A Comparative Study of Bag-of-Words and Bag-of-Topics Models of EO Image Patches

Reza Bahmanyar, Shiyong Cui, and Mihai Datcu, *Fellow, IEEE*

***Abstract*—The large volume of detailed land cover features, provided by high resolution Earth Observation (EO) images, has attracted considerable interest in the discovery of these features by learning systems. In this paper, we perform Latent Dirichlet Allocation (LDA) on the Bag-of-Words (BoW) representation of collections of EO image patches to discover their semantic level features, the so-called *topics*. To assess the discovered topics, the images are represented based on the occurrence of different topics, called *Bag-of-Topics (BoT)*. The value added by BoT to the BoW model of image patches is then measured based on existing human annotations of the data. In our experiments, we compare the classification accuracy results of BoT and BoW representations of two different remote sensing image datasets, a multi-spectral optical dataset and a Synthetic Aperture Radar (SAR) dataset. Experimental results demonstrate that BoT can provide a compact and semantically meaningful representation of data; it either causes no significant reduction in the classification accuracy or increases the accuracy by a sufficient number of topics.**

***Index Terms*—Bag-of-Words, Earth Observation, Latent Dirichlet Allocation, SAR images.**

I. Introduction

HIGH spatial resolution Earth Observation (EO) images represent land cover in much detail. This allows a better understanding of the contents of images by distinguishing more object categories (e.g. grass, buildings, roads). Exploring the full amount of detailed information requires the development of efficient learning systems which are able to provide relevant results to the users' understanding of the data. Although users understand images by recognizing their semantic level contents (objects or their parts), most of the current learning systems are based on a primitive representation of images. Recently, Bag-of-Words (BoW), a simplifying method used in natural language processing, has been shown to provide promising compact representations of images [1]. In BoW, primitive image features (e.g., color, texture, shape) are extracted for every local region using various feature extraction methods, such as rgbHist [2], Gabor [3], and Scale-Invariant Feature Transform (SIFT) [4]. Then, the distribution of the primitive features in the entire image collection is modeled by a dictionary of visual words. The visual words are usually generated by applying a clustering method (e.g., k-means) on a sample set of the primitive feature descriptors. Assuming each cluster center as a visual word, the feature descriptors are then assigned to their nearest cluster center. Finally, each image is represented by a histogram of its visual words. BoW has been successfully modified and applied to EO images, too ([5], [6], [7]). Representing images using BoW models does not provide good estimate of image semantics due to disregarding the statistical relations between the visual words. However, it has been shown in previous works ([8], [9], [10], [11]) that these relations can result in the discovery of objects and their parts in images. These works have used generative models such as probabilistic Latent Semantic Analysis (pLSA) [12] and Latent Dirichlet Allocation (LDA) [13] for the unsupervised discovery of object parts, so-called *topics*. The concepts behind the images are then represented by mixtures of the discovered topics. The authors of [8] investigate pLSA and LDA for object categorization and localization. They demonstrated the possibility of recognizing and localizing object categories by learning from unlabeled image collections. In [9], the authors used the topics obtained by pLSA in combination with a nearest neighbor classifier for scene classification. They showed that the statistical model discovered by pLSA is appropriate for the classification of datasets with multiple object categories in each image. In [10], it has been shown that pLSA-based image representation improves the retrieval performance on large-scale datasets due to the compact descriptions of the contents of images. Inspired by [14], the authors of [11] verified that the topics discovered by LDA outperform the ones obtained by pLSA in large-scale retrieval tasks owing to the completely generative probabilistic model provided by LDA. Various extensions of LDA have then been introduced for scene classification and segmentation ([15], [16]).

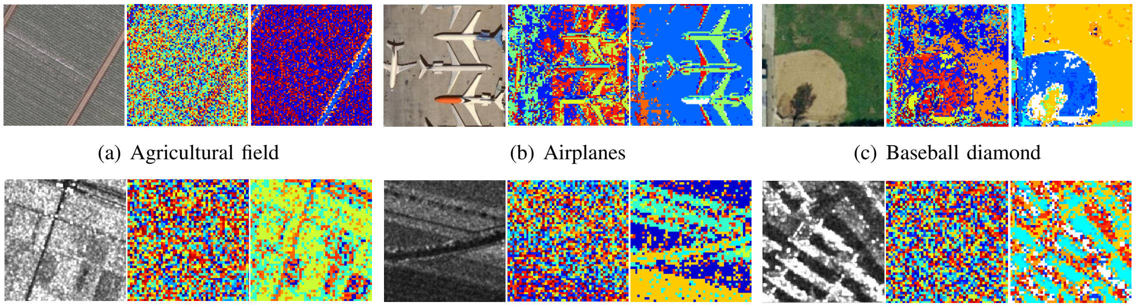
The high resolution of EO images has shifted the interest to patch level image analysis in recent years [17]. Conventionally, image patches are described by primitive features and a BoW model that represents the contents of the data at the signal level. In this paper, we study the discrimination of EO image patches using their semantic level representations, which we name *Bag-of-Topics (BoT)*, obtained from the BoW image models. To this end, EO image patches are represented using feature description methods such as Mean and Variance (MV) [7] and Gabor descriptors. Then LDA is applied to the BoW representation of the patches in order to discover the existing topics in the dataset. Finally, the contents of each image patch are represented as a mixture of the topics (BoT).

To evaluate the semantic level features described by the BoT model, a set of experiments are run on two annotated remote sensing image datasets, a multi-spectral optical patch collection and a Synthetic Aperture Radar (SAR) patch col-

R. Bahmanyar*, S. Cui, and M. Datcu* are with the Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Wessling, Germany (e-mail: gholamreza.bahmanyar@dlr.de; shiyong.cui@dlr.de; mihai.datcu@dlr.de).

*The authors are also affiliated with the Munich Aerospace Faculty, Munich, Germany.

Manuscript received July 23, 2014; latest revision January 20, 2015.



(a) Agricultural field (b) Airplanes (c) Baseball diamond

(d) Agricultural field (e) Highway (f) Urban area

Fig. 1. BoW and BoT representations of samples of multi-spectral (first row) and SAR (second row) EO images. For each sample, the first image is the original image. The second and the third images are BoW and BoT representations of the image, respectively. The various colors depict the visual words (in BoW) or the topics (in BoT). Dictionaries of 200 visual words generated from MV features are used. The BoT models of the images are made for 20 topics.

lection. In these experiments, a classification method (e.g., SVM[1]) is applied to the BoT and BoW representations of images. The accuracies and the run-times of the classifications are then compared for the two models. Experimental results demonstrate that BoT provides a compact representation of the data; however, it either causes no significant reduction, or in many cases, even increases the classification performance. While a compact representation improves the scalability of the learning systems by decreasing the computational effort, the semantic features (topics) are more discriminative.

The rest of the paper is organized as follows: Section II introduces the statistical topic models and LDA. Section III describes the semantic level representation of images. Experimental results and the efficiency of the BoT model are discussed in Section IV. Finally, Section V concludes the paper.

II. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a statistical generative model which has been developed to discover the topics occurring in text collections [13]. A topic is found by the occurrence of the words related to that topic, through all given text documents. This idea has been adapted to image analysis by assuming images as mixtures of visual patterns (topics) recurring through the entire corpus [18].

LDA is a three level directed graphical model. It assumes each image d as a combination of N_d visual words, $d = \{w_1, w_2, ..., w_{N_d}\}$ and each topic z_j as a distribution over a fixed dictionary of V visual words. In order to generate the n-th word (w_n) of the image d, a topic z_j is selected from the distribution $p(z_j|\theta_d)$ over a set of K topics, $Z = \{z_1, z_2, ..., z_K\}$, where θ_d is K-dimensional Dirichlet random variable corresponding to the image d. Then w_n is drawn from the multinomial distribution over the visual words in topic z_j, $p(w_n|z_j, \beta)$, where β is a matrix containing the distributions over the words for each topic. Thus, the word w_n is generated for the image d as follows:

$$p(w_n|\alpha, \beta) = \int p(\theta_d|\alpha) \left(\sum_{j=1}^{K} p(z_j|\theta_d) p(w_n|z_j, \beta) \right) d\theta_d, \tag{1}$$

where the parameters α and β determine the prior for Dirichlet distributions, and for a symmetric Dirichlet distribution, $p(\theta_d|\alpha)$ is computed as:

$$p(\theta_d|\alpha) = \frac{\Gamma(K\alpha)}{\Gamma^K(\alpha)} \prod_{j=1}^{K} \theta_{dj}^{\alpha-1}, \tag{2}$$

where $\Gamma(.)$ denotes the Gamma function.

In a learning phase, LDA finds the posterior distribution, i.e., the topic distribution of the images in the corpus. Due to the intractability of computing the posterior distribution, LDA uses approximation inference algorithms such as variational Expectation Maximization (EM) [13] to approximate it.

III. Semantic Level Feature Representation

In this section, we explain the representation of the semantic concepts of images using mixtures of topics. To this end, the primitive features of the images are extracted by feature extraction techniques such as MV and Gabor; each image is modeled as a BoW. In order to discover the existing topics, LDA is applied to the BoW models of the images. Then each image is represented by a simplex of the discovered topics, $p(\mathbf{z}|\theta_d, \alpha)$, as a vector in Euclidean space, where each element shows the occurrence of a particular topic in the image. We name this image representation a *BoT* model. Since the number of topics is usually much smaller relative to the number of visual words, BoT provides a more compact description of the images to be used in learning tasks such as classification. While the compact representation of the images helps to increase the scalability of the learning methods and reduces the computational time, semantically meaningful features lead to results being more relevant to the human understanding of the data. In this article, we evaluate the use of the BoT model in classifications of remote sensing images.

Figure 1 shows a comparison of BoW and BoT models in representing the contents of multi-spectral and SAR image patches. According to the samples, BoT can describe images with a few but semantically understandable topics; however, it is rather difficult to understand the semantics behind the signal level contents represented by BoW. For example, in Figure 1 (c), BoT describes the baseball diamond clearly by meaningful topics such as grass and sand, while the

[1] http://www.csie.ntu.edu.tw/ cjlin/libsvm

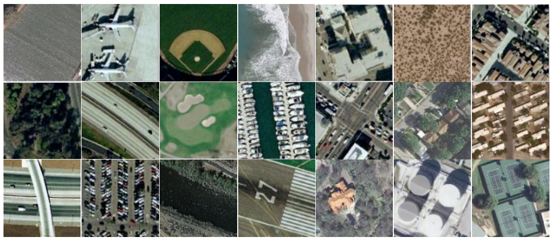


Fig. 2. The UCMerced-LandUse dataset contains 2100 images grouped into 21 land-use scenes: Agricultural, Airplane, Baseball Diamond, Beach, Buildings, Chaparral, Dense Residential, Forest, Freeway, Golf Course, Harbor, Intersection, Medium Density Residential, Mobile-home Park, Overpass, Parking Lots, River, Runway, Sparse Residential, Storage Tanks, Tennis Court.

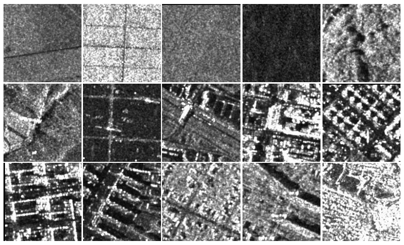


Fig. 3. Dataset of 3434 TerraSAR-X satellite images grouped into 15 classes. Top to bottom, left to right, the first two classes are Agricultural Fields. Then come Grass Fields, Water Surfaces, Forests, Mountains, Flooded Fields, Highways, Industrial Areas, and the rest are different kinds of Urban Areas.

semantics behind the visual words in the BoW model are hard to understand. The second row in Figure 1 shows that understanding the semantics behind the visual words is even more difficult for SAR data. In Figure 1 (e), for example, the BoW hardly represents any structure; however, a highway and its neighboring areas can be recognized in the BoT.

IV. Experiments and Discussion

In this section, we assess the semantic level descriptions of the two selected types of EO images, namely multi-spectral and SAR data. In order to analyze the images, we extract the primitive features to be used in generating the BoW models of the images. In the next step, LDA is applied to the BoWs in order to discover the latent structure behind the images as a set of topics (we use the LDA implementation of Blei[2]). Each image is then represented by a mixture of topics (BoT). Since the resulting topics are not unique, we run LDA three times for each experiment and average over the final results obtained by the three sets of topics. For evaluating the BoT representation, we use an SVM for classification. In order to generalize the task, we select randomly from every class 70 samples for training, 20 samples for parameter optimization, and the remainder for testing. The results are cross-validated by running 10 experiments and repeating each experiment 10 times. Finally, the accuracy and run-time are averaged over the experiments.

A. Datasets

In our experiments, we assess BoT models of two EO datasets. The first one is the *UCMerced-LandUse* dataset [6] which is a collection of 2100 multi-spectral image patches categorized into 21 classes of land-use scenes. Each class contains 100 images of 256×256 pixels from aerial cartography. The second dataset is called *15 TerraSAR-X Image Classes* dataset [19]. It contains 3434 SAR image patches of 160×160 pixels manually grouped into 15 non-equal size classes. Each class contains between 118 and 420 images. Figures 2 and 3 show some representative samples of these datasets.

B. Signal level image representation

For image analysis, their primitive features are extracted locally using MV and Gabor feature extraction techniques.

MV feature descriptors are obtained by computing the mean (μ) and variance (v) of the pixel values in local neighborhoods (non-overlapping windows of 3×3 pixels) of every image. Thus, the resulted feature vectors have two elements, $F_{MV} = [\mu \ v]$. These descriptors have been shown to achieve promising results in addition to their simple computation [17].

In a more complex scenario, Gabor descriptors are obtained by filtering a given image using Gabor filters [3]. These filters are linear band-pass filters generated by scaling and rotating a mother wavelet filter whose impulse responses are 2D modulated Gaussian functions. The Gabor feature vectors are then built by computing means (μ_{sr}) and standard deviations (σ_{sr}) of the response for S scales and R rotations, $F_{Gabor} = [\mu_{11} \ \sigma_{11} \ \mu_{12} \ \sigma_{12} \ \ldots \ \mu_{SR} \ \sigma_{SR}]$. These feature vectors have been shown to achieve promising results in texture analysis and EO tasks ([3], [19]). In our experiments, the features are extracted from local windows of 32×32 pixels with 50% overlap. The selection of $S = 3$ and $R = 6$ results in feature vectors of 36 elements.

In the next step, each image is represented by a BoW model of the primitive descriptors for various dictionary sizes (50, 100, 200, 300). To generate the dictionaries, k-means is applied to 10% of the feature vectors, selected randomly, where the cluster centers are considered as visual words. Each image is then modeled by a histogram of visual words obtained by assigning the feature vectors to their nearest visual words.

C. Semantic level image representation

In this step, LDA is applied to the BoW model of the images to discover the latent structure in each image collection. LDA represents the image structure as a set of topics. The semantic level of the topics is usually correlated to the number of topics discovered by LDA. More precisely, a small number of topics leads to general concepts (e.g., forest, urban area), while a larger number of topics provides more detailed contents (e.g., trees, buildings). Evaluating various numbers of topics allows us to asses the effects of different semantic levels in discriminating the image classes. Using the extracted topics, images are represented by mixtures of topics (BoT).

D. Results and discussion

In order to assess the value added by BoT to the BoW model, the performance of SVM in the classification of both representations of the EO datasets is measured. In our experiments, two primitive feature descriptors, namely MV

[2] http://www.cs.princeton.edu/ blei/lda-c/

and Gabor, are used for BoW generation. The horizontal lines plotted in Figures 4 and 5 depict the resulting classification accuracies and run-times, where the columns represent the results for various dictionary sizes. Then LDA is applied to each BoW model for different numbers of topics to build BoT models which are then used for classification; the results are also plotted in Figures 4 and 5.

Since the number of topics is usually smaller than the number of visual words, using BoT allows a compact representation of the data; it either causes no significant reduction in the performance or increases the classification accuracy by a sufficient number of topics. The BoT model can increase the discriminability of the descriptors, because concepts represented by topics are usually more descriptive than the contents described by visual words, as mentioned in Section III. For example, in Figure 4 (a), for 50 topics and for MV feature descriptors, BoT outperforms BoW; with a similar size of the BoW and BoT feature vectors, topics provide a more discriminable representation of the data. Moreover, BoT speeds up the classification by compacting the data representation. In Figure 4 (d), for example, the BoT (with 60 topics) built using the BoW (with 300 visual words) of MV feature descriptors obtains a higher accuracy than the BoW model and it is 15 times faster. According to Figure 5, for SAR data, BoT performs similarly to BoW; however, it is much faster and, therefore, more efficient than BoW.

Furthermore, comparing the two primitive feature descriptors indicates that the discriminability of the topics depends on the informativeness of the BoW model built upon primitive feature descriptors. For example, in Figures 4 and 5, since the discovered topics from MV are more discriminable than the Gabor topics, the BoW model of MV features results in higher classification accuracies than that of the Gabor descriptor.

Figure 6 shows how the dictionary size affects the visual words and the topics generated from the two datasets. It shows the classification accuracies and run-times versus dictionary size for the BoW model (the red solid curve) and the BoT models for various numbers of topics. As the results show, the performances usually improve sharply for small dictionary sizes, but they decrease for larger sizes. Furthermore, increasing the dictionary size brings about a higher dimensionality of the BoW descriptors which causes the run-time to increase dramatically. Since a larger number of visual words helps LDA to discover more descriptive topics; as a result, this leads to more discriminable descriptors, increasing the dictionary size usually speeds up the classification using BoT.

V. Conclusion

In this paper, LDA is applied to the BoW representation of two EO image patch collections to discover their semantic level features, so-called topics. Then, the patches are described as a mixture of the topics (BoT model). The BoT approach can be used in various learning scenarios such as image classification and retrieval. In this paper, it is evaluated in image classification by applying SVM to the BoT models of image patches. The results are then compared to the accuracies achieved by the BoW model. Experimental results demonstrate that semantic level features can provide comparable results to that of the BoW model; the description of data is much more compact in the BoT model. Consequently, BoT not only increases the scalability of learning systems, but also discriminates various image classes to a higher degree. In this paper, we show the effects of different number of topics in BoT on the classification performance. However, the selection of an optimized number of topics still deserves more detailed investigations.

Acknowledgment

The authors wish to thank Mr. Gottfried Schwartz who assisted in the proof-reading of this paper.

References

[1] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proc. Workshop on Statistical Learning in Computer Vision (ECCV)*, 2004, pp. 1–22.

[2] K. E. A. Van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1582–1596, 2010.

[3] B. S. Manjunath and W. Y. Ma, "Texture features for browsing and retrieval of image data," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 837–842, 1996.

[4] D. Lowe, "Object recognition from local scale-invariant features," in *Proc. IEEE International Conference on Computer Vision*, vol. 2, 1999, pp. 1150–1157.

[5] L. Zhao, P. Tang, and L. Huo, "Land-use scene classification using a concentric circle-structured multiscale bag-of-visual-words model," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, in print, 2015.

[6] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. ACM International Conference on Advances in Geographic Information Systems (GIS)*, 2010, pp. 270–279.

[7] M. Lienou, H. Maitre, and M. Datcu, "Semantic annotation of satellite images using Latent Dirichlet Allocation," *IEEE Geoscience and Remote Sensing Letters*, vol. 7, no. 1, pp. 28–32, 2010.

[8] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering object categories in image collections," in *Proc. International Conference on Computer Vision*, 2005.

[9] A. Bosch, A. Zisserman, and X. Muñoz, "Scene classification via pLSA," in *Proc. European Conference on Computer Vision (ECCV)*, 2006, pp. 517–530.

[10] R. Lienhart and M. Slaney, "PLSA on large scale image databases," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, 2007, pp. 1217–1220.

[11] E. Hörster, R. Lienhart, and M. Slaney, "Image retrieval on large-scale image databases," in *Proc. ACM International Conference on Image and Video Retrieval (CIVR)*, 2007, pp. 17–24.

[12] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *The Journal of Machine Learning*, vol. 42, no. 1-2, pp. 177–196, 2001.

[13] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[14] X. Wei and W. B. Croft, "LDA-based document models for ad-hoc retrieval," in *Proc. ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, 2006, pp. 178–185.

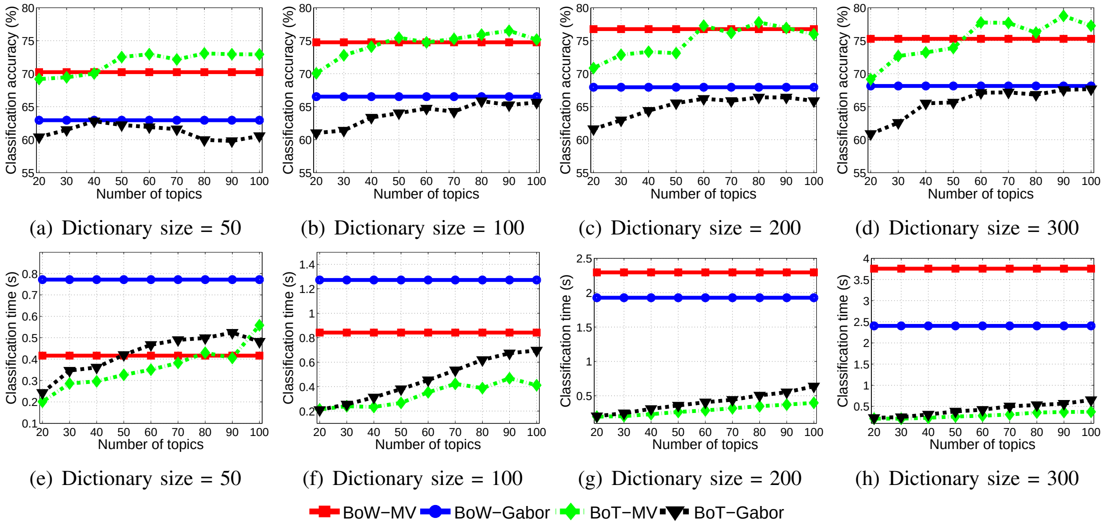
[15] L. Cao and L. Fei-Fei, "Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2007, pp. 1–8.

[16] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 524–531.

[17] S. Cui, C. O. Dumitru, and M. Datcu, "Ratio-detector-based feature extraction for very high resolution SAR image patch indexing," *IEEE Geoscience and Remote Sensing Letters*, vol. 10, no. 5, pp. 1175–1179, 2013.

[18] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

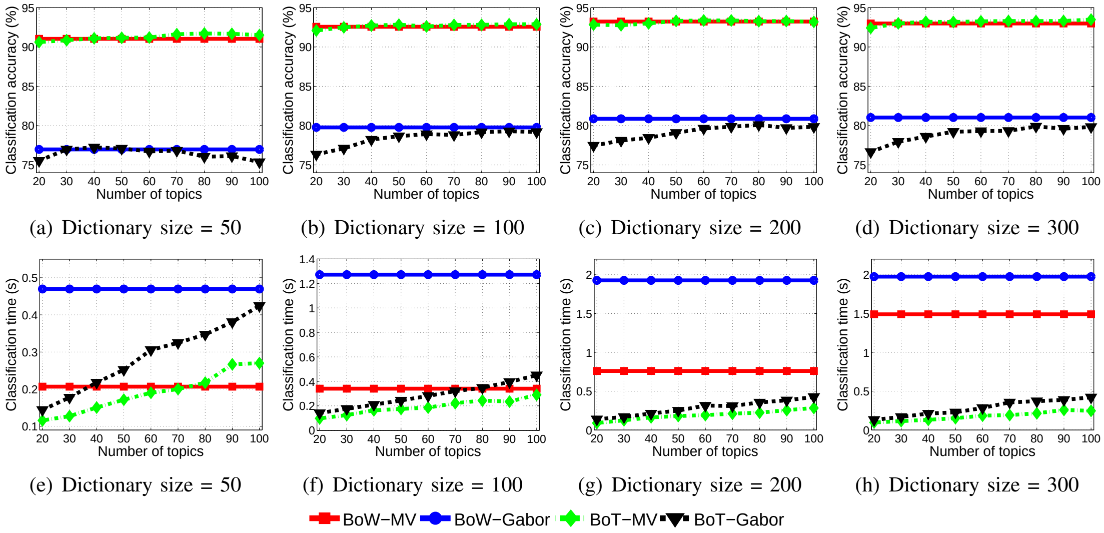
[19] C. Dumitru and M. Datcu, "Information content of very high resolution SAR images: Study of feature extraction and imaging parameters," *IEEE Trans. Geoscience and Remote Sensing,*, vol. 51, no. 8, pp. 4591–4610, 2013.



(a) Dictionary size = 50 (b) Dictionary size = 100 (c) Dictionary size = 200 (d) Dictionary size = 300

(e) Dictionary size = 50 (f) Dictionary size = 100 (g) Dictionary size = 200 (h) Dictionary size = 300

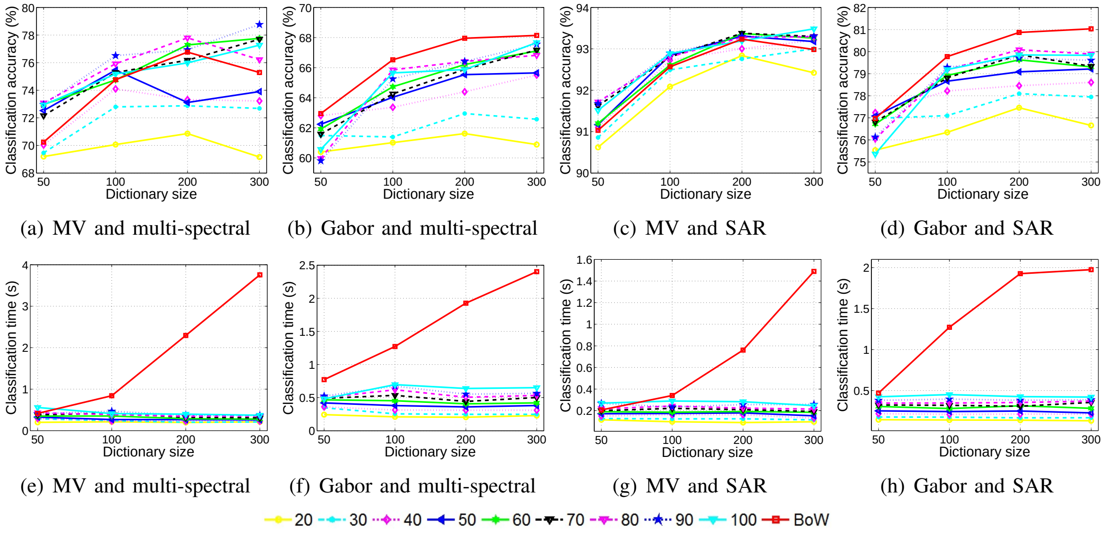
Fig. 4. Accuracy and run-time of the classification using BoW and BoT models for various dictionary sizes and different numbers of topics. In these experiments, SVM is applied to the *UCMerced-LandUse* dataset.



(a) Dictionary size = 50 (b) Dictionary size = 100 (c) Dictionary size = 200 (d) Dictionary size = 300

(e) Dictionary size = 50 (f) Dictionary size = 100 (g) Dictionary size = 200 (h) Dictionary size = 300

Fig. 5. Accuracy and run-time of the classification using BoW and BoT models for various dictionary sizes and different numbers of topics. In these experiments, SVM is applied to the *15 TerraSAR-X Image Classes* dataset.



(a) MV and multi-spectral (b) Gabor and multi-spectral (c) MV and SAR (d) Gabor and SAR

(e) MV and multi-spectral (f) Gabor and multi-spectral (g) MV and SAR (h) Gabor and SAR

Fig. 6. Assessing the dictionary size in the discriminability of the topics by comparing the classification accuracies and run-times for the *UCMerced-LandUse* and the *15 TerraSAR-X Image Classes* datasets.