Remote Sens. **2013**, *5*, 2275-2291; doi:10.3390/rs5052275

OPEN ACCESS

Remote Sensing

ISSN 2072-4292

www.mdpi.com/journal/remotesensing

Article

Tile-Level Annotation of Satellite Images Using Multi-Level Max-Margin Discriminative Random Field

Fan Hu [1], Wen Yang [1,*], Jiayu Chen [1] and Hong Sun [1,2]

[1] Signal Processing Laboratory, School of Electronic Information, Wuhan University, Wuhan 430072, China; E-Mails: hfmelizabeth@gmail.com (F.H.); jiayu.chen@ieee.org (J.Y.C.)

[2] TSI Department, TELECOM ParisTech, F-75013 Paris, France; E-Mail: hongsun@whu.edu.cn

* Author to whom correspondence should be addressed; E-Mail: yangwen@whu.edu.cn; Tel.: +86-27-6875-6356; Fax: +86-27-6875-6356.

Received: 1 March 2013; in revised form: 3 May 2013 / Accepted: 7 May 2013 / Published: 13 May 2013

Abstract: This paper proposes a multi-level max-margin discriminative analysis (M^3DA) framework, which takes both coarse and fine semantics into consideration, for the annotation of high-resolution satellite images. In order to generate more discriminative topic-level features, the M^3DA uses the maximum entropy discrimination latent Dirichlet Allocation (MedLDA) model. Moreover, for improving the spatial coherence of visual words neglected by M^3DA, conditional random field (CRF) is employed to optimize the soft label field composed of multiple label posteriors. The framework of M^3DA enables one to combine word-level features (generated by support vector machines) and topic-level features (generated by MedLDA) via the bag-of-words representation. The experimental results on high-resolution satellite images have demonstrated that, using the proposed method can not only obtain suitable semantic interpretation, but also improve the annotation performance by taking into account the multi-level semantics and the contextual information.

Keywords: satellite images annotation; topic model; MedLDA; multi-level max-margin; conditional random field

1. Introduction

Nowadays the information extraction and intelligent interpretation of high-resolution satellite images are frontier technologies in the remote sensing field. With the growing number of high-resolution satellite images, efficient content extraction and scene annotation that can help us quickly understand

the huge-size image are becoming more and more desirable. Given such a large data volume, manually based annotation tasks typically require a lot of human effort. Hence an effective interpretation method based on mid-level or high-level semantic is strongly required in remote sensing applications.

However, the low-level features (physical features), most of the time, cannot precisely represent the scene semantics of images, and consequently how to bridge the semantic gap is becoming the main issue to deal with. Recently there have been ever-growing interests in image annotation by using topic models, such as Probabilistic Latent Semantic Analysis (PLSA) [1,2], Latent Dirichlet Allocation (LDA) [3,4], which can map from low-level physical features to high-level semantic concepts, and essentially reduce the dimensionality of features. These generative probabilistic models were originally developed for text document modeling, which can generate an infinite sequence of samples according to the distribution of latent topics. It is assumed that each document is a mixture over latent topics and each topic is, in turn, a mixture over words from documents. The representation of latent topics can build a global information space, which is more reliant on content coherence than local description. Meanwhile, the computational efficiency based on approximate inference methods also makes the aspect models gain much popularity. It is necessary to build a corresponding relationship between the document and image for the application of these models from text domain into image domain. Conventionally, the whole image is treated as corpus and divided into rectangular tiles, which are regarded as documents. Each tile is further partitioned into multiple smaller patches. Local features extracted from patches are transformed by vector quantization into "visual words", and each tile is thus represented as a collection of words. Some researchers have demonstrated that aspect models provide an understanding of aerial images in an effective way. According to [5,6], a scene of a satellite image, modeled by LDA, is represented as a finite mixture over some underlying semantic classes. This discriminative representation leads to a satisfactory result on annotation performance of large satellite images.

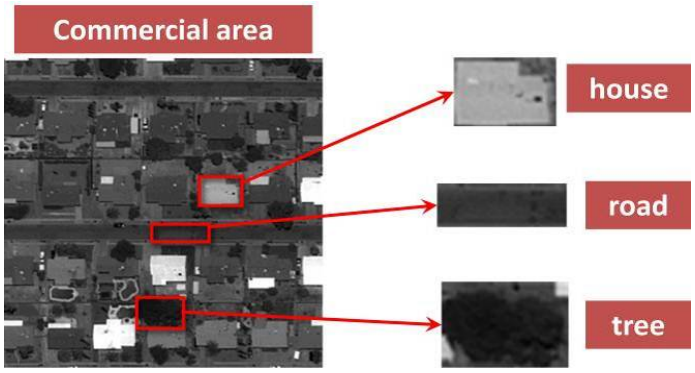
As we know, almost all kinds of topic models built on low-level features have their own limitation and serious drawbacks. Due to the independence assumption of each visual word in a tile and independence between tiles, these models frequently ignore the spatial relationship of adjacent regions and hence fail to capture important context information. In order to solve this kind of problem, many methods and algorithms have been proposed. The authors in [5] have introduced spatial information by cutting the patches in the large image with an overlap. Various extensions of the aspect models have been also designed, e.g. the spatial LDA (SLDA) model [7]. Different from LDA, the word-assignment of SLDA is a random hidden variable and the spatial information between visual words is encoded. In another method, the random field models such as the Markov Random Field (MRF), Conditional Random Field (CRF) have been employed for improving the spatial coherence of aspect models as well. Particularly, for the sake of describing the spatial relationship of latent topics, an MRF prior has been defined over hidden topic labels, which has been obtained by PLSA, and the experimental results of supervised and weakly supervised manner have demonstrated that the segmentation and recognition accuracy is obviously enhanced by the two complementary models detailed in [8,9].

In this paper, we present a method of annotation of satellite images based on the combination of a novel topic model and the CRF. According to [8], each latent topic in PLSA is regarded as one semantic class. However, such one-to-one mapping is inappropriate for the representation of local scene semantic in satellite image due to insufficient representation of complex scene information. It seems like more reasonable that one semantic class should contain several latent topics. As illustrated

in Figure 1, the randomly selected patch in a scene of a commercial area may consist of some objects, such as road, house, trees, *etc*., which could be represented in form of latent topics in aspect models. We are therefore motivated by the maximum entropy discrimination latent Dirichlet Allocation (MedLDA) model [10], which was originally proposed for regression and classification for text analysis and can train supervised models based on a max-margin principle. The discovery process in latent topics of this extension of LDA model is by way of optimizing an objective function with a set of margin constraints. The coupling of parameters and analysis of latent topics makes the representation of low-dimensional semantic vectors more suitable for a prediction task. Based on the MedLDA model, we propose a multi-level max-margin discriminative analysis (M^3DA) framework, which takes both coarse and fine semantics into consideration. Furthermore, we introduce the CRF model over the label inference in soft label fields generated by the multi-level max-margin discriminative topic model. In this way, the final label field is then optimized, since it takes into account the spatial information of neighboring areas and the local correlation between them is reinforced. The experimental results have shown the effectiveness and robustness of the proposed method for satellite image annotation.

The rest of the paper is organized as follows. Section 2 briefly introduces the MedLDA model and our proposed multi-level max-margin topic model. Section 3 talks about CRF, as well as the improved algorithm of the proposed model by combining the CRF. Section 4 gives an algorithm flowchart of our method on image annotation, and then shows the experimental results on two different satellite images. In Section 5 the discussion is presented with the future work discussed. Finally, Section 6 draws a conclusion for the paper.

Figure 1. Land-usage classes such as "Commercial area" often include several visually distinct kinds of image content. It is thus useful to associate several abstract visual "topics" to each class.



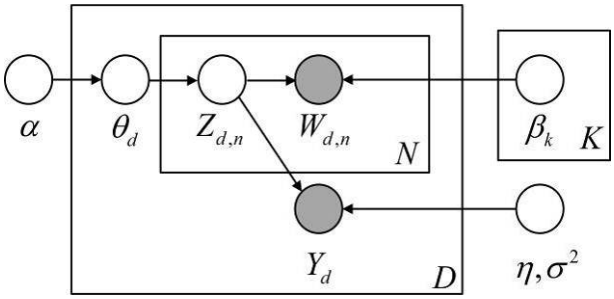
2. Multi-Level Max-Margin Discriminative Topic Model Based on MedLDA

In this section, an overview of MedLDA for classification is given. Then, multi-level max-margin discriminative topic model based on MedLDA is introduced. The MedLDA is a crucial part of our method, due to the appropriate latent semantic representation, which is usually difficult to handle in the annotation task of satellite images.

2.1. MedLDA Model

As explained in [10], MedLDA is derived from supervised topic models [11], depicted as a graphical model in Figure 2, that has introduced a response variable to LDA for each document. It allows the number of topics used to be decoupled from the number of classes. Meanwhile, the discriminative latent topics are still learned. Hence, it might be helpful to improve the overall accuracy. The experiments in [10] on text suggest that it works well and has a fast speed comparable to standard LDA. As a consequence, we attempt to apply this aspect model into satellite image annotation.

Figure 2. A graphical model representation of Supervised Latent Dirichlet Allocation.



MedLDA model is capable of processing both for regression and classification. Here we only briefly introduce the part of classification, which has been employed for the annotation task. Suppose each document is a sequence of N words w_n, denoted by $W = \{w_1, w_2,\ldots,w_N\}$ and the number of latent topics is K. The vector of response discrete variable in corpus D is y, where $y \in \{1,2,\ldots,M\}$. The generative process of MedLDA is the same as supervised topic models [11]:

(1) Draw topic proportions $\theta|\alpha \sim Dir(\alpha)$;

(2) For each of the N words w_n:

   (a) Draw a topic assignment $z_n|\theta \sim Multinomial(\theta)$;

   (b) Draw a word w_n from $P(w_n|z_n,\beta)$, a multinomial probability conditioned on the topic z_n, namely $w_n|z_n,\beta_{1:K} \sim Multinomial(\beta_{zn})$.

(3) Draw a response variable $y|z_{1:N},\eta,\sigma^2 \sim N(\eta^T \bar{Z},\sigma^2)$, where $\bar{Z} = 1/N \sum_{n=1}^{N} z_n$

Here $(\alpha,\beta,\eta,\sigma^2)$ are the unknown hyper parameters. We obtain the marginal distribution joined with the response variable y of a document:

$$P(y,W \mid \alpha,\beta,\eta,\sigma^2) = \int P(\theta \mid \alpha) \sum_{z_{1:N}} \left(\prod_{n=1}^{N} P(z_n \mid \theta) P(w_n \mid z_n, \beta_{1:K}) \right) P(y \mid z_{1:N}, \eta, \sigma^2) d\theta \quad (1)$$

The variational EM algorithm is adopted during the parameter estimation of supervised topic models, and the goal is to maximize the joint likelihood function $P(y,W|\alpha,\beta,\eta,\sigma^2)$ by learning a point estimate of η. Different from such learning method, the authors of MedLDA take a Bayesian-style approach to learn the distribution of parameters by max-margin principle due to intractability of the likelihood $P(y,W|\alpha,\beta,\eta,\sigma^2)$ (the normalization factor). Unlike fully generative topic models, a partially generative model on (θ,z,W) has been defined. The margin constraint is written as follows:

$$\min_{q,q(\eta),\alpha,\beta,\xi} L(q) + KL(q(\eta) \,||\, p_0(\eta)) + C\sum_{d=1}^{D}\xi_d$$
$$s.t.\ \ \forall d, y \neq y_d : \begin{cases} E[\eta^T \Delta f_d(y)] \geq 1 - \xi_d \\ \xi_d \geq 0 \end{cases} \tag{2}$$

$$L(q) = -E\left[\log p(\theta, z, W \mid \alpha, \beta)\right] - H(q(z, \theta)) \tag{3}$$

Here $L(q)$ is the variational upper bound of $-\log P(W|\alpha,\beta)$; $p_0(\eta)$ is a prior distribution over the parameters and $KL(p||q) \triangleq E_p[\log(p/q)]$ is the Kullback-Leibler (KL) divergence; C is a positive regularization constant;$\Delta f_d(y) = f(y_d, \bar{Z}_d) - f(y, \bar{Z}_d)$ and ξ are slack variables; $E[\eta^T \Delta f_d(y)]$ is the "expected margin" by which the true label y_d is favored over a prediction y; and $H(q)$ is the entropy of q. Because of the margin constraint in Equation (2) the model tries to learn a latent topic representation $q(\theta,z|\gamma,\phi)$ and a parameter distribution $q(\eta)$ both for the accurate prediction of training data and the proper explanation of data. During the parameter estimation the posterior distribution of the hidden variables is inferred, in which MedLDA is distinguished essentially from supervised topic models. After the distribution of $q(\eta)$ is learned, the label can be inferred as follows:

$$y^* = \arg\max_y E\left[F(y, z_{1:N}, \eta) \mid \alpha, \beta\right] \tag{4}$$

Here F is linear discriminant function and can be written as: $F(y,z_{1:N},\eta) = \eta^T f(y, \bar{Z})$, where $f(y, \bar{Z})$ is the feature vector. In the model, the process of latent topic discovery is integrated with max-margin principle by optimizing a single objective function with a set of margin constraints, which leads to a predictive topic representation.

2.2. Multi-Level Max-Margin Discriminative Topic Model

MedLDA could discover sparse and highly discriminative topical representation by exploiting the popular and potentially powerful max-margin principle. As we know, support vector machines (SVM) as a typical instance learned by the max-margin mechanism has been successfully applied to a wide range of discriminative problems such as image annotation and target recognition.

Formally, the linear SVM finds an optimal linear function by solving the following constrained optimization problem:

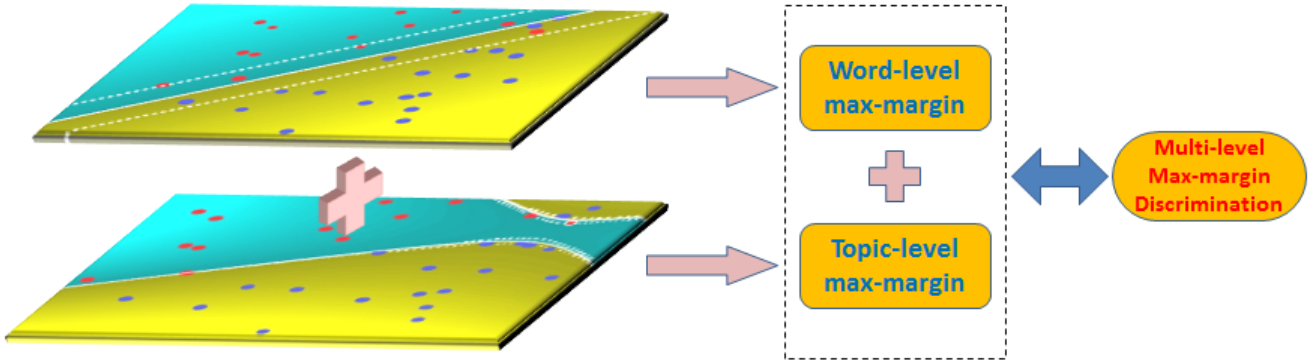
$$\min_{\eta,\xi,\xi^*} \ \frac{1}{2}\|w\|_2^2 + C\sum_{d=1}^{D}\xi_d$$
$$s.t. \forall d : y_d w^T f(x_d) \geq 1 - \xi_d, \xi_d \geq 0 \tag{5}$$

where $x_d \in X$are inputs/feature vectors of samples, which are the visual-word features in this paper; w is the parameter vector; ξ_d is a slack variable that tolerates some errors in the training data; y_d is the class label of samples; C is a positive regularization constant.

In our proposed framework, soft labels generated by MedLDA inference are essentially features of topical description. As a result of the complexity of large-scale image scenes, single level low dimensional topic feature may not discover effective semantic representations. In this paper, in consideration of inseparable cases result from the max-margin mechanism in both SVM and MedLDA, we construct a multiple soft label posterior as shown in Figure 3, which combine the word-level

feature generated by SVM and topic-level feature generated by MedLDA based on a bag of words representation (BOW). Our M^{3}DA topic model, which is described from two different feature levels that may make up each other effectively, could provide more discriminative labels. This improvement will be verified in the subsequent experiment.

Figure 3. Schematic diagram of multi-level max-margin discrimination.



3. M^{3}DA-Based Random Field

3.1. Conditional Random Field

Aforementioned topic models suffer from loss of spatial information in supervised classification. In order to complement the lost contextual information, some researchers have extended aspect models with MRF [8,12]. The resulting MRF aspect models, which usually build aspect models with MRF properties at a latent topic level, have shown significant boosts in classification performance over standard aspect models. Here we utilize CRF [13,14] to optimize the soft label field, which can directly model the posterior probability of classes. The basic formation of CRF can be written as:

$$P(x \mid y) = \frac{1}{Z}\exp\left\{-\left(\sum_i n_i\phi(x_i, y_i) + \sum_i \sum_{j\in N(i)} w_{ij}\varphi(x_i, x_j, y_i, y_j)\right)\right\} \tag{6}$$

Here, $\mathbf{x}$ and $\mathbf{y}$ denote the predictive class labels and observation image respectively. n_i and w_{ij} are the model parameters. ϕ and φ denote, respectively, the unary potential function and the dual potential function, which both describe the interrelations among basic elements in CRF. In our experiment, the unary potential is denoted by the soft probability, and meanwhile, these pairwise potentials are parameterized by the Potts model. Thus, the original CRF model could be transformed into the variational form as below, where σ is the smooth coefficient:

$$P(\mathbf{x} \mid \mathbf{y}) \propto \exp\left(\sum_i \log P(x_i \mid y_i) + \sum_i \sum_{j\in N(i)} \sigma \bullet [x_i = x_j]\right) \tag{7}$$

3.2. M^{3}DA-Based Random Field

In this section we describe the M^{3}DA-based Random Field (named as M^{3}DA-RF for short) approach for the semantic annotation of large satellite images.

The concept and category of semantics in image are beforehand defined. Then, a training set is built in the following steps. A large satellite image S to be annotated containing M semantic classes can be considered as a testing set consisting of a set of image patches S_d with equal size. It can be written as:

$$\bigcup_d S_d = S \tag{8}$$

The process of annotation can thus be regarded as a classification procedure where document S_d is labeled as semantic classes C_m, $m \in \{1,2,\ldots,M\}$. Since in the MedLDA model the order of words and documents in corpus is ignored, therefore we employ CRF into the label inference by introducing the contextual information, for the sake of improving the annotation task.

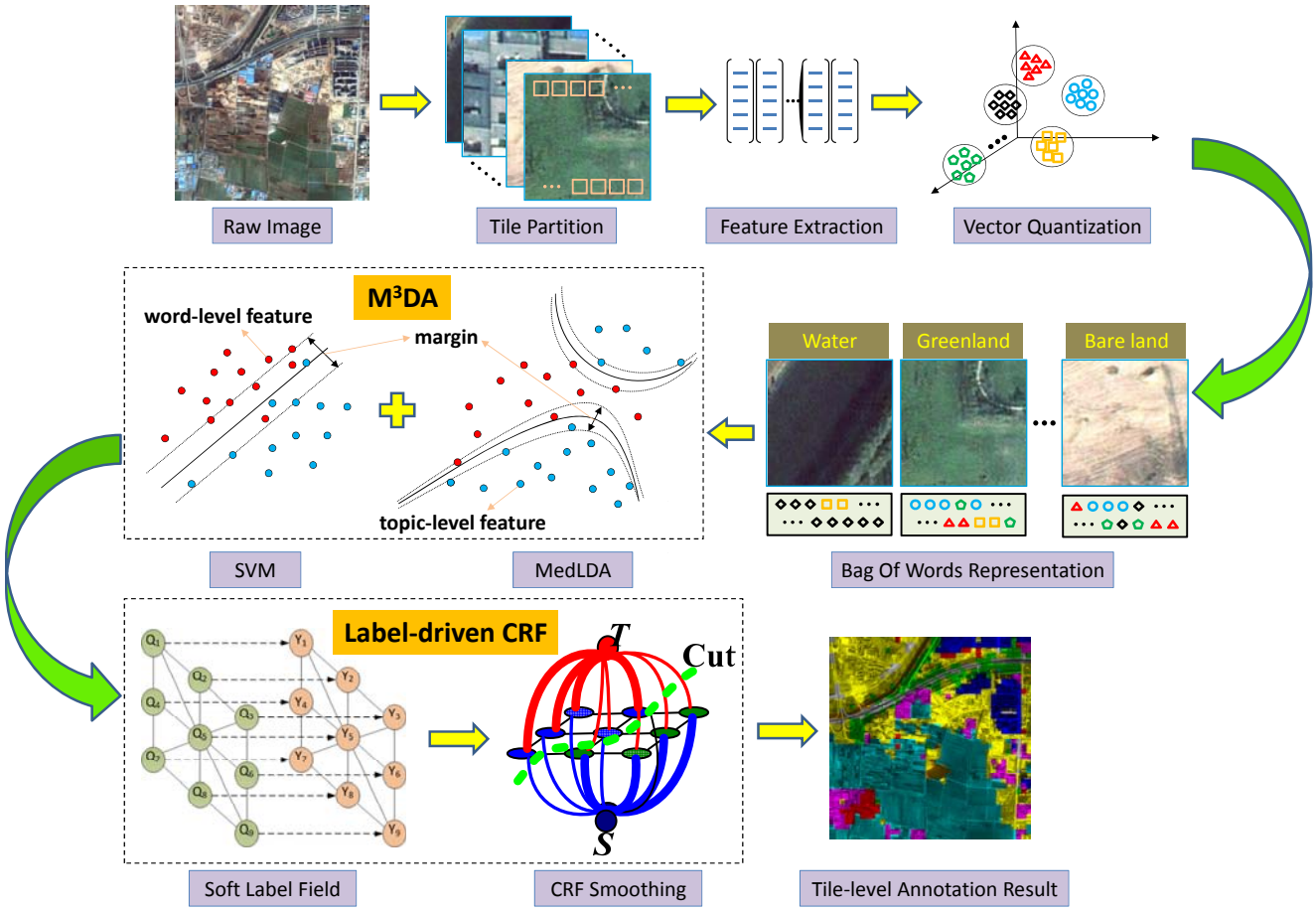
As we know, during the parameter estimation of the MedLDA model, it seeks for a latent topic representation $q(\theta,z|\gamma,\phi)$ and a parameter distribution $q(\eta)$ for the multi-class classification, so that it can, on one hand, as accurately as possible predict on the training set, while on the other hand also represent the data set well. In fact, the label inference of each test data is based on the statistic of discriminative latent topics. According to Equation (4), the K-dimensional vector of latent topics is transformed into M-dimension soft probability, and the final label is then inferred by MAP principle. Here we propose another algorithm for label inference, instead of MAP inferring. A CRF prior with eight-neighbor connectivity, which is also a vector with M-dimension and can be regarded as probability of each semantic class is introduced over a soft label field derived from our M^{3}DA topic model. Considering the relevance of surrounding areas, the optimization over soft label field is fulfilled by Graph Cut algorithm [15]. Then, predictive labels inferred by CRF will be smoothed and lead to a desirable annotation result compared to the ones without CRF inference. A remarkable smooth effect is presented in subsequent experiments.

4. Tile-Level Annotation Algorithm and Experimental Result Analysis

4.1. M^{3}DA-RF Based Tile-level Annotation Algorithm of Satellite Images

The flowchart of the M^{3}DA-RF based tile-level annotation algorithm is illustrated in Figure 4, and the pseudocode of this algorithm is shown in Algorithm 1. Here, the visual words are obtained through several steps, which are tile partition, feature extraction, vector quantization, and K-means clustering, in that order. Tiles and visual words represent documents and words respectively in the topic models. The set of visual words (bag-of-word representation) is then used to represent an image regardless of their spatial arrangement similar to how documents can be represented as an unordered set of words in text analysis. The image of BOW representation is handled in two ways simultaneously: by using the visual word histogram to train a SVM classifier, we get the scene class label distribution (the so-called soft label probability) of each tile; by training the MedLDA, we can also obtain the soft label distribution of each tile. Then by concatenating the two different soft label probabilities, the multiple class label posterior is generated, which is described with soft label field in the CRF manner in latter steps. This kind of combination is reasonable since soft probability P^{Med} and P^{SVM} are generated from two different feature levels (the former one is from a word-feature level and the latter one is from a topic-feature level).

Figure 4. Flowchart of the proposed M^{3}DA-RF based tile-level annotation algorithm.



Algorithm 1. Algorithm of M^{3}DA-RF Based Tile-level Annotation.

Input: original high-resolution image I^O

Output: the annotation image I^A

(1) divide I^O into uniform non-overlapping tiles $T=\{t_i\}_{i=1,2,\ldots,n}$ and each correspond to a true label $L^{TRUE}=\{l_i\}_{i=1,2,\ldots,n}$

(2) **for each** tile t_i

(a) divide t_i into small patches $\{p_j\}_{j=1,2,\ldots,m}$

(b) compute the SIFT features of each patch p_i and form the feature set S_i for tile t_i

(3) conduct vector quantization on the total feature set $S=S_1 \cup S_2 \cup \cdots \cup S_n$ and obtain visual word dictionary $D=\{w_i\}_{i=1,2,\ldots,h}$

(4) represent each tile t_i by the histogram of visual words h_i (BOW representation)

(5) **for each** tile t_i

(a) do MedLDA training and infer the soft label probability P_i^{Med}

(b) do SVM training and infer the soft label probability P_i^{SVM}

(c) get the joint soft label probability $P_i^{\text{Joint}} = P_i^{Med} \bigcup P_i^{SVM}$

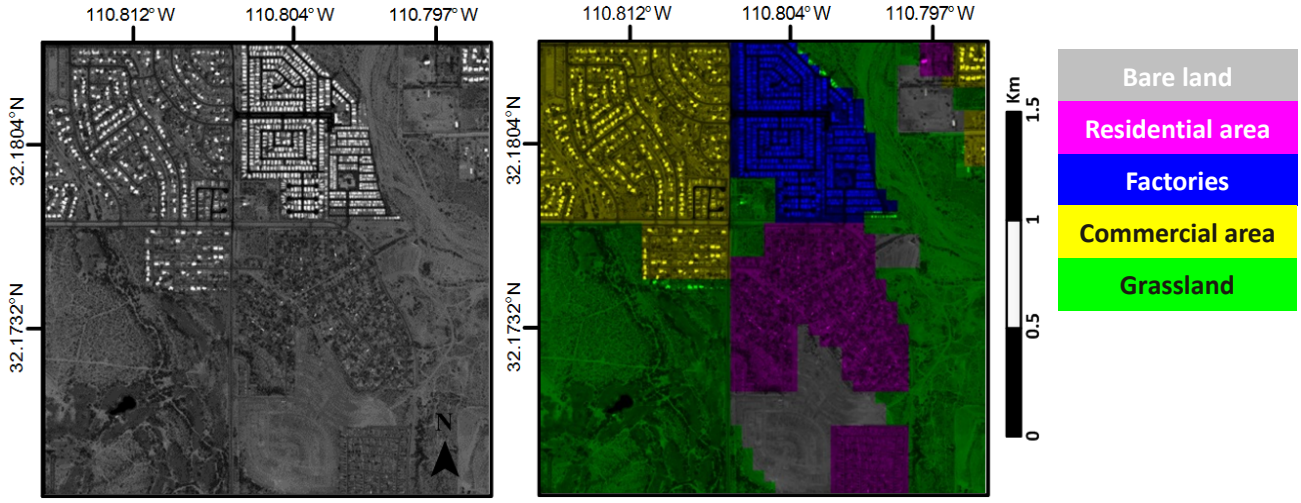
(6) construct the soft label field L^{SOFT} mapping from the CRF inferred by Graph Cuts

(7) return annotation label y_i of each tile and the annotation image $I^A = \{y_i\}_{i=1,2,\ldots,n}$

4.2. Experimental Data and Settings

In this section, we present the experimental results of two large high-resolution satellite images which are both acquired by GeoEye-1: one (image I) is taken from somewhere nearby the airport of Tucson in USA (shown in Figure 5), and the other (image II) is taken from the Majuqiao Town of southwest Tongzhou District in Beijing (shown in Figure 6). Here a series of experiments based on different methods are conducted to image I, but we do not spend much effort on carefully analyzing the experimental results. We deal with image II in depth and the experimental results will be interpreted qualitatively and quantitatively.

Figure 5. Original image I, to be annotated, and corresponding hand-labeled ground truth. (**a**) Original image (GeoEye-1). (**b**) Hand-labeled ground truth.



(**a**) (**b**)

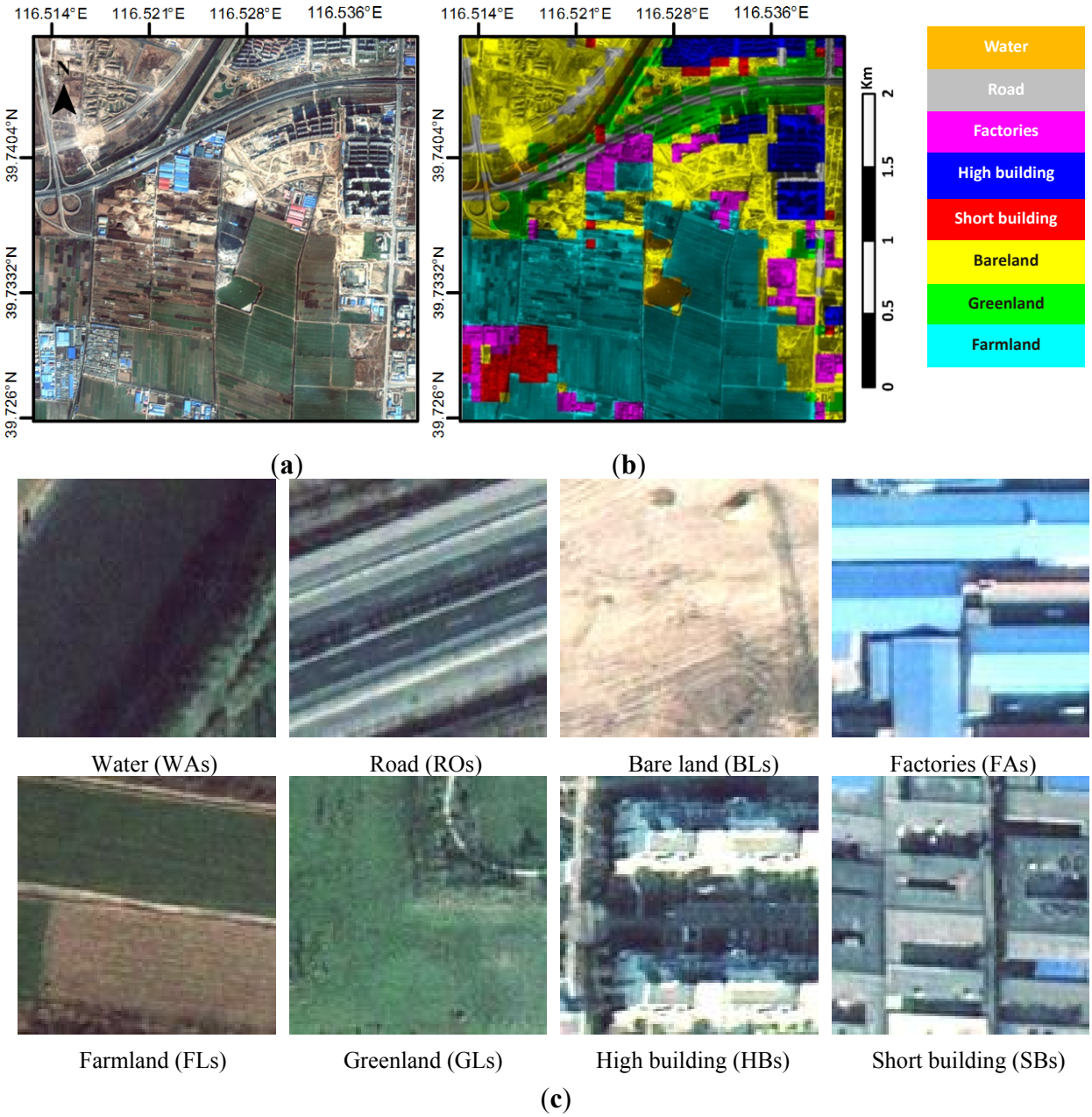
The size of image I (namely, Figure 5(a)) to be annotated is 4,000 × 4,000 pixels, which includes five semantic classes: residential area, bare land, factories, commercial area and grassland. And the large image is divided into 1,600 non-overlapping patches with size of 100 × 100 pixels, which are regarded as documents (tiles). We randomly choose 50% of each class as a training set and the remainder as a testing set. For generality, we only use a SIFT feature. We use K-means to quantize the descriptors, producing 300 clusters. The centroids are thus regarded as words. A word corresponds to a window with a size of 5 × 5, thus each document contains 400 words. The number of latent topics in MedLDA is fixed to 35 and, as well, we set σ = 0.5 empirically. Linear SVM is selected in our experiment because of its high computational efficiency as well as satisfying classification accuracy. Otherwise, the soft label field is optimized by utilizing Graph Cuts, and then we finally obtain the smoothed annotation result.

In the experiments, we have compared the performance based on original PLSA and LDA with our proposed method respectively. Furthermore we will find that the annotation performance that combines soft probability P^{SVM} and P^{Med} is better than single mode.

Identical experimental settings and workflow as mentioned above were conducted on image II with eight semantic classes: water (WAs), bare land (BLs), roads (ROs), factories (FAs), farmland (FLs), green land (GLs), high building (HBs, commercial building), short building (SBs, residential building),

with the number of topics varying from 10 to 100 and OPPSIFT features instead of SIFT features. Figure 6(c) shows one example of each class from the eight-class satellite scene.

Figure 6. Original image II to be annotated and examples of image II. (**a**) Original image (GeoEye-1). (**b**) Hand-labeled ground truth. (**c**) Example of each class in the eight-class satellite scene.



4.3. Annotation Results and Analysis

Annotation accuracy for each category is calculated as the ratio of the correctly annotated pixels to the total number of the category pixels, given in percentage with reference to the ground truth map.

According to the fixed experimental settings, we have done nine tests by employing different kinds of aspect models or the SVM with and without combining the CRF. In the BOW+SVM case, we especially test the accuracy and computational speed with linear kernel SVM and radial basis function

kernel SVM (RBF kernel, a kind of nonlinear kernel). Compared to the classification accuracy of 88% obtained by linear, RBF kernel achieves 90%, however the running time of the RBF kernel gains increases almost 250%. Therefore, given that the performance of linear kernel is acceptable, we choose linear kernel rather than nonlinear kernel. The entire annotation results and classification accuracy of our proposed M^{3}DA-RF method are illustrated in Figures 7 and 8, respectively. The annotation performance of our method outperforms those of other methods as expected. Due to the simplicity of scene structure, the accuracy of our method on image I reaches as high as 98.19%.

Figure 7. The annotation results of different methods for image I (number of topic is fixed to 35). (**a**) PLSA. (**b**) LDA. (**c**) MedLDA. (**d**) PLSA+CRF. (**e**) LDA+CRF. (**f**) MedLDA+CRF. (**g**) BOW+SVM. (**h**) BOW+SVM+CRF. (**i**) M^{3}DA-RF.

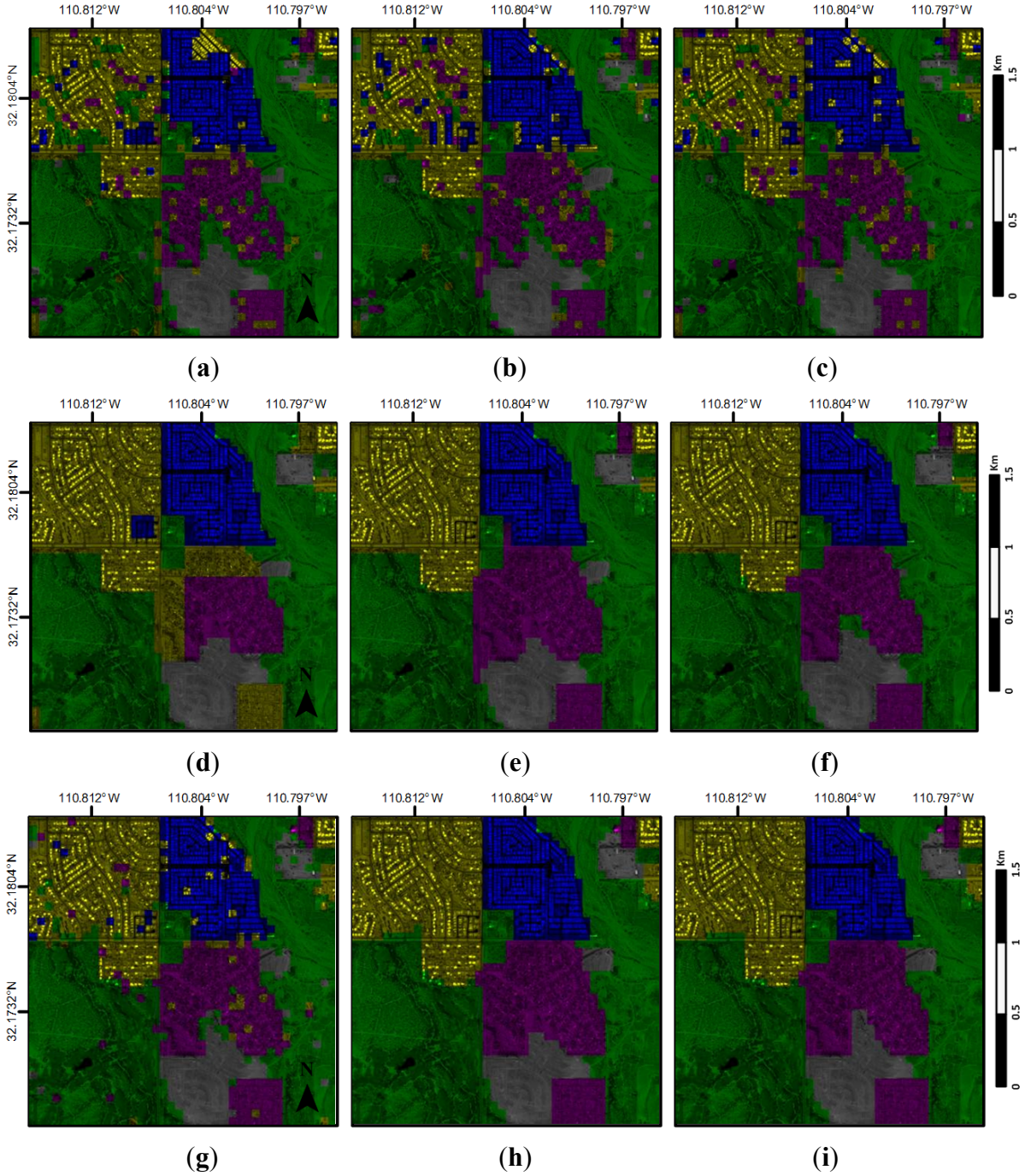
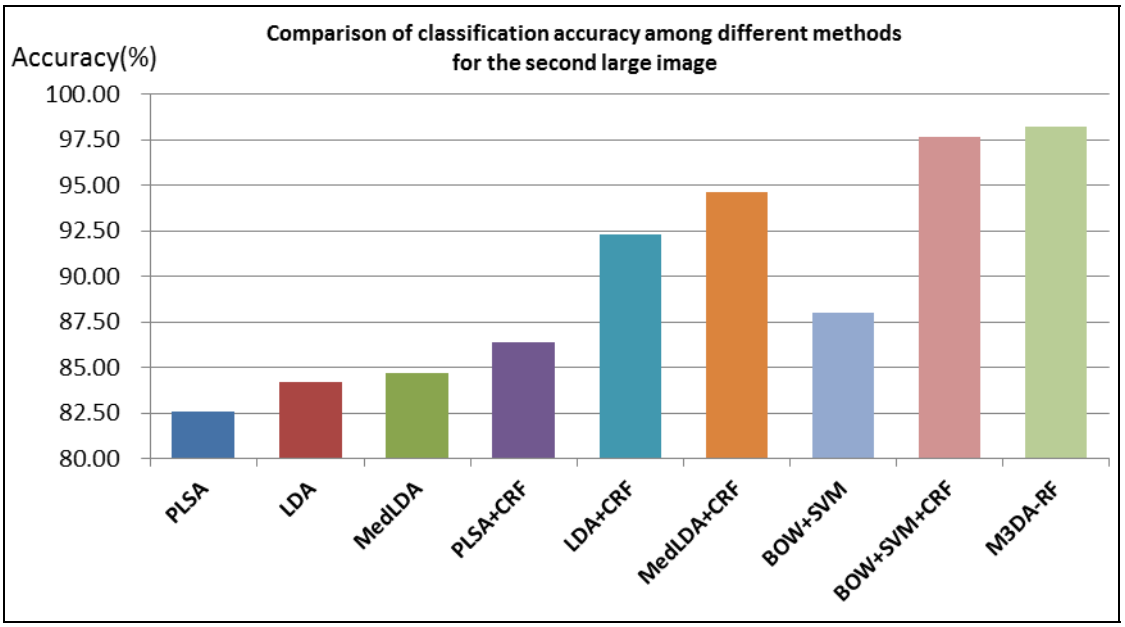
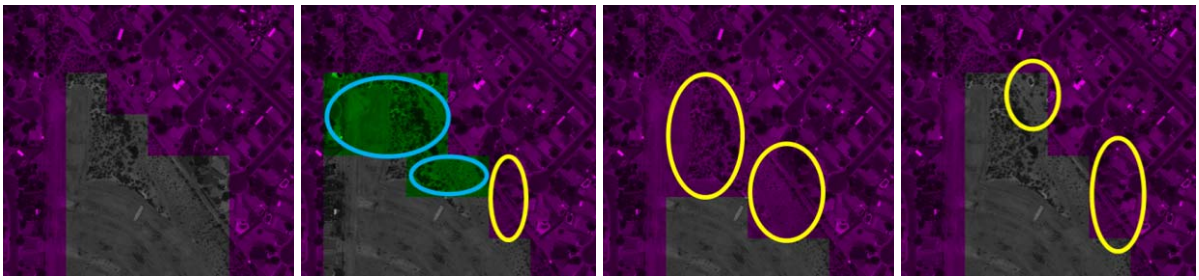


Figure 8. Comparison of annotation accuracy among different methods for image I.



In addition, for the sake of highlighting the effect of multiple soft label posterior probability, the partial enlarged view of M^{3}DA-RF is shown in Figure 9 (The yellow circles stand for the misclassification region between residential area and bare land, and the blue circles stand for the region that is misclassified as greenland). Compared to the other two methods which only utilize single soft posterior probability P^{Med} or P^{SVM}, the annotation result of M^{3}DA-RF gets much more close to the ground truth and produces less confusion than other semantic classes.

Figure 9. Partial enlarged view of different annotation results. (**a**) Ground truth. (**b**) MedLDA+CRF. (**c**) BOW+SVM+CRF. (**d**) M^{3}DA-RF.

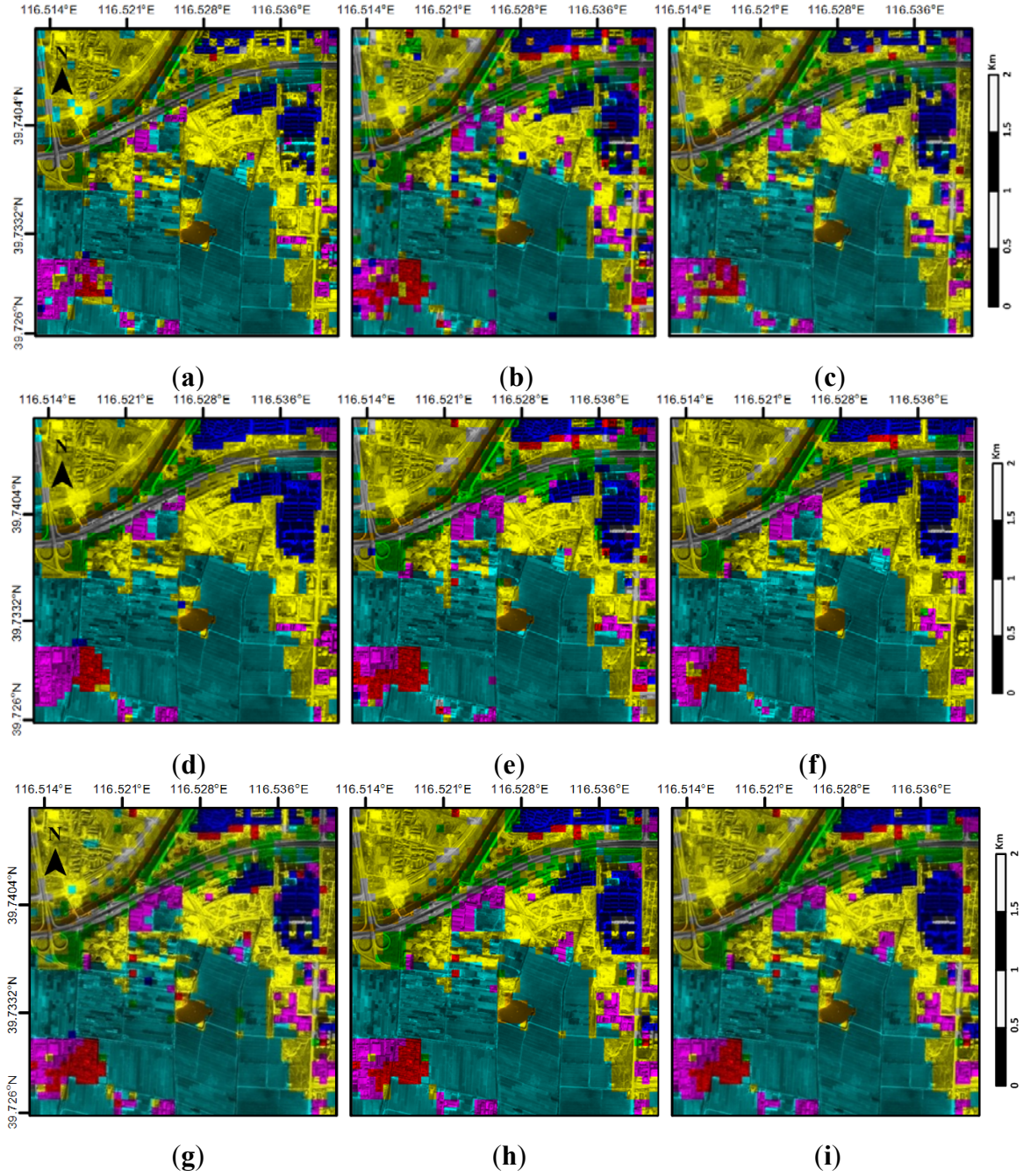


(**a**) (**b**) (**c**) (**d**)

Given that image II is a colorized image and has more complex scene structures as well as more semantic classes than image I, we put emphasis on dealing with image II. The annotation results of image II are shown in Figure 10. The results in the first row are obtained directly from three different topic models. It's not difficult to see that most of FLs, BLs, and HLs are labeled correctly. The satisfactory performances of these three semantic classes result from larger quantities of training samples and more recognizable structures. However, the confusions between SBs and FAs, GLs and ROs are obvious due to these semantic classes sharing a few similar topics. On the whole, results obtained from all three topic models, without considering spatial dependencies among labels, are rather noisy. For instance, in the upper-left area of the image, a few BLs are misclassified as FLs; in the

upper-middle area, some HBs are confused with other classes, which are more serious in PLSA and MedLDA than in LDA.

Figure 10. The annotation results of different methods for image II (number of topic is fixed to 35). (**a**) PLSA. (**b**) LDA. (**c**) MedLDA. (**d**) PLSA+CRF. (**e**) LDA+CRF. (**f**) MedLDA+CRF. (**g**) BOW+SVM. (**h**) BOW+SVM+CRF. (**i**) M^{3}DA-RF.



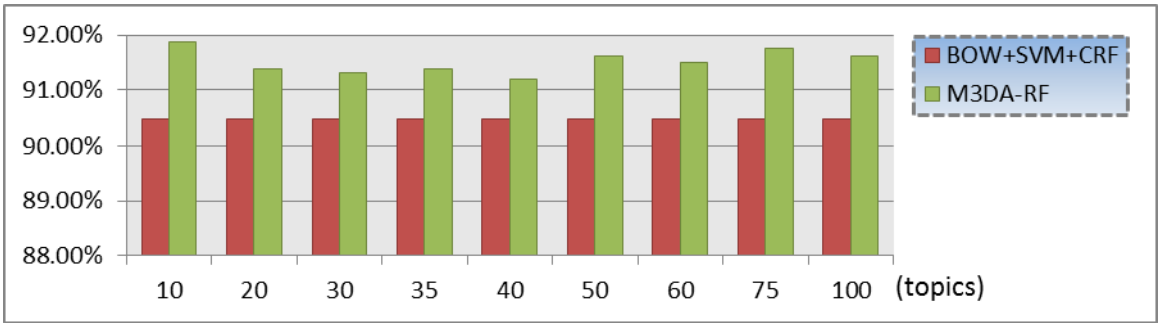
In order to take into account spatial contextual information, a soft label field described by CRF has been employed. The corresponding results are shown in the second row of Figure 10. As a result of the smooth effect, there exist just a small number of isolated patches in the annotation results and, hence, the results thereby appear to be much more homogeneous. Meanwhile, the classification accuracies shown in Table 1 of the three models smoothed by CRF have been improved compared to those without CRF.

Table 1. The overall accuracies of different methods.

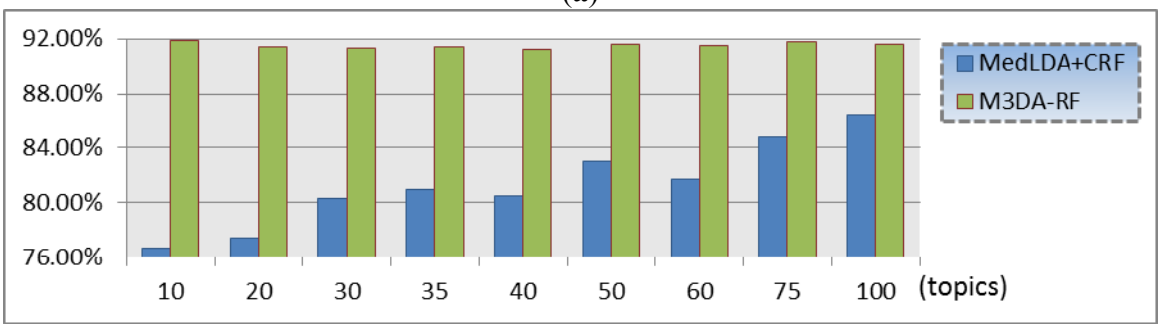
Method \ Topics	10	20	30	35	40	50	60	75	100
PLSA	68.06%	69.44%	71.38%	72.25%	73.5%	73.13%	73.69%	73.94%	74.44%
LDA	69.38%	73.13%	74.56%	76.13%	74.94%	75.94%	76.38%	77.94%	78.5%
MedLDA	71.4%	73.6%	76.4%	77.6%	79%	79.4%	80.1%	83.18%	83.93%
PLSA+CRF	72%	73%	75.75%	76.88%	76.94%	77.44%	78.125%	78.81%	78.81%
LDA+CRF	71.88%	78.18%	79.13%	80.06%	80.5%	81%	80.81%	82.31%	83.5%
MedLDA+CRF	76.69%	77.44%	80.31%	81%	80.5%	83%	81.69%	84.75%	86.44%
M^{3}DA-RF	91.88%	91.38%	91.31%	91.38%	91.19%	91.63%	91.5%	91.75%	91.63%

The overall accuracies of different methods with different topic numbers for image II are shown in Table 1. As mentioned above, combination of soft probability P^{Med} and P^{SVM} (multiple class label posterior) is reasonable and may make up inseparable instances, each from two different feature subspaces. The experimental results further validate our analysis. Our M^{3}DA-RF model shows better performance than MedLDA+CRF and BOW+SVM+CRF respectively, as shown in Figure 11. In addition, the classification accuracy is generally better with a larger number of topics, which is reflected in the first six groups of experiments in Table 1. Meanwhile the accuracy of our M^{3}DA-RF model leads to a relatively stable value with the growing number of topics, which is mainly because the word-level features (or the soft label probability P^{SVM}) play a dominant role in the multi-level max-margin discriminative feature space and the topic-level features, as a supplement, are just the minor feature components.

Figure 11. Annotation accuracies of three methods under the condition of different topic numbers. (**a**) Comparison between BOW+SVM+CRF and M^{3}DA-RF. (**b**) Comparison between MedLDA+CRF and M^{3}DA-RF.



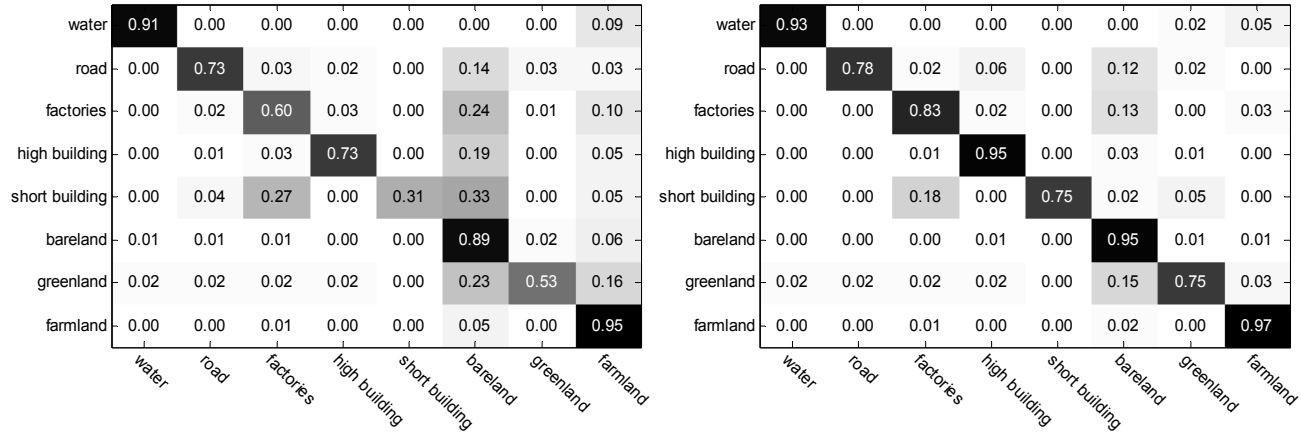
(**a**)



(**b**)

An overview performance of image II (the number of topic is set to 35) given by the confusion matrix of all eight semantic classes is presented in Figure 12. According to the result of our proposed method, each semantic class is considerably well preserved, especially SBs, FAs, GLs, and ROs, these four classes that are seriously misclassified in the former methods. The annotation results appear to be somehow serrated due to the rectangular cutting of patches. In order to eliminate the edge effect, we can conduct over-segmentation on the original image, and re-annotate the image with superpixels with our proposed method. This work will be done in the future.

Figure 12. Confusion matrix of semantic classes obtained by MedLDA and our proposed method. (**a**) MedLDA. (**b**) M^{3}DA-RF.



5. Discussion

In this work, we have attempted to improve the annotation performance of high-resolution satellite images from two different aspects. On the one hand, considering that the low-level features may not precisely represent the scene semantics of images, the MedLDA model [10], which is a powerful discriminative topic model, is employed to extract the high-level semantic features (also known as the topic-level features); on the other hand, topic models ignore the spatial neighborhood relationship because of the independence assumption of visual words, and hence we introduce the CRF for the purpose of strengthening the neighborhood coherence. Furthermore, due to the limitation of MedLDA in which only the topic-level features are available, whereas the word-level features are important in image annotation tasks, as well as are properly unobtainable, we propose the M^{3}DA framework, which takes both the coarse and fine semantics into consideration, to combine the topic-level features and word-level features together.

The experimental results shown in Figure 8 and Table 1 suggest that our proposed M^{3}DA-RF model performs better than the single MedLDA and other typical topic models [2-3], mainly because it can utilize discriminative features from different levels reasonably and reinforce the local correlation of neighboring area efficiently. Figure 9 and Figure 11 show that the advantage of M^{3}DA framework lies in less confusion among the different semantic classes in a feature combined multi-level max-margin fashion. Figure 7 and Figure 10 show that the M^{3}DA-RF model leads to more smooth and accurate annotation performance. Meanwhile the annotation accuracy of our M^{3}DA-RF model tends to a relatively stable value with the growing number of topics, which is mainly because the word-level

features (also known as the soft label probability P^{SVM}) play a dominant role in the multi-level max-margin discriminative feature space and the topic-level features are just helpful supplements.

The most related work to ours is detailed in [16], where the authors only exploit different types of feature representation and do not make full use of the contextual information that may be beneficial to annotation tasks. Some other related studies [1,5,9] have investigated the application of topic model in satellite images annotation task. These studies did not apply multi-level features into classification framework [5] and introduced spatial information by means of cutting large image into small patches with an overlap and [9] employed Markov random field for the sake of utilizing the contextual information in satellite images. However we suggest that the CRF model is more suitable for discriminant tasks like image annotation or scene classification. Therefore, our M^3DA-RF model not only exploits the features of different levels but also combines CRF model so as to obtain smoother and more precise annotation performance.

Otherwise, our proposed method is currently limited in the sense that the MedLDA and the CRF have not been jointly optimized, *i.e.*, the MedLDA is trained in fully supervised form using the training label for each tile, once MedLDA is fully trained, and then the CRF is trained using the MedLDA output probabilities as feature potentials. Because of the structure of this model, it should also be possible to combine the margin based training of the tile-level classifiers with the margin based training of the CRF layer into a single max-margin CRF with discriminatively trained topic model structure. As future work, we intend to envisage a coupled model in which both the MedLDA and the CRF are trained together in a variational max-margin framework.

6. Conclusion

In this paper, we focus on the semantic annotation of large high-resolution satellite image. Our proposed method multi-level max-margin discriminative analysis (M^3DA) can discover effective semantic representation and produce more discriminative class label posterior in the framework of multi-level max-margin discrimination. The semantic annotation performance is obviously improved by the combination with conditional random field (CRF) due to the consideration of contextual information, and meanwhile the proposed algorithm yields an average annotation accuracy of approximate 13.2% higher than the original maximum entropy discrimination latent Dirichlet Allocation (MedLDA) method. The experimental results on two satellite images, of quite different land covers, have demonstrated its robustness and effectiveness.

Acknowledgments

This work was supported in part by the National Key Basic Research and Development Program of China under Contract 2013CB733404 and the Chinese National Natural Sciences Foundation grants (NSFC) 61271401. The authors would like to specially thank Kan Xu for his helpful guidance on MedLDA.

Conflict of Interest

The authors declare no conflict of interest.

References

1. Yi, W.; Tang, H.; Chen, Y. An object-oriented semantic clustering algorithm for high-resolution remote sensing images using the aspect model. *IEEE Geosci. Remote Sens. Lett.* **2011**, *8*, 522–526.
2. Hofmann, T. Unsupervised learning by probabilistic latent semantic analysis. *Mach. Learn.* **2001**, *42*, 177–196.
3. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
4. Larlus, D.; Jurie, F. Latent mixture vocabularies for object categorization and segmentation. *Image Vis. Comput.* **2009**, *27*, 523–534.
5. Lienou, M.; Maitre, H.; Datcu, M. Semantic annotation of satellite images using latent dirichlet allocation. *IEEE Geosci. Remote Sens. Lett.* **2010**, *7*, 28–32.
6. Xu, K.; Yang, W.; Liu, G.; Sun, H., Unsupervised satellite image classification using markov field topic model. *IEEE Geosci. Remote Sens. Lett.* **2013**, *10*, 130–134.
7. Wang, X.; Grimson, E. Spatial Latent Dirichlet Allocation. In Proceedings of 21st Neural Information Processing Systems, Vancouver, BC, Canada, 3–8 December 2007; pp. 1577–1584.
8. Verbeek, J.; Triggs, B. Region Classification with Markov Field Aspect Models. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 18–23 June 2007; pp. 1–8.
9. Yang, W.; Dai, D.; Triggs, B.; Xia, G.-S. SAR-based terrain classification using weakly supervised hierarchical Markov aspect models. *IEEE Trans. Image Process.* **2012**, *21*, 4232–4243.
10. Zhu, J.; Ahmed, A.; Xing, E.P. MedLDA: Maximum Margin Supervised Topic Models for Regression and Classification. In Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, QC, Canada, 14–18 June 2009; pp. 1257–1264.
11. Blei, D.M.; McAuliffe, J.D. Supervised Topic Models. In Proceedings of 21st Neural Information Processing Systems, Vancouver, BC, Canada, 3–8 December 2007; pp. 121–128.
12. Zhao, B.; Li, F.; Xing, E. Image Segmentation with Topic Random Field. In Proceedings of 11th European Conference on Computer Vision, Crete, Greece, 5–11 September 2010; pp. 785–798.
13. Lafferty, J.; McCallum, A.; Pereira, F.C.N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of the 29th International Conference on Machine Learning, Williamstown, MA, USA, 28 June–1 July 2001; Volume 18, pp. 282–289.
14. Delong, A.; Osokin, A.; Isack, H.N.; Boykov, Y. Fast approximate energy minimization with label costs. *Int. J. Comput. Vis.* **2012**, *96*, 1–27.
15. Kolmogorov, V.; Zabin, R. What energy functions can be minimized via graph cuts? *IEEE Trans. Patt. Anal. Mach. Int.* **2004**, *26*, 147–159.
16. Wang, Y.; Mori, G. Max-Margin Latent Dirichlet Allocation for Image Classification and Annotation. In Proceedings of 22nd British Machine Vision Conference, Dundee, UK, 29 August–2 September 2011.