Semantic Annotation of Satellite Images using Latent Dirichlet Allocation

Marie Lienou, Henri Maître and Mihai Datcu

Abstract—In this paper, we are interested in the annotation of large satellite images, using semantic concepts defined by the user. This annotation task combines a step of supervised classification of patches of the large image and the integration of the spatial information between these patches. Given a training set of images for each concept, learning is based on the Latent Dirichlet Allocation model (LDA). The LDA-based image representation is obtained using simple features extracted from image words. By analogy to text terminology, a word is a window of pixels and a document corresponds to an image. Exploiting the capability of the LDA model to assign probabilities to unseen images, patches of the large image are classified into the semantic concepts using the Maximum Likelihood method. We conduct experiments on Panchromatic Quickbird images with 60 cm resolution. Integrating the spatial information between the patches, shows to improve the annotation performance.

Index Terms—Large image annotation, Latent Dirichlet Allocation, spatial information.

I. INTRODUCTION

HUGE quantities of image data have been delivered by satellite sensors during the last years, several terabytes being still acquired every day. This data volume makes the direct access to images more difficult. Moreover, manually-based tasks such as image indexing, are subjective and necessitate huge and expensive human effort. As a consequence, techniques to easily and selectively accede to images through their information content have become a highly desired goal for developing intelligent databases. In particular, effective and efficient access at semantic level has become the major step to allow users to benefit from such contents and meet their information needs.

In order to efficiently handle the tasks of image analysis, indexing and retrieval related to content information of remote sensing imagery, many techniques based on clustering algorithms were applied. This cluster analysis only exploits the similarities between the automatically extracted features representing the content of the images. However, the low-level features sometimes do not precisely represent the semantics of the images. Actually, the main problem in image semantic interpretation is the mapping from the low-level features to the high level semantic concepts (bridging the so called semantic gap). Indeed, the semantic analysis of a scene is highly dependent on prior knowledge coming from the user.

In this paper, we are interested in the annotation of large satellite images using semantic concepts, with a learning step where the examples for each concept are given by the user.

Several researchers have investigated techniques to automatically annotate images. One approach is to link a whole image to a set of keywords. In [1] for instance, the authors use a training set of images with annotations to produce a representation of a joint distribution linking images and words. Given an image, the words which have a high posterior probability are associated to the image. However, this approach does not indicate which image structure gave rise to which word [2]. Thus, some other methods rather link image parts (regions or tiles) to words. In particular, Mori et al [3] exploits a cooccurrence model in which they look at the cooccurrence of words with image tiles created using a regular grid. Duygulu and colleagues [4], as for them, use a translation model in order to associate a word to a region of the image. Actually, the images are segmented into regions, which are classified into region types using a variety of features. A mapping between regions types and keywords supplied with the images is then learned, using a method based on Expectation Maximisation algorithm.

On the other hand, some tools, initially developed for statistical text modeling in large document collections, have recently been used for semantic image annotation. Actually, generative probabilistic models such as probabilistic Latent Semantic Analysis (pLSA) [5] and Latent Dirichlet Allocation (LDA) [6] have been exploited for this purpose. Generative probabilistic models are random sources that can generate infinite sequences of samples according to a probability distribution. Using the bag-of-words assumption that the order of words in a document can be neglected, probabilistic Latent Semantic Analysis was one of the first methods that provided a probabilistic approach towards modeling text documents as mixture of intermediate (hidden) topics (or aspects). Given unlabeled training images, the probability distributions are estimated in a completely unsupervised manner. Latent Dirichlet Allocation extends pLSA model by treating the topic mixtures parameters as variables drawn from a Dirichlet distribution, thus defining a complete generative model which overcomes some limitations of pLSA.

The above generative models have also been applied to other image content analysis tasks such as scene classification [7], object recognition [8] and image retrieval [9]. Each image is modeled as a mixture of latent topics, each topic in turn, having a probability distribution based on the co-occurrence of visual words extracted from the image. The mixture of latent topics indicate the proportion of each object or scene.
type in the image. Thus this can be viewed as a concise representation of the image content, which can be further used for these various applications. Particularly, the pLSA model is used in [10] in order to annotate whole natural images, and Blei and Jordan in [11], use pairs of LDA modules to model relationships between images and their corresponding descriptive captions.

Various extensions of the LDA model have been proposed recently; among them, are the correlated topic model [12] and the Spatial LDA topic model (SLDA) [13]. In the latter case, the authors were motivated by the fact that the spatial information which is important in many computer vision problems, was neglected in the LDA model. Indeed, SLDA model encodes the spatial structure among visual words, this spatial information being used when designing the documents. Different from the LDA in which the partition of words into documents is known \textit{a priori}, the word-document assignment in SLDA is a random hidden variable. There is a generative procedure, where knowledge of spatial structure can be flexibly added as a prior, grouping visual words which are close in space into the same document.

Following this insight, we propose in this work, to make use of the LDA model and differently take into account the spatial information, in order to semantically annotate large satellite images. Actually, considering that documents correspond to patches of the large image, we exploit the LDA model to classify these documents into given semantic classes. The spatial information is first introduced in the process of cutting the documents with an overlapping, then in the majority vote after running the LDA.

This paper is organized as follows. In the next section, we describe the LDA model for images. Then section III describes the way we exploit the LDA model for the semantic annotation of large satellite images, including the classification task and the integration of the spatial information. Tests and results are presented in section IV, followed by a discussion. Conclusion and future scope of the work are discussed in the last section.

II. USING LATENT DIRICHLET ALLOCATION FOR IMAGE MODELING

As explained in [6], Latent Dirichlet Allocation is a generative probabilistic model for collections of discrete data. It is a three-level hierarchical model, in which documents of a corpus are represented as random mixtures over latent topics. Each topic is in turn, characterized by a distribution over words. In this work, since we are trying to apply techniques used in the text domain to images, we need to define an analogy between their respective terminologies:

- a word \(w_n\) corresponds to a segment (region) of the image or a window of pixels (tile)
- a document \(w\) is equivalent to an image,
- a corpus \(D = \{w_1, \ldots, w_M\}\) is an image dataset.

A. Image representation

Since the order of words in a document is ignored in the LDA model (\textit{bag-of-words} assumption), each image of the collection is represented as a sequence of \(N\) words. We need to define a vocabulary in order to reduce the size of image words in the collection. For this purpose, features are automatically extracted from each word of the image dataset, then a vector quantization is applied on the entire set of features. Each word then gets a single label (also called blob or visual word), which represents all the words within the same cluster. The number of visual words is the size of the vocabulary. Given the vocabulary, each image \(I\) is thus represented by a sequence of \(N\) visual words, where each word in the sequence is represented by its corresponding visual word. By counting the visual words occurrences, we get a word frequency vector for each image of the collection.

LDA models each word in a document (or image) as a sample from a mixture model, where the mixture components can be viewed as representations of “topics”. Thus, LDA leaves flexibility to assign a different topic to every observed word in a document. Each document is represented as a list of mixture proportions for the mixture components and thus reduced to a probability distribution on a fixed set of topics. The latent topic structure is learned without any usage of background knowledge.

B. LDA generative process

Each image \(I\) is a sequence of \(N\) visual words \(w_n\), denoted by \(w = \{w_1, w_2, \ldots, w_N\}\). LDA assumes the following generative process for each image in the collection:

1) Choose a K-dimensional Dirichlet random variable \(\theta \sim Dir(\alpha)\), where \(K\) is the number of topics in the collection

2) For each of word position \(n \in \{1, \ldots, N\}\) :

- Choose a topic \(z_n \sim Multinomial(\theta)\)
- Choose a word \(w_n\) from \(p(w_n | z_n, \beta)\), a multinomial probability conditioned on the topic \(z_n\).

The likelihood of an image \(w\) with such a model is given by:

\[
p(w | \alpha, \beta) = \int p(\theta | \alpha) \left( \prod_{n=1}^{N} p(z_n | \theta) p(w_n | z_n, \beta) \right) d\theta
\]  

Thus, the likelihood of the entire image collection is determined by the product of the marginal probabilities of single images. In learning, the goal is to find the corpus-level parameters \(\alpha\) and \(\beta\) such that the loglikelihood of the entire database is maximised. Unfortunately, (1) is not tractable for exact inference: the solution is to use approximate inference algorithms such as Laplace approximation, Gibbs sampling and variational expectation maximisation. In this work, variational inference, which makes use of Jensen’s inequality to obtain an adjustable lower bound on the loglikelihood, is employed. Contrary to the pLSA model, LDA makes it possible to assign probability to a document outside of the training corpus. Indeed, LDA generalizes easily to new documents, thus being a complete generative model.

III. LDA-BASED SEMANTIC IMAGE ANNOTATION

In this section, we describe our approach for performing the semantic annotation of large satellite images, using the LDA model.
Given $S$ semantic concepts defined by the user, for each concept, the user provides a set of images which will be used for learning. Given a large image $I$ to be annotated using the $S$ semantic concepts, we consider that the testing corpus corresponds to a set of image patches $I_d$ of equal size, such that:

$$\bigcup_d I_d = I$$  \hspace{1cm} (2)

Thus, the annotation of $I$ can be viewed as the classification of the documents $I_d$ into the $S$ semantic classes $C_s, s \in \{1, \ldots, S\}$. The classification task includes the visual words computation and the model generation. Since the LDA model also assumes that the order of documents in a corpus can be neglected [6], we take into account the spatial information between the documents in order to improve the annotation task.

### A. Image classification

The goal is to classify the patches extracted from the large image into $S$ classes. Features are automatically extracted from each word of the dataset (training set images and testing patches). In order to determine the optimal number of clusters for the vector quantization, we use an approach described in [14], which models the features as a Gaussian Mixture and uses the Minimum Description Length criteria to accede to the optimal complexity of the model. If we note $V$, the optimal number of clusters, which is the size of the vocabulary, then, the k-means algorithm is used to quantize the data into $V$ clusters, whose centers are the visual words.

Once the representation of images as sequences of visual words is obtained, the next step is to get a model for each class from its respective training set. But we need first to determine the number of topics for each class. This is done by computing the perplexity of a held-out test set. Used by convention in language modeling, perplexity is a measure of the ability of a model to generalise to unseen data, and is defined as the reciprocal geometric mean of the likelihood of a test corpus given the model. For a test set $D_{test} = \{w_1, \ldots, w_M\}$ of $M$ documents, perplexity is given by:

$$\text{perplexity}(D_{test}) = \exp \left( \frac{-\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{M} N_d} \right)$$  \hspace{1cm} (3)

This measure is monotonically decreasing in the likelihood of the test data, thus lower complexity score indicates better generalization performance.

Given their respective training sets, learning using LDA achieves a model that best represents the distribution of visual words for each class. Then for an image $w$ in the testing set, it is possible to assign a probability which corresponds to the likelihood of the image for a class. Classification is based on the Maximum Likelihood method, which assigns the image to the class which maximises the likelihood: $s^* = \arg\max_s p(w | \alpha_s, \beta_s)$.

### B. Spatial information

Different from [13] where the spatial information is modeled, in this work, the spatial information is introduced without any explicit model. Since we are trying to annotate a large image by classifying patches of the image using the LDA model, the spatial information between adjacent patches is lost. It is possible to conserve and exploit this information by cutting the patches (documents) in the large image with an overlapping. This introduces a redundancy which is exploited after the LDA process. Due to the overlapping, some parts of a document might belong to many adjacent documents, and thus, after the classification, they might be assigned to different classes. We use a majority vote for these common parts of documents, by ascribing them to the most represented class in their neighborhood. In this way, the large image is annotated more precisely.

### IV. Experiments and results

We describe the database used for experiments as well as the tests protocol. The results obtained for the semantic annotation of a large satellite image using our approach are further discussed.

### A. Data description

The experiments are performed on Panchromatic Quickbird images of Las Vegas, with 60 cm resolution.

The training images belong to five semantic classes: residential suburbs (RS), deserts (DS), commercial areas (CA), urban areas (UA) and golf fields (GF). Fig. 1 shows examples of such images. The training set for each class contains 40 images with sizes ranging from $150 \times 150$ to $500 \times 500$ pixels.

The large image to be annotated is of size $6000 \times 6000$ pixels (Fig. 3(a)). This image contains all the classes mentioned above, but also some other classes which have not been learned.

---

![Fig. 1. Examples of images belonging to the five semantic classes related to the concepts used in our experiments.](image-url)
(for instance the main roads). The testing set is made of patches from this image and its size depends on those of the patches and the overlapping.

B. Tests

In this work, a word corresponds to a window of pixels (tile), whose size is $10 \times 10$ pixels. The features automatically extracted from each word of the dataset are simple ones: mean and standard deviation. The parameters for learning are determined as described in section III-A. Using our data, the optimal number of visual words was 30 and the number of topics for residential suburbs, deserts, commercial areas, urban areas and golf fields was respectively 5, 2, 5, 4 and 5. Fig. 2 illustrates the perplexity versus the number of topics for residential suburbs. Since we have not estimated the number of topics for the whole set, but for each class separately, two topics from different models could be identical. In future work, we plan to estimate the number of topics for the whole set in order to be able to compare two different models.

![Fig. 2. Perplexity versus the number of topics for residential areas class. The perplexity is minimized when the number of topics is equal to 5.](image)

The models are evaluated by a cross-validation, using 80% of the images for learning and 20% for the tests. On average, we obtain 96.5% of good classification. An overview of the performance of the models is given by the confusion matrix presented in Table I. We see that there are few explainable confusions between some classes. For instance, some images which belong to the desert class are classified into golf fields and vice versa. This is not surprising since golf fields are mainly made of fields, deserts and lakes.

<table>
<thead>
<tr>
<th></th>
<th>RS</th>
<th>DS</th>
<th>CA</th>
<th>UA</th>
<th>GF</th>
</tr>
</thead>
<tbody>
<tr>
<td>RS</td>
<td>97.5</td>
<td>0</td>
<td>0</td>
<td>2.5</td>
<td>0</td>
</tr>
<tr>
<td>DS</td>
<td>2.5</td>
<td>95.0</td>
<td>0</td>
<td>0</td>
<td>2.5</td>
</tr>
<tr>
<td>CA</td>
<td>0</td>
<td>0</td>
<td>97.5</td>
<td>2.5</td>
<td>0</td>
</tr>
<tr>
<td>UA</td>
<td>2.5</td>
<td>0</td>
<td>0</td>
<td>97.5</td>
<td>0</td>
</tr>
<tr>
<td>TG</td>
<td>2.5</td>
<td>5.0</td>
<td>0</td>
<td>0</td>
<td>92.5</td>
</tr>
</tbody>
</table>

The testing set is composed of patches of the large image. The size of the patches is experimentally set to $150 \times 150$ pixels. Thus each patch or document contains 225 words. We perform two kinds of experiments on the test image: with and without overlapping. We then compare the two resulting annotated images in order to evaluate the importance of the spatial information in the annotation task. The evaluation of the results are visual, using Google maps as ground truth.

C. Results and discussions

We discuss here, the results obtained from the tests performed on the large image presented in Fig. 3(a). Fig. 3(b) shows the image annotated with the five semantic concepts, without taking the spatial information into account. So the $40 \times 40$ patches of the testing set are classified into the five classes. The results are globally interesting, especially when we consider the simple features used for the visual words computation. We however notice the confusions between deserts and residential suburbs. This is due to the fact that the latter class is mainly composed of houses, deserts and small green areas. In addition, since the test image contains more than the five learned classes, and since we have not defined a reject class for areas found in the test image which do not correspond to any kind of example images in the training set, some misclassifications occur. For instance, main roads are categorized into commercial areas, whereas lawns are included in golf fields. It is worth noting that the features used in this work are mean and standard deviation only, thus these misclassifications are not really surprising. In order to avoid such problems, we plan in future work to add a reject class or to define semantic concepts which cover all the possible areas types in the testing set.

In the second experiment, the patches of the testing set are obtained by cutting the large test image with an overlapping of 50 pixels. After the classification of the $118 \times 118$ patches, a majority vote is used to reconstruct the image. The results are quite satisfying as presented in Fig. 3(c). Compared to the previous test, the areas belonging to the different classes are less coarse and better delimited. As an example, the golf field area in the top of the image better corresponds to the ground truth in the second annotated image than in the first one. Until a certain threshold, the bigger is the overlapping, the finer is the annotation. Furthermore, there are less isolated patches in the second annotated image as the neighborhood relationships are introduced.

V. CONCLUSION

We have presented an approach allowing to semantically annotate large high resolution satellite images, using concepts defined by the user. This methods combines the Latent Dirichlet Allocation model allowing to classify patches of the large image into the given semantic classes, and the spatial information between these patches which improves the annotation performance. We observed that using simple features such as mean and standard deviation for the LDA-image representation can lead to satisfying results for large images annotation. We project to use more other features such as texture or to work with multispectral images in order to...
improve the performance of the learned LDA-models and thus, to better the annotation task.

REFERENCES


Fig. 3. Results of the annotation of a large image into the five semantic classes: urban areas (yellow), deserts (pink), commercial areas (green), golf fields (blue), and residential suburbs (no color). Compared to Fig. 3(b), in Fig. 3(c) the areas belonging to the different classes are less coarse and better delimited. In addition, there are less isolated patches as the neighborhood relationships are introduced.