OBJECT LOCALIZATION BASED ON SPARSE REPRESENTATION FOR REMOTE SENSING IMAGERY

Naoto Yokoya and Akira Iwasaki

Department of Advanced Interdisciplinary Studies, The University of Tokyo, Japan

ABSTRACT

In this paper, we propose a new object localization method named *sparse representation based object localization (SROL)*, which is based on the generalized Hough-transform-based approach using sparse representations for parts detection. The proposed method was applied to car and ship detection in remote sensing images and its performance was compared to those of state-of-the-art methods. Experimental results showed that the SROL algorithm can accurately localize categorical objects or a specific object using a small size of training data.

Index Terms— Sparse representation, object localization, generalized Hough transform

1. INTRODUCTION

The spatial resolution of optical remote sensing imagers has been improving particularly in the last decade, e.g., GeoEye, WorldView, and Pleiades series. Skysat series firstly launched on November 2013 enable the acquisition of movies with a 1-m ground sampling distance (GSD) from space. These advances in sensor technologies allow advanced image understanding and interpretation, such as object detection and localization. Fully automated object localization is required to analyze big data of high-spatial-resolution imagery.

Object detection has been actively studies in the field of computer vision. A joint use of local-feature extraction and classification based on machine learning algorithms is an effective approach for object detection. Various feature extraction methods, such as Haar-like feature, the scale invariant feature transform (SIFT), and histograms of oriented gradient (HOG), have been used for many tasks of object detection [1, 2, 3]. Support vector machine (SVM) and AdaBoost are well-known classifiers used for object recognition [1, 4].

An approach that uses the presence of parts for the object class and their structural relations has been getting attention for object detection. Agarwal and Roth proposed an approach for learning a sparse, part-based representation for object detection and showed its robustness to partial occlusion and background variation [5]. The Hough-transform-based method proposed by Leibe *et al.* learns the class-specific implicit shape model (ISM), which detect local appearances of categorical objects according to a codebook and localize objects considering their co-occurrence consistency by the generalized Hough transform [6]. Gall and Lempitsky proposed a class-specific Hough forest algorithm, which uses a random forest to discriminatively detect object parts and directly cast probabilistic votes about possible centers of the object to generate a Hough image [7].

Object detection of high-resolution remote sensing images has different characteristics compared to ground-shot images: objects are generally small relative to a GSD with cluttered backgrounds; rotation invariance is required, whereas scale invariance is not strongly required owing to the a fixed GSD for each imaging sensor and appearance changes are relatively small owing to limited pointing angles. Many researchers have worked on detection of specific object class in remote sensing imagery, such as car, ship, and airplane detection; however, many of them are adhoc and limited to specific uses. Lei *et al.* proposed an extended method of the Hough forest for object detection of remote sensing images [8]. The major improvement is to achieve rotation invariance by firstly detecting dominant gradient orientations and align local image patches.

The theory of sparse representation and compressed sensing has been getting attention in the areas of signal processing, computer vision and pattern recognition [9]. Sparse representations enable finding essential image patterns and have been used for a wide range of image processing applications [10]. In this paper, we present a new method, named sparse representation based object localization (SROL), to detect and localize categorical objects or specific objects in remote sensing imagery. Our approach uses sparse representations as local-feature detection of the generalized Hough-transform-based object localization. Parts of the object are detected by sparse representations of patches in an input image using pre-learned target and background dictionaries and the locations of the objects are determined by considering their structural relations using the generalized Hough voting. We adopt sparse representations for local-feature detection to deal with cluttered backgrounds, occlusion, and appearance changes of objects and to achieve a good performance with a small size of training data. Our experiments are performed on car and ship detection to demonstrate the effectiveness of the proposed method.

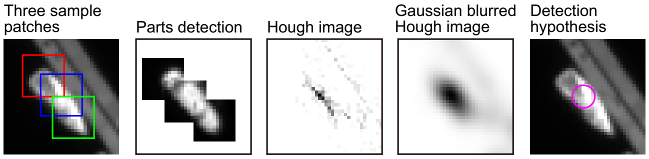


Fig. 1. Outline of object localization based on sparse representation.

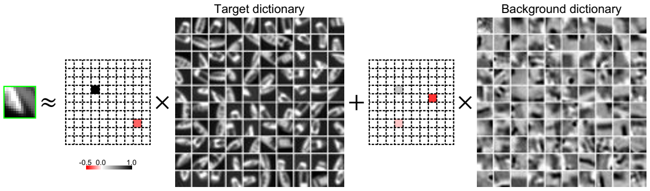


Fig. 2. Binary class labeled sparse representation of patch.

2. METHODOLOGY

Fig. 1 illustrates the outline of our proposed method. The SROL algorithm is composed of four steps: i) sliding window search to extract patch images in a test image; ii) parts detection via sparse representations of patch images; iii) Hough voting using offsets of detected parts of the target object; iv) finding local maxima in the Hough image.

2.1. Dictionary Construction

The fist task of any local-feature based approach is to detect object parts in a given image. Sparse representations can be used for this purpose. Any patch image is assumed to be represented as a sparse linear combination of *atoms*. A patch image $\mathbf{y} \in \mathbb{R}^P$ is formulated as

$$\mathbf{y} \approx \mathbf{D}\mathbf{x}, \tag{1}$$

where $\mathbf{D} \in \mathbb{R}^{P \times N}$ denotes the dictionary with each column vector representing an atom, $\mathbf{x} \in \mathbb{R}^N$ is the sparse coefficient vector, P is the number of pixels in the patch image, and N is the number of atoms. In the sparse representation, the number of nonzero values of $\mathbf{x}$ is assumed to be much smaller than P, i.e., $\|\mathbf{x}\|_0 \ll P$. Therefore, $\mathbf{x}$ is obtained by the following optimization

$$\min_{\mathbf{x}} \|\mathbf{x}\|_0 \text{ s.t.} \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2^2 \leq \epsilon \tag{2}$$

This optimization is known as an NP-hard problem; however, several techniques have been studied to approximately solve it, such as matching pursuit (MP) and its extensions [11, 12]. In the case of $\|\mathbf{x}\|_0 = 1$, the MP-based sparse representation acts as image matching using an Euclidean distance for similarity measurement.

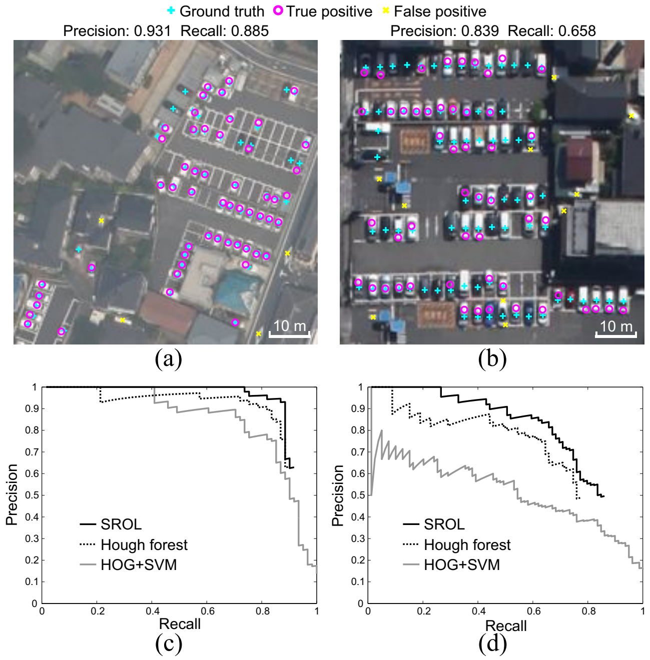
When atoms of the dictionary have class labels, i.e., *target* or *background*, sparse representations can be used for object part detection. We prepare the dictionary $\mathbf{D}$ as the horizontally stacked matrix $[\mathbf{D}_t \;\; \mathbf{D}_b]$, where $\mathbf{D}_t \in \mathbf{R}^{P \times N_t}$ is the target dictionary with the column vectors representing various parts images of the target and $\mathbf{D}_b \in \mathbf{R}^{P \times N_b}$ is the background dictionary with the column vectors representing atoms of the background. N_t and N_b are the numbers of atoms for the target and background dictionaries, respectively, and then $N = N_t + N_b$. Fig. 2 illustrates the binary class-labeled sparse representation of a patch for object part detection. Positive coefficients of the target-dictionary atoms, $x_k > 0$ ($k \leq N_t$), suggest the existence of the object parts. The first step of our approach aims to construct a structured dictionary.

Learning a dictionary directly from training data usually leads to better representation than using a predetermined dictionary. We adopt the K-SVD algorithm to build a reconstructive and compact background dictionary [13]. A large number of patch images that do not include the target objects are randomly sampled and K-SVD is applied to them to learn the background dictionary, which can reconstruct various patch patterns including cluttered backgrounds. In our implementation, we construct the background dictionary depending on the GSD of test images because patch patterns are conditional to the GSD.

Random sampling and atom selection are used to construct the target dictionary. We keep raw image patches because each atom of the target dictionary needs to have the offset of the patch relative to the center of the object for Hough voting. To achieve rotation-invariant object localization, we augment the training samples with rotated copies of the original images and obtain an initial redundant target dictionary by random sampling. Next, the number of atoms in the redundant dictionary is reduced by removing samples that have very similar other samples. The zero-mean normalized cross-correlation (ZNCC) is used to measure the similarities between all samples. The offset (δi_k, δj_k) of the patch center from the target center need to be linked to the patch image, i.e., the k^{th} atom in $\mathbf{D}_t$ ($k = 1, ..., N_t$). When the target have sufficient spatial features, it is possible to estimate its orientation. In this case, the orientation of the target (θ_k) is also linked to the patch image.

2.2. Object Localization with Sparse Representation Based Hough Voting

Sliding-window search for all patches of the test image is time-consuming. Interest point detectors have widely been used for the Hough-transform-based object detection; however, small sizes of objects in remote sensing images require lower level feature detector for sliding-window search. We use edges obtained by a Sobel operation for sliding-window search. In this case, the generation of the target dictionary can be effectively processed by sampling patches along edges to make consistency between patch images of the dictionary and test images.



(a) (b)

(c) (d)

Fig. 3. Car detection results and precision-recall curves for two test images.

Sparse representations are used to judge whether a patch image contains a part of the target or not dealing with changes of background and appearance of the target. For a patch with its center locating at (i, j) of the test image, a sparse representation $\mathbf{x}$ is derived solving (2) by MP. When $x_k > 0$ $(k \leq N_t)$, it suggests that the target may be located at $(i + \delta i_k, j + \delta j_k)$. Higher value of x_k indicates higher correlation between the given patch and the part of the target $\mathbf{d}_k$, and thus higher probability of the existence of the target. Therefore, we cast the vote of the value x_k to the location $(i + \delta i_k, j + \delta j_k)$, which is the generalised Hough transform. The Hough image can be obtained as a result of iterations of this process for all patches, which represents the existence probability of the target.

The target objects can be simply localized by returning the set of the local-maxima locations in the Gaussian-filtered Hough image. A Gaussian blur filter with its full width at half maximum setting to the size of the smaller side of the target object is useful to smooth the Hough image for effectively finding its local maxima.

3. EXPERIMENTAL RESULTS

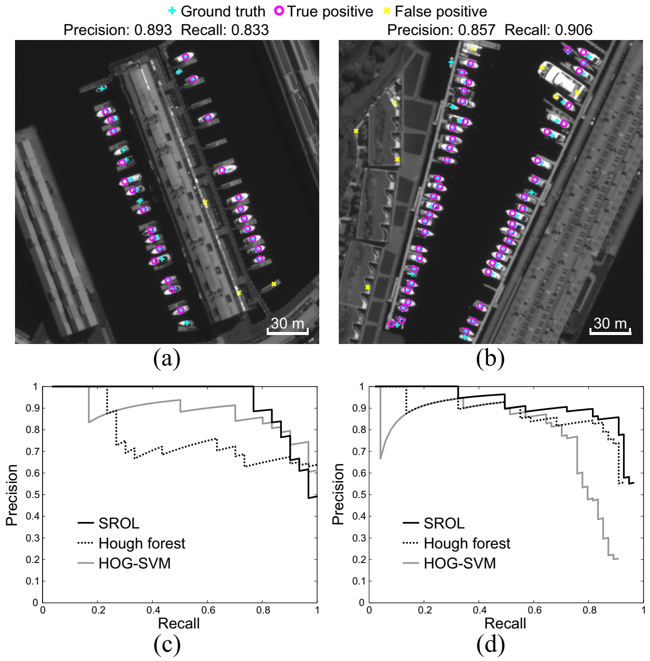
3.1. Car Detection

First, we demonstrate an experiment of car detection using airborne images. The RGB images were taken over urban areas of Tokyo Japan from approximately 1000 m altitude with the 0.2 m GSD on 11 August 2013. The images were converted to Lab color space and the lightness channel was used for processing. 36 training samples were used for this experiment. Detection results are considered as "true" when their Euclidean location errors from the ground truth are less than the average size of the objects. The patch is 13×13 pixel size. The proposed method is compared to the state-of-the-art methods, i.e., the joint use of HOG and SVM (HOG-SVM) and the rotation-invariant Hough forest [8]. We use the same window search and patch size for the Hough forest for a fair comparison. The patch size of HOG-SVM is set to 40×40 pixel size to include a whole object. The precision-recall curve is used to quantitatively evaluate their detection performances. Fig. 3 shows the car detection results of SROL and the comparison of precision-recall curves for two test images. As shown in Figs. 3(c)(d), SROL outperforms HOG-SVM and Hough forest for both images. One of the reasons for the low performance of HOG-SVM may be the small size of training data. Generally speaking, machine-learning-bade classifiers (e.g. SVM) require a large size of training data containing various appearances of the object class. The patch size of HOG-SVM is larger than those of SROL and Hough forest and thus requires more training samples to learn various effects of backgrounds. In this sense, the generalized Hough-transform-based approaches have the advantage to construct object detectors robust to cluttered backgrounds with a small size of training data. In addition, sparse representations can deal with appearance varieties of the object by the background dictionary using limited training samples, which result in robust detection of the parts compared to Hough forest. The robust detection of the parts and the integration of their co-occurrences enables the accurate detection of the object class.

As shown in Figs. 3(a)(b), many of false negatives are black cars. The main reason for this drawback is that there is not enough characteristic parts for black cars. It is much more challenging to distinguish black bodies and windows or dark backgrounds. In addition, less edges around black cars result in less searching windows for the proposed method, which is a critical issue to use co-occurrence of parts. Higher spatial resolution may be required to accurately detect such objects.

3.2. Ship Detection

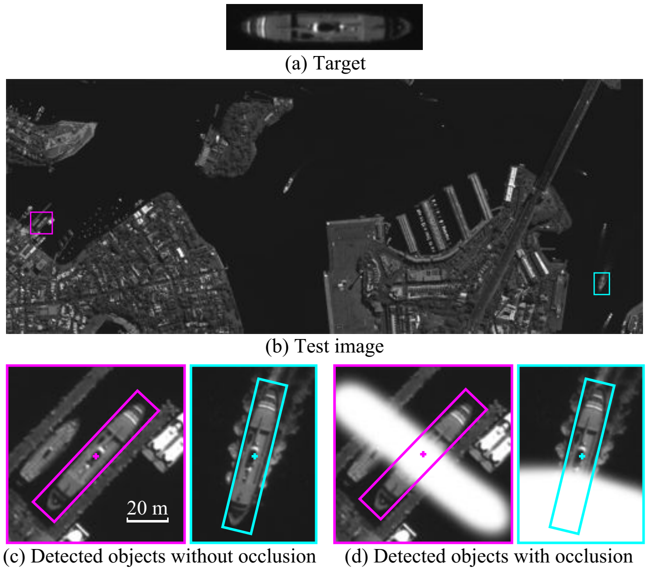
As a second illustration of the proposed method, we turn to ship detection examples. The test image was taken over Sydney by WorldView-2 on 21 August 2012 with the 0.5 m GSD in the panchromatic channel. The patch is 13×13 pixel size. 25 ships on the sea without no surrounding objects were used for training data and two subimages were selected for testing. Figs. 4(a)(b) show the ship detection results of SROL for the two test images and Figs. 4(c)(d) show the precision-recall curves comparing with HOG-SVM and Hough forest. Even though the test images include berths that clutter backgrounds relative to the training samples, the SROL method successfully detect the ships. HOG-SVM performed relatively better than the case of car detection because of the simpler backgrounds and the higher contrast between targets and back-



(a) (b)

(c) (d)

Fig. 4. Ship detection results and precision-recall curves for two test images.



(a) Target

(b) Test image

(c) Detected objects without occlusion (d) Detected objects with occlusion

Fig. 5. Ship identification results.

grounds. The precision value of SROL decreases later than those of other methods, which implies that SROL preferentially detect parts of the target object and return their locations owing to robust parts detection.

Finally, we show an example of identification of a specific object. Fig. 5(a) shows the target of this experiment and Fig. 5(b) is the test image that captured two ships manually recognized as the same type with the target. Since this ship has informative textures, we estimate its direction together with its location. Fig. 5(c) shows the detected two ships as the top two local maxima in the Hough image with each rectangle illustrating the estimated direction of the ship. To examine the robustness to occlusion, we added synthetic occlusion to the test image. As shown in Fig. 5(d), the SROL method can still localize the two ships as the 1^{st} and 5^{th} local maxima with the accurate locations and orientations. This experiment proves that the proposed method is also useful for target identification and it works well even with occlusion when informative textures of the target are visible.

4. CONCLUSION

We presented a novel object localization method based on sparse representation and demonstrated its effectiveness for remote sensing imagery. Parts of an object class or a specific target can be detected by sparse representations using the target and background dictionaries. This parts detection is integrated with the generalized Hough-transform-based object localisation. Experimental results on car and ship detection demonstrated a good performance of the proposed method compared to the state-of-the-art methods using a small size of training samples.

5. REFERENCES

[1] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *Proc IEEE Conf. on Computer Vision and Pattern Recognition*, 2001.

[2] D. G. Lowe, “Discriminative image features from scale-invariant keypoints,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[3] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *Proc IEEE Conf. on Computer Vision and Pattern Recognition*, 2005.

[4] M. Pontil and A. Verri, “Support vector machines for 3D object recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 6, pp. 637–646, 1998.

[5] S. Agarwal and D. Roth, “Learning a sparse representation for object detection,” In *Proc. ECCV*, pp. 113–130, 2002.

[6] B. Leibe, A. Leonardis, and B. Schiele, “Robust object detection with interleaved categorization and segmentation,” *Int. J. Comput. Vis.*, vol. 77, no. 1, pp. 259–289, May 2008.

[7] J. Gall and V. Lempitsky, “Class-specific Hough forests for object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1022–1029.

[8] Z. Lei, T. Fang, H. Huo, and D. Li, “Rotation-invariant object detection of remotely sensed images based on texton forest and Hough voting,” *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 4, pp. 1206–1217, 2012.

[9] D. L. Donoho, “Compressed sensing,” *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[10] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, “Robust face recognition via sparse representation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 210–227, 2009.

[11] S. G. Mallat and Z. Zhang, “Matching pursuits with time-frequency dictionaries,” *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, 1993.

[12] Y. Pati, R. Rezaiifar, and P. Krishnaprasad, “Orthogonal matching pursuit : recursive function approximation with application to wavelet decomposition,” in *Proc. Asilomar Conf. on Signals*, vol. 1, no. 40–44, 1993.

[13] M. Aharon, M. Elad, and A. M. Bruckstein, “The K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representation,” *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, 2006.