Object Detection Based on Sparse Representation and Hough Voting for Optical Remote Sensing Imagery

Naoto Yokoya, *Member, IEEE*, and Akira Iwasaki

***Abstract*—We present a novel method for detecting instances of an object class or specific object in high-spatial-resolution optical remote sensing images. The proposed method integrates sparse representations for local-feature detection into generalized-Hough-transform object detection. Object parts are detected via class-specific sparse image representations of patches using learned target and background dictionaries, and their co-occurrence is spatially integrated by Hough voting, which enables object detection. We aim to efficiently detect target objects using a small set of positive training samples by matching essential object parts with a target dictionary while the residuals are explained by a background dictionary. Experimental results show that the proposed method achieves state-of-the-art performance for several examples including object-class detection and specific-object identification.**

***Index Terms*—Hough transforms, object detection, sparse representations.**

I. Introduction

THE SPATIAL resolution of optical remote sensing imagers has been improving, particularly in the last decade, for example, the GeoEye, WorldView, and Pleiades series. The Skysat series, first launched in November 2013, enable the acquisition of movies with a 1-m ground sampling distance (GSD) from space. These advances in sensor technologies have led to advanced image understanding and interpretation; however, high-spatial-resolution optical remote sensed imagery contains a large amount of data and visual analysis by humans is time-consuming. Therefore, automated object detection is required to extract information from the data along with user interpretation.

Object detection has been actively studied in the field of computer vision. The 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 the use of Haar-like features, the scale-invariant feature transform (SIFT), and histograms of oriented gradients (HOG), have been used for many object detection tasks [1]–[3]. Neural networks, support vector machines (SVMs), and AdaBoost are well-known classifiers used for object recognition [1], [4], [5].

Manuscript received September 22, 2014; revised December 25, 2014; accepted January 26, 2015.

The authors are with the Department of Advanced Interdisciplinary Studies, University of Tokyo, Tokyo 153-8904, Japan.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTARS.2015.2404578

Psychological and physiological evidence for part-based representations in the brain has been reported in the literature [6]–[8], and certain computational theories of object detection are based on such representations [9], [10]. An approach that uses the presence of individual object parts and their structural relations has been receiving particular 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 [11]. Fergus *et al*. presented a probabilistic approach to learning and recognizing object class models as flexible constellations of parts using the shape, appearance, occlusion, and relative scale of the object [12]. In the Hough-transform-based object detection proposed by Leibe *et al*., a class-specific implicit shape model (ISM) is learned [13]; this model detects the local appearances of class objects in accordance with a codebook and localizes objects considering their spatial co-occurrence consistency by the generalized Hough transform [14]. Descriptors of points of interest in a test image are matched against the codebook and the matches cast probabilistic votes regarding possible centers of the object, which is referred to as Hough voting. The peaks in a Hough image that accumulates the votes from all parts represent detection hypotheses. 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 regarding possible centers of the object to generate the Hough image [15]. The Hough forest is one of the state-of-the-art methods for object detection.

Object detection in high-resolution remote sensing images has different characteristics from that in ground-shot images: objects are generally small relative to the GSD with cluttered backgrounds; rotation invariance is required, whereas scale invariance is not strongly required owing to the fixed GSD for each imaging sensor. Also, the changes in appearance are relatively small owing to the limited range of pointing angles. Many researchers have studied the detection of class objects in remote sensing imagery, such as cars, ships, and airplanes [16]–[19]; however, many of the methods they developed are *ad hoc* and limited to a specific use. Lei *et al*. proposed an extension of the Hough forest for object detection in remote sensing images [20]. Its main improvement was the achievement of rotation invariance by first detecting dominant gradient orientations and aligning local image patches.

One drawback of machine learning methods including the Hough forest is that they require a large set of training data

including many positive samples with various backgrounds in order to train classifiers that can accurately discriminate between objects and nonobjects. For some class objects, it may be expensive in terms of human and economic resources to collect many high-spatial-resolution remote sensing images and manually localize the objects with labels. In addition, for a specific target, it may not be feasible to prepare a large number of positive training samples since imaging sensors with very high spatial resolution do not observe a specific target as frequently. Therefore, it remains a challenging task to achieve good detection performance using a small set of positive training samples.

Recently, a great deal of attention has been paid to the theory of sparse representation and compressed sensing in the areas of signal processing, computer vision, and pattern recognition [21], [22]. Sparse representations enable essential image patterns to be found, even those with noise or occlusion, and have been used for a wide range of image-processing applications, such as image restoration, super-resolution, and face recognition [23]–[25]. For object detection in remote sensing, the sparse representation perspective has been applied to target detection in hyperspectral images [26], [27]. Class-specific sparse representations and discriminative learning have been shown to be effective in target or anomaly detection using spectral features [28].

In this paper, we present a novel method based on sparse representations and Hough voting (SR-Hough) for detecting instances of an object class or a specific object in remote sensing imagery. Our method integrates class-specific sparse image representations for local-spatial-feature detection into generalized-Hough-transform object detection. Object parts are detected by sparse representations of patches in an input image using prelearned target and background dictionaries, and the locations of candidate objects are determined by considering their structural relations using generalized Hough voting. We aim to efficiently detect target objects with a small set of positive training samples. Class-specific sparse representations for local-feature detection are expected to deal with a cluttered background, noises, partial occlusion, and various appearances of objects by representing essential object parts and residuals using the target and background dictionaries, respectively. Our experiments are performed on car, boat, and airplane detection as well as the identification of a specific ship to demonstrate the efficacy of the proposed method.

This paper is organized as follows. In Section II, we describe the methodology of the proposed method, mainly focusing on the introduction of sparse representations into the Hough transform framework. Experimental results are presented in Section III to verify the performance of the proposed method by comparison with state-of-the-art techniques for object detection. The conclusion is given in Section IV.

II. Methodology

Our proposed method can be divided into two phases: 1) a dictionary construction; and 2) a detection procedure for training and testing, respectively. Fig. 1 shows an outline of the detection procedure. The detection procedure of the SR-Hough algorithm is composed of four steps: 1) a sliding-window search to extract patch images in a test image; 2) detecting parts via the sparse representations of patches; 3) Hough voting using offsets of detected parts of the target object and 4) finding maxima in the Hough image. In this section, we first explain the dictionary construction for the sparse representation as the training phase since the second step of the detection procedure, i.e., detecting parts via sparse representations, is the main novelty of our method. Then, the entire detection procedure and implementation are described in detail.

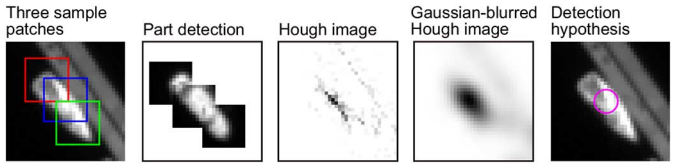

Fig. 1. Outline of the detection procedure.

A. Dictionary Construction

The first task of any local-feature-based approach is to detect object parts in a given image. Sparse representations can be used for this purpose by employing binary class-specific dictionaries. Each patch image is assumed to be a sparse linear combination of basis vectors called *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, and N is the number of atoms. The sparse representation assumes that the number of nonzero values of $\mathbf{x}$ is much smaller than P, i.e., $\|\mathbf{x}\|_0 \ll P$. Therefore, $\|\mathbf{x}\|$ is obtained by the following optimization:

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2^2 \quad \text{subject to} \quad \|\mathbf{x}\|_0 \leq T_0. \tag{2}$$

This optimization is known as an NP-hard problem; however, several techniques to approximately solve it, such as matching pursuit (MP) and its extensions, have been proposed [29], [30]. In the case of $\|\mathbf{x}\|_0 = 1$, the MP-based sparse representation can be used to perform template matching using an Euclidean distance as a similarity measurement.

When the atoms of the dictionary have class labels, i.e., *target* or *background*, sparse representations can be used for detecting parts of objects. We prepare the dictionary $\mathbf{D}$ as the horizontally stacked matrix $[\mathbf{D}_t\ \mathbf{D}_b]$, where $\mathbf{D}_t \in \mathbb{R}^{P \times N_t}$ is the target dictionary with the column vectors representing various parts of the target and $\mathbf{D}_b \in \mathbb{R}^{P \times N_b}$ is the background dictionary with the column vectors representing atoms of the background. N_t and N_b are the numbers of atoms in the target and background dictionaries, respectively, and then $N = N_t + N_b$. Fig. 2 illustrates the binary class-specific sparse representation of a patch used for detecting parts of objects. Positive coefficients of the target dictionary atoms, $x_k > 0$ ($k \leq N_t$), mean

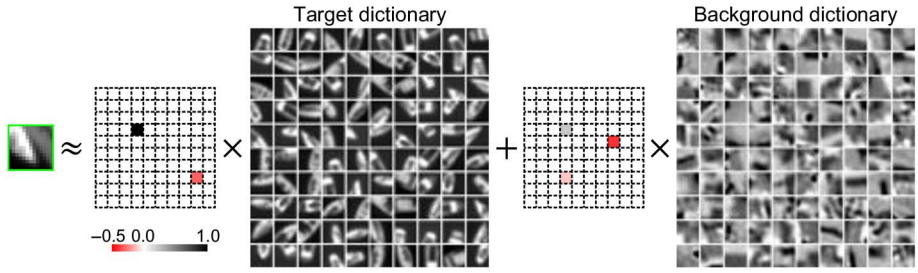

Fig. 2. Binary class-specific sparse representation of a patch.

the existence of object parts. The first step of our approach aims to construct a structured dictionary with atoms corresponding to the class labels so that only object parts can be detected as local features of the object.

Learning a dictionary directly from training data usually leads to better representation than using a predetermined dictionary, such as a wavelet or Gabor dictionary, and contributes to improved results in various applications [31], [32]. Many algorithms have been studied for dictionary learning using image examples. One of the most well-known algorithms is the K-SVD algorithm proposed by Aharon *et al*. [33], which solves the following optimization:

$$\min_{\mathbf{D},\mathbf{X}} \|\mathbf{Y}-\mathbf{D}\mathbf{X}\|_F^2 \quad \text{subject to} \quad \forall i,\ \|\mathbf{x}_i\|_0 \le T_0 \tag{3}$$

where $\mathbf{Y} \in \mathbb{R}^{P\times L}$ is a set of L sample images, with each column vector representing an example patch, and $\mathbf{X} \in \mathbb{R}^{N\times L}$ is a sparse matrix with the ith column vector $\mathbf{x}_i$ representing the sparse coefficient vector of the ith sample image. K-SVD iteratively alternates between sparse coding the examples based on the current dictionary and updating the dictionary atoms based on a singular value decomposition approach. Discriminative dictionary learning methods have recently been developed mainly for classification tasks [34]–[38]. In our method, each detected object part must have the offset of the patch relative to the center of the object to perform Hough voting. Therefore, we need to keep raw image patches only for the atoms in the target dictionary, while aiming at the construction of discriminative background and target dictionaries. This problem setting is not considered in common discriminative learning methods that do not keep raw image patches.

K-SVD is used to build a reconstructive and compact background dictionary. A large number (N_b^*) of patches that do not include the target objects are randomly sampled as negative training samples and we denote this set of raw patches as $\mathbf{D}_b^*$. Next, K-SVD is applied to them to learn the background dictionary $\mathbf{D}_b$, which can reconstruct various patch patterns including cluttered backgrounds. In our implementation, we construct background dictionaries depending on the GSDs of test images because the patch patterns depend on the GSD.

To construct the target dictionary for discriminating from the background dictionary, we adopt the random sampling of target patches and atom selection using discriminative criteria. To achieve rotation-invariant object detection, we augment the positive training samples with rotated copies of the original training images at 10° increments and obtain an initial redundant target dictionary. Next, the number of atoms in the redundant dictionary is reduced by removing samples that are very similar to other samples. Here, the less redundant target dictionary obtained is denoted as $\mathbf{D}_t^*$. The zero-mean normalized cross-correlation (ZNCC) [39] is used to measure the similarities between all samples.

A target patch that is more likely to be selected as an active atom for the sparse representation of a nontarget patch leads to FPs. Therefore, we calculate the sparse representations of background patches using MP as

$$\min_{\mathbf{X}} \|\mathbf{D}_b^*-\mathbf{D}\mathbf{X}\|_F^2 \quad \text{subject to} \quad \forall i,\ \|\mathbf{x}_i\|_0 \le T_0 \tag{4}$$

where $\mathbf{D}=[\mathbf{D}_t^*\ \mathbf{D}_b]$. All the coefficients of each atom are accumulated as $\sum_{m=1}^{N_b^*}\mathbf{X}_{nm}$. A high cumulative value for a target patch indicates that the patch easily generates FPs. Finally, target patches with cumulative coefficients smaller than a predefined threshold (e.g., 0.5 in this work) are selected as discriminative patches, which form a subset of $\mathbf{D}_t^*$. This subdictionary is used as $\mathbf{D}_t$ to make the target dictionary more discriminative and compact. The offset $(\delta i_k, \delta j_k)$ of the patch center from the target center must be linked to the patch, i.e., the kth atom in $\mathbf{D}_t$ ($k=1,\ldots,N_t$). When the targets have a sufficient number of spatial features, it is possible to estimate their orientations. In this case, the orientation of the target (θ_k) is also linked to the patch.

B. Sparse-Representation-Based Hough Voting

1) Patch Sampling: A sliding-window search of all patches in a test image is time-consuming. Interest point detectors, such as Harris and difference-of-Gaussian detectors [2], [40], have been widely used for Hough-transform-based object detection; however, the small sizes of objects in remote sensing images require a lower level-feature detector to perform a sliding-window search. Therefore, we use edges obtained by the Canny edge detector for the sliding-window search [41]. In this case, the target dictionary can be effectively generated by sampling patches along edges to ensure the consistency of patches between the dictionary and test images. The patch size (t) is defined relative to the object size as $t=k\sqrt{wh}$, where w and h are the width and height of the object, respectively, and k is a parameter. In this work, k is set to approximately 0.5 for all objects. The influence of this parameter on the detection performance is discussed in the experimental part.

2) Parts Detection: For a patch with its center located at (i, j) in the test image, the sparse representation $\mathbf{x}$ is obtained

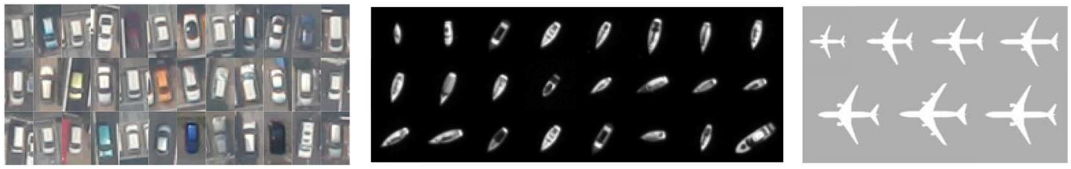

Fig. 3. Positive training samples of: (a) cars; (b) boats; and (c) airplanes.

by solving the optimization problem of (2) using the learned dictionary $\mathbf{D}$. In this work, we adopt MP to solve (2) owing to its mathematical simplicity and low computational cost. When there are positive coefficients of the target atoms in the sparse representation, only the kth atom ($k \leq N_t$) with the maximum coefficient is considered to be a part of the target, assuming that two or more objects do not exist in one patch. Since the relative center of the target is attached to each target atom, the positive coefficient x_k suggests that the target may be located at $(i + \delta i_k, j + \delta j_k)$.

3) Hough Voting: A higher value of x_k indicates a higher correlation between the given patch and the part of the target $\mathbf{d}_k$, and thus a higher existence probability 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 generalized Hough voting procedure using only the object center as parameters. The two-dimensional (2-D) Hough image can be obtained as a result of iterating this process for all patches, which represents the existence probability of the target.

4) Detection Hypothesis: The target objects can be simply detected by returning the set of locations where the maxima are greater than some threshold τ in the Gaussian-filtered Hough image. A Gaussian blur filter with its full width at half maximum set to the size of the smaller side of the target object is useful for smoothing the Hough image to effectively find its maxima. An alternative means of finding the maxima of the Hough image is the use of the mean-shift procedure or the iterative greedy maximum *a posteriori* inference technique [15], [20]. For a target with a sufficient number of spatial features, the orientation of a detected object can also be estimated as a weighted sum of the orientation vectors of the votes contributing to a detection hypothesis. The orientation is given by $\arctan\left(\frac{\sum_{l\in\psi} x_l \sin\theta_l}{\sum_{l\in\psi} x_l \cos\theta_l}\right)$, where ψ is the set of votes in the area surrounding a detection hypothesis and l is the index of the votes.

III. Experimental Results

In this section, we evaluate the proposed method by demonstrating four applications: three examples of object-class detection and one example of specific-object identification.

In the object-class detection, the proposed method is compared to state-of-the-art methods, i.e., the joint use of HOG and SVM (HOG-SVM) and the rotation-invariant Hough forest. For the Hough forest, grayscale images are used in this work instead of color-invariant gradient channels presented in [20] because our experiments include the use of panchromatic images and grayscale illustrations. A special case of the proposed method, which does not use Hough voting with the offset of the atoms in the target dictionary being zero, can be seen as a sparse-representation-based template matching (SR-TM) [42]. It is also included in the numerical evaluation to see the effectiveness of the generalized Hough transform. We use the same window search along edges for SR-Hough and the Hough forest to ensure a fair comparison. In contrast, a greedy search is adopted for HOG-SVM and SR-TM. The patch size is optimized for each method. The patch size for HOG-SVM and SR-TM is set larger than that of SR-Hough and the Hough forest to include a major part of an object. A detection result is counted as a true positive (TP) when its Euclidean location error from the ground truth is less than the width of the target object. Each target object can be detected only once and duplicate detections of the same object are counted as false positives (FPs). Target objects that are not detected are false negatives (FNs). A precision-recall curve is used to quantitatively evaluate the detection performance. Precision is defined as (TP)/(TP+FP) and recall is defined as (TP)/(TP+FN). The area under the precision-recall curve (AUPRC) is used to quantify the overall detection accuracy.

We demonstrate three examples of object-class detection: 1) car; 2) boat; and 3) airplane detection. Discriminative atom selection for the target dictionary is applied to only airplane detection since it does not improve the results of car and boat detection owing to the low-level spatial features of their patches. Estimation of the orientation is also conducted for only airplane detection because cars and boats do not have a sufficient number of spatial features to determine the front and rear.

A. Car Detection

First, we report an experiment on car detection using airborne images. RGB images were taken over urban areas of Tokyo, Japan, from an altitude of approximately 1000 m with a 0.2-m GSD on August 11, 2013. Three images that include 195 cars with various backgrounds are used for testing. The images were converted to lab color space and the lightness channel was used for processing. Fig. 3(a) shows the 36 positive samples used for training. The patch sizes are 9×9 pixels for SR-Hough and the Hough forest and 21×21 for HOG-SVM and SR-TM. 5200 positive and 10 000 negative patches are used for SR-Hough and the Hough forest. We set the size of the background and target dictionaries as $N_b = 50$ and $N_t = 3000$, respectively.

A comparison of the precision-recall curves is shown in Fig. 4(a) and the AUPRC is presented in Table I. As shown in these numerical evaluation, SR-Hough clearly outperforms the other methods. One of the reasons for the low performance

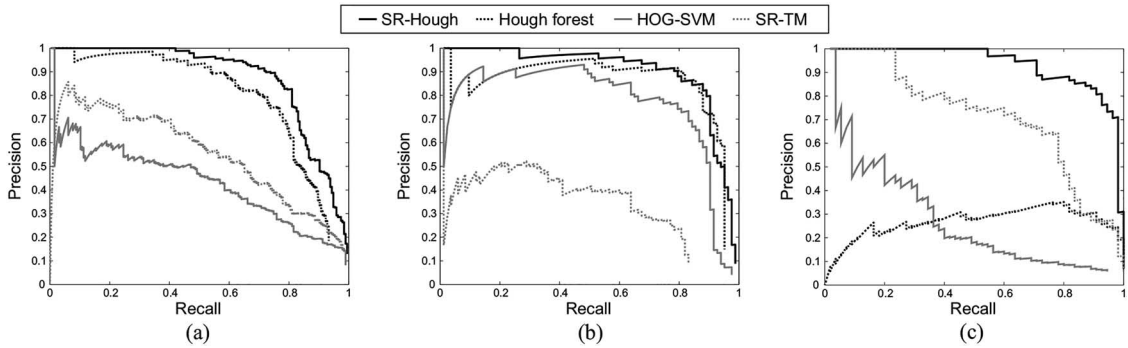

Fig. 4. Precision-recall curves obtained by SR-Hough, Hough forest, HOG-SVM, and SR-TM for: (a) car; (b) boat; and (c) airplane detection.

TABLE I
AUPRC OBTAINED BY SR-HOUGH, HOUGH FOREST, HOG-SVM, AND SR-TM FOR CAR, BOAT, AND AIRPLANE DETECTION

	Car	Boat	Airplane
SR-Hough	**0.869**	**0.888**	**0.918**
Hough forest	0.787	0.835	0.265
HOG-SVM	0.425	0.753	0.254
SR-TM	0.547	0.322	0.696

of HOG-SVM is that the amount of training data may not be sufficient for constructing an accurate classifier. Generally, machine-learning classifiers (e.g., SVM) require a large set of training data containing various appearances of an object class. Since the patch size of HOG-SVM is larger than those of SR-Hough and the Hough forest, more training samples are necessary to learn the various effects of backgrounds. In this sense, the Hough transform framework has the advantage of constructing object detectors that are robust to cluttered backgrounds using a small set of training data. The effectiveness of Hough voting is also proved by the advantage of SR-Hough compared with SR-TM. In addition, sparse representations can deal with the variations of appearance of an object by using a background dictionary with a limited number of training samples, which results in the robust detection of parts compared with the Hough forest. The robust detection of parts and the integration of their co-occurrences enable the accurate detection of class objects.

Fig. 5 shows the car detection results obtained by SR-Hough for the three test images at $\mathrm{precision} = 0.896$ and $\mathrm{recall} = 0.754$. As shown in Fig. 5, many of FNs are black cars. This is mainly because there are an insufficient number of characteristic parts for black cars. It is difficult even for humans to distinguish black bodies from windows or dark backgrounds in a local patch. In addition, the fewer visible edges around black cars result in fewer searching windows for the proposed method, which is a critical issue when using the co-occurrence of parts. Higher spatial resolution may be required to accurately detect such objects.

B. Boat Detection

As the second illustration of the proposed method, we turn to boat detection. The study image was taken over Sydney by WorldView-2 on August 21, 2012 with a 0.5-m GSD in the panchromatic channel. Two subimages of wharfs that include 74 boats were selected for testing. From the remainder of the image, 25 boats on the sea that were not surrounded by other objects, as shown in Fig. 3(b), and background areas were extracted and used for training. The patch sizes are 11×11 pixels for SR-Hough and the Hough forest and 21×21 for HOG-SVM and SR-TM. 4300 positive and 10 000 negative patches are used for SR-Hough and the Hough forest. We set the size of the background and target dictionaries as $N_b = 100$ and $N_t = 3000$, respectively.

Fig. 4(b) shows the precision-recall curves for the four method and their AUPRC is shown in Table I. SR-Hough and the Hough forest perform very well. Fig. 6 shows the boat detection results obtained by SR-Hough for the two test images at $\mathrm{precision} = 0.849$ and $\mathrm{recall} = 0.880$. Even though the test images include berths, which result in cluttered backgrounds compared with the positive training samples, the SR-Hough method successfully detected the boats. HOG-SVM performed relatively better than in the case of car detection because of the simpler backgrounds and the higher contrast between targets and backgrounds. The precision of SR-Hough decreases at a higher recall than those of other methods, which implies that SR-Hough preferentially detects parts of the target object and returns their locations owing to the robust detection of parts.

C. Airplane Detection

Thirdly, we present an experiment on airplane detection. The study images include 55 airplanes, which were taken over New Chitose Airport, Hokkaido, Japan, and Los Angels Airport, CA, US, by GeoEye-1 with a 0.5-m GSD in the panchromatic channel. This experiment demonstrates the effectiveness of the proposed method when illustrations are used for positive training samples. One issue in conventional object detection techniques based on feature extraction and classifiers is that they require a large set of training data with various appearances of the object class to construct an accurate classifier. However, it is time-consuming and expensive to collect such a large set of positive training samples from satellite images with very high spatial resolution. One benefit of remote sensing for object detection is that images are usually orthorectified, and thus the appearances of the object class do not change by as much as those in ground-shot images. If it is possible to detect instances of an object class with a small set of typical shape information,

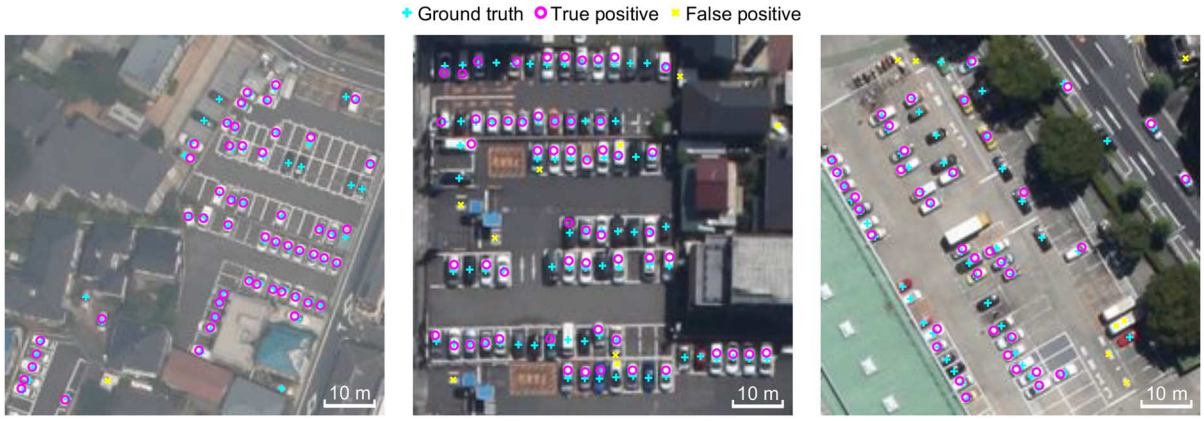

Fig. 5. Car detection results for three test images obtained by SR-Hough at $\text{precision} = 0.896$ and $\text{recall} = 0.754$.

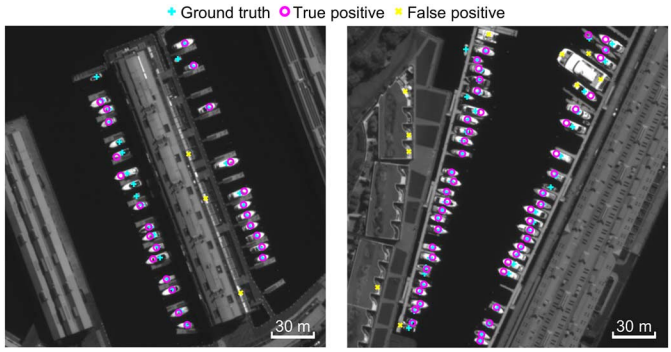

Fig. 6. Boat detection results for two test images obtained by SR-Hough at $\text{precision} = 0.849$ and $\text{recall} = 0.880$.

Fig. 7. Airplane detection results obtained by SR-Hough at $\text{precision} = 0.833$ and $\text{recall} = 0.909$.

such as illustrations, it will reduce the cost of collecting positive training samples.

Seven illustrations of airplanes, as shown in Fig. 3(c), were used as the positive training samples and background patches were randomly sampled from other images with the same GSD. The experiment was conducted with a 1-m GSD to reduce the computational cost. The patch sizes are 25×25 pixels for SR-Hough and the Hough forest, and 32×32 pixels for HOG-SVM and SR-TM. 1800 positive and 10 000 negative patches are used for SR-Hough and the Hough forest. We set the size of the

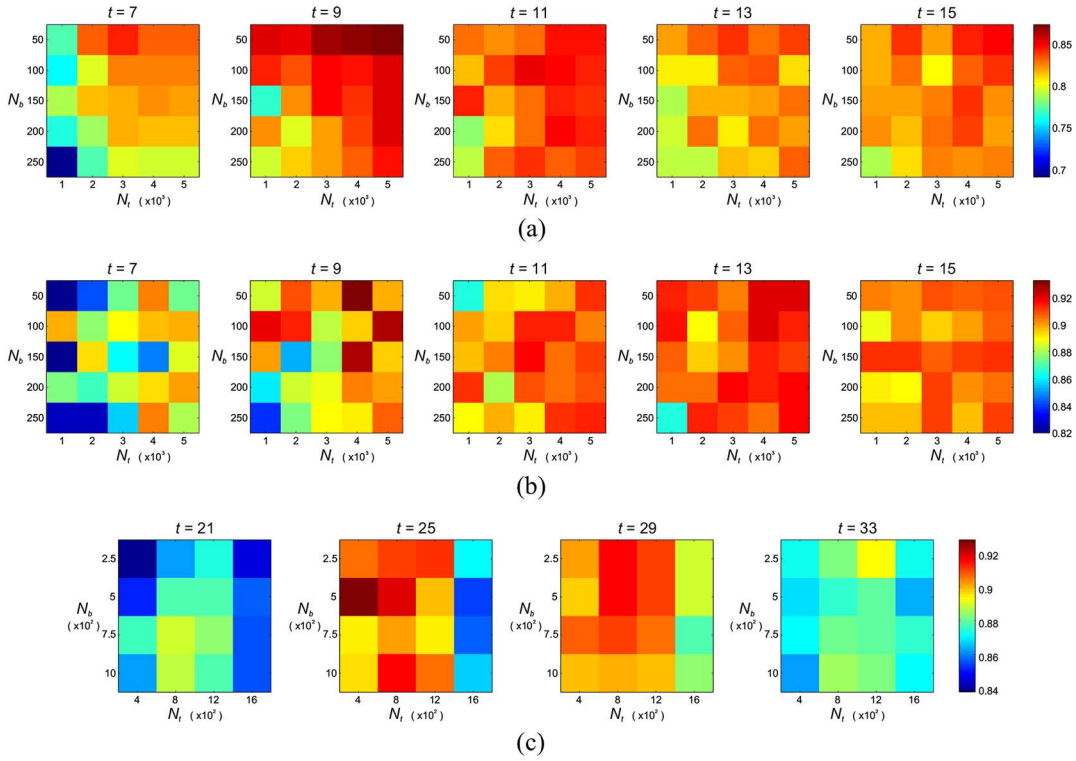

Fig. 8. AUPRC maps with respect to the patch size t for: (a) car; (b) boat; and (c) airplane detection. Each map is obtained by using various combinations of the number of atoms in background and target dictionaries denoted as N_b and N_t, respectively.

background and target dictionaries as $N_b = 500$ and $N_t = 800$, respectively. A comparison of the precision-recall curves for the different methods and the corresponding AUPRC is presented in Fig. 4(c) and Table I, respectively. Fig. 7 shows the detection result for SR-Hough at $\mathrm{precision} = 0.833$ and $\mathrm{recall} = 0.909$. Using only the seven illustrations, SR-Hough accurately detects airplanes and estimates their orientations even with cluttered backgrounds, such as shadows. SR-Hough and SR-TM clearly outperforms HOG-SVM and the Hough forest, which implies the advantage of the sparse-representation-based methods that the target dictionary can represent the essential local features of airplanes and the residual appearance can be explained by the background dictionary. HOG-SVM and the Hough forest failed to learn an accurate classifier, resulting in many FPs, which mainly appeared in areas of the airport with the spatial characteristics of a cross or a line. The proposed method has major potential because it works well with a small set of positive training samples, even with illustrations, in contrast to conventional methods for object detection.

TABLE II
COMPUTATION TIME (SEC) OF TRAINING AND TESTING PHASES FOR CAR, BOAT, AND AIRPLANE DETECTION. TOTAL IMAGE SIZE (MEGAPIXEL) IS PRESENTED FOR EACH OBJECT-CLASS DETECTION

Object	Car		Boat		Airplane	
Phase	Training	Testing	Training	Testing	Training	Testing
Image size	–	0.480	–	0.405	–	1.125
SR-Hough	96	224	109	157	275	1977
Hough forest	219	535	284	220	60	402
HOG-SVM	243	1206	95	1019	124	3124
SR-TM	100	4750	115	2662	179	3771

TABLE III
RATIOS OF CUMULATIVE VALUES CASTED IN TRUE POSITIONS TO THE TOTAL VOTING

	Car	Boat	Airplane
SR-Hough	**0.497**	0.492	**0.231**
Hough forest	0.408	**0.590**	0.098

D. Sensitivity of Parameters and Computation Time

We investigate the influence of parameters, such as the patch size (t), the number of atoms in the background and target dictionaries (N_b and N_t). Fig. 8 shows AUPRC maps with respect to the patch size, with each map being obtained by using various combinations of N_b and N_t. It visualizes the influence of the three parameters for: a) car; b) boat; and c) airplane detection.

For car detection, the AUPRC is large with $t = 9$ and 11, which correspond to $k \approx 0.45$ and 0.55. In each AUPRC map, the combination of the smaller N_b and the larger N_t results in the better performance. It may be because the positive training samples of cars contain some degree of background variety compared with those of boats and airplanes, and thus more sampling of patches for the target dictionary can result in accurate detection of object patches. In addition, since the spatial patterns of cars are simple in the GSD used in this experiment and the optimal patch size is small, the small number of N_b may be enough to avoid representing the target patch by only the background dictionary.

For boat detection, the AUPRC stably shows the best result with $t = 13$ ($k \approx 0.6$) followed by those with $t = 11$ and 15. In contrast to car detection, the smaller N_b not always results in

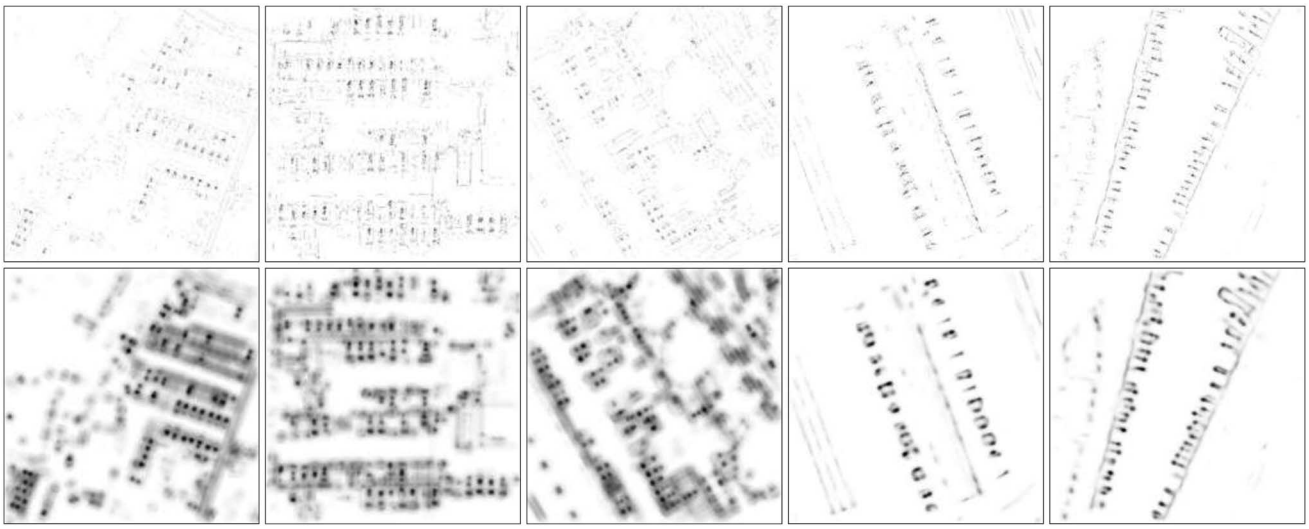

Fig. 9. Hough images for car detection (left three columns) and boat detection (right two columns) obtained by (top) SR-Hough and (bottom) Hough forest.

the better performance. This is because the backgrounds of the positive training samples are simple and the larger patch size requires some amount of N_b to represent various backgrounds.

For airplane detection, the AUPRC is large with $t = 25$ and 29 ($k \approx 0.4$ and 0.45). In contrast to car and boat detection, the AUPRC increases when the size of the atoms in the target dictionary (N_t) is reduced from the initial number of patch sampling. Especially, the detection performance is stably good with $N_t = 800$ and 1200. It proves the effectiveness of discriminative atom selection for the target dictionary. The smaller N_b not always results in the better performance. Some amount of N_b is required owing to the larger patch size to represent various backgrounds.

From these discussions mentioned above, the criteria for setting the parameters can be roughly considered as follows. The optimal patch size can be defined by setting $k \approx 0.5$. The size of the background dictionary (N_b) can be set to approximately t^2 and that of the target dictionary (N_t) can be reduced to the half size of the initial number of positive patches.

The computation time of training and testing is summarized in Table II. For all the methods, the testing phase takes more time than the training phase. The computation cost of HOG-SVM and SR-TM for testing is high owing to a greedy search. The computation time of the proposed method for testing is proportional to the patch size and the number of extracted test patches along edges, which is defined by the total image size and the spatial complexity. Accordingly, when the patch size is small, e.g., $t = 9$ for car detection and $t = 11$ for boat detection, the computation time of SR-Hough is reasonable compared with those of the Hough forest. In contrast, in the case of airplane detection with $t = 25$, the computation time is relatively high.

E. SR-Hough Versus Hough Forest

Here, we discuss the difference between SR-Hough and the Hough forest since the two methods are based on the generalized Hough-transform framework, which comprises the detection of parts and their co-occurrences. The detection performance depends on whether the peak in the Hough image is in the true position of the object. We evaluate the accuracy of Hough voting by investigating the ratio of the cumulative values casted in the true positions to the total voting. Table III shows the ratios for the three examples of object-class detection, and SR-Hough shows stable performance for the variety of objects and backgrounds compared with the Hough forest. This implies that the detection of parts by the proposed method is robust and accurate with a small set of positive training data owing to image decomposition by binary class-specific sparse representations. The class-specific sparse representations can accurately detect only parts with unknown backgrounds because the background effects may be flexibly explained by the background dictionary, whereas random forests may lead to the misclassification of unknown backgrounds as the classification performance is determined by the variety of training data.

Fig. 9 shows the Hough images for car and boat detection obtained by SR-Hough and the Hough forest on the top and bottom rows, respectively. The Hough images for SR-Hough show sparse distributions of votes and sharp peaks relative to those for the Hough forest. SR-Hough casts only one vote for each patch and the voting value reflects the degree of matching of object parts, i.e., when a patch mainly includes a part of the object, the voting value is close to 1; otherwise, it is close to 0. In contrast, the Hough forest casts multiple votes as a result of the bagging approach of random forests, and thus its voting values are distributed between 0.5 and 1.0. Since SR-Hough results in sparse voting and sharp maxima in Hough images, the blurring process of the Hough image is necessary to find maxima.

F. Ship Identification

Finally, we show an example of specific-object identification. Fig. 10(a) shows the target ship in this experiment and Fig. 10(b) shows the test image extracted from the study image used for boat detection. This image captures two ships manually

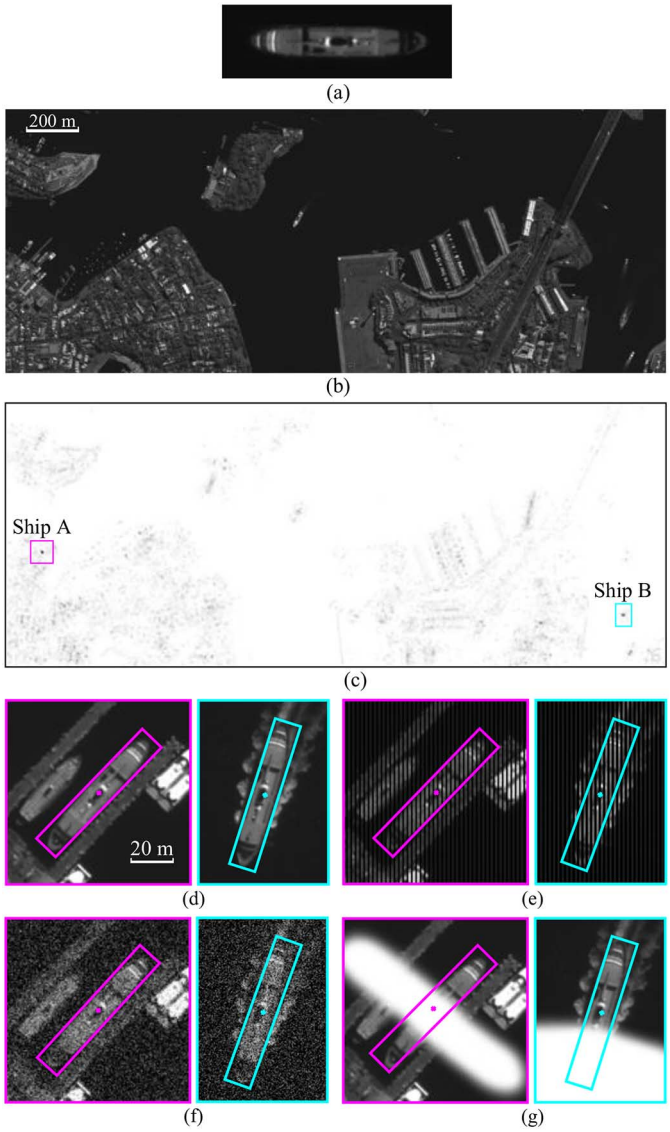

Fig. 10. (a) Target ship and (b) test image used for ship identification. (c) Gaussian-blurred Hough image obtained by SR-Hough with two rectangles indicating subimages of TPs (ships A and B). Enlarged subimages of detected objects for (d) original data and synthetic data with (e) stripe noise, (f) Gaussian noise, and (g) partial occlusion.

recognized as being of the same type as the target. Since this ship has informative spatial features, its orientation is estimated as well as its location. The experiment was conducted with a 1-m GSD and a patch size of 13×13 pixels to reduce the computational cost.

Fig. 10(c) shows a Gaussian-blurred Hough image with two rectangles located at the top two maxima. The corresponding enlarged subimages are shown in Fig. 10(d), where each rectangle indicates a detection hypothesis illustrating the estimated location and orientation of the ship. To examine the robustness of the proposed method against noise and partial occlusion, we added synthetic stripe noise, Gaussian noise, and a partial occlusion to the test image. As shown in Figs. 10(e)–(g), the SR-Hough method can still detect the two ships as higher-order maxima with accurate locations and orientations. Table IV shows the order of the maximum detected as the TP and the ratio of the maximum divided by that of the first FP. A comparison between the SR-Hough methods with and without discriminative atom selection for the target dictionary is presented to investigate the effectiveness of the discriminative dictionary construction. Many of the ratios are increased by discriminative atom selection, resulting in higher detection orders of the TPs. This experiment demonstrates that the proposed method is also useful for specific-object identification and that it works well even with noise and partial occlusion when the target has informative spatial features. Note that if the number of expected target is unknown, a large amount of FPs can be produced by the identification procedure. Therefore, additional user interpretation of the result is necessary in practical use.

TABLE IV
DETECTION ORDERS AND RATIOS OF THE MAXIMUM IN THE HOUGH IMAGE DIVIDED BY THAT OF THE FIRST FP WITH AND WITHOUT DISCRIMINATIVE ATOM SELECTION IN TARGET DICTIONARY CONSTRUCTION

Data	Random sampling				Discriminative selection			
	Ship A		Ship B		Ship A		Ship B	
	Order	Ratio	Order	Ratio	Order	Ratio	Order	Ratio
Original	1	2.065	2	**1.631**	1	**2.073**	2	1.617
Stripe noise	1	1.279	8	0.646	1	**1.485**	4	**0.795**
Gaussian noise	1	1.588	2	1.083	1	**1.877**	2	**1.208**
Occlusion	5	0.934	1	**1.352**	2	**1.025**	1	1.329

IV. CONCLUSION

We have proposed a novel method for object detection based on sparse image representations and demonstrated its effectiveness for remote sensing imagery. Parts of class objects or a specific object are detected by the sparse representation of each patch using learned target and background dictionaries. Whenever a part is detected, the center of the object is activated within the Hough transform framework so that the co-occurrence of parts can be used for object detection. The proposed method can efficiently detect instances of an object class or specific object with a small set of positive training samples since the essential object parts of the target are matched with target atoms while the residuals are explained by a background dictionary. We have shown that the proposed method leads to state-of-the-art object detection results in experiments on car, boat, and airplane detection as well as ship identification with a cluttered background, noise, and partial occlusion. Our future work includes determining the number and size of raw patches suitable for dictionary construction and further investigation of how to construct discriminative class-specific dictionaries.

REFERENCES

[1] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2001, pp. 511–518.
[2] D. G. Lowe, "Discriminative image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, 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. Comput. Vis. Pattern Recog.*, 2005, pp. 886–893.

[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, Jun. 1998.

[5] Y. Taigman, M. Yang, M. A. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 1701–1708.

[6] S. E. Palmer, "Hierarchical structure in perceptual representation," *Cognit. Psychol.*, vol. 9, pp. 441–474, 1977.

[7] E. Wachsmuth, M. W. Oram, and D. I. Perrett, "Recognition of objects and their component parts: Responses of single units in the temporal cortex of the macaque," *Cerebral Cortex*, vol. 4, pp. 509–522, 1994.

[8] N. K. Logothetis and D. L. Sheinberg, "Visual object recognition," *Annu. Rev. Neurosci.*, vol. 19, pp. 577–621, 1996.

[9] I. Biederman, "Recognition-by-components: a theory of human image understanding," *Psychol. Rev.*, vol. 94, pp. 115–147, 1987.

[10] S. Ullman, High-Level Vision: Object Recognition and Visual Cognition. Cambridge, MA, USA: MIT Press, 1996.

[11] S. Agarwal and D. Roth, "Learning a sparse representation for object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2002, pp. 113–130.

[12] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *Proc. Comput. Vis. Pattern Recog.*, 2003, vol. 2, pp. 264–271.

[13] 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.

[14] D. H. Ballard, "Generalizing the Hough transform to detect arbitrary shapes," *Pattern Recognit.*, vol. 13, no. 2, pp. 111–122, 1981.

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

[16] H. Grabner, T. T. Nguyen, B. Gruber, and H. Bischof, "On-line boosting-based car detection from aerial images," *ISPRS J. Photogramm. Remote Sens.*, vol. 63, pp. 382–396, 2008.

[17] T. Moranduzzo and F. Melgani, "Automatic car counting method for unmanned aerial vehicle images," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 3, pp. 1635–1647, Mar. 2014.

[18] C. Corbane, L. Najman, E. Pecoul, L. Demagistri, and M. Petit, "A complete processing chain for ship detection using optical satellite imagery," *Int. J. Remote Sens.*, vol. 31, no. 22, pp. 5837–5854, 2010.

[19] C. Zhu, H. Zhou, R. Wang, and J. Guo, "A novel hierarchical method of ship detection from spaceborne optical image based on shape and texture features," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 9, pp. 3446–3456, Sep. 2010.

[20] 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, Apr. 2012.

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

[22] M. Elad, Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing. New York, NY, USA: Springer, 2010.

[23] J. Mairal, M. Elad, and G. Sapiro, "Sparse representation for color image restoration," *IEEE Trans. Image Process.*, vol. 17, no. 1, pp. 53–69, Jan. 2008.

[24] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution as sparse representation of raw image patches," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2008, pp. 1–8.

[25] 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, Feb. 2009.

[26] Y. Chen, N. M. Nasrabadi, and T. D. Tran, "Sparse representation for target detection in hyperspectral imagery," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 3, pp. 629–640, Jun. 2011.

[27] Y. Zhang, B. Du, and L. Zhang, "A sparse representation-based binary hypothesis model for target detection in hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 3, pp. 1346–1354, Mar. 2015.

[28] B. Du and L. Zhang, "A discriminative metric learning based anomaly detection method," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 11, pp. 6844–6857, Nov. 2014.

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

[30] Y. Pati, R. Rezaiifar, and P. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with application to wavelet decomposition," in *Proc. Asilomar Conf. Signals*, 1993, vol. 1, pp. 40–44.

[31] R. Rubinstein, A. Bruckstein, and M. Elad, "Dictionaries for sparse representation modelling," *Proc. IEEE*, vol. 98, no. 6, pp. 1045–1057, Jun. 2010.

[32] J. Wright *et al.*, "Sparse representation for computer vision and pattern recognition," *Proc. IEEE*, vol. 98, no. 6, pp. 1031–1044, Jun. 2010.

[33] 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, Nov. 2006.

[34] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Discriminative learned dictionaries for local image analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2008, pp. 1–8.

[35] Q. Zhang and B. Li, "Discriminative k-SVD for dictionary learning in face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2010, pp. 2691–2698.

[36] M. Yang, X. F. L. Zhang, and D. Zhang, "Fisher discrimination dictionary learning for sparse representation," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 543–550.

[37] J. Mairal, F. Bach, and J. Ponce, "Task-driven dictionary learning," *IEEE Trans. Pattern Anal. Mach. Intell.* vol. 34, no. 4, pp. 791–804, Apr. 2012.

[38] Z. Jiang, Z. Lin, and L. S. Davis, "Label consistent K-SVD: Learning a discriminative dictionary for recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2651–2664, Nov. 2013.

[39] A. Rosenfeld and G. J. Vandenburg, "Coarse-fine template matching," *IEEE Trans. Syst. Man Cybern.*, vol. 7, no. 2, pp. 104–107, Feb. 1977.

[40] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. Alvey Vis. Conf.*, 1988, pp. 147–151.

[41] J. F. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8, no. 6, pp. 679–698, Nov. 1986.

[42] V. Boominathan and K. Sri Rama Murty, "Speaker recognition via sparse representations using orthogonal matching pursuit," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2012, pp. 4381–4384.

Naoto Yokoya (S'10–M'13) received the M.Sc. and Ph.D. degrees in aerospace engineering from the University of Tokyo, Tokyo, Japan, in 2010 and 2013, respectively.

From 2012 to 2013, he was a Research Fellow with Japan Society for the Promotion of Science, Tokyo, Japan. Currently, he is an Assistant Professor with the University of Tokyo. His research interests include image analysis and data fusion in remote sensing.

Akira Iwasaki received the M.Sc. degree in aerospace engineering and the Doctoral degree in engineering from the University of Tokyo, Tokyo, Japan, in 1987 and 1996, respectively.

He joined the Electrotechnical Laboratory, where he engaged in research on space technology and remote sensing system, in 1987. Currently, he is a Professor with the University of Tokyo.