Bag-of-Visual-Words Model for Classification of Interferometric SAR Images

Nazli Deniz Cagatay*, German Aerospace Center (DLR), nazli.kahyaoglu@dlr.de, Germany
Mihai Datcu, German Aerospace Center (DLR), mihai.datcu@dlr.de, Germany

Abstract

This work introduces a well-accepted image representation model in image analysis, namely the Bag-of-Visual-Words (BoVW), to interferometric SAR (InSAR) images. As the low-level local features, Gabor- and fractional Fourier transform (FrFT)-based feature descriptors are used. The supervised classification results with BoVW-Gabor and BoVW-FrFT features are compared to those with global Gabor and global FrFT features. Although the global Gabor features are better than the global FrFT features, by the implementation of BoVW model, FrFT outperforms Gabor features. Also, the classification performances of different baseline acquisitions for the same scenes are compared. For each baseline, the mean and individual class accuracies are improved by using BoVW-FrFT features.

1 Introduction

Today, automated and fast exploration of large databases is important in many fields. Earth observation is one of them, where many space- and airborne images pile up every day. Among various imaging technologies, SAR and InSAR are special candidates for scene classification.

On the other hand, recent developments in image processing are consistently introduced to remote sensing field. BoVW model is one of such successful techniques used in target detection, object and scene classification, etc.

In this work, BoVW model is adapted for classification of InSAR images by paying a special attention to the feature extraction step, as the acquisition geometry of such images plays an important role in the interpretation of the image content. The framework of BoVW model and its implementation for InSAR images are given in Section 2 and Section 3, respectively. The experimental database and the results are presented in Section 4 and Section 5.

2 Bag-of-Visual-Words Model

A text document can be represented as an orderless collection of words from a dictionary by using the frequencies of these words. This model is referred to as *Bag-of-Words (BoW) model*, which is a prevalent method in document analysis, such as document classification. The *Bag-of-Visual-Words (BoVW) model* is an adaptation of BoW model to image analysis, where each image is represented as an orderless collection of *visual words* from a *visual dictionary* by a histogram of these visual words. The BoVW image representation consists of 5 steps:

Step 1. Local feature detection: The images are divided into overlapping or non-overlapping patches using a regular grid as in dense sampling [1]. Also, the local features can be detected by random sampling or by a more complex detector, such as interest point detector [2]. Without loss of generality, these detected feature regions will be referred to as *local patches* in the rest of this paper.

Step 2. Local feature extraction: For the local patches detected in Step 1, the low-level feature descriptors are extracted. The most common local feature in literature is the scale-invariant feature transform (SIFT) [3]. Also, there are studies which use very simple statistics such as mean and standard deviation [4], and even the pixel values themselves [1] within a small local patch.

Step 3. Dictionary learning: Generally, an unsupervised clustering algorithm, such as K-means clustering, Gaussian mixture models [5] or random forests [6], is used to learn the visual dictionary using all local features from all images in the database. The cluster centroids obtained are referred to as the *visual words* in the visual dictionary.

Step 4. Feature coding: Once the visual dictionary is learned, the next step is to assign each local feature to a visual word in the visual dictionary. This can be done by hard voting, soft coding or Fisher coder.

Step 5. Feature pooling: In the last step, pooling is performed to generate a fixed length image representation. Max-pooling and sum-pooling are the common methods.

BoVW feature is an intermediate model relying on the low-level features. The well-known bottleneck of this model is the lack of global spatial relation of the local features within an image, or in other words, its inability of re-localizing objects in the image. Nevertheless, this model has already proved itself as a successful method in image retrieval thanks to its fixed-length feature descriptor and ability to discriminate local objects in an image.

3 BoVW Model for InSAR Images

In this work, the BoVW model is used for classification of single-look complex (SLC) SAR and InSAR images. For this purpose, in the local feature detection step, non-overlapping dense sampling is used with different local patch sizes. The visual dictionary is learned by means of K-means clustering for different number of clusters, and the local features are assigned to the nearest visual words using hard voting. Hard voting is followed by sum-pooling, which is equivalent to computing a histogram. The histogram of visual words for an image constitutes the fixed-length BoVW feature descriptor, which will later represent that image in a supervised classifier.

In the local feature extraction step of BoVW model, two different low-level features, namely the *Gabor-based* and *FrFT-based* features, which have been previously used for classification of SLC SAR and InSAR images [7], [8], are computed. The details of these low-level features are presented in the following Sections 3.1 and 3.2.

3.1 Low-Level Gabor Features

The first local feature used in this work is the Gabor-based feature descriptor, which is a well-known multiscale approach capturing the textural information [9]. It is important to note that, due to the sub-aperture decomposition of SLC SAR images in case of wavelet-based representation such as Gabor filter banks, the amplitude of the local patches for SLC SAR images is used [7].

In order to construct a local feature descriptor, Gabor filter banks with 3 scales and 4 orientations are implemented. Then, the 3 log-cumulants (*log-mean, log-variance and log-skewness*) of the full Gabor response (i.e., the real and imaginary parts of Gabor filtered local patch) are appended [7], [9]. The feature descriptor of length 72 for the $(i,j)^{th}$ local patch looks as follows:

$$\begin{aligned} F_G\left(i,j\right) = \big[\, & \mu^R_{1,1},\ \mu^I_{1,1},\ \sigma^R_{1,1},\ \sigma^I_{1,1},\ \gamma^R_{1,1},\ \gamma^I_{1,1}, \\ & \dots\ \mu^R_{S,O},\ \mu^I_{S,O},\ \sigma^R_{S,O},\ \sigma^I_{S,O},\ \gamma^R_{S,O},\ \gamma^I_{S,O}\,\big]\,, \end{aligned} \tag{1}$$

where $\mu^R_{S,O}$, $\sigma^R_{S,O}$, $\gamma^R_{S,O}$ and $\mu^I_{S,O}$, $\sigma^I_{S,O}$ and $\gamma^I_{S,O}$ are the log-cumulants computed from the real and imaginary parts of the Gabor filtered local patch for scale $S \in \{1,2,3\}$ and orientation $O \in \{1,2,3,4\}$.

3.2 Low-Level FrFT Features

FrFT is the generalization of standard Fourier Transform and it decomposes the signal into chirps as in (2) where α is the transform angle [10].

$$\begin{aligned} F^{\alpha}\left(\xi\right) &= A_{\alpha} \cdot exp(j\pi\xi^2 \cot\alpha) \\ &\quad \cdot \int \exp[j\pi(-2x\xi \csc\alpha + x^2 \cot\alpha)] f(x) dx \\ A_{\alpha} &= \frac{exp[-j(\pi\,\mathrm{sgn}(\sin\alpha)/4 - \alpha/2)]}{|\sin\alpha|^{1/2}} \end{aligned} \tag{2}$$

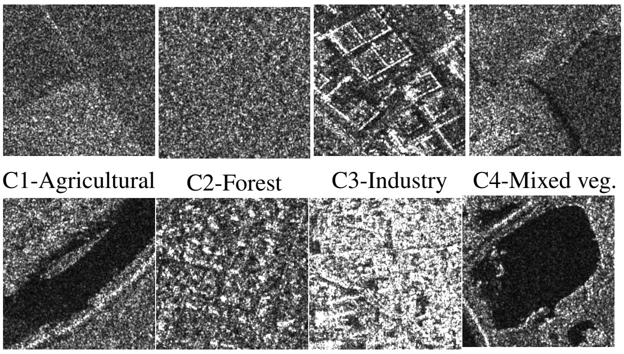
In [11], the FrFT is designated as an appropriate multiscale approach for SLC SAR images, where the scaling is performed in the phase by controlling the transform angle. Hence, as the second local feature, FrFT-based feature descriptors are computed from 9 images transformed with different angles equally spaced between 0 and π.

As in Gabor-based features, the 3 log-cumulants (μ^R_α, σ^R_α, γ^R_α and μ^I_α, σ^I_α, γ^I_α) are computed from the full response of local patches in the FrFT domain for each transform angle $\alpha \in \{0, 0.125\pi, 0.25\pi, \dots, \pi\}$, and then, appended to form the feature descriptor of length 54 [8]:

$$\begin{aligned} F_{FrFT}\left(i,j\right) = \big[\, & \mu^R_0,\ \mu^I_0,\ \sigma^R_0,\ \sigma^I_0,\ \gamma^R_0,\ \gamma^I_0, \\ & \dots\ \mu^R_\pi,\ \mu^I_\pi,\ \sigma^R_\pi,\ \sigma^I_\pi,\ \gamma^R_\pi,\ \gamma^I_\pi\,\big] \end{aligned} \tag{3}$$

4 Experimental Image Database

The database used in this work is composed of 3 sets of StripMap SLC SAR and InSAR images acquired by TanDEM-X mission with 3 different effective baselines (88.2m, 128.4m, 229.6m) over Toulouse, France. The size of an image is 200 x 200 pixels corresponding to an area of about 400m x 340m on ground. There are 400 images from 8 classes representing different scenes consisting of natural and man-made structures. These classes are exemplified by a sample image in **Figure 1**.



C1-Agricultural C2-Forest C3-Industry C4-Mixed veg.

C5-Riverside C6-Urban 1 C7-Urban 2 C8-Waterbody

Figure 1: Classes in the InSAR image database.

5 Experimental Results

The local patch sizes (AW) used in this work are 10, 20 and 40. Then, the corresponding total numbers of local patches per image are 400, 100 and 25, respectively.

The choice of the visual dictionary size (K) is important for the performance of the method. If the dictionary size is too small, the visual words do not represent all the local patches. On the other hand, a too large dictionary will be vulnerable to noises, yielding very similar local patches to be assigned to different visual words. Also, the dictionary size determines the final BoVW feature descriptor length. Usually, the dictionary size varies from several hundreds to thousands and tens of thousands [12]. However, in this work, since the database is relatively small, smaller dictionary sizes are used ($K = 20, 30, 40, 50$).

For the supervised classification of images, k-nearest neighbor (KNN) classifier is used with Euclidean distance and $k = 1$. The 4% of the images are used as the training set. The classification experiment is repeated 100 times, each with a different set of training samples, and the average accuracies are presented.

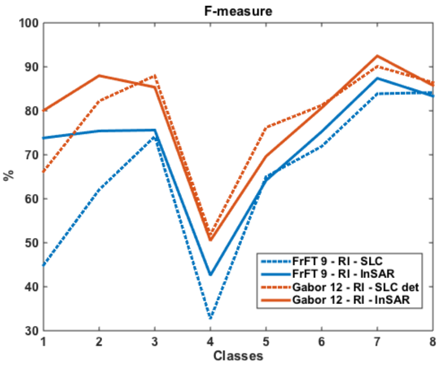
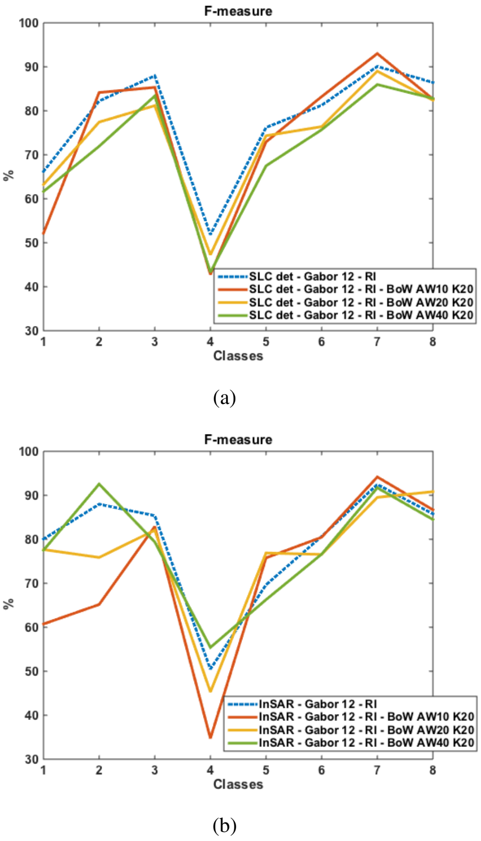


Figure 2: Performance of global Gabor and global FrFT features for SLC/SLC detected and InSAR data.

First, the global Gabor and FrFT features extracted from the *whole image* are compared for SLC (or SLC detected, in case of Gabor features) and InSAR images [7], [8]. The individual class accuracies are presented in **Figure 2**. As it can be seen from this figure, the global Gabor features are more successful for this database than the global FrFT features. Also, for both features, the use of InSAR improves the accuracies especially for C1 and C2.

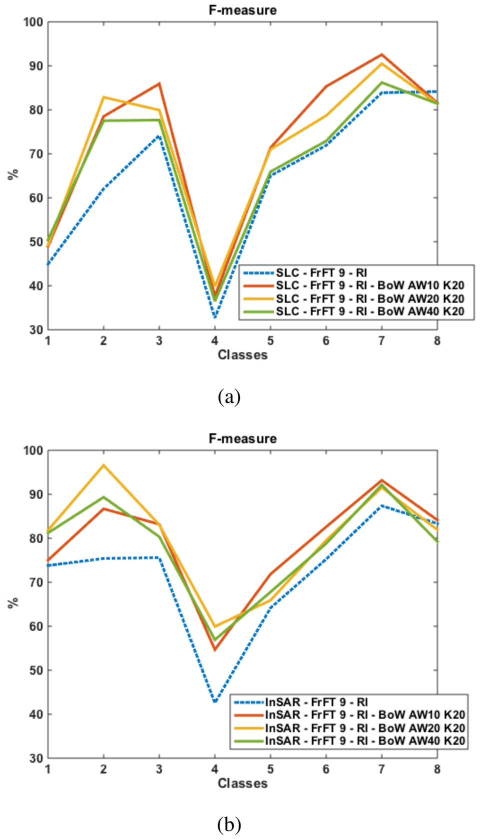
Next, the BoVW model is implemented based on the low-level Gabor features for different local patch sizes. It can be seen in **Figure 3(a)** and **Figure 3(b)** that the BoVW model degrades the accuracies for both SLC detected and InSAR images compared to the global Gabor features.



(a)

(b)

Figure 3: Performance of global Gabor and BoVW-Gabor features for (a) SLC detected and (b) InSAR data.



(a)

(b)

Figure 4: Performance of global FrFT and BoVW-FrFT features for (a) SLC and (b) InSAR data.

On the other hand, the implementation of BoVW model improves the performance for FrFT features for both SLC SAR and InSAR images for almost all classes in the database as shown in **Figure 4(a)** and **Figure 4(b)**.

In order to make a comparison of BoVW-Gabor and BoVW-FrFT features for InSAR images, BoVW model is implemented for all local patch and dictionary sizes mentioned above. Then, the best features of each group are selected and plotted in **Figure 5**.

For the BoVW-FrFT, the overall best result is obtained with $AW = 20$ and $K = 30$. The mean accuracy (MA) for this feature is 81.79%. On the other hand, for the BoVW-Gabor, it is hard to find an overall best feature for all classes. For instance, it is observed that the BoVW-Gabor with $AW = 40$ and $K = 20$ improves the classification performance of classes C1, C2 and C4 only, and the accuracies drop for other classes. Nevertheless, the BoVW-Gabor with $AW = 20$ and $K = 50$ can be chosen as the overall best feature descriptor considering the highest mean accuracy of 78.55% and relatively balanced individual class accuracies.

As it can be seen in Figure 5, although the global Gabor features give better classification results for InSAR images than the global FrFT features, the FrFT-based features outperform the Gabor-based ones by the implementation of BoVW model on these low-level features.

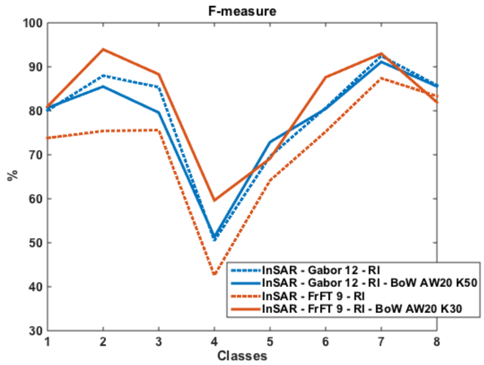


Figure 5: Performance of Gabor-based and FrFT-based features with and without BoVW model for InSAR data.

Finally, for the BoVW-FrFT features from InSAR images with $AW = 20$ and $K = 20$, the effect of baseline (BL) is summarized in **Table 1**. The mean and individual class accuracies show that BoVW is quite successful for almost all classes. The overall improvement in mean accuracy for all BLs is 10-11%. For small BL, the implementation of BoVW model improves the accuracies almost equally except for C5 and C8. For medium BL, the improvement of natural classes (C1, C2, C4) is significant, while for large BL, the biggest classification improvements are observed for urban-like classes (C3, C6, C7). That is, the spatial discriminability of BoVW model together with the strong interferometric signature of urban structures for large BL results in a good classification performance.

Table 1: Effect of baseline on classification accuracy

Class index	**Small BL**		**Medium BL**		**Large BL**	
	FrFT	**BoVW-FrFT**	**FrFT**	**BoVW-FrFT**	**FrFT**	**BoVW-FrFT**
C1	77.54	**85.41**	73.82	**81.81**	73.22	**79.36**
C2	82.21	**92.24**	75.42	**96.60**	88.83	**98.30**
C3	68.47	**81.77**	75.61	**83.07**	72.02	**90.45**
C4	52.22	**64.60**	42.58	**59.94**	57.53	**60.81**
C5	**70.52**	70.28	64.16	**65.90**	62.17	**65.76**
C6	69.61	**80.94**	75.26	**79.53**	66.31	**88.84**
C7	79.42	**89.88**	87.39	**91.55**	83.42	**90.58**
C8	**88.03**	83.88	**83.34**	81.92	**83.45**	80.57
MA	73.78	**81.26**	72.54	**80.04**	73.57	**81.82**

6 Conclusions

In this work, the BoVW model, which is a well-known image representation method, is adapted to SLC SAR and InSAR images with Gabor and FrFT being the low-level features. The classification results show that although the global Gabor features are better than the global FrFT features, by the implementation of BoVW model, FrFT outperforms Gabor features. Moreover, the method is assessed for 3 different baselines, and for each baseline, the FrFT-based BoVW feature is found to improve the mean accuracy by 10-11%. Also, the spatial discriminability of BoVW model together with the strong interferometric signature of urban structures for large baseline results in good classification performance. In the light of the initial results, this work will be extended to a larger database and the low-level local features will be further improved.

References

[1] S. Cui, G. Schwarz, and M. Datcu. Image Classification: No Features, No Clustering. In *Proc. of ICIP*, Sep. 2015.

[2] K. Mikolajczyk and C. Schmid. Scale and Affine Invariant Interest Point Detectors. *Int. J. Comput. Vision*, 60(1):63–86, Jan. 2004.

[3] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vision*, 60(2):91–110, Nov. 2004.

[4] M. Lienou, H. Maitre, and M Datcu. Semantic Annotation of Satellite Images Using Latent Dirichlet Allocation. *IEEE Geosci. Remote Sens. Lett.*, 7(1):28–32, Jan. 2010.

[5] B. Fernando, E. Fromont, D. Muselet, and M. Sebban. Supervised Learning of Gaussian Mixture Models for Visual Vocabulary Generation. *Pattern Recognition*, 45(2):897–907, Feb. 2012.

[6] F. Moosmann, B. Triggs, and F. Jurie. Fast Discriminative Visual Codebooks Using Randomized Clustering Forests. In *Proc. of NIPS*, pages 985–992, 2006.

[7] J. Singh and M. Datcu. SAR Image Categorization with Log Cumulants of the Fractional Fourier Transform Coefficients. *IEEE Trans. Geosci. Remote Sens.*, 51(12):5273–5282, Dec. 2013.

[8] N. D. Cagatay and M. Datcu. FrFT Based Scene Classification of Phase Gradient InSAR Images and Effective Baseline Dependence. *IEEE Geosci. Remote Sens. Lett.*, 12(5):1131–1135, May 2015.

[9] B. S. Manjunath and W. Y. Ma. Texture Features for Browsing and Retrieval of Image Data. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18(8):837–842, Aug. 1996.

[10] H. M. Ozaktas, B. Barshan, D. Mendlovic, and L. Onural. Convolution, Filtering, and Multiplexing in Fractional Fourier Domain and Their Relation to Chirp and Wavelet. *J. Opt. Soc. Am.*, A(11):547–559, Feb. 1994.

[11] M. Datcu and J. Singh. Phase-Scale Analysis of Complex Valued SAR Images. In *Proc. of EUSAR*, pages 1121–1124, June 2014.

[12] J. Yang, Y. G. Jiang, A. Hauptmann, and C. W. Ngo. Evaluating Bag-of-Visual-Words Representation in Scene Recognition. In *Proc. of ACM SIGMM MIR*, pages 197–206, Sep. 2007.