Target Recognition in SAR Image via Keypoint based Local Descriptor—Foundation

Ganggang Dong and Jocelyn Chanussot, Fellow, IEEE

***Abstract*—This paper considers target characterization and recognition in radar images with keypoint-based local descriptor. Most of the preceding works rely on the global features or raw intensity values, and hence produce the limited recognition performance. Moreover, the global features are sensitive to the real-world sources of variability, such as aspect view, configuration, and incidence angle changes, clutter, articulation, and occlusion. Keypoint-based local descriptor was developed as a powerful strategy to address invariance to contrast change and geometric distortion. This property inspires us to investigate whether the family of local features are relevant for radar target recognition. Most of the preceding works typically devote to finding the correspondences between a collected image and a reference one. The representative applications include image register and change detection. Little work was pursued to target recognition in SAR images. This is because the huge number of local descriptors resulting from radar images make the computational cost and memory consumption unacceptable. To handle the problems, this paper develops two families of methods. The proposed methods are used to achieve target recognition by means of local descriptors. Our first solver refers to building multiple linear regression models, and addresses the problem by the theory of sparse representation. The second scheme rebuilds a new feature by the feature quantization skill, from which the inference can be drawn. Multiple comparative studies are pursued to verify the performance of detectors and descriptors popularly used. The source code was publicly released on https://ganggangdong.github.io/homepage/.**

Keywords*—Target recognition, SAR, keypoint, local descriptor, sparse representation, feature quantization, classification.

I. Introduction

WITH the development of integrated circuit and manufacturing technology, the resolution of synthetic aperture radar in range and azimuth is capable to achieve target recognition. However, images taken from various sensor platforms are too huge to be handled by analyst timely. This situation produces an urgent need for automatic image interpretation [1], [2]. Though many works have been initiated to provide a baseline knowledge of target scattering characteristics, automatic target recognition in radar image is still far from being solved due to the complicated imaging condition. It is nevertheless worth to be addressed because this technology gives great potential for the civil use as well as the military application.

The typical system for radar target recognition is usually composed of three separate phases, prescreening, discrimination, and classification [3]. The work mechanism is shown in Fig. 1, where the input is a scenery of radar image, and the output are the target types. The prescreening stage produces the candidate targets by examining the amplitude of radar signal pixel by pixel [4]. The discrimination stage locates the candidate accurately and generates the orientations [5]. The natural clutter false alarms are rejected by texture features. The classification stage predicts the identifications.

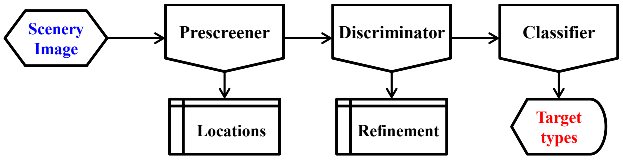


Fig. 1. The mechanism of automatic target recognition from radar image.

A. Background

This paper considers the classification subsystem. The input is a chip image cropped from the scenery, while the output is the identification. The chip image inputted usually contains a single target, radar shadow, and background, as visually shown in Fig. 2. In the conventional methods, the prediction of target type is achieved by feeding a designed feature into a trained classifier. The recognition performance is therefore dependent on the representation and classification scheme.

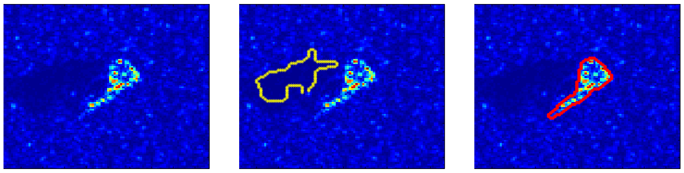


Fig. 2. Illustration of target and radar shadow for radar chip image.

1) Feature Representation: The signature information that distinguishes one target from another is fundamentally determined by the interactions between the incident radar waveform and target physical structure. For radar imagery, the backscattered signal results from multiple scattering mechanisms, *e.g.*, direct backscatter, single-direction double bounce, return-direct multi-bounce, and high-order multiple bounce. The unique imaging mechanism makes feature extraction from radar image

This works was supported by the National Natural Science Foundation of China under Grant 61601481, and by the Hunan Provincial Science Foundation of China under Grant 2016JJ3023.

Dr. Ganggang Dong is with National Laboratory of Radar Signal Processing, Xidian University, Xi'an, China, and with School of Electronic and Information Engineering, Xi'an, China (e-mail: dongganggang@nudt.edu.cn).

Prof. Jocelyn Chanussot is with with the engineering school on Energy, Water and Environment (ENSE3) of the Grenoble Institute of Technology, Grenoble, France (e-mail: jocelyn.chanussot@gipsa-lab.grenoble-inp.fr).

doi:10.20944/preprints201805.0116.v1

much more difficult. The present approaches can be reviewed as follows.

Intensity values. The early methods achieve target recognition by the raw intensity values or the enhanced image. The distance between two sets of pixel values (or enhanced values) are used to make the decision. L. Novak *et al.* propose a new super-resolution image-processing technique that enhances SAR image resolution [6]. The enhanced image is fed into a MSE-classifier to predict the target type. Q. Zhao and J. Principle design a support vector machine, into which the raw intensity values are pushed directly [7]. J. Thiagarajan propose to feed the randomly projected coefficients of image into sparse representation-based classifier [8]. This kind of methods are dependent on the trained classifier more and on feature less.

Projection coefficients. This kind of methods refer to constructing a linear subspace by the intensity vectors. A feature representation is then defined by the projected coefficients, such as principal component analysis, independent component analysis. The physical meaning is therefore ambiguous. M. Liu *et al.* present a statistical model embedding the locality preserving property to extract the maximum amount of desired information from the data [9]. Y. Huang *et al.* propose to preserve the global and local discriminative information based on the tensor representation [10]. Z. Cui *et al.* generate the pattern feature of SAR image by a variant of non-negative matrix factorization [11].

Filter banks. In some works, the feature is defined by a set of filter banks, such as Fourier transform, Gabor filters, and the analytic signal. R. Patnaik and D. Casasent achieve target recognition by the correlation patter recognition strategy [12], [13]. A set of correlation filters are first generated by Fourier transformed coefficients of the training. The decision is made according to the correlation response between the query and the generated filters. G. Dong *et al.* develop a new method for target recognition [14]–[16]. The target signature information is characterized by an extended analytic signal, the monogenic signal [17], [18]. Sparse representation modeling is built to implement target classification.

Geometric features. The geometric feature means the shape, edge, size of target or radar shadow. This family of features are mainly dependent on the fine segmentation of radar image, which is still an open problem now. J. Park *et al.* discriminate target from clutter by some designed geometric features, such as the minimum projected length, the contrast of the projected length, the energy of the projected length in the frequency domain [19]. J. Zhu *et al.* introduce a famous trick of computer vision, shape context, to exploit the distinguishing characters of the ship targets from radar image [20]. They jointly consider the topology and intensity of scattering points of ship.

Statistical feature. Some researchers employ the image statistics for target recognition, such as the various statistical moment. J. Singh and M. Datcu utilize a chirplet-derived transform and fractional Fourier transform to generate a compact feature descriptor for single-look SAR images [21]. The statistical response resulting from the projections on different planes of the joint time-frequency space is easy to be analyzed. M. Anoon and G. Rezai-rad generate a representation by the Zernike moments [22]. The resulting feature is of linear transformation invariance and robustness in the presence of the noise. The similar thought was employed in [23]. P. Bolourchi *et al.* propose a feature descriptor by Radial Chebyshev moment, a discrete orthogonal moment with some distinctive advantages [24].

Learned feature. Target recognition via the learned feature is a recent research hotspot. S. Chen *et al.* propose to learn the hierarchical features from massive training data by convolutional neural networks [25]. The similar thoughts can be found in [26]–[28]. S. Deng *et al.* introduce stacked autoencoder for target classification [29]. The reshaped image is specified as the visible layer, while the latent states are used to classification. Though performed well, this family of features are computationally unattractive. In addition, they present a heavy demand for the hardware configuration.

Scattering center model. Radar energy backscattered from the object contains the key information that distinguishes one target from another. An intuitive idea is to define a feature by the scattering center models. L. Potter and R. Moses pursue the preliminary studies [1]. They present a framework for feature extraction predicated on parametric models for the radar returns. The developed models are motivated by the scattering behavior predicted by the geometrical theory of diffraction. J. Zhou *et al.* propose a global scattering center model established offline using range profiles at multiple viewing angles, with which features at different target poses can be conveniently predicted. B. Ding *et al.* introduce a new statistics-based metric to measure the distance between the attributed scattering center models [30]. This family of features are difficult to be flexibly generalized.

Though multiple schemes were presented previously, feature extraction is still far more to be solved, especially for the real-word applications. The invariance to the real-world sources of variability should be further studied.

2) Classification Learning: The extracted feature is used to determine the class to which a detected target belongs by the knowledge learned from the training. It is a typical application of patter recognition in radar image. The current approaches are reviewed as follows. ***KNN*** is of the most fundamental and explicit strategy. It predicts the class membership according to the similarity between the probe and the gallery. The key is how to define an appropriate distance metric for the designed feature. ***Kernel-based Classifier***. This family of method projects the original data into an implicit feature space whose dimension can be as high as possibly or even infinite. The class separability can be then enhanced. The most representative is support vector machine learning [7]. ***Regression analysis*** skills are popularly used in neural network configuration [25], [29]. It models the relationship between a scalar dependent variable and a set of explanatory variables. The response is the target type, whose output is the probability of the class label taking one each of the possible values. Sparse representation-based classification is a specialty of regression analysis [32].

B. Contributions

Though many studies were pursued over the years, most of them rely on the global feature, resulting in limited performance. In addition, this kind of feature are sensitive to the real-world source of variability. To handle the problems, this paper considers keypoint-based local descriptor for target recognition. Two families of methods, following the thought of sparse representation and feature quantization, are developed. The pipeline is displayed in Fig. 3. To our knowledge, the relevance of keypoint-based scheme in target recognition has been investigated seldom. We aim at studying to which extent the local descriptor can improve the recognition performance. We intend to open a new door for target recognition under the non-literal conditions. Our contributions therefore include:

- the comprehensive review of the preceding works on feature extraction from radar image,
- the tune of keypoint-based local descriptor for radar target recognition,
- the development of two families of schemes to implement target classification with the local descriptor,
- the evaluation of proposed strategy with multiple comparative studies.

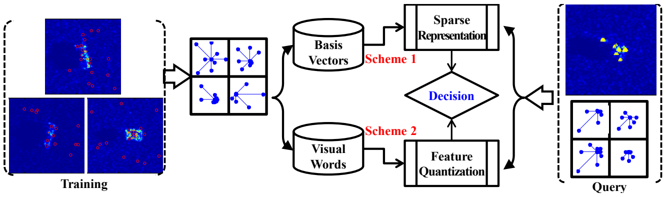


Fig. 3. Pipeline of proposed framework.

C. Organization

The rest of this paper is organized as follows. Section II reviews the representative keypoint detectors and local descriptors. Section III develops two families of methods for target recognition. Section IV verifies the proposed schemes with multiple comparative experiments. Section V concludes this paper.

II. Keypoint-based Local Descriptor

Most of the previous studies achieve target recognition by the global feature. They are not effective to the non-literal conditions. The recent development on machine learning prove that the local region description could produce very powerful cues. Compared to the global feature, the keypoint-based local descriptor is much more robust to real-world source of variability. This property motivates us to achieve target recognition with the local descriptor. This section provides a simple review of related studies from two aspects, the detection of keypoint and the representation of local feature.

A. The Detection of Keypoint

Keypoint detection refers to checking image pattern which differs from the immediate neighborhood. The fashion of representation could yield a high repeatability, *i.e.*, the keypoints can be extracted reliably and are often found again at the similar locations in other images of the same object or scene. For radar chip image, the keypoints are undoubtedly located in the target imaging region, or radar shadow. The popularly used approach to keypoint detection includes difference of Gaussian (DoG), Harris corner detector, Hessian blob-like structure detector, and the variants.

1) Difference of Gaussian: DoG is an approximation of Laplace of Gaussian [33], and much faster to compute. A scale-space $\mathcal{S}(x, y, \sigma)$ is first built by convolving the image $I(x, y)$ with a Gaussian low-pass filter $\mathcal{G}(x, y, \sigma)$ parameterized by a standard deviation σ,

$$\mathcal{S}(x, y, \sigma_s) = \mathcal{G}(x, y, \sigma_s) * I(x, y).$$

where σ_s is a function of index $s = \{0, 1, \ldots, s_{max}\text{-}1\}$. DoG images is the difference between two successive layers

$$\mathcal{D}(x, y, \sigma_s) = \mathcal{S}(x, y, \sigma_{s+1}) - \mathcal{S}(x, y, \sigma_s).$$

To achieve scale-invariance, image pyramids are usually built. The number of octave is determined by the size of image. Image in the higher-order octave is obtained by downsampling the one of the previous octave in a factor of 2. Keypoint is defined as the local extrema in 3-dimension space (x, y, σ). It is checked by comparing every pixel to the eight neighboring pixels in the current scale and the nine pixels in the scales above and below. If a pixel is larger or smaller than all of its neighbors, it is accepted as a preliminary keypoint candidate. The implementation flow of DoG is pictorially shown in Fig. 4.

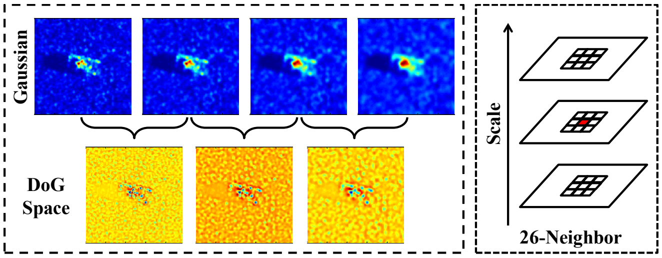


Fig. 4. The illustration of DoG detector. The pixel (in red) is compared to the eight neighbors in current scale, and the nine neighbors in the previous and next scales.

2) Harris Detector: Harris corner detector is probably one of the earliest method for keypoint detection [34]. It is based on the eigenvalues of second-moment matrix (or autocorrelation). The candidates with low contrast are then filtered by thresholding the values of the built matrix,

$$\mathcal{M}(x, y, \sigma) = \det(\mathcal{H}_a(x, y, \sigma)) - \lambda \cdot trace(\mathcal{H}_a(x, y, \sigma))^2.$$

The Harris matrix is

$$\mathcal{H}_a(x, y, \sigma) = \begin{bmatrix} \mathcal{S}_x^2(x, y, \sigma) & \{\mathcal{S}_x \cdot \mathcal{S}_y\}(x, y, \sigma) \\ \{\mathcal{S}_y \cdot \mathcal{S}_x\}(x, y, \sigma) & \mathcal{S}_y^2(x, y, \sigma) \end{bmatrix}$$

where $\mathcal{S}_x$, $\mathcal{S}_x$ are the convolution of the Gaussian first-order derivative $\frac{\partial}{\partial x}\mathcal{G}$, $\frac{\partial}{\partial y}\mathcal{G}$ with image $I(x, y)$. The parameter λ is to balance the determinant and trace of Harris matrix, and usually set between 0.04∼0.06.

To achieve scale-invariance, Lindeberg introduced the concept of automatic scale selection [35]. He propose to assignment the detected interest points with their own characteristic scale. Mikolajczyk and Schmid further present a refined strategy, Harris-Laplace detector, by which scale-invariant feature detectors with high repeatability can be created [36]. The location of keypoint is selected by the determinant, while the scale is determined by the Laplacian operator.

3) Hessian Detector: Hessian detector defines the keypoints as the ones localized in space at the local maxima of the Hessian determinant and in scale at the local maxima of the Laplacian-of-Gaussian. For 2-D function $I(\mathbf{x}), \mathbf{x} = [x, y]$, the second-order Taylor's expansion is expressed as

$$I(\mathbf{x}_0 + \triangle\mathbf{x}) \approx I(\mathbf{x}_0) + \triangle\mathbf{x}^T \nabla I(\mathbf{x}_0) + \triangle\mathbf{x}^T \mathcal{H}(\mathbf{x}_0)\triangle\mathbf{x}$$

The Hessian matrix at each point location is

$$\mathcal{H}_e(x, y, \sigma) = \begin{bmatrix} \mathcal{S}_{xx}(x, y, \sigma) & \mathcal{S}_{xy}(x, y, \sigma) \\ \mathcal{S}_{yx}(x, y, \sigma) & \mathcal{S}_{yy}(x, y, \sigma) \end{bmatrix}$$

where $\mathcal{S}_{xx}$, $\mathcal{S}_{xy}$, $\mathcal{S}_{yy}$ are the convolution of Gaussian second-order derivative $\frac{\partial^2}{\partial x^2}\mathcal{G}$, $\frac{\partial^2}{\partial x \partial y}\mathcal{G}$, $\frac{\partial^2}{\partial y^2}\mathcal{G}$ with image $I(x, y)$. The blob-like structures are detected by the determinant of Hessian matrix and the trace of Hessian matrix (Laplacian),

$$\det(\mathcal{H}_e) = \mathcal{S}_{xx}\mathcal{S}_{yy} - \lambda' \cdot \mathcal{S}_{xy}\mathcal{S}_{yx}$$

where λ' is a weight parameter. For the 9×9 filter in size and $\sigma = 1.2$, the weight parameter is approximated as

$$\frac{\|\mathcal{S}_{xy}(1.2)\|_F \|\mathcal{D}_{xx}(9)\|_F}{\|\mathcal{S}_{xy}(1.2)\|_F \|\mathcal{D}_{xy}(9)\|_F} = 0.912 \backsimeq 0.9$$

where $\mathcal{D}_{xx}$, $\mathcal{D}_{xy}$, $\mathcal{D}_{yy}$ are the approximations of $\mathcal{S}_{xx}$, $\mathcal{S}_{xy}$, $\mathcal{S}_{yy}$[1]. By assigning each detected keypoints its own characteristic scale, the scale invariance can be achieved.

To boost the computational efficiency, H. Bay *et al.* also present a fast version of Hessian-Laplace detector, Fast-Hessian [37], [38], where the integral image is used to circumvent image derivative operations.

4) Features from Accelerated Segment Test: E. Rosten and T. Drummond propose a novel efficient approach to corner detection, features from accelerated segment test (FAST) [39], [40]. The designed segment test criterion operates by considering a circle of sixteen pixels around the keypoint candidate, as illustrated in Fig. 5.

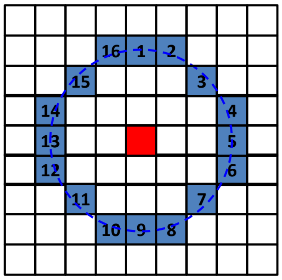


Fig. 5. Illustration of FAST. The pixel in red is the center of a candidate corner, while the 16 pixels in circle are considered in corner detection.

A corner p is defined if there exists a set of n contiguous pixels in the circle which are all brighter than the intensity of the candidate pixel I_p plus a threshold τ, or all darker than $I_p - \tau$. The parameter of n is set as twelve, with which a very large number of candidates can be excluded. Considering the computational efficiency, only four neighborhood pixels at $\{1, 5, 9, 13\}$-site are tested. Therefore, if a keypoint is detected, then at least three of these values should all be brighter than $I_p + \tau$ or smaller than $I_p - \tau$. For a candidate p, these sixteen sites on the circle are noted as $s_1, s_2, \ldots, s_{16} \in \mathcal{N}_p$. The pixels at those sites can be categorized as one of three states,

$$I_{s_i} \Rightarrow \begin{cases} \text{darker} & I_p \leqslant I_p - \tau \\ \text{similar} & I_p - \tau < I_p < I_p + \tau \\ \text{brighter} & I_p \geqslant I_p - \tau \end{cases}$$

The detected keypoints are then refined by non-maximal suppression trick.

B. Feature Representation around Keypoint Detected

Have detected keypoint, a pair of pixel coordinates, another key issue is how to characterize the neighborhood around the point, *i.e.*, define an invariant feature descriptor.

1) SIFT: SIFT may be the most popularly used descriptor [41]. It is defined as the histogram of gradient orientation weighted by magnitude and a Gaussian window. The dominant orientation is estimated to achieve rotational invariance.

For image pyramid $\mathcal{S}(x, y, \sigma)$, the gradient magnitude and orientation are computed at all scales and octaves,

$$\text{M}agnitude \Rightarrow \sqrt{\mathcal{S}_x^2 + \mathcal{S}_y^2}$$

$$\text{O}rientation \Rightarrow \arctan\big(\frac{\mathcal{S}_y}{\mathcal{S}_x}\big)$$

where $\mathcal{S}_x = \frac{\partial \mathcal{S}}{\partial x}$ and $\mathcal{S}_y = \frac{\partial \mathcal{S}}{\partial y}$ are the partial derivatives along x- and y-axis directions. The gradient orientation weighted by magnitude and a Gaussian window is used to produce a 3-D histogram. The first two dimensions correspond to the spatial location, and the additional dimension to gradient orientation. Each pixel within the local region contributes to the histogram depending on the location, gradient orientation and magnitude. Image gradient computed around every keypoint is integrated to the 3D histogram, resulting $2 \times 2 \times 2$ bins, each of which is incremented by gradient magnitude multiplied by a weight inversely related to distance between the location and keypoint.

The peak value of histogram and those ones larger than 80% peak are defined as the dominant orientation. A square neighborhood around each point with a size depending on the scale is cropped. It is inversely rotated by the dominant orientation. To assure the scale invariance, image gradients are calculated at the same scale to which the keypoint belongs. Local descriptor is obtained by concatenating the histograms, producing a 128-element-vector, as shown in Fig. 6.

2) GLOH: GLOH is an extended version of SIFT [42], in which a new quantization of location is developed. It defines the local patch as a log-polar grid with 3 bins in radial direction and 8 bins along the angular direction, and hence results in 8+8+1=17 bins of location. The radius for the outer circulars

[1] The detail derivation can be found in the preceding works [37].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 May 2018 doi:10.20944/preprints201805.0116.v1

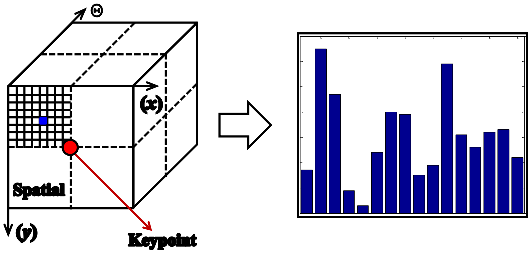


Fig. 6. The generation of SIFT descriptor.

are set according to the task at hand. The central circle regio is handled as a single bin, while the outer circular region are divided into 8 bins equally distributed along angula direction, in each of which the gradient orientation weighte by magnitude is quantized into 8 levels. The generation o histogram is similar to SIFT. The local descriptor is obtaine by concatenating the histograms of all sectors. GLOH refine the division of location, and hence results in an improvement on performance, as proved in [42]. The comparison of SIFT and GLOH are pictorially shown in Fig. 7.

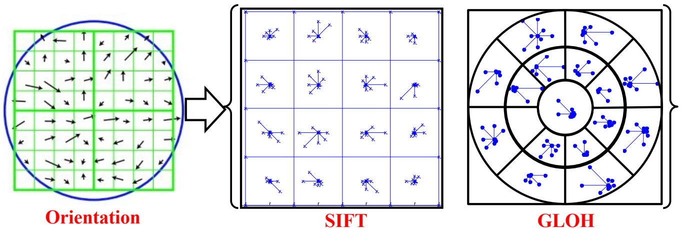


Fig. 7. Square grid for SIFT and log-polar sector for GLOH.

3) Adaptive Binning Strategy: The conventional descriptor usually divides the neighborhood around keypoint into some fixed cells, *e.g.*, 4×4 grids for SIFT, 1+8+8+8 sectors for GLOH, in each of which the gradient orientation is quantized. The levels of quantization for gradient orientation is also fixed. This mode of quantization may reduce the discriminative power. A. Sedaghat and H. Ebadi propose to separate the inner circle and the outer circular regions into different radial sectors. The gradient orientations within each sector is quantized into different levels [43].

For each keypoint detected, its neighborhood is divide into n non-overlapping circular rings, $\mathcal{R}_1$, $\mathcal{R}_2$, …, $\mathcal{R}_n$, as similar to GLOH. Each circular region $\mathcal{R}_i$ is then divided into N_i cells equally distributed along the angular direction $\{\mathcal{R}_i(j)\}_{j=1}^{N_i}$. The gradient orientations in $\mathcal{R}_i(j)$ are further quantized into k_i levels of histogram, *i.e.*, the level of quantization is different from cell to cell, as shown in Fig. 8. The descriptor is obtained by combining the histograms. The dimension of feature descriptor is $\sum_i N_i k_i$.

4) DAISY: Engin Tola *et al.* develop a new local descriptor, DAISY [44], [45]. It circumvents the weighted sums of gradient norms by convolutions of the gradients in specific directions with Gaussian filters.

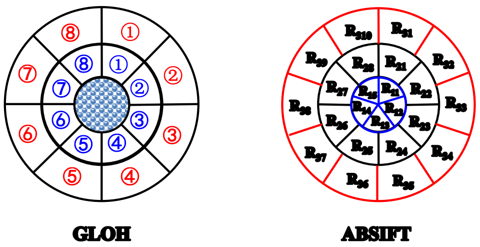


Fig. 8. Illustration of GLOH and AB-SIFT. GLOH only divides the outer circular regions along angular directions, while AB-SIFT separates both the inner circle and the outer circulars into several sectors.

For chip image $I(x,y)$, a set of orientation maps can be defined, $\mathcal{O}_i(x,y) = \left(\frac{\partial I}{\partial i}\right)^+$, where i is the quantized direction, and $(\cdot)^+ = \max\{\cdot, 0\}$. The orientation maps are convolved with Gaussian kernel parameterized by σ to generate different size of local regions,

$$\mathcal{O}_i^\sigma(x,y) = \mathcal{O}_i(x,y) * \mathcal{G}_\sigma(x,y).$$

DAISY is defined as a vector whose entries are the coefficients resulting from the convolved orientation maps located on concentric circles centered on the location,

$$h_\sigma(x,y) = [\mathcal{O}_1^\sigma(x,y), \mathcal{O}_2^\sigma(x,y), \ldots, \mathcal{O}_N^\sigma(x,y)].$$

Considering M circular cells, the descriptor is obtained by concatenating the normalized vectors

$$\begin{gathered}[h_{\sigma_1}^T(x,y), h_{\sigma_1}^T(d_1(x,y,R_1)), \ldots, h_{\sigma_1}^T(d_n(x,y,R_1))\\ h_{\sigma_2}^T(d_1(x,y,R_2)), \ldots, h_{\sigma_2}^T(d_n(x,y,R_2))\\ \vdots\\ h_{\sigma_M}^T(d_1(x,y,R_M)), \ldots, h_{\sigma_M}^T(d_n(x,y,R_M))]\end{gathered}$$

where $d_j(x,y,R_i)$ denotes the sites with R_i distance from (x,y) along the j-direction. The orientations are quantized into n levels.

5) SURF: H. Bay *et al.* present an improved version of Hessian detector, Fast-Hessian, by which a local descriptor, Speeded-Up Robust Features is defined [37], [38]. They first detect blob-like structure with the Hessian matrix, *i.e.*, a second-order derivative of Gaussian filtered image. The derivative operation is achieved by means of integral image. The local descriptor is defined as the distribution of first-order Haar wavelet response.

Given image $I(x,y)$, the Haar-like wavelet response is calculated along the x- and y- axis direction in a circular neighborhood around the keypoint. The wavelet response weighted with a Gaussian centered at the keypoint is represented as vectors in a space. The horizontal response strengthens along the abscissa, while the vertical response strengthens along the coordinate. The Haar-like wavelet responses within a sliding orientation window covering an angle of $\frac{\pi}{3}$ are summed to estimate the dominant orientation. The local descriptor is defined on a square region centered around the keypoints, and oriented along the dominant orientation. The region is further split up into smaller 4×4 square cells, in each of which two simple features at 5×5 spaced sampling points are computed.

doi:10.20944/preprints201805.0116.v1

Denote by h_x, h_y the Haar wavelet response in horizontal and vertical directions. The local descriptor is formed as

$$\Big[\sum h_x, \sum h_y, \sum |h_x|, \sum |h_y|\Big]^T.$$

III. Classification

Keypoint-based local descriptor is initially developed to find the correspondences between a pair of images. The representative application includes image register and change detection. Though studied widely, little work is devoted to radar target recognition. The local descriptors resulting from radar images may be the order of millions, and hence could not be handled as usual. To solve the problem, this paper proposes two methods. The first casts the recognition problem as one of classifying among multiple linear regression models [32]. The latter produces a single new feature for each image by encoding the local descriptors.

Given N labeled images $I_1, I_2, \ldots, I_N$ from K distinct classes, the task of target recognition is to predict the class membership of query using the knowledge learned from the labeled ones. For image I_i, we extract the local descriptors $\mathcal{V}_i^1$, $\mathcal{V}_i^2$, …, $\mathcal{V}_i^{n_i}$ around n_i keypoints. The total number of local descriptors available for training is $n = \sum_i n_i$. Target recognition refers to predict the class identity of I_q according to its local descriptors $\mathcal{V}_q^1$, $\mathcal{V}_q^2$, …, $\mathcal{V}_q^{n_q}$.

A. Solver 1: Sparse Representation

Our first solver proposes to build multiple linear regression models with the local descriptors. The descriptors available for training play the role of regressor, while the one of query is the response. The regression coefficients is obtained by optimizing ℓ_1-norm minimization. The theory of sparse representation offers the key to address the problem [32]. We first represent the descriptors extracted from query $\mathcal{V}_q^i$ as a linear combination of those resulting from the training,

$$\begin{aligned} \mathcal{V}_q^i =& \mathcal{V}_1^1\alpha_1^1 + \mathcal{V}_1^2\alpha_1^2 + \cdots + \mathcal{V}_1^{n_1}\alpha_1^{n_1} + \\ & \mathcal{V}_2^1\alpha_2^1 + \mathcal{V}_2^2\alpha_2^2 + \cdots + \mathcal{V}_2^{n_2}\alpha_2^{n_2} + \cdots \\ & \mathcal{V}_N^1\alpha_N^1 + \mathcal{V}_N^2\alpha_N^2 + \cdots + \mathcal{V}_N^{n_N}\alpha_1^{n_N} \end{aligned} \tag{1}$$

where $[\alpha_1^1,\ \alpha_1^2,\ \ldots,\ \alpha_2^1,\ \alpha_2^2,\ \ldots,\ \alpha_N^1, \alpha_N^2, \ldots, \alpha_N^{n_N}]$ are the regression coefficients. This problem is incapable to be handled directly due to the huge number of descriptors, making the computation and memory unacceptable. A feasible method is to represent the query only by the related descriptors, and ignore the remaining. Following this thought, this paper develops a prescreener, with which the descriptors unrelated are filtered out. The preserved samples are used to represent the query. The key issue is therefore to design the prescreener.

For each descriptor of query $\mathcal{V}_q^i$, this paper first computes the linear correlation response with all of the local descriptors available for training,

$$r_i = (\mathcal{V}_q^i)^T[\mathcal{V}_1^1, \mathcal{V}_1^2, \ldots, \mathcal{V}_1^{n_1}, \ldots, \mathcal{V}_N^1, \mathcal{V}_N^2, \ldots, \mathcal{V}_N^{n_N}].$$

The (dis)similarity between a pair of local descriptors is measured by the Euclidean distance metric. The other measurement could play a similar role. To filter the redundancy out, only the most correlated descriptors are kept, while the remaining are ignored. We sort the correlation response r_i in a descending order, and hold the former descriptors $\mathcal{P} = \mathbf{b}_1$, $\mathbf{b}_2$, …, $\mathbf{b}_L$. They are employed as the basis vectors to represent the descriptor of query

$$\mathcal{V}_q^i = \mathbf{b}_1\alpha_1 + \mathbf{b}_2\alpha_2 + \cdots + \mathbf{b}_L\alpha_L = \mathcal{P}\alpha \tag{2}$$

where $\alpha_1, \alpha_2, \ldots, \alpha_L$ are the weights. The number of atoms in $\mathcal{P}$ is much smaller than the ones available for training, $L \ll n$. The computational cost is then greatly alleviated. Notedly, the dictionary $\mathcal{P}$ is different from descriptor to descriptor.

According to the theory of sparse representation, we expect that most of the entries for α are zero except those associated with the real class identity of query. This is realized by optimizing ℓ_0-norm minimization problem. Thanks to the recent development of compressed sensing [46], the solution of ℓ_0-norm minimization is equal to the one of ℓ_1-norm minimization if α is parsimonious enough,

$$\min_{\alpha} \|\alpha\|_1 \quad s.t. \quad \|\mathcal{V}_q^i - \mathcal{P}\alpha\|_2^2 < \varepsilon \tag{3}$$

where $\|\cdot\|_1$ sums the absolute value of entries. It can be further converted to the unconstrained optimization problem,

$$\min_{\alpha} \big\{\|\alpha\|_1 + \lambda\|\mathcal{V}_q^i - \mathcal{P}\alpha\|_2^2\big\} \tag{4}$$

where the parameter λ balances the fidelity and the sparsity. The optimal representation $\hat{\alpha}$ is used to calculate the reconstruction error

$$e_i^j = \mathcal{V}_q^i - \delta_j(\mathcal{P})\hat{\alpha}_j, \quad j = 1, 2, \ldots, K \tag{5}$$

where the function $\delta_j(\mathcal{P})$ is designed to select those atoms associated with the j-th class, and $\hat{\alpha}_j$ is the corresponding weight coefficients. The overall residual is obtained by accumulating (5) over the descriptors, $\mathbf{e} = \big[\sum_i^{n_q} e_i^1, \sum_i^{n_q} e_i^2, \ldots, \sum_i^{n_q} e_i^K\big]$. The class membership of query I_q is estimated by finding the minimum overall reconstruction error $\arg\min_j \Big\{\sum_i e_i^j\Big\}$.

Since only a small portion of descriptors related are employed to build the multiple linear regression models, the proposed method has the advantage of simplicity and computational efficiency.

B. Solver 2: Feature Quantization

Different from the first method, we propose another solver by feature quantization. It treats an image as a collection of unordered descriptors extracted from the local region, and quantizes them over the "visual words" [47]. A new compact histogram is produced by a predefined codebook for semantic classification.

1) Bag of Visual Words (Bow): BoW represents the local descriptors by a set of visual words. A codebook composing of visual words is first generated by a batch of descriptors. They can be the overall descriptors available, or randomly selected from the training set. Denote by $\mathcal{V}_{(1)}$, $\mathcal{V}_{(2)}$, …, $\mathcal{V}_{(n')}$ the batch of feature selected from the training set. K-means clustering algorithm is employed to generate a codebook $\min_{\mathcal{C}} \sum_{i=1}^{n'} \min_{j=1,\ldots,m} \|\mathcal{V}_{(i)} - \mathcal{C}_j\|^2$, where $\mathcal{C}_1$, $\mathcal{C}_2$, …, $\mathcal{C}_m$

are the clustering centers. The prototype of BoW commits each descriptor to the nearest atom. The hard assignment is too restrictive, and hence produces a coarse reconstruction. Yang *et al.* [48] propose to relax the constraint by sparse coding (SC), enforcing the representation to be with a small number of nonzero entries,

$$\min_{\mathcal{C},\alpha} \sum_{j=1}^{n'} \{\|\mathcal{V}_{(j)} - \mathcal{C}\alpha\|_2^2 + \lambda\|\alpha\|_1\} \tag{6}$$

Wang *et al.* [49] present another trick, locality-constrained linear coding (LLC) by projecting the descriptor into its local-coordinate system with a locality constraint,

$$\min_{\mathcal{C},\alpha} \sum_{j=1}^{n'} \|\mathcal{V}_{(j)} - \mathcal{C}\alpha\|_2^2 + \lambda\|d_j \otimes \alpha\|_2^2 \tag{7}$$

where $\otimes$ denotes the element-wise multiplication, and

$$d_j = \exp\left(\frac{[d(\mathcal{V}_{(j)}, \mathcal{C}_1), d(\mathcal{V}_{(j)}, \mathcal{C}_2), \ldots, d(\mathcal{V}_{(j)}, \mathcal{C}_m)]^T}{\sigma}\right)$$

is a constraint composing of the distance to atoms.

2) Fisher Vectors (FV): Since building an universal and compact vocabulary seems irreconcilable, an alternative idea is to depart the generation of codebook. F. Perronnin *et al.* propose to apply Fisher kernels for image categorization [50], [51]. The core is to characterize a signal with a gradient vector derived from a probability density function which models the generation process of the signal. Gaussian Mixture Models which approximates the distribution of image features is usually employed.

Denote by $\mathcal{V} = [\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_J]$ a set of local descriptors available, and $\Theta = \{\omega_i, \mu_i, \Sigma_i\}_{i=1}^{N}$ the GMM parameters to be estimated, corresponding to weight, mean, and covariance matrix. Each Gaussian distribution represents a word of visual vocabulary. Under an independence precondition, it is capable to produce

$$\mathcal{F}(\mathcal{V}|\Theta) = \sum_{j=1}^{J} \log p(\mathcal{V}_j|\Theta) = \sum_{j=1}^{J} \log(\sum_{i=1}^{N} \omega_i p_i(\mathcal{V}_j|\Theta)) \tag{8}$$

where the component $p_i(\cdot)$ is the i-th Gaussian distribution

$$p_i(x|\Theta) = \frac{1}{\sqrt{(2\pi)^D |\Sigma_i|}} \exp(-0.5(x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i)).$$

Assuming the diagonal covariance matrices, $\sigma_i^2 = diag(\Sigma_i)$, only the derivatives $\frac{\partial \mathcal{F}(\mathcal{V}|\Theta)}{\partial \omega_i}, \frac{\partial \mathcal{F}(\mathcal{V}|\Theta)}{\partial \mu_i}, \frac{\partial \mathcal{F}(\mathcal{V}|\Theta)}{\partial \sigma_i}$ are considered. This leads to the representation which captures the average first and second order differences between local features and each of the GMM centers,

$$\Phi_k^{(1)} = \frac{1}{N\sqrt{\omega_k}} \sum_{p=1}^{N} \alpha_p(k) \left(\frac{\mathcal{V}_p - \mu_k}{\sigma_k}\right)$$

$$\Phi_k^{(2)} = \frac{1}{N\sqrt{2\omega_k}} \sum_{p=1}^{N} \alpha_p(k) \left(\frac{(\mathcal{V}_p - \mu_k)^2}{\sigma_k^2} - 1\right)$$

where $\alpha_p(k)$ is the soft assignment weight of the p-th feature to the k-th Gaussian distribution. The new feature is obtained by stacking the difference $[\Phi_1^{(1)}, \Phi_1^{(2)}, \ldots, \Phi_N^{(1)}, \Phi_N^{(2)}]$.

IV. Experiments and Discussions

This paper develops two kinds of methods to implement target classification by local descriptor. The proposed strategy is validated on MSTAR SAR images, a database collected by a 10 GHz SAR sensor with 1 × 1-foot resolution in range and azimuth. Images of four military vehicles, BMP2, T72, BTR70, and T62 are employed, among which BMP2 and BTR60 are armored personnel carriers, while T72 and T62 are main-battle tanks. BMP2 and T72 have several variants with the structural modifications, noted by the series number, SN_9563, SN_9566, SN_c21 for BMP2, SN_132, SN_812, SN_s7 for T72. BTR60 and T62 are of single configuration. The standard, SN_9563 and SN_132 taken at a 17° depression angle are used for training, while the remaining collected at a 15° depression angle comprise the testing set. Significant changes of configuration and depression angle are present, as detailed in TABLE I. The original images are of around 128×128 pixels in size, and standardized as 96×96 pixels by cropping the center patches. All experiments are performed on Matlab 2015a.

TABLE I. The number of aspect view images for BMP2, BTR70, T62, and T72.

Depr.	BMP2	T72	BTR60	T62	*Total*
17° (**Gallery**)	233 (SN_9563)	232 (SN_132)	256	299	1020
15° (**Probe**)	196 (SN_9566) 196 (SN_c21)	195 (SN_812) 191 (SN_s7)	195	273	1246

A. The Detection of Keypoint

We first evaluate the performance of representative detectors. We aim to studying whether these methods could seek keypoint from radar image, and whether the local descriptors around the keypoints could exploit target signature information.

We provide a set of instance on keypoint detection. Fig. 9 draws the detection maps obtained using DoG, Hessian, Harris-Laplace, and Hessian-Laplace[2]. We found that the keypoints detected are mainly located in target imaging region, corresponding to the local scattering centers. The number of keypoints produced by DoG is much more than the other detectors. Keypoints generated by Hessian-Laplace detector are more than Hessian. Those points located in main gun of tank have also been detected.

Fig. 10 shows the detection chart of FAST on the same image[3]. The threshold value τ is set as 0.25, 0.2, 0.15, and 0.1. With the threshold value decreased, the number of keypoints detected are increased. The smaller the threshold, the more the number of keypoints detected. Most of these keypoints are located in target imaging region, similar to the previous maps. The small threshold value usually produces some keypoints in the site of speckle.

[2] **VLFeat** toolkit is used to implement the detectors.

[3] The function ‘fast9’ is employed. More information can be found in the homepage at http://mi.eng.cam.ac.uk/~er258/work/fast.html.

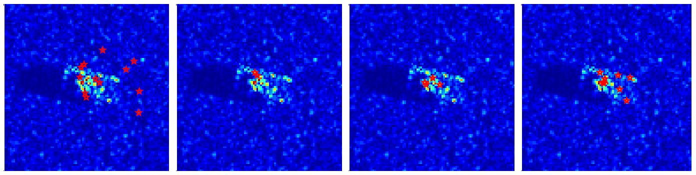


Fig. 9. The detection maps obtained by DoG, Hessian, Harris-Laplace, and Hessian-Laplace. The keypoints are marked by pentagram in red.

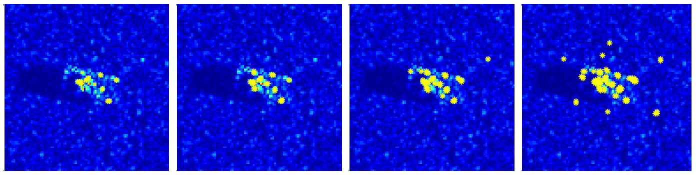


Fig. 10. Keypoints detected by FAST with four different threshold values. The keypoints detected are noted by asterisk in yellow.

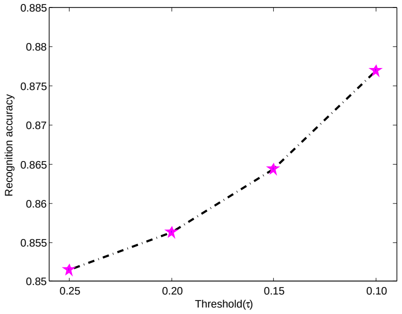
The average number of keypoints detected from the whole training set is then given. To study whether the number of keypoint is related to target pose, the aspect view is divided into four ranges, as detailed in TABLE II. We found that Harris-Laplace detector produces the least number of keypoints consistently, while FAST detector always seeks the most number of points. For BMP2, the number of detected points produced from Angle$_3$ is much more than the remaining angle range. For BTR60, T72, and T62, the number of detected points is irregular. Hence, we could come the conclusion that the number of keypoints detected is not related to target pose.

TABLE II. THE NUMBER OF KEYPOINTS DETECTED FROM FOUR CLASSES OF RADAR IMAGES. ANGLE$_1$, ANGLE$_2$, ANGLE$_3$, ANGLE$_4$ REFERS TO IMAGES WITH POSE FALLEN IN THE RANGE OF $0° \sim 90°$, $91° \sim 180°$, $181° \sim 270°$, $271° \sim 360°$.

Detector		DoG	Hessian	HarrisL	HessianL	FAST^1	FAST^2
BMP2	*Angle$_1$*	14.84	12.98	6.07	19.72	31.77	42.59
	Angle$_2$	14.73	14.65	6.87	22.18	36.07	48.02
	Angle$_3$	**18.06**	**17.69**	**8.51**	**26.33**	**40.06**	**52.57**
	Angle$_4$	17.23	12.49	6.41	18.31	28.95	39.05
	Overall	16.14	14.31	6.90	21.43	33.95	45.24
BTR60	*Angle$_1$*	15.34	19.87	8.79	30.52	40.47	55.24
	Angle$_2$	14.88	17.96	8.46	27.39	41.57	55.13
	Angle$_3$	14.92	18.73	8.53	31.67	48.41	63.53
	Angle$_4$	18.05	18.95	8.75	29.19	44.68	59.81
	Overall	15.87	18.91	8.64	29.66	43.63	58.28
T72	*Angle$_1$*	15.40	16.63	7.50	21.63	37.18	48.21
	Angle$_2$	17.16	16.72	8.09	23.13	39.74	51.40
	Angle$_3$	16.65	16.28	7.15	22.43	39.25	51.93
	Angle$_4$	13.65	13.22	6.57	18.25	31.80	42.77
	Overall	15.75	15.77	7.36	21.43	37.08	48.65
T62	*Angle$_1$*	12.45	21.55	11.38	37.45	56.07	72.19
	Angle$_2$	12.43	19.12	10.43	38.08	55.30	71.14
	Angle$_3$	13.09	22.38	11.16	41.78	62.65	79.33
	Angle$_4$	13.50	24.10	12.06	38.43	58.79	75.71
	Overall	12.87	21.78	11.25	39.00	58.29	74.69

We further quantitatively assess the detectors on recognition performance. The local descriptors extracted from the training images are used to predict the class membership of query. Our first proposed solver is employed to implement classification.

We first evaluate FAST detector by SIFT descriptor. There is a free parameter to be determined, the threshold value τ. We test four different values, 0.25, 0.2, 0.15, and 0.1. The recognition accuracy as a function of threshold value is drawn in Fig. 11. The recognition performance is inversely proportional to the threshold value. The bigger the threshold, the lower the recognition rate. This is because FAST with smaller threshold value usually results in much more number of keypoints than the one with big threshold. Simultaneously, the computational cost (CPU-Time) is increased sharply with the decrease of threshold. The CPU-Time for FAST with 0.10-threshold is even 8 times longer than FAST with 0.25-threshold. Therefore, it is needed to balance the computational cost and recognition accuracy.



τ	0.25	0.20	0.15	0.10
CPUTime	490s	792s	1465s	4015s

Fig. 11. Recognition accuracy over the threshold value τ (FAST).

All of the detectors are compared with in Fig. 12. SIFT and DAISY descriptors are employed to achieve target classification. We can see that DoG detector tends to perform better than Harris and Hessian detector when SIFT descriptor is used to represent the local pattern. The most likely reason for this difference is that the Laplacian detector tends to extract two or three times more keypoints per image than Harris-Laplace (verified in TABLE II), and hence produce a richer representation. On the contrary, the performance obtained using FAST, Hessian-Laplace, and Hessian detectors are much better than DoG and Harris-Laplace detectors when DAISY descriptor is employed. The results prove that the performance of detector is related to the choice of descriptor. Hence, the further comparison of descriptors is needed.

B. *The Representation of Local Pattern*

The detected keypoint serves the generation of local descriptor, from which target classification can be achieved. This section devotes to the verification of local descriptor.

For AB-SIFT descriptor, there are three parameters to be determined, the level of radial quantization, the number of angular quantization, and the bin of histogram. To evaluate the effect of these parameters, we pursue a set of experiments. Two detectors, Hessian and DoG, are evaluated. The neighborhood around the detected keypoint is represented by several adaptive

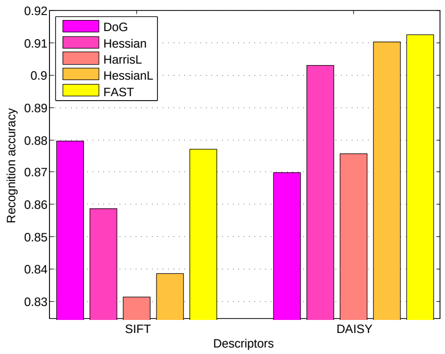


Fig. 12. Recognition performance obtained using different detectors.

binning histograms. The experimental results are given in TABLE III[4], where GLOH descriptor is employed as the baseline (fixed binning). The results prove that the recognition performance can be improved by adaptive binning compared to fixed fashion of quantization. The 2-level radial quantization provides a much poor performance. Some improvement is achieved when the level of radial quantization is increased from 3 to 4, especially when DoG detector is employed. The result also proves that the performance of local descriptor is related to the choice of detector. Considering efficiency and accuracy, the angular $\{6, 6, 6, 6\}$ and the histogram $\{6, 6, 6, 6\}$ are configured.

TABLE III. The performance of AB-SIFT under different settings.

Radial	*Angular*	*Histogram*	*Dim.*	*Hessian*	*DoG*
2	{6,10}	{10,6}	120	0.8467	0.8483
	{8,12}	{8,6}	136	0.8563	0.8571
	{10,14}	{6,6}	144	0.8579	0.8491
3	{4,6,8}	{10,8,6}	136	0.8612	0.8860
	{5,8,10}	{8,6,4}	128	0.8740	0.8965
	{6,8,12}	{8,6,4}	144	0.8740	0.8957
4	{4,6,8,10}	{8,6,4,4}	136	0.8876	0.8989
	{5,6,8,10}	{6,6,4,4}	138	0.8892	0.8965
	{6,6,8,8}	{6,6,4,4}	136	**0.8949**	0.8949
	{6,6,6,6}	{6,6,6,6}	144	0.8860	**0.9045**
GLOH	{1,8,8}	{8,8,8}	136	0.8740	0.8892

DAISY involves four parameters, radius, the number of ring, the division of sector, and the level of quantization. We set the rings and sectors as 3 and 8, and change the radius and the quantization level. Hessian detector is employed to seek the keypoints. Effect of the parameters on performance are tabulated in TABLE IV[5]. As can be seen, the recognition accuracy has been improved with the number of bins increased. The bigger the number of bins, the better the performance. Meanwhile, the dimension of feature is also increased. DAISY$_{20}$ even results in 500-D feature, greatly higher than the preceding descriptors. As for the radius of rings, it is proportional to the recognition accuracy. The longer the radius, the better the performance. The performance reaches a plateau when the radius is beyond 20. To make a balance between efficiency and accuracy, we set the radius as 20, and the number of bins as 8. The running times of different settings are reported. We can see that the computational cost is acceptable.

TABLE IV. The performance of DAISY over the radius and quantization level.

Bins	DAISY$_4$	DAISY$_8$	DAISY$_{12}$	DAISY$_{16}$	DAISY$_{20}$
Dim.	100	200	300	400	500
Radius$_{16}$	0.8740	0.8909	0.8884	0.8900	0.8909
Radius$_{18}$	0.8836	0.9005	0.8997	0.8989	0.9005
Radius$_{20}$	0.8892	0.9005	0.9021	0.9021	0.9037
Radius$_{22}$	0.8886	0.8997	0.8989	0.8997	0.9013
Times	82s	137s	205s	262s	344s

Fig. 13 compares with all of the local descriptors. Hessian-Laplace and Harris-Laplace detectors are employed to search the keypoints. The neighborhood around the detected keypoints are characterized by the representative descriptors, SIFT, GLOH, AB-SIFT, DAISY, and SURF[6]. In the prototype of SIFT, the neighborhood is quantized into a 4×4 square grids, and the gradient angle is quantized into 8 orientations, resulting in a 128D descriptor. The implementation of GLOH is slightly different from the original work [42]. We assign a log-polar location grid with 3 bins in the radial direction and 8 bins in the angular direction. The central bin is not divided in angular directions. The gradient orientations are further quantized into 8 bins, generating a 17×8=136D feature. SURF quantizes the neighborhood around keypoint into a 4×4 square grids, in each of which four Haar-like wavelet responses are extracted, resulting in a 64D representation.

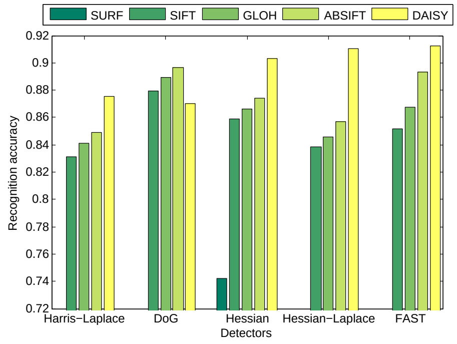


Fig. 13. Recognition performance obtained using various descriptors.

As can be seen, the recognition performance may be relevant to whatever the keypoint detector is considered. For DoG detector, the performance obtained using DAISY is poorer than SIFT and GLOH. For the remaining detectors, Hessian, Hessian-Laplace, and Harris-Laplace, DAISY outperforms the

[4] The source code of AB-SIFT was not available. We implement it by only changing the log-polar sector and the level of quantization. In each cell, the generation of histogram is similar to previous work.

[5] DAISY is a dense descriptor to wide-baseline stereo. We first compute the descriptors pixel by pixel, and then extract the ones in the site of keypoints. It is detailed in http://cvlab.epfl.ch/~tola.

[6] The source codes and more details can be found at Chris Evans's homepage, http://www.chrisevansdev.com/opensurf/.

other local features, with a gain greater than 3%∼8% over than SIFT and GLOH. SURF generates the lowest recognition accuracy, even 11.63%, 12.40%, and 15.81% lower than SIFT, GLOH, and DAISY when Hessian detector is employed. This result can be attributed to the fashion of keypoint detection and feature representation. To boost the computational efficiency, SURF circumvents image convolutions by means of integral images. Fig. 14 provides a pair of detection maps generated by 'Fast-Hessian' and the prototype of Hessian detector. As can be seen, the keypoints detected by 'Fast-Hessian' are irregularly scattered in the whole image. Contrarily, the keypoints produced by Hessian detector are mainly concentrated in target imaging region. They are representative to reflect the target scattering phenomenology. Furthermore, SURF represents the local pattern by the Haar-like wavelet response, $h_x, h_y, |h_x|, |h_y|$, resulting in a 64-D feature. The discriminative ability is limited in comparison to SIFT, GLOH, and DAISY, whose representations are 128-, 136-, and 200-dimension. The experimental results also prove that the approximation of convolution with integral image is not effective to radar image due to the multiplicative noise.

We can see that AB-SIFT tends to perform better than GLOH, while GLOH always outperforms SIFT. It is not surprising that SIFT performs poorly than GLOH, since GLOH employs a much finer division of location, log-polar sectors. As for the performance rank between DAISY and the remaining descriptors, it depends on the choice of detector. Overall, the combination of FAST detector with DAISY descriptor is the preferable choice in terms of recognition accuracy.

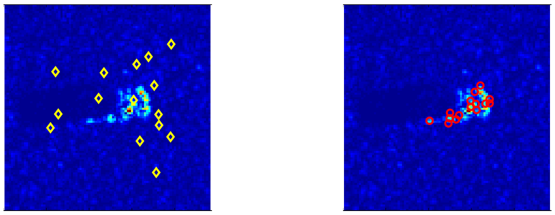


Fig. 14. The detection map produced by 'Fast-Hessian' and Hessian detectors. Keypoints defined by 'Fast-Hessian' are marked by diamond in yellow, while the ones produced by Hessian are noted by circle in red.

C. Classification

This paper proposes two kinds of methods to target classification. The effect of related factors on recognition performance is studied. Our first proposed method, sparse representation over neighbor descriptors is abbreviated as NSR.

1) Sparse Representation: Our first proposed scheme refers to building multiple linear regression models. The regression coefficients are obtained according to the thought of sparse representation. Different from the preceding works, where the query (feature) is directly represented by the whole training set (features), this paper develops a prescreener procedure, with which only the nearest neighbor descriptors are kept. The local descriptors far away from the query are ignored. The related factors therefore include the number of neighbor descriptor L and the regularization parameter λ. To study their effect, we perform two sets of experiments. Hessian detector is used to search keypoints, while SIFT and GLOH are evaluated.

Fig. 15 draws the recognition accuracy across the number of neighbor descriptor (the prescreener). We tune the number of neighbor descriptor from [20, 40, 60, 80, 100, 120, 160, 200]. The recognition accuracies are slightly varied with the number of neighbors changed. For SIFT, the best recognition rate is obtained using sparse representation over 60-nearest-neighbor descriptors, while the best performance for GLOH descriptor is produced by sparse representation over 120-nearest-neighbor descriptors. The recognition performance reaches the plateau when the number of neighbor is bigger than 60 and less than 120. To draw a balance, 120-nearest-neighbor strategy is employed to build a dictionary, over which the descriptors of query can be represented.

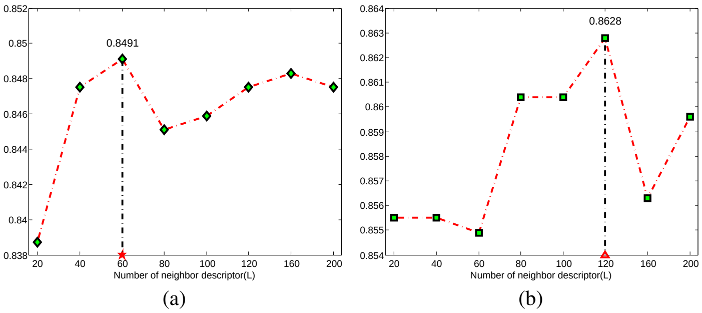


Fig. 15. Recognition accuracy across the number of neighbor descriptor. (a) SIFT, (b) GLOH.

Fig. 16 plots the recognition performance as a function of regularization parameter. The results are similar to the above experiments. The recognition accuracy is varied when the regularization parameter is changed. The best recognition rate for SIFT is obtained by 0.18, while the best performance for GLOH is produced by 0.16. The performance is robust for both two descriptors when the parameter value is beyond 0.1 and below 0.2. To obtain a tradoff, this paper set the regularization parameter as 0.18.

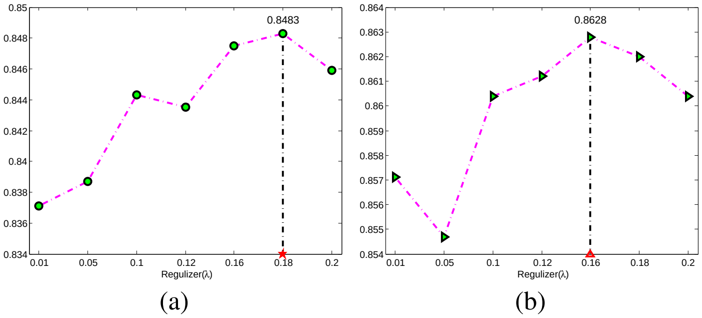


Fig. 16. Recognition accuracy across the regularization parameter. (a) SIFT, (b) GLOH. Performance plateaus when λ is bigger than 0.1.

2) Encoding: Our second proposed scheme involves feature quantization. Two encoding tricks, BoW and FV are employed. The local descriptors resulting from a chip image are encoded to define a new single feature. The new defined feature is fed into a trained discriminative classifier. We evaluate several different experimental settings. Hessian detector is used to produce the keypoint, while DAISY descriptor is employed to characterize the local pattern.

BoW. We verify two tricks, LLC [49] and SC [48], popularly studied in the preceding works. We manually change the number of neighbors from 4 to 20 (LLC). The results are displayed in Fig. 17, where the performance obtained using SC is given as the baseline. We can see the recognition performance is changed with the size of codebook increased. For 1024-atom codebook, the recognition accuracy obtained using SC is much better than all settings of LLC. On the contrary, the recognition rate obtained using linear coding is better than sparse coding when 3072-atom codebook is generated. For 2048-atom codebook, the recognition rate for SC is better than LLC(10) and LLC(20), and poorer than LLC(4), LLC(5), LLC(6), and LLC(8). The recognition performance is proportional to the size of codebook. For locality-constrained linear coding, the number of neighbors plays an important role. The little neighbor (5) produces the better performance. In addition, linear coding is computationally more attractive than sparse coding.

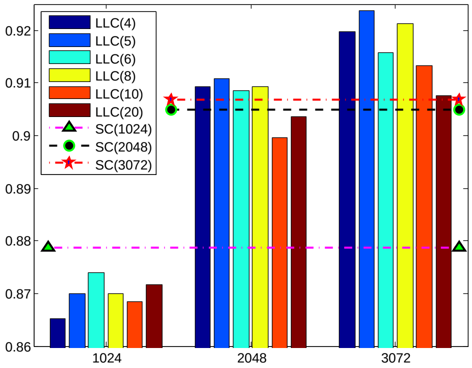


Fig. 17. Performance obtained using LLC and SC.

FV. Fisher vectors approximates the distribution of low-level features with Gaussian mixture model. The related factor to be decided is the number of Gaussian components. We change the number of Gaussian components from 32 to 144. Since the parameters, mean, covariance, and prior are randomly initialized, the recognition accuracies are not deterministic. We implement FV with the same setting repeatedly for 10 times. The recognition performance as a function of the number of Gaussian components is shown in Fig. 18, in which NSR and BoW are employed as the baseline. Two descriptors, SIFT and DAISY are assessed. We found the recognition performance is different for two descriptors when the number of Gaussian components is changed. For SIFT, FVs with all of Gaussian components perform poorly than NSR and BoW. FV with 96-Gaussian-component produces the better and robust performance. For DAISY, most settings of FV perform better than NSR, and poor than BoW with 3072-atom-codebook. Again, FV with 96-Gaussian-component achieves the better and robust performance than the remaining settings.

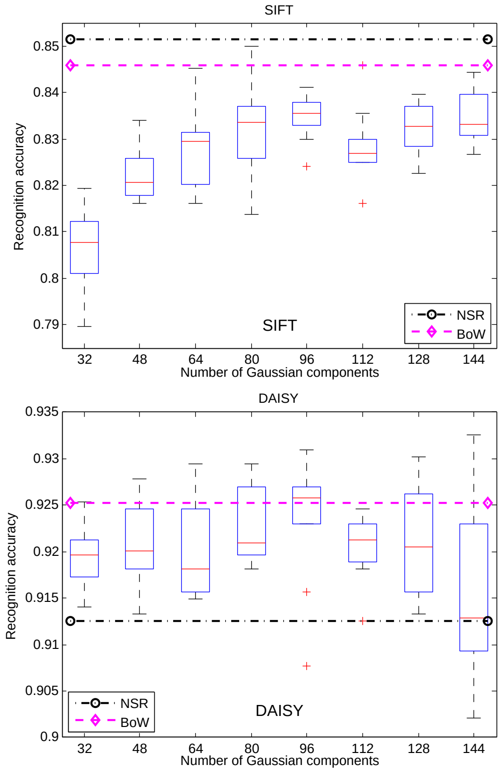


Fig. 18. Recognition accuracy of FV across the number of Gaussian components. Performance plateaus when λ is bigger than 0.1.

The experimental results prove that 120-nearest-neighbor is appropriate for NSR, while 2048- and 3072-atom-codebook is suitable for BoW. NSR and BoW are the preferable classifier for SIFT, GLOH, and AB-SIFT, while FV is more appropriate for DAISY.

D. The Validation of Recognition Performance

This paper achieves radar target recognition by keypoint-based local descriptor. Two kinds of methods are proposed to implement classification. The effect of related factors are studied previously. The recognition performance of proposed strategy is validated. State-of-the-art global features are employed as the baseline. Support vector machine learning (SVM) is popularly studied over the years [7]. Sparse representation-based classification (SRC) is a recently developed method [8], [32]. Both of them input the raw pixel values for classification[7]. Furthermore, the preceding works achieve distortion and translation invariance by Fourier transformed spectrum. The representative includes optional tradeoff synthetic discriminant function (OTSDF) [52]. Another family of filter banks, the monogenic signal, has also been used for target classification [14]–[16] (MSRC). These methods are employed

[7] The feature dimension is reduced by downsampling, principal component analysis, and random projection.

to compared with the proposed strategy. The experimental results are given in Fig. 19.

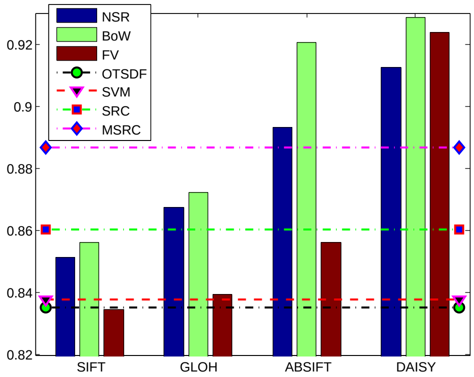


Fig. 19. Comparison to the preceding works.

From Fig. 19, we found the recognition performance obtained using four kinds of local descriptors are sharply different. For SIFT, GLOH, and AB-SIFT, NSR performs better than FV, and poorly than BoW. Differently, the recognition accuracy obtained using FV with DAISY is better than NSR, and poorly than BoW. On the other hand, DAISY with NSR, BoW, and FV, outperforms all baseline algorithms, even 2.57%, 3.85%, and 3.79% better than the main competitor, MSRC. Similarly, AB-SIFT with NSR and BoW also performs better than the baselines. The performance obtained by SIFT is poor than SRC and MSRC, and better than SVM and OTSDF. The recognition accuracy for GLOH is poorer than MSRC, and better than SVM, SRC, and OTSDF. The results prove that the local descriptor could achieve comparable or even better performance compared to the global feature.

V. Conclusion

This paper considers radar target recognition with keypoint-based local descriptor. We develop two kinds of schemes to implement classification. Multiple comparative experiments are performed, from which several different combinations of detectors and descriptors are evaluated. The experimental results prove:

- keypoint-based local descriptor could be fully tuned to target recognition,
- it is important to configure the related factors appropriately,
- the proposed strategy could achieve comparable or even better performance than the preceding studies,
- the proposed strategy provides great potential for target recognition under the non-literal conditions,
- the local descriptors can be further refined according to the imaging mechanism of radar.

However, some issues are needed to be further considered. We plan to study whether the advantage of our proposed strategy will persist under the non-literal conditions.

We design a prescreener procedure for sparse representation. It makes the computational cost and memory consumption acceptable. The linear correlation response is employed to measure the (dis)similarity. An important future research direction is to develop the specific metric. The study on the measurement of similarity has been noticed in [53], [54] and more recently explored in [55]. The further research in target recognition is yet to be uncovered.

On the other hand, this paper verifies the generic model of local descriptor by some fundamental experiments. The mechanism of radar imaging is not yet considered in the phase of detection, or representation. We believe the performance can be improved if the specificity of SAR image has been exploited. Moreover, addressing the problem of target recognition under the less constrained conditions is another interesting direction for future work.

References

[1] L. Potter and R. Moses, “Attributed scattering centers for SAR ATR,” *IEEE Trans. Image Process.*, vol. 6, no. 1, pp. 79–91, Jan. 1997.

[2] C. Olson and D. Huttenlocher, “Automatic target recognition by matching oriented edge pixels,” *IEEE Trans. Image Process.*, vol. 6, no. 1, pp. 103–113, Jan. 1997.

[3] L. Novak, G. Owirka, and C. Netishen, “Performance of a high-resolution polarimetric SAR automatic target recognition system,” *Lincoln Lab J.*, vol. 6, no. 1, pp. 11–23, 1993.

[4] A. Banerjee, P. Burlina, and R. Chellappa, “Adaptive target detection in foliage-penetrating SAR images using alpha-stable models,” *IEEE Trans. Image Process.*, vol. 8, no. 12, pp. 1823–1831, Dec. 1999.

[5] D. Kreithen, S. Halversen, and G. Owirka, “Discriminating targets from clutter,” *Lincoln Lab J.*, vol. 6, no. 1, pp. 25–52, 1993.

[6] L. Novak, G. Owirka, and A. Weaver, “Automatic target recognition using enhanced resolution SAR data,” *IEEE Trans. Aerosp. Electron. Syst.*, vol. 35, pp. 267–278, 1999.

[7] Q. Zhao and J. Principe, “Support vector machines for SAR automatic target recognition,” *IEEE Trans. Aerosp. Electron. Syst.*, vol. 37, no. 2, pp. 643–654, Apr. 2001.

[8] J. Thiagarajan, N. Karthikeyan, K. Peter, P. Knee, A. Spanias, and V. Berisha, “Sparse representation for automatic target classification in SAR images,” in *Int’l Sym. Communcitaion, Control and Signal Processing*, 2010, pp. 1–4.

[9] M. Liu, Y. Wu, P. Zhang, Q. Zhang, Y. Li, and M. Li, “SAR target configuration recognition using locality preserving property and gaussian mixture distribution,” *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 2, pp. 268–272, Mar. 2013.

[10] X. Huang, H. Qiao, and B. Zhang, “SAR target configuration recognition using tensor global and local discriminant embedding,” *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 2, pp. 222–226, Feb. 2016.

[11] Z. Cui, Z. Cao, J. Yang, J. Feng, and H. Ren, “Target recognition in synthetic aperture radar images via non-negative matrix factorisation,” *IET Radar, Sonar & Navigation*, vol. 9, no. 9, pp. 1376–1385, Dec. 2015.

[12] R. Patnaik and D. Casasent, “MINACE filter classification algorithms for ATR using MSTAR data,” in *Automatic Target Recognition XV, Proc. SPIE*, vol. 5807, Aug. 2005, pp. 100–111.

[13] ——, “MSTAR object classification and confuser and clutter rejection using MINACE filters,” in *Automatic Target Recognition XVI, SPIE*, vol. 6234, Aug. 2006, pp. 1–13.

[14] G. Dong, N. Wang, and G. Kuang, “Sparse representation of monogenic signal: with application to target recognition in SAR images,” *IEEE Signal Process. Lett.*, vol. 21, no. 8, pp. 952–956, Aug. 2014.

[15] G. Dong and G. Kuang, “Classification on the monogenic scale space: Application to target recognition in SAR image,” *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2527–2539, Aug. 2015.

doi:10.20944/preprints201805.0116.v1

[16] G. Dong, G. Kuang, N. Wang, and W. Wang, "Classification via sparse representation of steerable wavelet frames on grassmann manifold: Application to target recognition in SAR image," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2892–2904, Jun. 2017.

[17] M. Felsberg and G. Sommer, "The monogenic signal," *IEEE Trans. Signal Process.*, vol. 49, no. 12, pp. 3136–3144, 2001.

[18] ——, "The monogenic scale-space: A unifying approach to phase-based image processing in scale space," *J. Math. Imag. Vis.*, vol. 21, no. 1, pp. 5–26, 2004.

[19] J. Park, S. Park, and K. Kim, "New discrimination features for SAR automatic target recognition," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 3, pp. 476–480, May 2013.

[20] J. Zhu, X. Qiu, Z. Pan, Y. Zhang, and B. Lei, "An improved shape contexts based ship classification in SAR images," *Remote Sensing*, vol. 9, no. 2, 2017.

[21] J. Singh and M. Datcu, "SAR image categorization with log cumulants of the fractional Fourier transform coefficients," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 12, pp. 5273–5282, Dec. 2013.

[22] M. Amoon and G.-A. Rezai-rad, "Automatic target recognition of synthetic aperture radar (SAR) images based on optimal selection of Zernike moments features," *IET Computer Vision*, vol. 8, pp. 77–85, 2014.

[23] X. Zhang, Z. Liu, S. Liu, D. Li, Y. Jia, and P. Huang, "Sparse coding of 2D-slice Zernike moments for SAR ATR," *International Journal of Remote Sensing*, vol. 38, pp. 412–431, 2016.

[24] P. Bolourchi, H. Demirel, and S. Uysal, "Target recognition in sar images using radial Chebyshev moments," *Signal Image & Video Processing*, vol. 11, p. 10331040, 2017.

[25] S. Chen, H. Wang, F. Xu, and Y.-Q. Jin, "Target classification using the deep convolutional networks for SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 8, pp. 4806–4817, Aug. 2016.

[26] J. Ding, B. Chen, H. Liu, and M. Huang, "Convolutional neural network with data augmentation for SAR target recognition," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 3, pp. 364–368, March 2016.

[27] N. Wang, Y. Wang, H. Liu, Q. Zuo, and J. He, "Feature-fused SAR target discrimination using multiple convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1695–1699, Oct. 2017.

[28] D. Malmgren-Hansen, A. Kusk, J. Dall, A. A. Nielsen, R. Engholm, and H. Skriver, "Improving SAR automatic target recognition models with transfer learning from simulated data," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 9, pp. 1484–1488, Sept. 2017.

[29] S. Deng, L. Du, C. Li, J. Ding, and H. Liu, "SAR automatic target recognition based on Euclidean distance restricted autoencoder," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 7, pp. 3323–3333, Jul. 2017.

[30] B. Ding, G. Wen, X. Huang, C. Ma, and X. Yang, "Target recognition in synthetic aperture radar images via matching of attributed scattering centers," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 7, pp. 3334–3347, July 2017.

[31] L. Novak, G. Owirka, and W. Brower, "Performance of 10- and 20-target MSE classifiers," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 36, pp. 1279–1289, Oct. 2000.

[32] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[33] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int'l J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[34] C. Harris and M. Stephen, "A combined corner and edge detector," p. 10.5277, Aug. 1988.

[35] T. Lindeberg, "Feature detection with automatic scale selection," *Int'l J. Comput. Vis.*, vol. 30, no. 2, pp. 79–116, 1998.

[36] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," *Int'l J. Comput. Vis.*, vol. 60, no. 1, pp. 63–86, 2004.

[37] H. Bay, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 404–417.

[38] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[39] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 430–443.

[40] E. Rosten, R. Porer, and T. Drummond, "Faster and better: A machine learning approach to corner detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 105–119, Jan 2010.

[41] D. Lowe, "Object recognition from local scale-invariant features," in *Proc. IEEE Int'l Conf. Comput. Vis. (ICCV)*, 1999, pp. 1150–1157.

[42] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615–1630, Oct. 2005.

[43] A. Sedaghat and H. Ebadi, "Remote sensing image matching based on adaptive binning SIFT descriptor," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 10, pp. 5283–5293, Oct. 2015.

[44] E. Tola, V. Lepetit, and P. Fua, "A fast local descriptor for dense matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[45] ——, "DAISY: An efficient dense descriptor applied to wide-baseline stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 815–830, May 2010.

[46] E. Candes and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies," *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5406–5425, 2006.

[47] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 2169–2178.

[48] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1794–1801.

[49] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constraint linear coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3360–3367.

[50] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, June 2007, pp. 1–8.

[51] F. Perronnin, J. Snchez, and T. Mensink, "Improving the Fisher kernel for large-scale image classification," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 1–14.

[52] R. Singh and B. Kumar, "Performance of the extended maximum average correlation height filter and the polynomial distance classifier correlation filter for multiclass SAR detection and classification," in *Algorithms for SAR Imagery IX*, vol. 4727. SPIE, 2002, pp. 265–279.

[53] E. Levina and P. Bickel, "The Earth Mover's distance is the mallows distance: Some insights from statistics," in *Proc. IEEE Int'l Conf. Comput. Vis. (ICCV)*, 2001, pp. 251–256.

[54] Y. Rubner, C. Tomasi, and L. Guibas, "The Earth Mover's distance as a metric for image retrieval," *Int'l J. Comput. Vis.*, vol. 40, no. 2, pp. 99–121, 2000.

[55] H. Ling, "An efficient Earth Movers distance algorithm for robust histogram comparison," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 5, pp. 840–853, May 2007.