
Ruben Fernandez-Beltran, Juan M. Haut, Student Member, IEEE, Mercedes E. Paoletti, Student Member, IEEE, Javier Plaza, Senior Member, IEEE, Antonio Plaza, Fellow, IEEE, and Filiberto Pla

Abstract—The generative semantic nature of probabilistic topic models has recently shown encouraging results within the remote sensing image fusion field when conducting land cover categorization. However, standard topic models have not yet been adapted to the inherent complexity of remotely sensed data, which eventually may limit their resulting performance. In this scenario, this paper presents a new topic-based image fusion framework, specially designed to fuse synthetic aperture radar (SAR) and multi-spectral imaging (MSI) data for unsupervised land cover categorization tasks. Specifically, we initially propose a Hierarchical Multi-modal probabilistic Latent Semantic Analysis (HMPLSA) model that takes advantage of two different vocabulary modalities, as well as two different levels of topics, in order to effectively uncover inter-sensor semantic patterns. Then, we define a SAR and MSI data fusion framework based on HMPLSA in order to perform unsupervised land cover categorization. Our experiments, conducted using three different SAR and MSI datasets, reveal that the proposed approach is able to provide competitive advantages with respect to standard clustering methods and topic models, as well as several multi-modal topic model variants available in the literature.

Index Terms—Image fusion, probabilistic latent semantic analysis, synthetic aperture radar (SAR), multi-spectral imaging (MSI).

I. INTRODUCTION

The current expansion of different Earth observation programs provides excellent opportunities to conduct interdisciplinary research in many relevant application domains, such as Earth monitoring [1]–[3], contingency management [4], [5], climate warming [6], [7] and security applications [8]. In this context, one of the most pressing challenges is how to effectively combine the complementary information acquired by different sensors in order to deal with the data integration requirements of all these applications. Precisely, the image fusion field aims at combining remotely sensed data with complementary nature [9]. From early years, synthetic aperture radar (SAR) and high-resolution multi-spectral imaging (MSI) data have particularly exhibited potential synergies, as they fundamentally represent the Earth’s surface in a complementary manner. That is, whereas MSI instruments capture information about chemical characteristics of materials, SAR sensors quantify the scattering properties of the objects in the scene, which makes image fusion a very useful tool to relieve individual sensor limitations [10].

A. SAR and MSI image fusion

Broadly speaking, three different kinds of SAR and MSI fusion techniques can be found in the literature, depending on the data integration level [11]: decision, pixel and feature-level methods.

- Decision level-based methods estimate a separate predictor for each individual sensor and they eventually generate a final output by fusing all these independent results. For instance, Waske et al. present in [12] a fusion model based on two independent levels, with a first one to separately classify the input SAR and MSI data using a support vector machine (SVM) classifier, and a second one to combine both independent results. Other authors propose more elaborated decision schemes to uncover the correlation between SAR and MSI classification results. It is the case of the work presented in [13], where Mazher et al. propose a novel decision-based fusion technique which initially computes the posterior probability for each land-cover class and then generates the fused result using a maximum a posteriori approach.

- Pixel-based fusion methods aim at directly combining image pixels from different sensors in order to generate a fused result with improved spatial-spectral data. It is the case of the work presented in [14] which proposes a pan-sharpening algorithm based on independent component analysis and an adaptive curvelet-based fusion rule. Alternative pixel-based fusion methods use different kinds of image transformations to generate the fused result, such as the work presented in [15], which uses the principal component analysis decomposition approach for combining SAR and MSI image pixels, and a genetic algorithm together with a SVM-based classifier.

- Finally, feature-based fusion methods pursue to integrate attributes extracted from multiple modalities in order to generate a joint data representation that considers features from several sensors. For instance, Zhang et al. study in [16] the fusion of four different gray-level optical features together with polarimetric SAR data. Furthermore, the work presented in [17] proposes a data fusion approach which concatenates SAR features over both polarizations, and three different optical features based on information theory radiometric descriptors, local homogeneity texture information and time-frequency
B. Current limitations and trends

All the aforementioned techniques have shown to be effective under particular circumstances [20]. Whereas decision-based fusion models allow combining any kind of data from different instruments, the performance for SAR and MSI imagery may become rather limited because the data are independently analyzed and the fusion process is eventually defined as a post-classification one, which may be difficult to design. In this regard, pixel-based methods provide a more general fusion scheme because they generate a multi-sensor enhanced image. However, the speckle noise typically present in SAR imagery makes this approach usually inefficient for SAR and MSI data fusion [11]. In contrast, feature-based models overcome some of these limitations by uncovering a higher-level image representation, which associates correlative features from SAR and MSI sensors. In other words, features extracted from SAR images can provide discriminative object information to reduce some of the optical uncertainty of MSI imagery, and this is precisely the rationale behind feature-based SAR and MSI fusion methods. Some recent research lines take advantage of so-called semantic features in order to generate such high-level characterization using probabilistic topic models [21], [22]; nonetheless additional research is still required to adapt the standard architectures of these generative models to the particular requirements of the remote sensing image fusion field [23].

C. Semantic features based on topic models

In general, probabilistic topic models [24] are a kind of generative statistical models which provide methods to represent data collections according to their hidden semantic feature patterns. As a result, these models have been successfully used to provide data with a higher level of semantic understanding in many different application domains, such as text categorization [25], vocabulary reduction [26], image segmentation [27], object recognition [28] or even video retrieval [29], [30]. Within the remote sensing field, topic models have also shown a growing potential in image fusion tasks, due to their effectiveness to manage different data modalities at higher abstraction levels. For instance, this is the case of the work presented in [21] where Zhong et al. propose a topic-based fusion approach which concatenates three complementary kinds of spatial and spectral features to conduct remote sensing scene classification. In [31], Zhu et al. also present another relevant topic-based fusion approach which integrates spectral, texture and SIFT (Scale-Invariant Feature Transform) features by using a novel sparse semantic topic model framework. Despite the contrasted performance of all these methods, their corresponding fusion schemes are constrained by the use of standard topic models with a single modality, because the remote sensing data fusion problem has an intrinsic multi-modal nature.

Recently, Bahmanyar et al. presented in [22] a novel SAR and MSI fusion approach based on a multi-modal topic model to jointly manage multiple sources of data. Specifically, Bahmanyar’s work conducts multi-sensor land-cover classification using a visual bag-of-words (vBoW) characterization scheme and a multi-modal variant of the standard Latent Dirichlet Allocation (LDA) model [32], which makes use of two different vocabularies to represent SAR and MSI data modalities, respectively. Even though this recent LDA-based fusion approach has been shown to outperform individual single modality data, LDA is not the only type of topic model available in the literature and, besides, standard topic models’ architectures have not yet been specially adapted to deal with the inherent complexity of remotely sensed SAR and MSI data. As a result, improving the design of different kinds of probabilistic topic models for fusing SAR and MSI remotely sensed data still remains an open problem.

D. Contributions of this work

With all the previous considerations in mind, this paper proposes a new topic model which has been specifically designed to effectively fuse and categorize SAR and MSI remotely sensed data from an unsupervised perspective. First, we deeply analyze the performance of the two main topic model families, i.e. the Latent Dirichlet Allocation (LDA) [32] and probabilistic Latent Semantic Analysis (pLSA) [33] models within the remote sensing SAR and MSI data fusion field. Second, we propose a novel pLSA-based topic model, called Hierarchical Multi-modal pLSA (HMpLSA), which integrates two different features to cope with the special complexity of remotely sensed data. On the one hand, the proposed model integrates two divergent vocabularies to jointly manage SAR and MSI data modalities. On the other hand, HMpLSA makes use of a hierarchical latent space to project the input data onto a high dimensional space useful to uncover highly descriptive multi-modal semantic patterns. Third, we define an image fusion framework based on the proposed topic model in order to conduct unsupervised land cover categorization. Finally, we conduct an experimental comparison where different LDA and pLSA versions are tested with respect to the proposed HMpLSA model.

The remainder of the paper is organized as follows. Section II presents general background on probabilistic topic models. In section III, the proposed HMpLSA topic model is defined. Section IV describes an image fusion framework based on the proposed topic model that we use to conduct unsupervised land cover categorization. Section V presents our experimental results, in which several LDA and pLSA unimodal and multi-modal variants are tested over two different SAR and MSI datasets. Section VI discusses the obtained results. Finally, section VII concludes the paper with some remarks and hints at plausible future research lines.
II. BACKGROUND ON PROBABILISTIC TOPIC MODELS

From a practical perspective, topic models [24] represent probabilistic graphical models aimed at uncovering the hidden structure of a data collection. That is, given a corpus of documents \( D = \{d_1, d_2, ..., d_M\} \) characterized in a particular word-space \( W = \{w_1, w_2, ..., w_N\} \), latent topic algorithms estimate two probability distributions: the description of topics in words, \( p(w|z) \), and the description of documents in topics, \( p(z|d) \). In the literature, it is possible to find two main different topic model families depending on the generative process nature, one based on pLSA and another one based on LDA.

On the one hand, pLSA [33] defines a semi-generative data model by introducing a single latent random variable \( z \) to associate documents \( (d) \) and word polysemy occurrences \( (w) \). Under pLSA assumptions, documents are considered model parameters because they set topic mixtures and, at the same time, they are generated by topics, which eventually makes pLSA-based models particularly memory-demanding and prone to over-fitting. On the other hand, LDA [22] proposes a more general scheme by using two different Dirichlet distributions, one to model documents \( \theta \sim Dir(\alpha) \) and another one to model topics \( p(w|t, \beta) \sim Dir(\beta) \). However, \( \alpha \) and \( \beta \) hyper-parameters have to be initially estimated by iterating over the document collection, which makes LDA performance highly sensitive to relatively small collections [34], [35]. As a result, pLSA-based models are typically recommended when the amount of available information is reduced, considering the complexity of the problem [36].

Within the remote sensing field, the inherent complexity of SAR and MSI imagery generally makes that the amount of available information over a specific region of interest is rather limited to effectively conduct unsupervised land cover categorization [37], which makes pLSA a suitable generative architecture to design remote sensing fusion models.

III. HIERARCHICAL MULTI-MODAL PLSA

A. Proposed model

Starting from the asymmetric formulation of the standard pLSA scheme [33], the proposed Hierarchical Multi-modal probabilistic Semantic Analysis (HMpLSA) model (see Fig. 1) introduces two different innovations to deal with the remotely sensed SAR and MSI data fusion problem: (i) a multi-modal vocabulary nature, and (ii) a two-level latent topic architecture. On the one hand, HMpLSA makes use of two diverging vocabulary modalities by introducing \( w_s \) and \( w_m \) random variables, which represent SAR and MSI observable data. Note that this multi-modal nature allows the proposed approach to uncover common patterns across SAR and MSI sensors, which is a key factor to effectively associate correlative features from an image fusion perspective. On the other hand, the proposed HMpLSA model also defines two different levels of topics in order to increase the abstraction level of the uncovered multi-modal patterns. Note that the visual uncertainty of remotely sensed data is one of the most important problems when it comes to unsupervised land cover categorization, and the use of hierarchical architectures has recently shown to be very helpful to relieve this problem in other application domains, e.g. supervised image classification [38]. In particular, the hidden random variables \( z_s \) and \( z_m \) represent the first level of specific topics for SAR and MSI modalities, respectively. Over these two variables, a converging hidden unit \( z_c \) is used to generate a second level of common topics, which is useful to unfold a higher level of semantic patterns. Fig. 1 depicts a HMpLSA graphical model representation, where shaded nodes represent visible random variables. Regarding the HMpLSA generative process, it can be summarized by the following steps:

1) A document \( d \) is chosen from \( p(d) \) probability distribution.
2) For each one of the \( N_d \) words in the document \( d \),
   a) A hierarchical categorical topic \( z_c \) is chosen according to conditional distribution \( p(z_c|d) \) that expresses documents in high-level common multi-modal topics.
   b) A pair of SAR-topics \( z_s \) and MSI-topics \( z_m \) are chosen according to conditional distribution \( p(z_s, z_m|z_c) \), which encapsulates the relation between both levels of topics.
   c) Finally, words \( w_s \) and \( w_m \) are sampled according to the conditional distributions \( p(w_s|z_s) \) and \( p(w_m|z_m) \) which express SAR and MSI topics in their corresponding observable vocabularies.

Unlike other latent topic-based fusion approaches available in the literature [21], [22], [31], the proposed HMpLSA model defines a novel pLSA-based architecture which is able to work on width and depth, simultaneously. That is, whereas the two vocabulary modalities (dual-width) allow uncovering feature patterns from SAR and MSI data, the second level of topics (dual-depth) generates a high level semantic representation of the data that is able to gather inter-sensor correlative feature patterns. Precisely, this dual effect constitutes the main rationale behind the use of HMpLSA for fusing SAR and MSI data to conduct unsupervised land cover categorization. Initially, SAR-topics \( (z_s) \) and MSI-topics \( (z_m) \) generate an primary semantic representation of the SAR and MSI data. Then, the hierarchical categorical latent topic space \( (z_c) \) is uncovered to generate a multi-modal document characterization with a higher abstraction level for land cover analysis. That is, \( z_s \) and \( z_m \) project the SAR and MSI data onto a higher-dimensional fusion space in order to capture fine common semantic patterns. Then, \( z_c \) fuses these patterns to conduct unsupervised land cover categorization by fixing the number of categorical topics to the number of land cover categories.
B. Model relaxation

Another important contribution of our work is the development of a two-step model relaxation procedure in order to reduce the computational cost of the proposed model. Note that the two different levels of hidden random variables used by HMpLSA generate an additional degree of freedom with respect to the standard pLSA. Therefore, each variable marginalization over the posterior distribution requires to evaluate the Cartesian product between $z_c$ and \{ $z_s, z_m$ \}, which may become rather expensive in terms of computational cost as the number of input documents increases. In order to relieve this cost, we propose to estimate the HMpLSA parameters by using the following two sequential steps:

1) Learning SAR and MSI topics (HMpLSA-1): In the first step (Fig. 2a), the proposed model is simplified to uncover the primary SAR and MSI latent topic spaces using a single level of hidden units. That is, the hidden random variable $z_c$ is initially omitted to approximate $p(w_s|z_s)$ and $p(w_m|z_m)$ model parameters directly from the observable input documents $d$. Specifically, we make use of the Expectation-Maximization (EM) algorithm [39] to maximize the HMpLSA-1 complete log-likelihood function:

$$
\ell^1_c = \sum_d \sum_{w_s} n(w_s, d) \log \left( p(d) \sum_z p(w_s|z) p(z|d) \right)
+ \sum_d \sum_{w_m} n(w_m, d) \log \left( p(d) \sum_z p(w_m|z) p(z|d) \right),
$$

where $M$ is the total number of documents, $N$ represents the SAR and MSI vocabulary size, $n(w_s, d)$ contains document word-counts, and $K$ is the number of considered topics, which is defined to project documents into a higher dimensional space ($K \gg N$). Specifically, the EM algorithm works in two stages: (i) E-step [Eqs. (2)–(3)], where the expected value of the likelihood is computed given the current estimation of the parameters, and (ii) M-step [Eqs. (4)–(7)], where the new optimal values of the parameters are calculated according to the current setting of the hidden variables.

$$
p(z_s|w_s, d) = \frac{p(w_s|z_s)p(z_s|d)}{\sum_z p(w_s|z_s)p(z_s|d)} \quad (2)
$$

$$
p(z_m|w_m, d) = \frac{p(w_m|z_m)p(z_m|d)}{\sum_z p(w_m|z_m)p(z_m|d)} \quad (3)
$$

More specifically, the EM process is performed as follows. First, Eqs. (4)–(7) are randomly initialized. Then, the E-step and the M-step are alternated until the model converges. As convergence conditions, we use a $10^{-2}$ stability threshold in the difference of the log-likelihood [Eq. (1)] between two consecutive iterations or a maximum number of 1000 EM iterations.

2) Learning hierarchical categorical topics (HMpLSA-2): In the second step, we relax the proposed model to learn the hierarchical categorical topic space from the previously estimated parameters. As Fig. 2a shows, $w_s$ and $w_m$ random variables are omitted from the model and, accordingly, $z_s$ and $z_m$ nodes become observable. The rationale behind this model simplification is based on using the previous $p(z_s|d)$ and $p(z_m|d)$ parameter estimations as the input word-document distribution for HMpLSA-2, i.e. $n(z_s, z_m, d) \approx p(z_s|d)p(z_m|d)$. Then, it is possible to uncover the proposed multi-modal fused space $p(z_c|d)$ using only a single level of topics. Like in the first step, we estimate HMpLSA-2 parameters by using the EM algorithm to maximize the complete log-likelihood function:

$$
\ell^2_c = \sum_d \sum_{w_s,w_m} n(z_s, z_m, d) \log \left( p(d) \sum_z C p(z_c|z) p(z_c|d) \right),
$$

where $C$ represents the number of categorical topics, which is set to the number of categories considered to conduct unsupervised land cover categorization. Eventually, the E-step and M-step procedures are defined according to Eq. (9) and Eqs. (10)–(11), respectively.

$$
p(z_c|z_s, z_m, d) = \frac{p(z_s, z_m|z_c)p(z_c|d)}{\sum_z p(z_s, z_m|z_c)p(z|d)} \quad (9)
$$
After the model convergence (considering a $10^{-6}$ stability threshold and a maximum of 1000 EM iterations), the $p(z_c|d)$ parameter provides the fused representation of the input data, which jointly models SAR and MSI high level feature patterns.

IV. PROPOSED IMAGE FUSION FRAMEWORK

The hierarchical multi-modal pLSA-based SAR and MSI fusion scheme presented in this work consists of the three steps (see Fig. 3): (i) image characterization, (ii) HMpLSA-based image fusion and (iii) land cover categorization.

A. Image characterization

The first step of the proposed latent topic-based fusion model is based on defining the corresponding image characterization framework in order to enable the use of topic models over SAR and MSI remote sensing imagery. Specifically, we use the visual-bag-of-words (vBoW) approach [40], which consists of a three-step procedure. First, SAR and MSI data products are tiled into $32 \times 32$ image patches, which define topic model documents, i.e. $d$. Second, the k-means clustering algorithm [41] is globally applied over each image modality to build the corresponding SAR and MSI visual vocabularies. In particular, we consider a total number of $N = 50$ clusters (visual words) and vectorized $3 \times 3$ image patches with one-pixel of overlapping as clustering local primitive features. Eventually, the local primitive features (vectorized $3 \times 3$ image patches) within each topic model document (a $32 \times 32$ image patch) are encoded in a single histogram of visual words by accumulating the number of local features into their closest clusters. After this step, we obtain a collection of documents $D = \{d_1, d_2, ..., d_M\}$ described in both SAR and MSI vocabularies, i.e. $d_i = \{n(w_i^1, d_i), n(w_i^m, d_i)\}$ for $j, k \in [1, 2, ..., 50]$.

B. HMpLSA-based image fusion

The second step consists of applying the proposed HMpLSA model over a document collection in order to obtain the corresponding topic-based fused characterization, that is, the $p(z_c|d)$ probability distribution. According to the aforementioned two-step model relaxation (described in subsection III.B), HMpLSA-1 is initially used to uncover the two independent semantic representations for SAR and MSI modalities, i.e. $p(z_s|d)$ and $p(z_m|d)$, using $K = 1000$ topics. Then, these two semantic characterizations are fused by HMpLSA-2 in order to estimate $p(z_c|d)$, which jointly models SAR and MSI feature patterns in a single data distribution of $C$ categorical topics. Note that the proposed model relaxation aims at reducing HMpLSA computational cost to the standard pLSA one, and $C$ needs to be set to the number of land cover categories.

$$p(z_s, z_m|z_c) = \frac{\sum_d n(z_s, z_m, d)p(z_c|z_s, z_m, d)}{\sum_{z_s, z_m} \sum_d n(z_s, z_m, d)p(z_c|z_s, z_m, d)}$$ (10)

$$p(z_c|d) = \frac{\sum_{z_s, z_m} n(z_s, z_m, d)p(z_c|z_s, z_m, d)}{\sum_{z_s, z_m} \sum_d n(z_s, z_m, d)p(z_c|z_s, z_m, d)}$$ (11)

After the model convergence (considering a $10^{-6}$ stability threshold and a maximum of 1000 EM iterations), the $p(z_c|d)$ parameter provides the fused representation of the input data, which jointly models SAR and MSI high level feature patterns.

C. Land cover categorization

After the input SAR and MSI data products have been characterized and fused according to the proposed HMpLSA model, the third step is based on providing a land cover categorization for documents based on $p(z_c|d)$. In particular, we assume that each of the $C$ uncovered categorical topics represents a land cover class. Then, documents are categorized according to the topic with the highest probability value in $p(z_c|d)$, i.e. arg max$_{z_c} p(z_c^*|d)$. It should be noted that the uncovered categorical topics (unsupervised categories) need to be sorted according to the available ground-truth information in order to guarantee the use of the same labels for assessment purposes. In order to establish this correspondence, the cosine similarity function is computed between the uncovered categorical topics and the centroids of the ground-truth classes [42]. Then, each topic is tagged with the closest ground-truth land-cover label.

V. EXPERIMENTS

The experimental part of our work aims at validating the performance of LDA [32], pLSA [33], DpLSA (Dual-Depth pLSA) [42], MMLDA (multi-modal LDA) [43], MpLSA (multi-modal pLSA) [44], [45] and the proposed HMpLSA model in the task of unsupervised land cover categorization within the SAR and MSI image fusion field. Additionally, we also test the performance of K-Means [41] and BIRCH [46] clustering algorithms over the concatenated SAR and MSI modalities as a baseline fusion result. Subsection V.A introduces the three different datasets used in the experiments and subsection V.B shows the obtained unsupervised land cover categorization results. Note that all the methods considered in this work follow the same unsupervised experimental protocol as the proposed approach.

A. Datasets

Three different datasets (Munich, Berlin and Rome), made up of paired Sentinel-1 (SAR) and Sentinel-2 (MSI) data products, have been selected for the experimental part of the work (see Fig. 3). On the one hand, Munich and Berlin data products have been downloaded from the German Earth Observation Center website (https://gool/qTEpEQ), where ground-truth land-cover information is also available for assessment purposes. Specifically, four different land cover types are provided: ‘Agriculture’, ‘Building’, ‘Forest’ and ‘Water’. On the other hand, the Rome dataset has been obtained from the available European training data of the 2017 IEEE
GRSS Data Fusion Contest (https://goo.gl/C4y82I) which is based on Sentinel-2 and Landsat-8 imagery. In particular, we adapted this dataset to a SAR and MSI data fusion context by downloading the corresponding Sentinel-1 product from the Copernicus Open Access Hub (https://goo.gl/uXmPXL). Regarding the ground-truth data, the class label information available for the contest has been simplified by fusing the 17 available classes [47] into the four aforementioned categories: 'Agriculture' (15-16), 'Forest' (11-14), 'Building' (1-10) and 'Water' in blue).

1) Munich: The first dataset [22] comprises coupled Sentinel-1 and Sentinel-2 data products collected over the City of Munich (Germany). In particular, these images were acquired on September 29 and 30, 2016, respectively, and they cover the Earth surface between the (48.33°N, 11.06°E) upper left coordinates and (47.77°N, 11.78°E) lower right coordinates. Regarding the product features, the Sentinel-1 image is a Level-1 ground-range-detected SAR product which has been taken in interferometric wide swath mode with 10.13-m ground sampling distance (GSD) and geometrically rectified. Additionally, the Sentinel-1 data was preprocessed using the Refine Lee speckle filtering of the SNAP (Sentinel Application Platform) toolbox. The Sentinel-2 product is a multi-spectral Level-2A reflectance image generated from the corresponding geometrically rectified Level-1C product using the Sen2Cor software. For the experiments, only the highest resolution Sentinel-2 bands (10-m GSD), i.e. B2 (blue), B3 (green), B4 (red) and B8 (infra-red), have been considered. Besides, the Sentinel-1 product has been accordingly re-sampled to 10-m GSD. Eventually, both resulting Sentinel-1 and Sentinel-2 images have been co-registered obtaining a final size of 5596 × 6031 and 5596 × 6031 × 4 pixels, respectively.

2) Berlin: The second dataset [22] contains two Sentinel-1 and Sentinel-2 data products acquired over Berlin (Germany), which were captured on May 26 and 27, 2017, respectively, and cover the area between the (52.78°N, 12.45°E) upper left coordinates and (52.26°N, 13.67°E) lower right coordinates. Analogously to the Munich dataset, both Sentinel-1 and Sentinel-2 data products have been processed following the aforementioned procedure in order to obtain a final size 8149 × 5957 and 8149 × 5957 × 4 pixels, respectively.

3) Rome: The third dataset [47] is also made up of two Sentinel-1 and Sentinel-2 data products captured over Rome (Italy), on September 2 and 3, 2016, respectively, and cover the area between the (42.12°N, 12.26°E) upper left coordinates and (41.73°N, 12.71°E) lower right coordinates. Both Sentinel-1 and Sentinel-2 data products have been processed following the aforementioned process to obtain a final size of 3200 × 3200 and 3200 × 3200 × 4 pixels, respectively.

B. Results

Tables I-III provide a quantitative assessment of the unsupervised land cover categorization experiments for Munich, Berlin and Rome datasets. In particular, four different quality metrics are computed to evaluate the results, i.e. accuracy, precision, recall and f-score. For each one of these metrics, we show in rows the result for each individual ground truth category, as well as the corresponding average value. In columns, we provide Sentinel-1 (SAR), Sentinel-2 (MSI) and fusion (SAR+MSI) results, where tables’ headers indicate the method used in each case. Note that the SAR and MSI column blocks correspond to single-modal experiments, whereas SAR+MSI contains the multi-modal fusion results. It should be noted that each table cell contains the average percentage and the corresponding standard deviation value obtained after five runs of the indicated topic models. Additionally, the best result for each metric and column block is highlighted in bold font. Regarding the computational time, the last row in each table reports the average computational time in seconds of the corresponding experiments using a hardware environment made of an Intel(R) Xeon(R) CPU E5-2640 at 2.50GHz and 189GB of RAM.

In addition to the quantitative evaluation provided by the four considered metrics, we also provide the visual results of the land cover categorization experiments as a qualitative evaluation of the tested methods. Specifically, Figs. 5-7 show the corresponding land cover categorization maps for Munich, Berlin and Rome datasets, using the following order: (a) ground-truth data, (b) standard LDA with SAR data, (c) standard pLSA with SAR data, (d) Dual-Depth pLSA with SAR data, (e) standard LDA with MSI data, (f) standard pLSA with MSI data, (g) Dual-Depth pLSA with MSI data, (h) K-means with SAR+MSI data, (i) BIRCH with SAR+MSI data, (j) Multi-Modal LDA with SAR+MSI data, (k) Multi-modal pLSA with SAR+MSI data, and (l) the proposed HMpLSA with SAR+MSI data.
### TABLE I
**QUANTITATIVE ASSESSMENT OF THE UNSUPERVISED LAND COVER CATEGORIZATION RESULTS FOR MUNICH DATASET.**

<table>
<thead>
<tr>
<th>CATEGORY</th>
<th>SENTINEL-1 (SAR)</th>
<th>SENTINEL-2 (MSI)</th>
<th>FUSION (SAR+MSI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LDA</td>
<td>pLSA</td>
<td>DpLSA</td>
</tr>
<tr>
<td>Agriculture</td>
<td>81.80±0.03</td>
<td>81.51±0.03</td>
<td>81.62±0.06</td>
</tr>
<tr>
<td>Forest</td>
<td>77.41±0.28</td>
<td>74.51±0.09</td>
<td>78.65±0.17</td>
</tr>
<tr>
<td>Building</td>
<td>86.59±0.13</td>
<td>85.72±0.07</td>
<td>88.04±0.10</td>
</tr>
<tr>
<td>Water</td>
<td>99.34±0.00</td>
<td>99.53±0.00</td>
<td>99.32±0.02</td>
</tr>
<tr>
<td>AVG</td>
<td>86.28±0.11</td>
<td>85.27±0.04</td>
<td>86.91±0.05</td>
</tr>
</tbody>
</table>

**ACCURACY**

<table>
<thead>
<tr>
<th>CATEGORY</th>
<th>Agriculture</th>
<th>Forest</th>
<th>Building</th>
<th>Water</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>Agriculture</td>
<td>79.76±0.35</td>
<td>75.05±0.04</td>
<td>78.47±0.28</td>
<td>79.13±0.29</td>
</tr>
<tr>
<td></td>
<td>Forest</td>
<td>84.33±0.26</td>
<td>87.44±0.04</td>
<td>83.49±0.25</td>
<td>81.79±16.36</td>
</tr>
<tr>
<td></td>
<td>Building</td>
<td>32.83±0.80</td>
<td>27.40±0.49</td>
<td>43.38±0.98</td>
<td>78.18±1.26</td>
</tr>
<tr>
<td></td>
<td>Water</td>
<td>87.72±0.08</td>
<td>87.72±0.11</td>
<td>89.36±0.33</td>
<td>74.50±37.09</td>
</tr>
<tr>
<td></td>
<td>AVG</td>
<td>71.16±0.26</td>
<td>69.42±0.10</td>
<td>73.67±0.3</td>
<td>72.85±14.83</td>
</tr>
</tbody>
</table>

**RECALL**

<table>
<thead>
<tr>
<th>CATEGORY</th>
<th>Agriculture</th>
<th>Forest</th>
<th>Building</th>
<th>Water</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>Agriculture</td>
<td>79.11±0.10</td>
<td>77.81±0.04</td>
<td>78.68±0.10</td>
<td>71.36±0.15</td>
</tr>
<tr>
<td></td>
<td>Forest</td>
<td>71.33±0.19</td>
<td>69.78±0.08</td>
<td>72.47±11</td>
<td>83.27±10.81</td>
</tr>
<tr>
<td></td>
<td>Building</td>
<td>48.54±0.85</td>
<td>42.50±0.56</td>
<td>58.28±0.75</td>
<td>58.61±0.37</td>
</tr>
<tr>
<td></td>
<td>Water</td>
<td>91.10±0.05</td>
<td>91.04±0.06</td>
<td>91.07±0.27</td>
<td>71.81±36.61</td>
</tr>
<tr>
<td></td>
<td>AVG</td>
<td>72.57±0.32</td>
<td>70.28±0.22</td>
<td>78.12±0.27</td>
<td>70.22±14.87</td>
</tr>
</tbody>
</table>

**F-MEASURE**

<table>
<thead>
<tr>
<th>CATEGORY</th>
<th>Agriculture</th>
<th>Forest</th>
<th>Building</th>
<th>Water</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>Agriculture</td>
<td>112.67 (s)</td>
<td>35.72 (s)</td>
<td>218.93 (s)</td>
<td>71.78 (s)</td>
</tr>
<tr>
<td></td>
<td>Forest</td>
<td>71.38 (s)</td>
<td>71.78 (s)</td>
<td>217.72 (s)</td>
<td>36.03 (s)</td>
</tr>
<tr>
<td></td>
<td>Building</td>
<td>71.53 (s)</td>
<td>71.78 (s)</td>
<td>217.72 (s)</td>
<td>36.03 (s)</td>
</tr>
</tbody>
</table>

---

**VI. DISCUSSION**

According to the results reported in Tables I, there are several points which deserve to be mentioned regarding to the considered topic models’ performance in the task of unsupervised land cover categorization. When considering single-modal data, i.e. Sentinel-1 (SAR) and Sentinel-2 (MSI) columns, it is possible to observe that pLSA obtains a better result than LDA, with the exception of Munich and Rome SAR data. The higher complexity of the MSI imagery, together with the fact that some land cover categories are significantly unbalanced, make that pLSA can take advantage of using the document collection as model parameters in order to uncover more descriptive semantic patterns than LDA. Note that LDA’s Dirichlet hyper-parameter estimation may become rather inaccurate when considering a limited number of documents, like in the case of Berlin’s ‘Water’ category, where experiments show that LDA performance substantially decreases. Despite the general good performance of pLSA with single SAR and MSI data, we can see that DpLSA, which is a dual-depth pLSA version, can only one vocabulary has been taken into account, is able to consistently achieve even a superior result. That is, DpLSA exploits the two different levels of topics to extract more informative high-level patterns than regular LDA and pLSA, which eventually leads to a notewor-
TABLE II
QUANTITATIVE ASSESSMENT OF THE UNSUPERVISED LAND COVER CATEGORIZATION RESULTS FOR BERLIN DATASET.

<table>
<thead>
<tr>
<th>CATEGORY</th>
<th>SENTINEL-1 (SAR)</th>
<th>SENTINEL-2 (MSI)</th>
<th>FUSION (SAR+MSI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LDA</td>
<td>pLSA</td>
<td>DpLSA</td>
</tr>
<tr>
<td>Agriculture</td>
<td>77.33±0.31</td>
<td>81.13±0.03</td>
<td>78.98±0.49</td>
</tr>
<tr>
<td>Forest</td>
<td>71.72±0.15</td>
<td>74.39±0.23</td>
<td>85.31±0.54</td>
</tr>
<tr>
<td>Building</td>
<td>65.99±0.44</td>
<td>76.90±0.15</td>
<td>84.86±0.86</td>
</tr>
<tr>
<td>Water</td>
<td>95.43±0.11</td>
<td>91.13±0.02</td>
<td>88.14±0.56</td>
</tr>
<tr>
<td>AVG</td>
<td>77.82±0.13</td>
<td>80.89±0.09</td>
<td>84.32±0.15</td>
</tr>
</tbody>
</table>

Fig. 6. Qualitative assessment of the unsupervised land cover categorization results for Berlin dataset.

doing the comprehensive analysis of the considered fusion schemes, i.e. fusion (SAR+MSI) columns, the proposed model achieves a remarkable metric improvement with respect to the baseline K-Means and BIRCH clustering methods as well as the MMLDA and MpLSA topic models. On average, HMPLSA obtains 90.25% (accuracy), 75.06% (precision), 76.28% (recall) and 72.72% (f-score) metric results, whereas MMLDA average results are 84.17% (accuracy), 64.36% (precision), 61.23% (recall) and 61.41% (f-score) and the corresponding MpLSA ones are 88.19% (accuracy), 71.02% (precision), 74.22% (recall) and 69.81% (f-score). Even though the three considered multi-modal topic models (i.e. MMLDA, MpLSA and HMPLSA) are able to obtain an important improvement with respect to the baseline clustering algorithms (i.e. K-Means and BIRCH) and the corresponding single-modal topic models (i.e. LDA, pLSA and DpLSA), the proposed approach exhibits the best overall performance, obtaining an average precision improvement of 10.70) and 4.04 points over MMLDA and MpLSA, respectively. Like in the single-modal case, pLSA-based models show a better performance than LDA-based ones; however the two levels of topics defined by the proposed HMPLSA model provide an additional performance advantage to fuse SAR and MSI data. According to the qualitative results in Figs. [5][6], HMPLSA obtains the most accurate land cover categorization maps. For instance, it is possible to see that HMPLSA [Fig. [5][h]-[i]] substantially reduces the amount of noise present in Munich’s ‘Building’ category with respect to MMLDA and MpLSA [Figs. [5][h]-[i]]. In the case of Berlin, HMPLSA shows the
### TABLE III
Quantitative assessment of the unsupervised land cover categorization results for Rome dataset.

<table>
<thead>
<tr>
<th>CATEGORY</th>
<th>SENTINEL-1 (SAR)</th>
<th>SENTINEL-2 (MSI)</th>
<th>ROME</th>
<th>FUSION (SAR+MSI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LDA</td>
<td>pLSA</td>
<td>DPpLSA</td>
<td>LDA</td>
</tr>
<tr>
<td>Agriculture</td>
<td>66.11±0.29</td>
<td>67.63±0.12</td>
<td>66.31±0.28</td>
<td>55.91±0.22</td>
</tr>
<tr>
<td>Forest</td>
<td>65.05±0.85</td>
<td>70.41±0.80</td>
<td>73.99±0.35</td>
<td>85.90±0.17</td>
</tr>
<tr>
<td>Building</td>
<td>85.29±0.22</td>
<td>84.58±0.12</td>
<td>84.45±0.11</td>
<td>62.75±0.12</td>
</tr>
<tr>
<td>Water</td>
<td>93.89±0.30</td>
<td>90.82±0.49</td>
<td>91.19±0.08</td>
<td>82.99±0.39</td>
</tr>
<tr>
<td>AVG</td>
<td>77.58±0.10</td>
<td>78.36±0.09</td>
<td>78.99±0.04</td>
<td>71.89±0.12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CATEGORY</th>
<th>SENTINEL-1 (SAR)</th>
<th>SENTINEL-2 (MSI)</th>
<th>ROME</th>
<th>FUSION (SAR+MSI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Agriculture</td>
<td>Forest</td>
<td>Building</td>
<td>Water</td>
</tr>
<tr>
<td>ACCURACY</td>
<td>76.99±0.78</td>
<td>74.36±0.43</td>
<td>72.24±0.38</td>
<td>64.98±0.29</td>
</tr>
<tr>
<td>PRECISION</td>
<td>61.26±1.30</td>
<td>69.48±1.10</td>
<td>71.07±0.56</td>
<td>55.95±0.58</td>
</tr>
<tr>
<td>RECALL</td>
<td>35.81±1.41</td>
<td>23.73±1.77</td>
<td>18.59±0.49</td>
<td>35.01±1.12</td>
</tr>
<tr>
<td>TIME</td>
<td>46.00±0.34</td>
<td>44.56±0.27</td>
<td>44.74±0.15</td>
<td>31.09±0.32</td>
</tr>
</tbody>
</table>

---

Fig. 7. Qualitative assessment of the unsupervised land cover categorization results for Rome dataset.

The best land cover estimation because it obtains a significantly better result than MMLDA and it also reduces a significant proportion of the appearing in MpLSA ‘Water’ category. Finally, the proposed approach also obtains the most accurate visual result in the Rome dataset according to the available ground-truth information.

In all the conducted experiments, the proposed HMpLSA model has shown to provide a competitive advantage with respect to the rest of the tested methods, because it combines two important features to effectively deal with the remote sensing image fusion problem: a multi-modal nature and a hierarchical latent feature space. On the one hand, the problem of fusing SAR and MSI imagery raises the challenge of dealing with more complex data to uncover inter-sensor feature patterns useful to conduct land cover categorization. In this regard, HMpLSA makes use of two different vocabulary modalities to integrate SAR and MSI features acquired by different sensors within the same probabilistic characterization space. In other words, the multi-modal nature of the proposed approach allows HMpLSA to inherently associate correlative features from SAR and MSI modalities, which eventually enhances the amount of information available in the fused characterization space with respect to the single-modal approach. On the other hand, the proposed model also provides two different levels of...
topics which pursue to uncover high level semantic patterns over the previously extracted multi-modal features. That is, the first level of topics (SAR and MSI topics) aim at independently extracting the semantic representations of the input data. Then, the second level of topics (categorical topics) is used to fuse the multi-modal data over these semantic representations in order to conduct the data fusion at a higher abstraction level and, consequently, to generate more descriptive multi-modal patterns.

Precisely, the main advantage of the proposed approach over MMLDA and MpLSA lies on its ability to simultaneously combine these two different aspects within the same fusion scheme: the multi-modal nature of multi-sensor features and two different levels of hidden random variables to introduce a higher abstraction level. First, the dual vocabulary allows the proposed approach to uncover multi-modal patterns which intrinsically relate features from different sensors. Second, the two levels of topics are aimed at uncovering the unsupervised categories over a semantic space instead of over the original feature space. That is, the hierarchical architecture of the proposed approach makes that the unsupervised categories are based on multi-modal feature patterns instead of raw features from different sensors (like in the case of MMLDA and MpLSA), which eventually leads to significant noise reduction, especially in the most challenging scenarios. Despite the proposed approach potential, it may still have some limitations on the use of arbitrary taxonomies due to the own constrains of the vBoW encoding procedure [23]. Note that the proposed approach represents SAR and MSI images as a collection of documents via the vBoW approach using a specific image patch size, i.e. $32 \times 32$. Therefore, ground-truth land cover categories are required to be sufficiently general to make sense at this spatial level, otherwise the vBoW image characterization, used as the proposed approach input, may become unrepresentative for the expected detail level.

VII. CONCLUSIONS AND FUTURE WORK

This paper has proposed a new latent topic-based image fusion framework specially designed to fuse SAR and MSI remotely sensed data for land cover categorization tasks. Initially, we have defined a new topic model, called HMpLSA, which makes use of two different vocabulary modalities as well as two different levels of topics in order to deal with the complex nature of remote sensing data. Then, we present a SAR and MSI data fusion framework to effectively perform unsupervised land cover categorization. Our experimental comparison, conducted using two different Sentinel-1 and Sentinel-2 datasets, reveals that the proposed approach is able to provide competitive performance with respect to standard pLSA and LDA topic models, and also to several multi-modal topic model variants available in the literature.

One of the first conclusions that arises from this work is the potential of the pLSA architecture to fuse SAR and MSI remotely sensed data. In general, pLSA-based models have shown to obtain better land cover categorization results than LDA because they can take advantage of the use of input documents as model parameters in order to uncover more descriptive semantic patterns with a limited amount of data. Another important conclusion is based on the fact that multi-modal models provide a remarkable precision gain with respect to their corresponding single-modal counterparts. Nonetheless, the most relevant conclusion of the work is related to the effectiveness of the proposed HMpLSA-based fusion approach to effectively fuse SAR and MSI remote sensing data. Whereas other existing topic models only rely on the multi-modal component, the proposed approach also integrates two different levels of topics in order to provide a higher level representation space of multi-modal patterns which have shown to be useful to reduce the uncertainty in unsupervised land cover categorization tasks.

Finally, our future work will be aimed at the following directions: (i) extending HMpLSA to incorporate sparsity constrains for multi-modal semantic patterns, (ii) implementing automatic tools to find out the ideal number of topics in HMpLSA first latent space, and (iii) studying the effect of using different quantization techniques over the initial characterization space.

ACKNOWLEDGMENT

This work has been supported by Generalitat Valenciana (APOSTD/2017/007), the Spanish Ministry (FPU14/0212-FPU15/02090, ESP2016-79503-C2-2-P,TIN2015-63646-C5-5-R), and Junta de Extremadura (GR15005). We gratefully thank Dr. Reza Bahmanyar for his helpful assistance.

REFERENCES


