LETTER

Utilizing Shape-Based Feature and Discriminative Learning for Building Detection

Shangqi ZHANG[†], Haihong SHEN[†a)], *Nonmembers*, *and* Chunlei HUO[††], *Member*

SUMMARY Building detection from high resolution remote sensing images is challenging due to the high intraclass variability and the difficulty in describing buildings. To address the above difficulties, a novel approach is proposed based on the combination of shape-specific feature extraction and discriminative feature classification. Shape-specific feature can capture complex shapes and structures of buildings. Discriminative feature classification is effective in reflecting similarities among buildings and differences between buildings and backgrounds. Experiments demonstrate the effectiveness of the proposed approach.

key words: *building detection, shape-specific feature, metric learning, feature classification*

1. Introduction

Building detection from remote sensing images has numerous practical applications, *e.g.*, urban planning and disaster management, *etc*. Compared with low-to-medium resolution images, high resolution remote sensing images are more promising for practical applications since the improved spatial resolution provides more details about buildings. However, detecting buildings from high resolution remote sensing images is more challenging. The reasons are summarized as follows:

First, due to the complexities of buildings in shapes, structures and materials, uniformly defining "buildings" is difficult. In the literature, feature-based approaches are widely used for building detection. Sirmacek *et al.* [1] extracted SIFT feature to detect buildings from high resolution satellite images. It performs well in building locations detection, but fails to extract building boundaries accurately. The reason is illustrated in Fig. 1. For building SIFT descriptor, a circle is constructed at the boundary without considering building shapes, which indicates points C and F are similar. Using the obtained SIFT feature for building detection, the two points will be classified as "buildings" or "non-buildings" simultaneously. Moreover, this wrong result can lead to inaccurate building boundaries. For this reason, it is necessary to capture building shapes and structures for feature description.

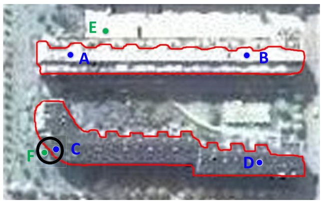

Fig. 1 Illustration of intraclass dissimilarities among buildings and interclass similarities between buildings and backgrounds. Blue Points: buildings. Green Points: backgrounds.

Second, the spectral resolution is not improved simultaneously with the spatial resolution. It is because the employment of sensors with the improved spatial resolution simplifies the problem of mixed pixels, but increases the internal spectral variability (intraclass variability) of the same class and decreases the spectral variability between different classes (interclass variability). As illustrated in Fig. 1, points A, B, C and D indicate buildings, while points E and F indicate roads. We can see the differences between buildings are even higher than that between buildings and backgrounds. Traditional method [2] only utilizes a simple classifier (i.e., SVM) for feature classification. Owing to the low interclass discrimination, points B and E will be classified into the same group. Therefore, measuring the similarities or dissimilarities between two data points to improve the separability is indispensable.

As for the first difficulty, several state-of-the-art approaches considered shapes and structures for building detection [3], [4]. Compared to the first difficulty, the second one is rarely solved. In this paper, we propose a novel approach aiming at addressing the above two difficulties. The novelty of the proposed method lies in the combination of the following two aspects:

1. Shape-specific feature is extracted using the method proposed by Khan *et al.* [5] for capturing complex shapes and structures of buildings.
2. Based on discriminative feature classification, similarities between buildings and differences between buildings and backgrounds are acquired to improve the overall separation.

Manuscript received June 26, 2016.
Manuscript revised November 2, 2016.
Manuscript publicized November 18, 2016.

[†]The authors are with the School of Information Engineering, China University of Geosciences (Beijing), Beijing 100083, China.
[††]The author is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China.
a) E-mail: haihongshen11@163.com (Corresponding author)

DOI: 10.1587/transinf.2016EDL8138

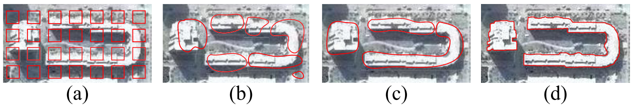

Fig. 2 Segmentation evolution of shape-specific feature with piece-wise smooth Mumford-Shah model. (a) iteration=1; (b) iteration=50; (c) iteration=80; (d) convergence.

2. Feature Extraction and Discriminative Classification

2.1 Shape-Specific Feature Extraction

The feature should be representative to capture complex shapes and structures of buildings. For this reason, this paper employs the shape-specific feature (SSF) [5], which is tolerant to multicolor rooftops, sensitive to various shapes, and efficient in computation. In detail, SSF extraction consists of the following three steps. For approximating the real shapes, alternative iterations between step 2) and step 3) are performed until convergence. Figure 2 shows the iterative process.

1) **Feature Initialization**. The existing local features aggregate oriented gradients across the textured regions. This leads to ambiguity in detecting building shapes and boundaries. To address this problem, this step is to define an initial shape-specific feature which gathers spectral features only from homogeneous regions. For convenience, we let Ω be the domain of the image, and $R \subset \Omega$ is the initialized region (*e.g.* the region within the red boundaries in Fig. 2 (a)). The feature $\mathbf{S}$ is constructed by gathering spectral features and oriented gradients in the neighborhood of each pixel x inside R, where $\mathbf{S} = (\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_Q)$, $\mathbf{s}_x$ $(1 \leq x \leq Q)$ is the feature at pixel x, Q indicates the size of the image (pixel numbers). For each feature $\mathbf{s}_x$, it is a normalized vector consisting of N $(N = m{\times}n, m, n \geq 1)$ components, where $\mathbf{s}_x = (s_{11}, \ldots, s_{1n}, \ldots, s_{m1}, \ldots, s_{mn})^T$. Each feature component s_{ij} $(1 \leq i \leq m, 1 \leq j \leq n)$ is defined as follows:

$$\begin{cases} s_{ij}(x) - \alpha_i \Delta s_{ij}(x) = P_j(x) & x \in R \\ \nabla s_{ij}(x) \cdot B = 0 & x \in \partial R \end{cases} \tag{1}$$

where Δ denotes the Laplacian operator, ∇ denotes the gradient operator, B is the unit outward normal to the boundary of R (∂R), $\alpha_i > 0$ denotes the scale of neighborhoods in feature computation, and $P_j(\cdot)$ is the point-wise function, which includes oriented gradients of the image.

The above problem can be solved by simply using the scale-space [6] defined by the PDE, which is the minimizer of the following function:

$$E(s) = \int_R (P_j(x) - s(x))^2 dx + \alpha_i \int_R |\nabla s(x)|^2 \, dx \tag{2}$$

Therefore, s_{ij} is a smoothing of P_j and α_i controls the degree of smoothing. Using the scale-space defined by PDE, feature S can finally be computed.

At the first iteration ($\tau = 1$, where τ indicates the number of iterations), we get the initial feature $\mathbf{S}_1 = (\mathbf{s}_{1,1}, \mathbf{s}_{1,2}, \ldots, \mathbf{s}_{1,Q})$ by (1). The initial feature $\mathbf{S}_1$ is difficult to capture shapes of buildings exactly. In order to capture the real shapes and structures, we use an iterative strategy to refine $\mathbf{S}$. First, shape-specific feature is incorporated into Mumford-Shah energy [7] to make segmentations. Then, based on the new region, feature $\mathbf{S}$ is updated by (1). The refinement process is implemented by iterating the following two steps.

2) **Region Updating**. Region updating is to verify whether segmentations are close to the real shapes of buildings. By incorporating feature $\mathbf{S}_\tau$ extracted at the τth iteration into Mumford-Shah energy, we get renewed energy E_τ for segmentations and updated region R_τ. By aggregating data from the new region R_τ, updated feature $\mathbf{S}_{\tau+1}$ is computed in the next step.

3) **Feature Updating**. Feature updating is to capture the real shapes and structures of buildings progressively. To this end, with the renewed region R_τ, shape-specific feature is reconstructed using (1). For example, at the $(\tau+1)$th iteration, feature $\mathbf{S}_{\tau+1} = (\mathbf{s}_{\tau+1,1}, \mathbf{s}_{\tau+1,2}, \ldots, \mathbf{s}_{\tau+1,Q})$ is computed and can be used for further region updating in step 2).

With the increase of iteration, the segmentation results are more close to the real shapes and structures of buildings [Fig. 2 (b), (c)]. In each iteration, we use gradient decent to update the Mumford-shah energy E, so the algorithm will always converge to a local minima depending on the initialization used for segmentation. When the convergence is reached [Fig. 2 (d)], we get final shape-specific feature $\mathbf{S}_c = (\mathbf{s}_{c,1}, \mathbf{s}_{c,2}, \ldots, \mathbf{s}_{c,Q})$ and optimized segmentation results. The final feature $\mathbf{S}_c$ is a $(N \times Q)$ matrix with N feature dimension and is applied to the next subsection for feature classification.

The extraction of feature $\mathbf{S}_c$ is implemented by traversing all the pixels and all the feature dimensions,
so the computation complexity is: $O(N \times Q)$. Using an 12 core processor, it costs about 18 seconds until convergence.

2.2 Discriminative Feature Classification

Due to the low spectral resolution of remote sensing images, many state-of-the-art classification approaches fail to reflect similarities among buildings and differences between buildings and backgrounds. The discriminative feature classification (DFC) scheme is proposed to address the aforementioned difficulty. In this scheme, discriminative metric learning [8] is adopted to learn a distance metric. Based on this metric, we can acquire similarity or dissimilarity between each pair of two data points to improve the k-NN classification. Based on the improved classification, buildings can be detected with high accuracy. In detail, DFC consists of the following three steps.

1) **Couples Construction**. In DFC scheme, discriminative metric learning usually considers a set of constraints imposed on the couples of training data. Therefore, in this step, couples should be constructed firstly using train-

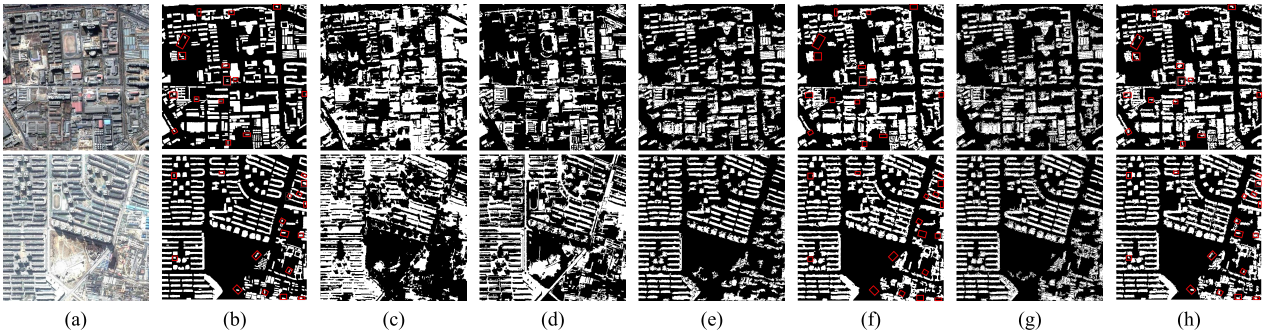

Fig. 3 Visual comparison on dataset 1 (top) and dataset 2 (bottom). (a) original image; (b) ground truth; (c) Feature Analyst; (d) eCognition; (e) DAISY+SVM; (f) SSF+SVM; (g) DAISY+DFC; (h) SSF+DFC. Some buildings are marked by red rectangular boxes.

ing data. For each training sample $\mathbf{s}_x$, $(v_1 + v_2)$ couples are constructed by selecting v_1 nearest similar neighbors and v_2 dissimilar neighbors from training data. Moreover, we let r, the label of the couple to be 1 if the two samples belong to the same class, $r = -1$ if the two samples belong to different classes. As an illustration, for each training sample $\mathbf{s}_x$, by selecting v_1 similar couples and v_2 dissimilar couples, we obtain $(v_1 + v_2)$ couples: $\underbrace{(\mathbf{s}_x, \mathbf{s}_{x,1}), \ldots, (\mathbf{s}_x, \mathbf{s}_{x,v_1})}_{v_1 \quad similar \quad couples}, \underbrace{(\mathbf{s}_x, \mathbf{s}_{x,v_1+1}), \ldots, (\mathbf{s}_x, \mathbf{s}_{x,v_1+v_2})}_{v_2 \quad dissimilar \quad couples}.$ With all such couples constructed from all training samples, we obtain a couple set: $W = \{P_1, \ldots, P_{Q_d}\}$, $P_d = (\mathbf{s}_{d,1}, \mathbf{s}_{d,2})$ is one component of the couple set W, r_d is the label of P_d, where $d = 1, 2, \ldots, Q_d$, and Q_d denotes the size of W. Then, constrains imposed on the couple set W can be used for discriminative metric learning in the next step.

2) **Discriminative Metric Learning**. The discriminative metric can be efficiently used for measuring the distance between two samples. In order to achieve competitive classification accuracies, we learn the metric on the couple set W to penalize both large distances between samples with the same label and small distances between samples with different labels. The metric is learned by solving the following minimization problem:

$$\min_{M,t,\rho_d} \frac{1}{2} \|M\|_F^2 + C \sum_d \rho_d \tag{3}$$

$$s.t.\ \ r_d \left((\mathbf{s}_{d,1} - \mathbf{s}_{d,2})^T M (\mathbf{s}_{d,1} - \mathbf{s}_{d,2}) + t \right) \geq 1 - \rho_d \tag{4}$$

$$\rho_d \geq 0 \quad \forall d \tag{5}$$

This is a standard SVM-like model, where $\|\bullet\|_F$ denotes the Frobenius norm, ρ_d is the slack variable enabling to deal with the permitted errors, C is the regularization parameter which controls the generalization capabilities, $P_d = (\mathbf{s}_{d,1}, \mathbf{s}_{d,2})$ is the couple constructed from training data, r_d is the label of P_d, and t is the bias. By solving the above minimization problem, the metric matrix M is obtained for classification improvement in step 3).

3) **k-NN Classification**. The goal of improved k-NN classification is to improve classification performance. It can be achieved by measuring the distance between each pair of two samples. For this purpose, we introduce a kernel decision function $y(A)$ [9], where $A = \left(\mathbf{s}_i, \mathbf{s}_j\right)$ is the test couple of two samples. One sample of A is from testing data, and the other is from training data. The decision function $y(A)$ operates on couples from W to learn the distance between A, and t is the bias. The definition of $y(A)$ is as follows:

$$y(A) = sgn\left(\left(\mathbf{s}_i - \mathbf{s}_j\right)^T M \left(\mathbf{s}_i - \mathbf{s}_j\right) + t \right) \tag{6}$$

Using (6), we can tell whether two samples of A have the same class label. In the same way, all such similarities or dissimilarities can be obtained. Then, with the known labels of training samples, the labels of testing samples are acquired with high accuracy.

3. Experiments and Discussion

3.1 Experiment Description

To validate the effectiveness of the proposed approach, a series of experiments were conducted on various data sets. For space limitation, only results on two data sets (DS1 and DS2) are analyzed in this paper. The images are shown in Fig. 3 (a), which are pan-sharpened MS images (960 × 960 pixels) and were acquired by QuickBird 2 over Beijing in 2002. To verify the effectiveness of the novel approach (SSF+DFC), five related methods are proposed for performance comparison:

1) Feature Analyst [10]. Feature Analyst is a object-oriented system. It uses Automated Feature Extraction (AFE) application to extract and classify target features and recognize objects in complex scenes.

2) eCognition [11]. eCognition is a segmentation-based classifier that uses fuzzy reasoning techniques to extract buildings or other objects.

3) DAISY+SVM. DAISY [12] feature is obtained from each pixel. Then we utilize the trained SVM model, whose parameters are selected by 5-fold cross-validation.

4) SSF+SVM. In this approach, SSF is extracted from each pixel. And the same training data as SSF+DFC are used to train SVM for classification.

Table 1 Performance Comparison

DS	Method	Recall	Precision	F-score
DS1	Feature Analyst	72.95	47.71	57.69
	eCognition	51.39	50.87	51.08
	DAISY+SVM	74.13	73.22	73.67
	SSF+SVM	92	80.86	86.07
	DAISY+DFC	80.85	77.01	78.89
	SSF+DFC	**94.22**	81.46	87.38
DS2	Feature Analyst	66.70	45.59	54.16
	eCognition	56.13	36.84	44.49
	DAISY+SVM	80.81	78.69	79.74
	SSF+SVM	92.77	80.94	86.45
	DAISY+DFC	86.35	79.39	82.72
	SSF+DFC	**94.71**	81.37	87.53

5) DAISY+DFC. In this scheme, DAISY feature is extracted the same as DAISY+SVM, based on which an application of DFC leads to results.

For fair comparison, DAISY+SVM and DAISY+DFC are based on the same training data, which are extracted from DAISY feature. In the same way, SSF+SVM and the proposed approach use the same training data set, which are extracted from SSF.

In this paper, we extract 40000 training samples from 921600 samples randomly for each data set. In addition, the feature dimension is 128. For evaluation, manually labeled ground truths are given for both two data sets. In each experiment, Recall, Precision and F-score are used for evaluating the performance.

3.2 Experiment Results and Analyses

For DS1 and DS2, results of different approaches are shown in Fig. 3. And quantitative comparisons of different approaches are listed in Table 1. Based on Feature Analyst [Fig. 3 (c)] and eCognition [Fig. 3 (d)], many buildings are missed and some backgrounds are wrongly classified as buildings. The reason is that for Feature Analyst, spectral and textural features are not sufficient to detect buildings from backgrounds. eCognition sometimes contain irregular or jagged segments, which leads to poor building extractions. DAISY+SVM [Fig. 3 (e)] and DAISY+DFC [Fig. 3 (g)] perform poor in building boundaries. The reason mainly lies in the limitation of DAISY feature, which aggregates oriented gradients across the boundaries and ignores shapes. For comparison, SSF+SVM and SSF+DFC achieve better performance. From Table 1, the Recall (%) of SSF+DFC is about 2% higher than SSF+SVM, which means SSF+DFC can correctly detect more buildings. For visual comparison, in ground truths [Fig. 3 (b)], some buildings are marked by red rectangular boxes. But in results of SSF+SVM [Fig. 3 (f)], these buildings fail to be extracted and the red rectangular boxes are empty. In contrast, SSF+DFC can successfully extract and mark them [Fig. 3 (h)]. Therefore, SSF+DFC performs the best.

From Table 1, we can conclude that SSF+DFC is superior to other methods in terms of Recall (%), Precision (%) and F-score (%). The advantage of the proposed approach is mainly taken from the combination of shape-specific feature extraction and discriminative feature classification. The extracted SSF is effective in capturing complex shapes and structures of buildings. Based on discriminative feature classification, similarities between buildings and differences between buildings and backgrounds are acquired to improve the accuracy of building detection.

4. Conclusion

In this paper, shape-specific feature extraction and discriminative feature classification are combined for building detection from remote sensing images. With this combination, the aforementioned two difficulties can be addressed in a simple and efficient way, and improved performances are thus obtained. Future work will focus on adding DEM and LIDAR data [13] into our scheme to improve the accuracy further.

References

[1] B. Sirmacek and C. Ünsalan, "Urban-area and building detection using sift keypoints and graph theory," IEEE Trans. Geos. Rem. Sens., vol.47, no.4, pp.1156–1167, 2009.
[2] G. Cheng, J. Han, P. Zhou, and L. Guo, "Multi-class geospatial object detection and geographic image classification based on collection of part detectors," ISPRS Jour. Phot. Rem. Sens., vol.98, pp.119–132, 2014.
[3] P. Saeedi and H. Zwick, "Automatic building detection in aerial and satellite images," IEEE Int. Conf., Contr., Aut., Rob. Vis., pp.623–629, 2008.
[4] C.S. Fraser, E. Baltsavias, and A. Gruen, "Processing of ikonos imagery for submetre 3d positioning and building extraction," ISPRS Jour. Phot. Rem. Sens., vol.56, no.3, pp.177–194, 2002.
[5] N. Khan, M. Algarni, A. Yezzi, and G. Sundaramoorthi, "Shape-tailored local descriptors and their application to segmentation and tracking," IEEE Conf. Comp. Vis. Pat. Rec., pp.3890–3899, 2015.
[6] T. Lindeberg, "Detecting salient blob-like image structures and their scales with a scale-space primal sketch: a method for focus-of-attention," Int. Jour. Comp. Vis., vol.11, no.3, pp.283–318, 1993.
[7] D. Mumford and J. Shah, "Optimal approximations by piecewise smooth functions and associated variational problems," Com. pur. app. mat., vol.42, no.5, pp.577–685, 1989.
[8] F. Wang, W. Zuo, L. Zhang, D. Meng, and D. Zhang, "A kernel classification framework for metric learning," IEEE Tran. Neur. Net. Lear. Sys., vol.26, no.9, pp.1950–1962, 2015.
[9] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet, "More efficiency in multiple kernel learning," Proc. of the int. conf. Mac. learn., pp.775–782, 2007.
[10] J. Blundell and D. Opitz, "Object recognition and feature extraction from imagery: The feature analyst approach," Int. Arc. Phot., Rem. Sens. Spat. Inf. Scien., vol.36, no.4, p.C42, 2006.
[11] M. Baatz, U. Benz, et al., "Ecognition user guide," Defi. Imag. Gmb., 2001.
[12] E. Tola, V. Lepetit, and P. Fua, "Daisy: An efficient dense descriptor applied to wide-baseline stereo," IEEE Trans. Pat. Anal. Mac. Intel., vol.32, no.5, pp.815–830, 2010.
[13] Y. Ichikawa, D. Deguchi, I. Ide, et al., "A study on features for pedestrian detection using a low resolution lidar," IEICE Technical Report, vol.114, no.230, pp.7–12, 2014.