Multi-Sensor Earth Observation Image Classification Based on a Multimodal Latent Dirichlet Allocation Model

Reza Bahmanyar, Daniela Espinoza-Molina, and Mihai Datcu, *Fellow, IEEE*

Abstract

Many previous researches have already shown the advantages of multi-sensor land-cover classification. Here, we propose an innovative land-cover classification approach based on learning a joint latent model of Synthetic Aperture Radar (SAR) and multispectral satellite images using multimodal Latent Dirichlet Allocation (mmLDA), a probabilistic generative model. It has already been successfully applied to various other problems dealing with multimodal data. For our experiments, we chose overlapping SAR and multispectral images of two regions of interest. The images were tiled into patches and their local primitive features were extracted. Then each image patch is represented by SAR and multispectral Bag-of-Words (BoW) models. The BoW values are both fed to the mmLDA, resulting in a joint latent data model. A qualitative and quantitative validation of the topics based on ground truth data demonstrate that the land-cover categories of the regions are correctly classified, outperforming the topics obtained by using individual single modality data.

Index Terms

Image fusion, Land-cover classification, Multimodal Latent Dirichlet Allocation, Multispectral images, Synthetic Aperture Radar images.

I. Introduction

The regular acquisition of Sentinel-1 and Sentinel-2 satellite data and the availability of their images have considerably increased the number of Earth Observation applications using the complementary information of the images acquired by the two different sensors. While Sentinel-1 offers Synthetic Aperture Radar (SAR) images with medium and high resolution [1], Sentinel-2 utilizes a multispectral

R. Bahmanyar, D. Espinoza-Molina, and M. Datcu are with the Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Wessling, Germany (e-mail: reza.bahmanyar@dlr.de; daniela.EspinozaMolina@dlr.de; mihai.datcu@dlr.de).

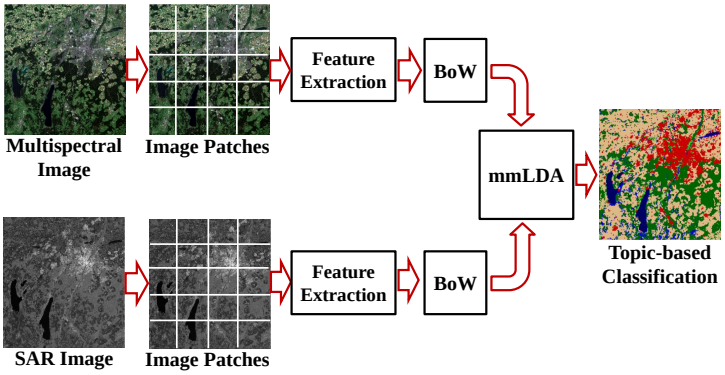


Fig. 1. An overview of the proposed multi-sensor land cover classification approach using mmLDA.

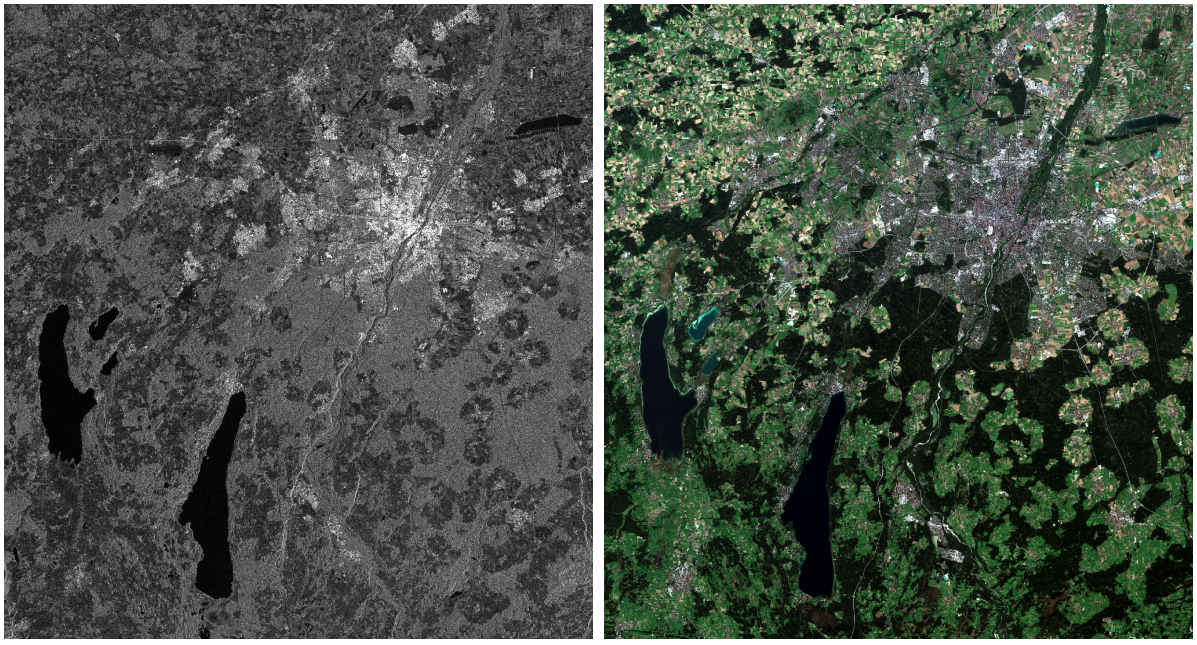
camera instrument capable of acquiring data in 13 different spectral bands at different ground sampling distances (GSDs), ranging from 10 to 60 m [2]. Using the information provided by both sensors allows us to compensate for the limitations of one sensor with the other one. For example, SAR images allow us to distinguish elements with strong microwave backscattering characteristics (e.g., typically found in man-made structures), while the differences between various natural scenes are more obvious in multispectral optical images than in SAR images. Many previous works have proposed the fusion of information from multiple sensors at different levels such as pixel level, feature level, and decision level. For example, the authors of [3] fused SAR and optical satellite images to improve the classification of four different land-cover types. They developed a work-flow based on the Principal Component Analysis technique for fusing data modalities at the pixel level. Then a Genetic Algorithm was used for primitive feature extraction and a Support Vector Machine (SVM) was employed for classification. The authors then showed that the classification accuracy achieved by their approach is higher as compared to the results obtained by using only data from a single sensor. Although a pixel level image fusion leads to an information gain, it is slow when working with large amounts of data. As an alternative, Wen and Li [4] proposed an image fusion technique for target detection based on extracting edges as primitive features from optical and SAR images. The extracted edges are then used to compute the contour affinity and contour correlations which are used as parameters for their fusion method. After that, an affinity value is computed and used as a threshold for the feature fusion process. They also showed an improvement in the detection accuracy as compared to using a single source of data. Further, the authors of [5] developed an algorithm based on Bayes' theorem, which is used to compute the correlated probabilities of three different sensors (two optical and one SAR sensor) in order to classify five different types of land-cover. This method performs data fusion at the decision level. Firstly, it classifies three different images of a target scene using an SVM, resulting in three classifications of the scene (one per sensor) in the form of class-wise conditional probabilities. The classifications are then combined to compute the posterior probability

for each class at every pixel position. After that, the cross-conditional dependencies (or correlations) between the three classifications are computed and later used as weighting parameters. The authors showed the superior classification accuracy resulting from the fusion of the three data sources.

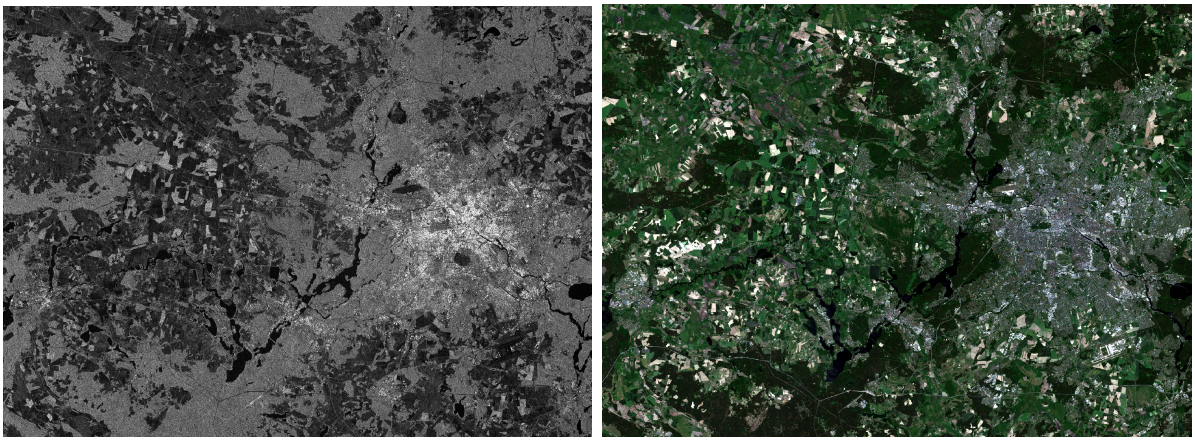
Inspired by the advantages of multi-sensor land-cover classification presented in previous works [3], [4], [5] and taking into account that multispectral and SAR sensors record very different parameters, we propose a fusion-free land-cover classification approach which exploits a joint latent model of satellite images acquired by multiple sensors (e.g., the SAR images of Sentinel-1 and the multispectral images of Sentinel-2). To this end, we employ a multimodal Latent Dirichlet Allocation (mmLDA) [6], [7], a generative probabilistic model. MmLDA is an extension of the widely-used Latent Dirichlet Allocation (LDA) [8] for discovering the latent structure of data collections containing documents with multiple modalities (e.g., text, images). MmLDA represents the latent data structure as a set of so-called topics and models each document by a probability distribution over the topics. This model has been shown to achieve promising joint models of data from different modalities (e.g., textual, visual, cognitive data) dealing with various problems such as the language grounding problem [6], [7], [9], [10]. Although generative models (e.g., LDA) have been applied to various remote sensing problems (e.g., [11], [12]), to the best of our knowledge, multimodal generative models have not yet been used in multi-sensor land-cover classification.

In our proposed approach, the co-aligned SAR and multispectral satellite images of a target scene are tiled into smaller patches so that the corresponding image patches cover the same area on ground. Fig. 1 shows an overview of the proposed approach. Then for each image patch its local primitive features are computed resulting in a set of feature vectors. After that a clustering technique (e.g., k-means) is separately applied to the feature vectors of all image patches of both sensors in order to create two visual word dictionaries. Using the dictionaries, the image patches are then represented by two Bag-of-Words models. Next, considering each pair of corresponding image patches as a document, they are fed into an mmLDA model. This model discovers the structure behind the images as a combination from two sets of topics thus avoiding a direct fusion of incommensurable data.

A qualitative and quantitative validation of the topics based on ground truth images (Open Street Maps [13] and Urban Atlas [14]) demonstrates that their semantics correspond to the existing land-cover categories in the target regions. Since the topics are estimated based on the statistics of the image patterns, they are not sensitive to rotation and shift in the images. In order to evaluate the complementarity effect of the two data types, we compare the topics to the set of topics obtained by applying LDA individually to each data type. The results demonstrate that the combined topics outperform the ones based on single data modality (e.g., 94% versus 87% of classification accuracy) in discriminating the land-cover categories.



(a) Munich (b) Munich



(c) Berlin (d) Berlin

Fig. 2. Illustrations of the regions of interest used in our experiments. (a,c) The SAR images, (b,d) compositions of B2, B3, and B4 bands of the multispectral images (the images are not to scale.)

II. Description of the Dataset

For our experiments, we selected two datasets taken over Germany. The first one includes a Sentinel-1B and a Sentinel-2A image of Munich acquired on September 30th and 29th, 2016, respectively. These images were then trimmed to extract a common region of interest. This region is located between the (48.33° N, 11.06° E) upper-left coordinates and (47.77° N, 11.78° E) lower-right coordinates which covers Munich city and the main nearby lakes. The second dataset includes a Sentinel-1B and a Sentinel-2A image of Berlin acquired on May 26th and 27th, 2017, respectively. This region is located between the (52.78° N, 12.45° E) upper-left coordinates and (52.26° N, 13.67° E) lower-right coordinates which covers Berlin city and the main nearby water bodies.

The Sentinel-1B images are geometrically rectified Level-1 ground range detected SAR products taken in interferometric wide mode with 10.13 m GSD. Figs. 2. (a,c) show quick-looks of the trimmed Sentinel-1B images. In contrast, the Sentinel-2 images are geometrically rectified multispectral Level-1C reflectance products. For our experiment, we use the important spectral bands B2, B3, and B4 (corresponding to blue, green, and red) and the Near Infrared (NIR) spectral band B8 which all have 10 m GSD. Figs. 2. (b,d) display the compositions of B2/B3/B4 for the regions of interest. Then the Sentinel-1B images were re-sampled to have the identical GSD (10 m) as the Sentinel-2 images and all images were co-registered. The final size of the images became 5596×6031 pixels for the Munich and 8149×5957 pixels for the Berlin dataset.

In order to evaluate the results, we created a ground truth dataset based on Open Street Maps [13] and Urban Atlas [14]. These data comprise the categories "Built-up", "Agriculture field", "Forest", and "Water body" according to the main land-cover types in the regions of interest. Figs. 5 (a,e) illustrate the different categories of the ground truth data.

III. Land-Cover Classification Using mmLDA

Latent Dirichlet allocation (LDA) [8] is a generative probabilistic model describing the latent structure of data collections as sets of topics. In order to jointly model multiple data modalities (e.g., text, images), LDA has been extended to multimodal latent Dirichlet allocation (mmLDA) [6] and later adapted to discrete visual feature spaces [7].

In our paper, we adapt the implementation of mmLDA presented in [7] to model the joint latent structure of multispectral and SAR satellite images and discover the existing semantic categories as sets of topics.

A. Data Preparation and Feature Extraction

To apply mmLDA, the multispectral and SAR images of a target scene are preprocessed as described in Section II. Then the images are tiled into small non-overlapping patches ("documents" in [8]) so that the corresponding patches from the two images cover the same area on the ground. In our experiments, a patch size of 32×32 pixels was selected as a compromise between small patches keeping the semantic analysis simple and bigger patches capturing the spatial context of objects.

After that, the local primitive features of each image patch are generated. In our experiments, we simply vectorize a window of 3×3 pixels around every pixel (according to [15]) resulting in a set of nine-dimensional feature vectors. For the multispectral image patches, the feature vectors are generated for each band separately and the resulting vectors are concatenated to form 36-dimensional feature vectors.

Next, from each image patch set (i.e., multispectral and SAR patches), 1% of the generated feature vectors are randomly sampled and a clustering technique (e.g., k-means) is applied to them in order to

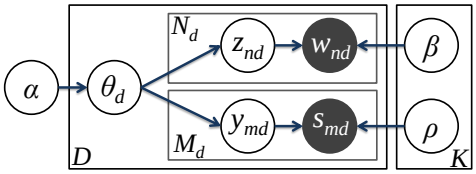


Fig. 3. The plate notation of mmLDA applied to our multimodal images.

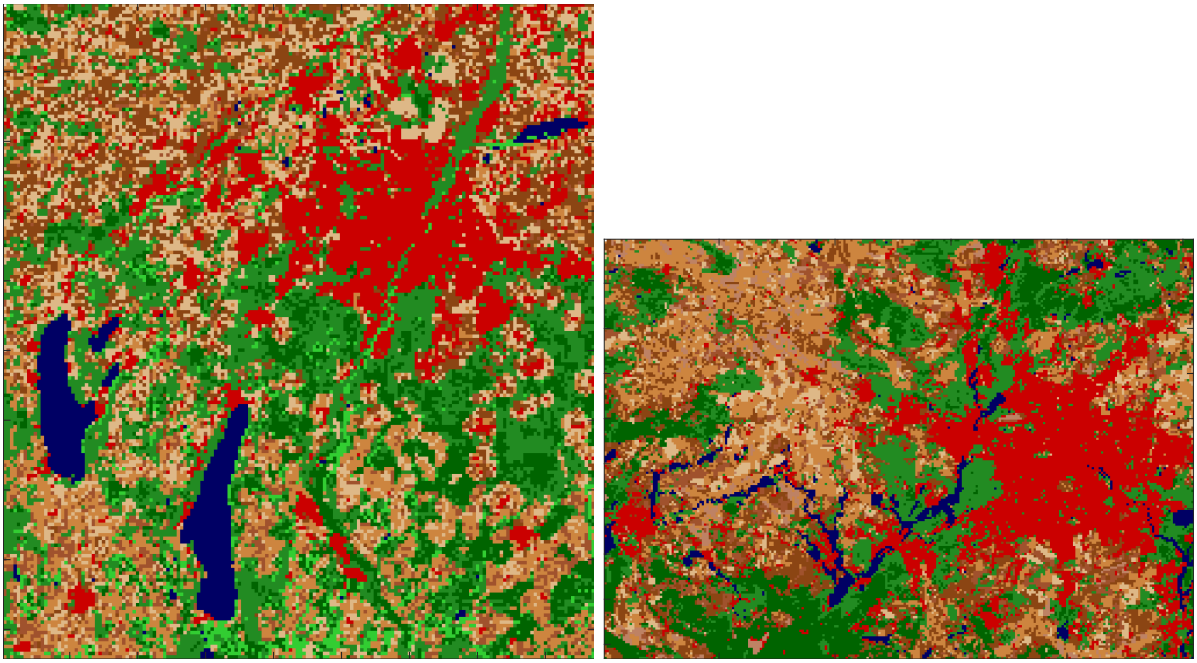
generate separate multispectral and SAR visual word dictionaries. The number of k-means clusters for both data types was empirically set to 50. Using the resulting dictionaries, two BoW models for each patch are generated. Each image patch is then represented as sequences of N_d and M_d multispectral and SAR visual word-tokens, respectively, $\mathbf{w}_d = \{w_{d1}, w_{d2}, ..., w_{dN_d}\}$ and $\mathbf{s}_d = \{s_{d1}, s_{d2}, ..., s_{dM_d}\}$, where the visual word-tokens are drawn from the corresponding dictionaries (the words of [8] are our cluster centers).

B. Joint Modeling of Multimodal Images by mmLDA

In the next step, mmLDA is applied to the BoW models of the patches in order to discover the latent structure as a set of K topics, where the topics are multinomial distributions over the multispectral visual words (denoted by β) and the SAR visual words (denoted by ρ). Then mmLDA models each pair of patches d as a distribution of the K topics defined by a K-dimensional Dirichlet random variable vector θ_d. Fig. 3 represents the plate notation of mmLDA applied to the multispectral and SAR BoW models of the image patches, where D denotes the collection of patches on the ground corresponding to the available image patches.

In its generative process, mmLDA updates for each patch pair its BoW patch parameters by choosing $\theta_d \sim Dir(\alpha)$, where α is the K-dimensional Dirichlet distribution's prior parameter vector. In our experiments, we consider an asymmetric α as in [7]. After that, for each multispectral visual word-token, mmLDA samples a topic for the topic-token z_{nd} from θ_d. Then a visual word is picked for w_{nd} from the multinomial probability distribution over the multispectral visual word dictionary conditioned on the selected topic, $p(w_{nd}|z_{nd}, \beta)$. Likewise, for each SAR visual word-token, a topic is sampled for the topic-token y_{md} from θ_d and then for s_{nd} a visual word is drawn from the multinomial probability distribution over the SAR visual word dictionary conditioned on the selected topic, $p(s_{md}|y_{md}, \rho)$.

In order to estimate the model parameters α, β, and ρ, and inferring the posterior distributions θ, mmLDA employs a variational Expectation Maximization algorithm [7]. Using the model parameters,



(a) Munich (b) Berlin

Fig. 4. Visualization of the topics discovered by applying mmLDA jointly to the multispectral and SAR image data. The topics are shown by different colors. Red illustrates the "Built-up" topic. Different shades of brown depict various "Agriculture field" topics. Different shades of green refer to various "Forest" topics. The topic "Water body" is illustrated in blue.

the joint distribution of the multispectral and SAR visual words for each patch is:

$$p(\mathbf{w}_d, \mathbf{s}_d, \theta_d, z, y) = p(\theta_d, \alpha).\left(\prod_{n=1}^{N_d} p(z_{nd}|\theta_d).p(w_{nd}|z_{nd}, \beta)\right). \left(\prod_{m=1}^{M_d} p(y_{md}|\theta_d)p(s_{md}|y_{md}, \rho)\right). \quad (1)$$

In order to classify the image patches based on the K discovered topics, to each image patch of the scene its most frequent topic is assigned as:

$$patch\_label_d \leftarrow \arg\max_j \{p(\theta_{jd}|\alpha)\},\ j \in [1, K]. \quad (2)$$

Since the topic distributions within the patches are learned jointly from multispectral and SAR images, we assume that the obtained classification results are derived based on the complementary information provided by both sensor types. These classification results are then assigned to real-world semantic categories. In our case, we concentrate on four final semantic categories (cf. Table I); therefore, we merge topics with similar semantics into one of our four main categories.

IV. Experimental Results and Discussion

In this section, we assess the classification results of mmLDA applied jointly to medium-resolution multispectral and SAR image data of the target scenes by comparing them to our ground truth data.

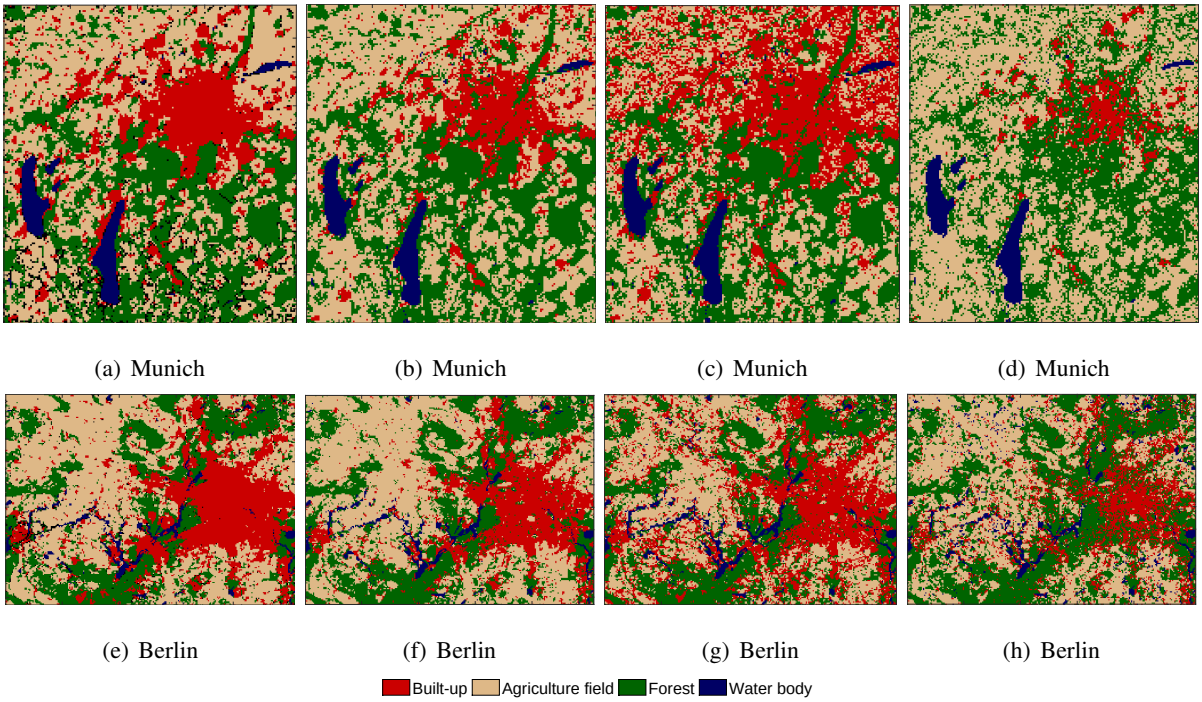


Fig. 5. Color illustration of the selected land-cover categories. (a,e) show the ground truth data, (b,f) represent the discovered categories using a joint model learning of multispectral and SAR images. The discovered categories using only the multispectral images and only the SAR images are demonstrated in (c,g) and (d,h), respectively.

Furthermore, we compare the results to the classification results obtained by applying LDA separately to the two image data types. For the individual experiments, we fed mmLDA with only one BoW model of the image patches (either $\mathbf{w}$ or $\mathbf{s}$), and used the same setup as for the joint experiment. In addition, we assessed our visual impressions about the consistency of the results. According to the results, the discovered topics in the experimental settings refer to one of the four main land-cover categories, namely “Built-up”, “Agriculture field”, “Forest”, and “Water body”.

In Fig. 4, the topics discovered by applying mmLDA jointly to the two data modalities of the test scenes are illustrated in different colors, where nine and ten topics were used for Berlin and Munich, respectively, based on an educated guess. In this figure, the topics referring to the “Built-up” and “Water body” categories are illustrated in red and blue, respectively. Additionally, different shades of brown represent the topics referring to different types of agricultural fields. Moreover, different shades of green demonstrate the topics corresponding to various forest types (e.g., sparse/dense forest, riparian forest.) A visual inspection of the resulting topics using external information (e.g., Google Earth/Maps) showed that they are consistent with the land-cover of the test scenes to a large degree.

Fig. 5 allows us to visually compare the topics obtained by using multiple data modalities to the ones achieved based on a single data modality and to the ground truth data. For the sake of comparability, the topics referring to the same main land-cover category were merged. Figs. 5 (b,f) show the land-cover classification based on the joint model learning of multispectral and SAR images,

TABLE I

QUANTITATIVE METRICS USED FOR THE EVALUATION OF FOUR DIFFERENT LAND-COVER CATEGORIES

	Munich			Berlin		
	Multisp.+SAR	Multisp.	SAR	Multisp.+SAR	Multisp.	SAR
	Accuracy (%)					
Built-up	**91.1**	82.0	87.1	**92.4**	84.8	82.6
Agriculture field	**85.5**	77.5	77.0	**91.8**	82.6	84.6
Forest	**89.1**	**89.8**	76.7	**94.0**	91.5	82.8
Water body	**99.5**	**99.3**	**99.4**	**99.5**	98.7	98.0
Overall	**91.3**	87.1	85.0	**94.4**	89.4	87.0
	Precision (%)					
Built-up	80.6	50.0	**91.2**	**92.0**	67.3	72.3
Agriculture field	**82.7**	**82.4**	70.6	**92.2**	90.7	89.6
Forest	**80.2**	**81.8**	62.0	**87.0**	81.7	64.2
Water body	89.7	88.1	**90.2**	**85.4**	68.2	58.9
Overall	**83.3**	75.6	78.5	**89.2**	77.0	71.2
	Recall (%)					
Built-up	66.0	**76.6**	30.0	**75.9**	**76.0**	48.7
Agriculture field	**85.3**	62.5	82.5	**90.4**	70.4	76.3
Forest	**88.2**	**88.2**	72.8	**92.6**	89.9	88.2
Water body	**96.1**	**96.0**	93.2	97.4	**99.1**	90.6
Overall	**83.9**	80.8	69.6	**89.1**	83.9	76.0

Figs. 5 (c,g) present the classification results obtained by only using the multispectral images, and Figs. 5 (d,h) depict the SAR-based classification results. In addition, we computed some quantitative metrics such as Precision, Recall, and Accuracy [16] to estimative the performance of the classification results which are presented in Table I. This table contains a detailed evaluation of the methods for each land-cover category as well as the overall classification performances for the two datasets. It can be seen that the joint model-based classification has the highest overall quality demonstrating that the joint model information contributes significantly to the classification quality.

According to Table I, for the "Built-up" category, for example, the precision of the classifications based on SAR data is significantly higher than those of based on the multispectral data (91.2% for SAR/Munich and 72.3% for SAR/Berlin, 50% for multispectral/Munich, and 67.3% for multispectral/Berlin), whereas the opposite holds for the recall results (see Table I). In other words, using the SAR data, most of the areas are correctly assigned to the "Built-up" category; however, a large amount of built-up areas was not identified. As the results show, our approach with the two data modalities led to classifications with high precision, recall, and accuracy. This can also be seen in Fig. 5. While the low precision of multspectral-based classifications can be corroborated in Figs. 5 (c,g), where there are many red spots showing a misclassification, using the joint model learning, the precision yields 80.6% and 92%.

In the case of "Agriculture field", the SAR-based classification reports a lower recall of 62.5% and 70.4% as compared to the multispectral-based classifications with 82.5% and 76.3% of recall; however, our approach improves the classification recall to 85.3% and 90.4%. Additionally, the joint model learning improves the classification precisions and consequently their accuracies. The "Forest"

category is well discriminated based on multispectral data and the joint model results, where for Berlin a better performance is obtained by the joint model learning.

In the case of the "Water body" category, for the Munich dataset, it is well recognized in the three cases (SAR, multispectral, and joint model) yielding an accuracy of $> 99\%$. However, for the Berlin dataset, where the water bodies are not large, the precision and recall of the multispectral-based classifications are higher than those of the SAR-based results, while the joint model learning outperforms both of them.

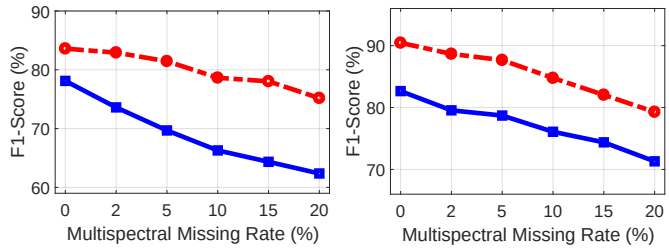
To evaluate how joint model learning allows SAR data to compensate for the limitations in multispectral images (e.g., cloud effects), we randomly masked different areas on the multispectral images and compared the classification performance of the joint model to the model learned based on the multispectral images. More precisely, patches of 100×100 pixels of zero values were randomly spread over the multispectral image covering different degrees of the image area (2%, 5%, 10%, 15%, 20%). Then the previously learned models were used for image classification through the mmLDA and LDA inference procedure, and their performances were compared using the F1-Score metric [16]. As the results in Fig. 6 show, although increasing the missing data rate decreases the F1-Score of both models, the joint model still outperforms the individual one. Moreover, in Munich, the performance of the joint model decreases more slowly that that of the other model.

V. Conclusion

The advantages of fusing data from various sensors have been shown in previous researches. In this paper, we presented a joint model learning approach based on an mmLDA for multi-sensor land-cover classification. For our experiments, we used two datasets of a SAR image acquired by Sentinel-1B and a multispectral image from Sentinel-2A taken over Munich and Berlin. Through a quantitative and a qualitative comparison of the results with ground truth data, we demonstrated that the obtained topics using the joint model of the images resulted in topics which are corresponding to the actual land-cover categories being contained in the regions of interest. We then compared the results to the topics obtained by using only single modality data. We observed that the joint model classification allows us to detect land-cover types which cannot be detected by individual data modalities. In addition, a quantitative evaluation showed an increment in the accuracy by using the joint model of the modalities as compared to using a single modality. Furthermore, the joint model used the complementarity of both modalities as it obtains its performance from the best modality for each class and helped compensate the limitations of a single sensor.

Acknowledgments

R. B. was supported by the German Academic Exchange Service (DAAD). We also thank G. Schwarz for helpful hints.



(a) Munich (b) Berlin

Fig. 6. Multispectral data missing impact on the classification results of the joint model (red) and the model learned using the multispectral images (blue.)

REFERENCES

[1] ESA. (2013, September) Sentinel-1 user handbook. https://sentinel.esa.int/documents/247904/685163/Sentinel-1_User_Handbook.

[2] ——. (2015, July) Sentinel-2 user handbook. https://sentinel.esa.int/documents/247904/685211/Sentinel-2_User_Handbook.

[3] C. Sukawattanavijit, J. Chen, and H. Zhang, "GA-SVM algorithm for improving land-cover classification using SAR and optical remote sensing data," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 3, pp. 284–288, January 2017.

[4] X. Wen and C. Li, "Feature-level image fusion for SAR and optical images," in *Proc. ICISCE*, December 2012, pp. 1–4.

[5] A. Mazher and P. Li, "A decision fusion method for land-cover classification using multi-sensor data," in *Proc. EORSA*, July 2016, pp. 145–149.

[6] D. M. Blei and M. I. Jordan, "Modeling annotated data," in *Proc. ACM/SIGIR*, July 2003, pp. 127–134.

[7] P. Das, C. Xu, R. F. Doell, and J. J. Corso, "A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching," in *Proc. IEEE CVPR*, June 2013, pp. 2634–2641.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, March 2003.

[9] M. Andrews, G. Vigliocco, and D. Vinson, "Integrating experiential and distributional data to learn semantic representations." *Psychological Review*, vol. 116, no. 3, pp. 463–498, 2009.

[10] S. Roller and S. S. im Walde, "A multimodal LDA model integrating textual, cognitive and visual modalities," in *Proc. EMNLP*, October 2013, pp. 1146–1157.

[11] L. Shen, L. Wu, Y. Dai, W. Qiao, and Y. Wang, "Topic modelling for object-based unsupervised classification of VHR panchromatic satellite images based on multiscale image segmentation," *Remote Sensing*, vol. 9, August 2017.

[12] R. Bahmanyar, S. Cui, and M. Datcu, "A comparative study of Bag-of-Words and Bag-of-Topics models of EO image patches," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, pp. 1357–1361, June 2015.

[13] OpenStreetMap. (2017, July) Planet dump retrieved from http://planet.osm.org & http://www.openstreetmap.org .

[14] European-Environmental-Agency. (2017, July) Urban Atlas. https://www.eea.europa.eu/data-and-maps/data/urban-atlas.

[15] S. Cui, G. Schwarz, and M. Datcu, "Remote sensing image classification: No features, no clustering," *IEEE J. Sel. Topics Appl. Earth Observ. in Remote Sens.*, vol. 8, pp. 5158–5170, November 2015.

[16] D. M. W. Powers, "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation," *J. Mach. Learn. Tech.*, vol. 2, pp. 37–63, 2011.