mathbf{P}_A(y)) \geq c_A \end{cases}, \quad (6)$$

$$\tau_4(\mathbf{P}; x, y) = \begin{cases} 1 & (\mathbf{P}_B(x) + \mathbf{P}_B(y)) < c_B \\ 0 & (\mathbf{P}_B(x) + \mathbf{P}_B(y)) \geq c_B \end{cases}, \quad (7)$$

where $P_A(x)$ is the intensity of point x in A channel of patch **P**, $P_B(x)$ is the intensity of point x in B channel of patch **P**, c_A is the color threshold of A channel, and c_2 is the color threshold of B channel. In our experiments, c_A and c_B are both set to 255.

The feature is defined as a vector of $4 \times n$ binary tests

$$f_n(\mathbf{P}) = \sum_{1 \leq i \leq n} 2^{i-1} \tau(\mathbf{P}; x_i, y_i). \quad (8)$$

Here we use a Gaussian distribution around the center of the patch, and choose $n = 256$.

4. EXPERIMENTS

4.1 Test Library

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 0.2m per pixel.

All aerial orthophotos are segmented using the simple linear iterative clustering (SLIC) method (Achanta et al., 2012). Therefore, we can obtain thousands upon thousands intersections of superpixels. Figure 1 shows the an original input orthophoto and Figure 2 shows the orthophoto after SLIC superpixel segmentation. The density of seeds is 1600 pixels.

Figure 1. Original input orthophoto

Figure 2. Orthophoto after SLIC superpixel segmentation

In each test, we randomly select 500 intersections, and generate patches with the center of the selected intersections. These patches are to be used for classification to evaluate the performance of the descriptor.

4.2 Experimental Results

Each patch is at first described using SURF and BUT. Figure 3 shows the original patch and the corresponding BUT descriptor. We also draw the orientation in the original patch. Then SURF descriptor is classified by a classifier which is pre-trained using a standard random forest (Ho, 1998), and BUT descriptor is classified by a binary classifier which is pre-trained using a binary version of random forest.

Experiments show that the classification accuracy of BUT can achieve 97.4%, which is 4.6% higher than SURF. Moreover, the time cost of classification using BUT is only 62.6% of SURF.

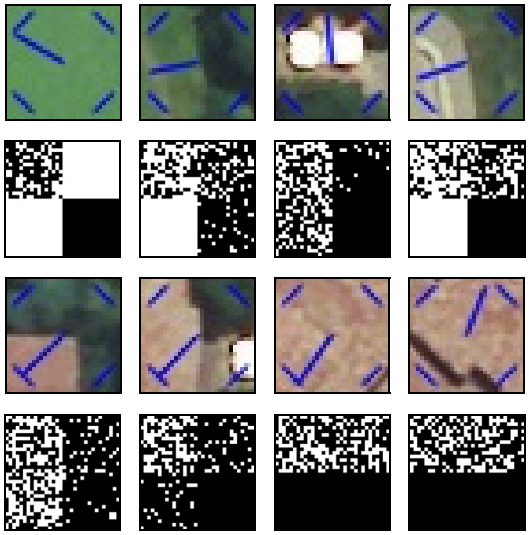

Figure 3. Original patch and BUT descriptor

5. CONCLUSIONS

Benefiting on the superiority of binary description and fully use of color information, the proposed hybrid BUT descriptor is compact but distinctive. In future, we will develop a semi-automated tool based on the proposed descriptor to facilitate the labelling stage for those kind of approaches based on deep learning model.

ACKNOWLEDGEMENTS

This research was supported by the Key Laboratory for Earth Observation, National Administration of Surveying, Mapping and Geoinformation of China (K2015009), the Chongqing Postdoctoral Science Foundation (Xm2015014), and the Opening Fund of Key Laboratory of Inland Waterway Regulation Engineering, Ministry of Communications (NHHD-201503). The test library is offered by the Chongqing Geomatics Center.

REFERENCES

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

Alahi, A.; Ortiz, R.; Vandergheynst, P., 2012. FREAK: Fast retina keypoint. *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 510-517.

Bay, H.; Ess, A.; Tuytelaars, T.; Van Gool, L, 2008. Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 110(3), pp. 346-359.

Bruzzone, L., Demir, B., 2014. *Land use and land cover mapping in Europe: Practices and trends*. Springer, Dordrecht, pp. 127-143.

Calonder, M.; Lepetit, V.; Ozuysal, M.; Trzcinski, T.; Strecha, C.; Fua, P., 2012. BRIEF: Computing a local binary descriptor very fast. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(7), 1281-1298.

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 Systems With Applications*, 58(2016), pp. 130-142.

Ghamisi, P., Dalla, M., Benediktsson, J., A., 2015. A survey on spectral spatial classification techniques based on attribute proles. *IEEE Transactions on Geoscience and Remote Sensing*, 53(5), pp. 2335-2353.

Ho, T., K., 1998. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20 (8), pp. 832-844.

Khurana, M., Wadhwa, V., 2015. Automatic building detection using modified grab cut algorithm from high resolution satellite image. *International Journal of Advanced Research in Computer and Communication Engineering*, 4(8), pp. 158-164.

Lecun, Y., Bengio, Y., Hinton, G., E., 2015. Deep learning. *Nature*, 521, pp. 436-444.

Leutenegger, S.; Chli, M.; Siegwart, R., Y., 2011. BRISK: Binary robust invariant scalable keypoints. *Proceedings of the 2011 IEEE International Conference on Computer Vision*, pp. 2548-2555.

Li, Z.; Gong, W.; Nee, A., Y., C.; Ong, S., K., 2009a. Region-restricted rapid keypoint registration. *Optical Express*, 17(24), pp. 22096-22101.

Li, Z.; Gong, W.; Nee, A., Y., C.; Ong, S., K., 2009b. The effectiveness of detector combinations. *Optical Express*, 17(9), 7407-7418.

Lowe, D., G., 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2), pp. 91-110.

Passino, K., M., 2002. Biomimicry of bacterial foraging for distributed optimization and control. *IEEE Control Systems Magazine*, 22(3), pp. 52-67.

Rosin, P., L., 1999. Measuring corner properties. *Computer Vision and Image Understanding*, 73(2), pp. 291-307.

Rother, C., Kolmogorov, V., Blake, A., 2004. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics*, 23(3), pp. 309-314.

Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G., 2011. ORB: An efficient alternative to SIFT or SURF. *Proceedings of the 2011 IEEE International Conference on Computer Vision*, pp. 2564-2571.

Vakalopoulou, M., Karantzalos, K., Komodakis, N., Paragios, N., 2015. Building detection in very high resolution multispectral data with deep learning features. *Proceedings of IEEE International Geoscience and Remote Sensing Symposium*, pp. 1873-1876.

Volpi, M., Tuia, D., 2017. Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 55(2), pp. 881-893.