

IMAGE CLASSIFICATION: NO FEATURES, NO CLUSTERING

Shiyong Cui, Gottfried Schwarz, Mihai Datcu

Remote Sensing Technology Institute (IMF)
German Aerospace Center (DLR)
Münchener Straße 20, 82234 Wessling
shiyong.cui, gottfried.schwarz, mihai.datcu@dlr.de

ABSTRACT

In this paper, we consider the problem of satellite image classification, in which feature extraction is a critical step. One of the most prevalent methods is the Bag-of-Words (BoW) feature representation, which attains state-of-the-art performance in many applications. It has five steps: feature detection, local feature extraction, dictionary learning, feature coding, and feature pooling. In this paper, we focus on the second and third step. We propose a simple yet efficient feature extraction method within the BoW framework. It has two main advantages. Firstly, this method does not need any complex local feature extraction; instead, it uses directly the pixel values from small windows as low level features. Secondly, instead of using a time-consuming clustering algorithm for dictionary learning, a random dictionary is built and applied to feature space quantization. An extensive experimental evaluation has been performed and compared with other feature extraction methods. It is demonstrated that our feature extraction method is quite competitive for optical and SAR satellite image classification.

Index Terms— Bag-of-words (BoW), Dictionary learning, Feature extraction, Image classification, Unsupervised feature learning.

1. INTRODUCTION

One of the fundamental challenges in Earth observation is to explore the large volume of data, in which image classification plays an important role. In image classification, feature extraction is a critical step. Traditionally, image classification relies on hand-crafted features that try to capture the essence of different visual patterns. In recent years, feature learning approaches have gained significant interest as a way of representing images. In this paper, we focus on feature extraction for satellite image classification.

From the beginning of the twenty-first century, prominent advances in texton and local feature extraction have been witnessed, leading to the Bag-of-Words (BoW) method for feature extraction [1]. Since then, within this framework, a large variety of methods have been proposed for solving various problems, for instance, image classification, image retrieval, and object recognition. The BoW technique has been recently introduced also to the remote sensing community for image annotation [2], object classification [3], target detection [4] and land use classification [5] and it has already proven its discrimination power in image classification. The BoW method consists of five main components: feature detection, local feature extraction, dictionary learning, feature coding, and feature pooling. All elements have been investigated with a lot of effort. Specifically, in this paper, we focus on local feature extraction and dictionary learning.

As the BoW feature vector is an intermediate feature depending on low level features, distinctive local features should be carefully designed. The scale invariant feature transform (SIFT) [6] method is one of the most widely used methods for local feature detection and extraction. For texture images, local rotation-invariant features are more preferable for image classification. Pixel values in a local patch instead of local filter responses are proposed for texture classification in [7] claiming that compact local patches can achieve better performance than a texton distribution of the filter responses. Based on this work, a random projection [8] was applied to reduce the dimensionality of the local feature vectors and a significant improvement in classification accuracy was shown. However, it was observed that the random projection of the local features is not rotation-invariant; thus, a sorted random projection of five local features [9] was developed by the same authors, who claimed to achieve significant improvements compared with the method of [8]. Although a large effort has been devoted to local feature extraction, it is still not well understood what is essentially important for local features.

Another problem is about dictionary learning. The dictionary is usually learned by various unsupervised clustering algorithms, such as k -means clustering, Gaussian mixture models [10], and random forests [11]. However, these methods are usually very time-consuming, although they can be applied offline. In the case of large datasets, it is prohibitively time-consuming to learn a dictionary. The goal of dictionary learning is to find a universal codebook for feature coding. This universal codebook does not necessarily coincide with the actual cluster centers. We show that a random dictionary, constructed by a random selection of some local descriptors in the feature space, can achieve similar performance to that of a dictionary learned by an unsupervised clustering method.

To address these two problems, we propose a simple yet efficient method for image classification. This method does not need any complex local feature extraction and any unsupervised method for dictionary learning. We show that pixel values in very small patches have sufficient information for classification and a random dictionary constructed by a random selection of some local descriptors gives better performances for our datasets. The major contributions of this paper are two-fold:

1. We use the pixel values in a very compact local neighborhood, e.g., taken from a 3×3 window and a column-wise conversion into a vector of elements (“Vectorized Patch”), as low level features for the BoW method.
2. Instead of unsupervised dictionary learning, we randomly select some feature points and use them as our dictionary.

These two contributions are evaluated through comparisons with other state-of-the-art methods for satellite image classification.

2. BOW FEATURE EXTRACTION

In this section, we first present the general framework of BoW feature extraction and then, we propose our method for feature extraction and dictionary learning.

2.1. BoW Feature Extraction Framework

The framework of BoW feature extraction shown Fig. 1 is composed of five steps, which are feature detection, local feature extraction, dictionary learning, feature coding, and feature pooling. Suppose we have a dataset of N images $I_i, i = 1, \dots, N$, the first step is to sample a collection of patches from the images in the database. This is done by dense sampling in our case. The second step is to extract local feature vectors $\mathbf{x}_i^j \in \mathbb{R}^D, j = 1, \dots, M$ from all patches. The third one is learning a dictionary $\mathbf{D} = (\mathbf{d}_1, \dots, \mathbf{d}_K) \in \mathbb{R}^{D \times K}$ with K words using all local feature vectors. Normally, this is done by a time-consuming unsupervised learning method, such as k -means clustering or a Gaussian mixture model. The elements \mathbf{d}_k in a dictionary are the centers of the clusters. The next step is to find a dictionary-based representation $\mathbf{v} = [v_1, \dots, v_K]$ for each previously extracted local feature vector \mathbf{x} . This is usually solved by hard feature assignment. Hard assignment assigns a single label, i.e., the index of the nearest neighbor in the dictionary, to each local feature vector \mathbf{x} . Formally, it is defined as:

$$v_k(\mathbf{x}) = \begin{cases} 1 & \text{if } k = \min \|\mathbf{x} - \mathbf{d}_i\|^2 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

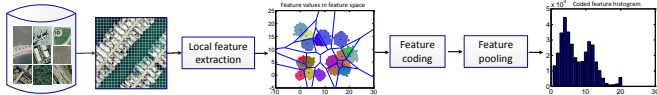


Fig. 1. The framework of Bag-of-Words feature extraction.

Thus, the final descriptor representation $\mathbf{v} = [v_1, \dots, v_K]$ has only one non-zero element. The last step is to do the sum-pooling¹ of all local feature vectors extracted from one image $\mathbf{v}_i = \text{sum}(\mathbf{v}_i^1, \dots, \mathbf{v}_i^M)$.

2.2. Two Problems and Our Method

The two important problems in the BoW method we focus on are as follows.

1. What local features should be extracted? How about the patch size and the patch sampling strategy?
2. What method should we use for dictionary learning?

The first problem is about local feature extraction. To solve this, we have to carefully consider the patch size and the patch sampling strategy, which are practically related. If the patch size is quite large, the dimensionality of the local features is very high [8], which makes a subsequent unsupervised dictionary learning time-consuming, thus infeasible for large scale databases. In addition, there would be large overlaps among patches if the patch size is large. This could potentially degrade the feature space. We will compare regular dense sampling and random sampling in Section 3.2.

¹Sum-pooling is equivalent to computing the histogram in the case of hard feature assignment.

Local features that can be extracted with minimum computational effort are preferable. There are many local features that have been proposed in the literature. We analyzed several discriminative local features and found that the vectorized pixel values of very small patches, e.g., defined by windows of 3×3 pixels, provide enough information for discrimination. We demonstrate that this simple local feature vector can achieve a rather promising performance for image classification. The main advantages are its simplicity and the low computational cost.

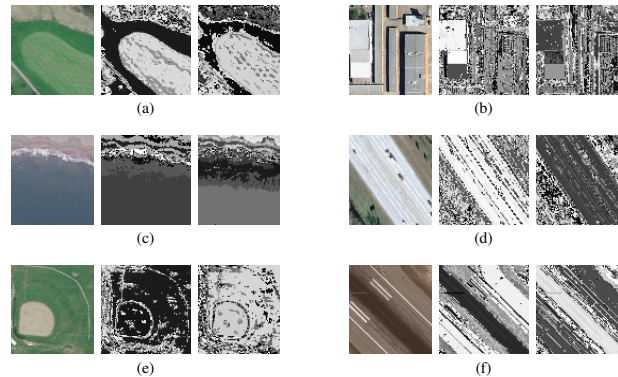


Fig. 2. Comparison of vector quantization using a random dictionary and k -means clustering on the UCMerced land use dataset. One example is given for each class. The first color image is an example from each class. The second and third images in each group are the dictionary entries using a random dictionary and a k -means dictionary with the same size of 200 entries.

The second problem is about the method for dictionary learning that is usually solved by various unsupervised clustering algorithms. However, in the case of large datasets, this step is prohibitively time-consuming, although it can be applied offline. The goal of dictionary learning is to find a universal codebook for feature coding. We found that this universal codebook is not necessarily coincident with the actual cluster centers. In contrast, a random dictionary collected by a random selection of some local descriptors in the feature space, can lead to similar results as that obtained by an unsupervised clustering method. This is demonstrated in Fig. 2. Here, we use the pixel values of a 3×3 vectorized patch as a local feature vector; the patches are sampled regularly from the given images. Then we compare the results of vector quantization using k -means with the results of a random dictionary. From the results of vector quantization, we see that a random dictionary can achieve similar performance as k -means. This point is very important, because the time-consuming procedure for dictionary learning is avoided. Thus, it makes BoW applicable and scalable for large databases. Another advantage of this method is that we do not have to load all the features into memory. Only the random dictionary is needed to be loaded into memory. Thus, the memory requirements are significantly reduced. This is very important for large datasets because they probably will not fit into memory in many cases.

3. DATASETS AND EXPERIMENTS

3.1. Datasets and Setup

Two datasets were used for evaluation. The first one is composed of 15 classes of altogether 3434 TerraSAR-X sub-scenes with a size

of 160×160 pixels and a pixel spacing of about 3 m. Example images are shown in Fig. 3. The second is the UCMerced land use dataset [5]². It comprises 21 classes and each class has 100 images. Example images from each class are shown in Fig. 2. The classifier used by us is a C -SVM [12] with a kernel function $\chi^2(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^K \frac{2x_i y_i}{x_i + y_i}$. The parameter C is empirically set to 1000. The classification accuracy is measured in 20 test runs and we show their average accuracy.

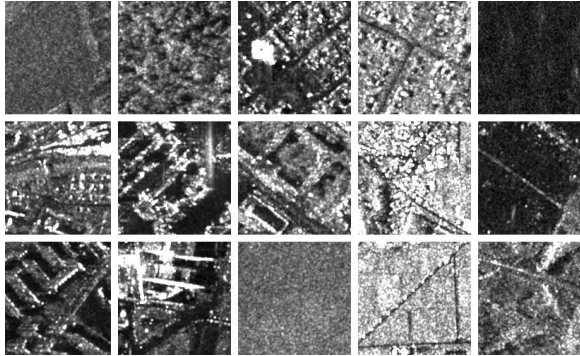


Fig. 3. Example images of 160×160 pixels from 15 classes of 3434 TerraSAR-X images being used for evaluation. The number of images in each class ranges from 118 to 430.

3.2. Evaluation of Local Features

Two evaluations are performed in this section. In the first evaluation, different window sizes (from 3×3 to 21×21 pixels) are used for patch sampling. We compare regular dense patch sampling (with and without overlap) with random sampling of differently sized patches while keeping the number of patches fixed. In case of random sampling, the row and column positions of the patches are determined by drawing random samples from a uniform distribution. The results of the first evaluation is shown in Fig. 4(a). The number next to each point on the curve is the number of patches having been sampled from an image. We can observe that the accuracy decreases as the patch size increases for both sampling methods. In addition, regular sampling with overlap performs better than that without overlap. There is no much difference between regular and dense sampling as long as the entire image can be fully covered. In the second evaluation, we show a comparison of local features using different dictionary sizes. We choose the methods proposed in [9], namely SRP Global, SRP Square, SRP Circular, SRP Radial-Diff, and SRP Angular-Diff, for comparison since they perform quite well. The resulting classification accuracy versus dictionary size is shown in Fig. 4(b). It can be clearly seen that for a sufficiently large dictionary with more than 200 entries the performances of the SRP Global, SRP Square, and SRP Circular options are not much different from our baseline method. However, our method performs much better than the SRP Radial-Diff and SRP Angular-Diff options for all dictionary sizes.

3.3. Evaluation of Random Dictionaries

In this section, we compare random dictionary learning and k -means dictionary learning in terms of classification accuracy. Two evalua-

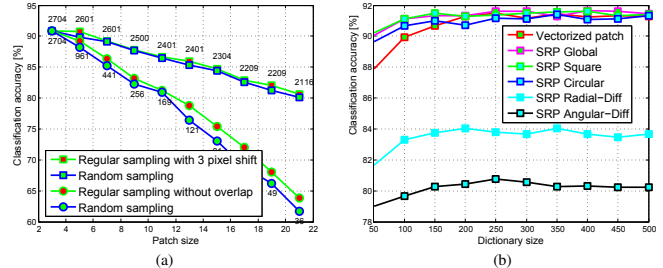


Fig. 4. Evaluation of local features: (a) Comparison of regular sampling and random sampling with the same number of patches but different patch sizes; (b) comparison of different local feature extraction methods.

tions are performed. In the first evaluation, we use vectorized patches of 3×3 pixel windows as low level feature vectors. The elements in the random dictionary are randomly selected from all the local feature vectors. The classification accuracy versus dictionary size is shown in Fig. 5(a). We can clearly see that a random dictionary and the one learned using k -means perform similarly. In the case of large dictionaries, a random dictionary is even better. This is very important for practical applications as dictionary learning using k -means is usually quite time consuming and may become prohibitively slow for large datasets. In the second evaluation, we change the number of training samples while keeping fixed the patch size of 3×3 pixels and the dictionary size of 200 entries. The classification accuracy versus the number of training samples is shown in Fig. 5(b). We can clearly see that they have almost the same performances. From these two evaluations, we see that it is not necessary to spend time with learning a dictionary using unsupervised learning methods.

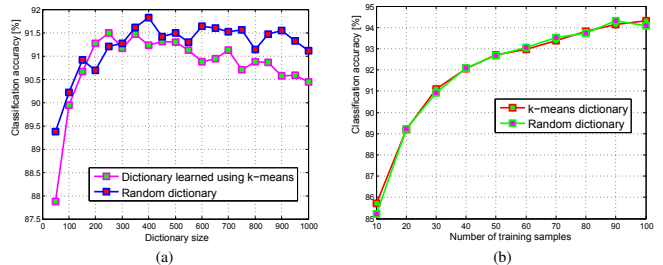


Fig. 5. Comparison of a random dictionary with a dictionary learned using k -means: (a) using different dictionary sizes; (b) using different numbers of training samples.

3.4. Comparison using SAR images

In the last experiment, we compare the BoW method using vectorized pixels of a 3×3 patch and a random dictionary with state-of-the-art feature extraction methods, namely Gabor feature extraction, GLCM feature extraction, wavelet feature extraction, and feature extraction based on Short-Time Fourier Transform (STFT), Quadrature Mirror Filters (QMF) and Fractional Fourier Transform (frFT).

- Two sets of Gabor texture features: the mean and the variance of the sub-bands [16], as well as the log-mean and the log-variance of the sub-bands [17]. The number of scales and orientations are set to 4 and 6.
- GLCM texture features [18]: we set the number of quantization levels to 32. The number of orientations is set to 4 and the number of shifts ranges from 1 to 4. Twenty statistics are computed from a co-occurrence matrix.

²<http://vision.ucmerced.edu/datasets/landuse.html>

Table 1. Accuracy comparison with previously reported accuracies on the UC Merced dataset.

Method	BoW [13]	SPMK [14]	SPCK [13]	SPCK+ [13]	SPCK++ [13]	UFL [15]	Color Histogram [5]	Our Method
Accuracy	71.86%	74.00%	73.14%	76.05%	77.38%	81.67%	81.19%	87.67%

- Two sets of wavelet features: the means and the variances of the sub-bands of a non-decimated 2D wavelet transform (NDWT) and a dual tree complex wavelet transform (DTCWT) [19] with 3 levels as well as the log-means and log-variances. For NDWT, a Daubechies filter is applied, while in DTCWT, near-symmetric 13,19-tap filters are being used for the first level and Q-Shift 14,14-tap filters are employed for all higher levels.
- STFT features [20]: the mean and variance, the spectral centroid and the spectral flux in horizontal and vertical direction of the short-term Fourier transform.
- QMF features [21]: the mean and variance of all the sub-bands in the pyramid with 3 levels.
- Features based on Fractional Fourier transform (frFT) [17] are the log-moment and log-variance of all sub-bands. The number of angles is set to 18.

The classification accuracies of our feature extraction methods (including some logarithmic versions of known methods) for all 15 SAR image classes are shown in Fig. 6. It can be clearly seen that the BoW method using vectorized patch pixels and a random dictionary performs significantly better than all other methods and has an average accuracy of more than 90%. In contrast, the average accuracies of all the other methods are lower than 90%. Log-Gabor and log-DTCWT have similar performances next to BoW, followed by frFT that performs better than all the remaining methods. In addition, we can see that the logarithmic versions of Gabor, NDWT, and DTCWT perform better than their linear counterparts. The STFT method lies far behind; the reason for it could be the lower dimension of its feature vector.

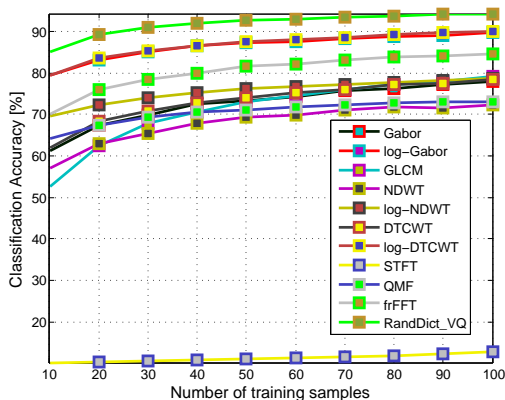


Fig. 6. Comparison of the BoW method with state-of-the-art feature extraction methods.

3.5. Comparison using the UC Merced land use dataset

In this section, we evaluate our method on the UC Merced land use dataset and compare our method with other state-of-the-art methods. The two methods we choose for comparison are spatial pyramid co-occurrence [5] [13] and the unsupervised feature learning method [15]. The spatial pyramid co-occurrence method extends the spatial pyramid kernel, which is a concatenation of the BoW feature vectors of all patches on a multi-resolution grid. In contrast, the unsuper-

vised feature learning method [15] follows a conventional procedure of unsupervised feature learning, which comprises two steps, namely dictionary learning and feature coding. Both methods have been evaluated on the UC Merced dataset. We follow the same experimental setup for both methods. 80 images from each class of the dataset are randomly selected as training data and the remaining data are used as test data. For our method, we employ the vectorized pixel values from a 3×3 local window as low level feature vectors and use a random dictionary. All classifications are performed in 20 test runs and we present their average accuracy. Then we compared our results with other methods. The average accuracy of all classes is 86.42%, as shown in Table. 1. The accuracy of our method is 5% better than the best one reported in [15]. In addition, our method is much simpler in terms of both computational effort and memory requirements.

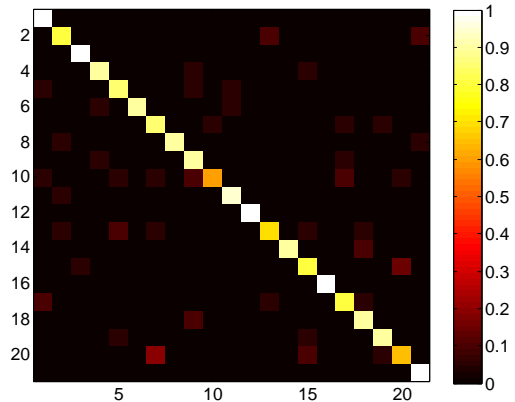


Fig. 7. Classification confusion matrix of our proposed method.

4. CONCLUSION

In this paper, we focus on remote sensing image classification, including both optical and SAR images. We propose a simple yet quite effective method in the BoW framework. It has two main contributions. The first contribution is that our method does not need to extract any low level features using some complex algorithms during a pre-processing step, which normally requires a certain amount of computational effort; instead, vectorized pixel values from very small local windows yield a superior performance for our dataset. The second contribution is that a random dictionary can achieve the same performance as one learned via clustering, which is usually a very time-consuming step. In the case of large datasets, this clustering step can make a method infeasible for large datasets. We performed an extensive investigation of the BoW method and these two contributions have been demonstrated in detail. In addition, we give clear answers to some other relevant but critical questions about BoW feature extraction. These two advantages over conventional methods not only significantly reduce the computational burden but also decrease the memory requirements, thus making the BoW method applicable and scalable for large satellite image databases.

5. REFERENCES

- [1] J. Sivic and A. Zisserman, "Video Google: A Text Retrieval Approach to Object Matching in Videos," in *Proc. Ninth IEEE International Conference on Computer Vision, ICCV '03*, Washington, DC, 2003, vol. 2, pp. 1470–1477.
- [2] M. Lienou, H. Maitre, and M. Datcu, "Semantic Annotation of Satellite Images Using Latent Dirichlet Allocation," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 1, pp. 28–32, Jan. 2010.
- [3] S. Xu, T. Fang, D. Li, and S. Wang, "Object Classification of Aerial Images With Bag-of-Visual Words," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 2, pp. 366–370, Apr. 2010.
- [4] H. Sun, X. Sun, H. Wang, Y. Li, and X. Li, "Automatic Target Detection in High-Resolution Remote Sensing Images Using Spatial Sparse Coding Bag-of-Words Model," *IEEE Geosci. Remote Sens. Lett.*, vol. 9, no. 1, pp. 109–113, Jan. 2012.
- [5] Y. Yang and S. Newsam, "Bag-Of-Visual-Words and Spatial Extensions for Land-Use Classification," in *Proc. 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, GIS '10*, New York, NY, 2010, pp. 270–279.
- [6] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
- [7] M. Varma and A. Zisserman, "A Statistical Approach to Material Classification Using Image Patch Exemplars," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 2032–2047, Nov. 2009.
- [8] L. Liu and P. Fieguth, "Texture Classification from Random Features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 3, pp. 574–586, Mar. 2012.
- [9] L. Liu, P. Fieguth, D. Clausi, and G. Kuang, "Sorted random projections for robust rotation-invariant texture classification," *Pattern Recognition*, vol. 45, no. 6, pp. 2405–2418, Jun. 2012.
- [10] B. Fernando, E. Fromont, D. Muselet, and M. Sebban, "Supervised learning of Gaussian mixture models for visual vocabulary generation," *Pattern Recognition*, vol. 45, no. 2, pp. 897–907, Feb. 2012.
- [11] F. Moosmann, B. Triggs, and F. Jurie, "Fast Discriminative Visual Codebooks using Randomized Clustering Forests," in *Advances in Neural Information Processing Systems (NIPS19)*, Vancouver, B.C., Canada, 2006, pp. 985–992.
- [12] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 27:1–27:27, 2011, Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [13] Y. Yang and S. Newsam, "Spatial Pyramid Co-occurrence for Image Classification," in *Proc. 2011 International Conference on Computer Vision ICCV 11*, Washington, DC, 2011, pp. 1465–1472.
- [14] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. Conference on Computer Vision and Pattern Recognition, CVPR06*, Washington, DC, 2006, vol. 2, pp. 2169–2178.
- [15] A. M. Cheryadat, "Unsupervised Feature Learning for Aerial Scene Classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 439–451, Jan. 2014.
- [16] B. S. Manjunath and W. Y. Ma, "Texture Features for Browsing and Retrieval of Image Data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 8, pp. 837–842, Aug. 1996.
- [17] 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. 1–10, Dec. 2013.
- [18] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural Features for Image Classification," *IEEE Trans. Syst., Man, Cybern.*, vol. 3, no. 6, pp. 610–621, Nov. 1973.
- [19] I. W. Selesnick, R. G. Baraniuk, and N. C. Kingsbury, "The dual-tree complex wavelet transform," *IEEE Signal Process. Mag.*, vol. 22, no. 6, pp. 123–151, Nov. 2005.
- [20] A. Popescu, I. Gavut, and M. Datcu, "Complex SAR image characterization using space variant spectral analysis," in *Proc. IEEE Radar Conference, RADAR 08*, Rome, Italy, 2008, pp. 1–4.
- [21] E. P. Simoncelli and E. H. Adelson, "Non-Separable Extensions of Quadrature Mirror Filters to Multiple Dimensions," in *Proceedings of the IEEE*, Apr. 1990, vol. 78, pp. 652–664.