A NOVEL LAND COVER MAPPING ALGORITHM BASED ON RANDOM FOREST AND MARKOV RANDOM FIELD MODELS

T. Kasetkasem[1], P. Aonpong[2], P. Rakwatin[3], T. Chanwimaluang[4], and I. Kumazawa[5]

[12]Department of Electrical Engineering, Kasetsart University, Bangkok, Thailand 10900
[3]GISTDA, Bangkok Thailand 10210
[4]National Electronics and Computer Technology Center, Pathumthani, Thailand 12120
[5]Tokyo Institute of Technology, Yokohama, 226-8503, Japan
Email:fengtsk@ku.ac.th[1], ozzarnavy@gmail.com[2], preesan@gistda.or.th[3],
thitiporn.chanwimaluang@nectec.or.th[4] and kumazawa@isl.titech.ac.jp[5]

ABSTRACT

This paper proposes a new land cover mapping algorithm that combines the strengths of random forest (RF) with a Markov random field (MRF) model. The idea is to transform the observed data into the decision domain of weak classifiers inside an RF. Due to how RF are trained, these decisions can be considered to be independent from each others, and therefore the joint probability density function in the decision domain can be both easily and accurately estimated. For a decision vector from RF, and under an MRF model, the optimum land cover map is iteratively searched. The performances of the proposed algorithm were evaluated using a real remote-sensing image, and we found that the resulting land cover maps are more accurate than most traditional classifiers in all sizes of training samples.

Index Terms— Random forest, Markov random field, land cover mapping, image classification

1. INTRODUCTION

Land cover classification is one of the most important applications of remote sensing data. There are a lot of land cover mapping algorithms proposed in the literature [1]-[2]. They can be broadly categorized into two groups, namely, parametric and non-parametric classifiers. The parametric classifiers depend on various statistical models and provide very accurate maps if the underlying model represents the actual data. For instance, Markov random field models have been used extensively in several land cover mapping problems [1]-[2] since these models can accurately capture the class dependency among neighboring pixels in the image. However, one major drawback of parametric classifiers is that the performance can be severely degraded if the actual data do not agree with the assumed model, especially in characterizing the probability distribution of observed spectral colors. In contrast, the non-parametric approaches [3],[4] where images are classified through the minimization of a heuristic function, seem to outperform the parametric classifiers when the observed data is complex. Among the non-parametric classifiers, the random forest (RF) [4] developed by combining several weak decision trees together to form a forest seem to provide superior performance. The random forest (RF) can be applied to various types of applications. such as data classification, image processing and so on. Even though the RF has produced very promising results in the past, it becomes very complicated to incorporate the class dependency among neighboring pixels into the forest.

As a result, this paper will focus on how to combine the strength of the RF in the discrimination of spectral colors from different land cover classes, and the MRF in characterizing class dependency among neighboring pixels. To achieve this goal, we use the RF to transform the observed image into a new discrete domain where an observed intensity vector is replaced by a decision vector from all the weak classifiers inside the RF. Due to how RFs are trained, the joint probability density function (PDF) of a decision vector can be approximated by the multiplication of the marginal ones. Next, the MRF model is employed to derive a new energy function under the maximum *a posteriori* (MAP) criteria, where the mean field [5] approximation is employed to find the optimum solution.

2. PROBLEM STATEMENT

Let Y be the observed image having $M \times N$ pixels and X be the corresponding land cover map (LCM) of the same size. Next, let $\mathcal{S}$ denote the sets of all the sites (i.e., pixels) belonging to both the observed image and the LCM. The observed image is usually represented in a vector form, such as, $\mathbf{y}(s) \in R^B$, for one pixel $s \in \mathcal{S}$ where $R = \{1, 2, \ldots, D\}$ denotes the set of possible digital numbers (e.g., intensity val-

This is work is supported in part by the Kasetsart University Research and Development Institute.

ues) and B is the number of spectral bands. We note here that, in most practical scenarios, R is assumed to be the set of all real numbers. Furthermore, let $x(s) \in \Lambda$ be the configuration (i.e., class label) of LCM at s, where $\Lambda = \{1, \ldots, L\}$ is the set of land cover class labels with L as the number of classes. It is further assumed that the LCM has the MRF property which can mathematically be represented as

$$\Pr\left(x(s)\,|X(\mathcal{S}\backslash s)\right) = \Pr\left(x(s)\,|X(N_s)\right) \tag{1}$$

where $\mathcal{S}\backslash s$ is the set of all pixels in $\mathcal{S}$ excluding a pixel s, and N_s is a set of neighboring pixels of a pixel s. In the context of land cover classification, this property implies that the same land cover class is more likely to occur in connected regions or patches than isolated pixels. It has been shown in [2] that the marginal PDF of X takes the form of a Gibbs distribution, i.e.,

$$\Pr\left(X\right) = \frac{1}{Z}\exp\left[-\sum_{C\in\mathcal{S}} V_C(X)\right] \tag{2}$$

where Z is a normalizing constant, C is a clique, and $V_C(X)$ is a Gibbs potential function. A clique is a singleton or any subsets whose two distinct elements are mutual neighbors [1]. In this paper, we define C as a set of neighboring pixel pairs, i.e., $C = \{s, r\}$, where s and r are neighboring pixels.

We further assume that when the LCM is given, the observed vectors, $\mathbf{y}(s)$, from different pixels are statistically independent, and depend on the underlying class label of a given pixel. Hence, we can write the conditional PDF of the observed image given LCM as

$$\Pr\left(Y\,|X\right) = \prod_{s\in\mathcal{S}} \Pr\left(\mathbf{y}(s)\,|x(s)\right) \tag{3}$$

In many papers,[1]-[2], $\Pr\left(\mathbf{y}(s)\,|x(s)\right)$ is assumed to follow some simple statistical models that can be described by few parameters (such as mean vectors and covariance matrices). One of the most popular models is Gaussian. In practice, the observed data often do not follow the assumed distribution, and the resulting land cover maps are clearly sub-optimum. A better approach is to replace the statistical model with the histograms. However, in order to estimate histograms accurately, a large number samples must be obtained, which may not be feasible in most situations.

As a result, we transform the observed data into a new discrete space. Let W be the transformed image of Y into a new discrete space where $w\left(s\right) = f\left(\mathbf{y}(s)\right) \in \Gamma$. Here, we assume that $\Gamma = \{1, \ldots, J\}$ where J is the number of possible values of $w(s)$. Note that, the transformation may not be one-to-one, and therefore, some information is lost after transformation. However, the goal is to classify an image which also maps an observed image in an even smaller space. As a result, the information can be lost as long as the transformed vectors associated with different land cover classes are clearly separable. Next, we assume that we have sufficient samples to accurately estimate the histograms of w for each land cover class. Hence, the conditional PDF of W given X can be written as

$$\Pr\left(W\,|X\right) = \prod_{s\in\mathcal{S}} \Pr\left(w(s)\,|x(s)\right) \tag{4}$$

3. OPTIMUM SOLUTION

In this paper, the classifier based on the maximum *a posteriori* (MAP) criterion selects the most likely LCM among all possible LCMs given the transformed image. The resulting probability of error is minimum among all other classifiers. The MAP criterion is expressed as

$$X^{opt} = \arg\max_{X}\left[\Pr\left(X\,|W\right)\right] \tag{5}$$

By using Bayes' rule, substituting Eq. (2) and (4) into Eq. (5), and using the fact that $\Pr(W)$ does not depend on X, the optimum criteria becomes

$$X^{opt} = \arg\min_{X}\left[E\left(X, W\right)\right] \tag{6}$$

where

$$E\left(X, W\right) = \sum_{C\in\mathcal{S}} V_C(X) - \sum_{s\in\mathcal{S}} \log\left(\Pr\left(w(s)\,|x(s)\right)\right) \tag{7}$$

4. PROPOSED ALGORITHM

The critical step of solving Eq.(6) is to find a good transformation between W and Y such that the discriminations of w between two or more classes are preserved as much as possible, and the histograms of w can be accurately estimated. Here, the concept of the RF [4] that creates groups of weak decision trees is employed, since decision trees are easy and efficient to train, and the training process does not require any statistical model.

In the RF, K decision trees are trained from different random subsets of training samples where the sizes of the random subsets are smaller than the total number of training samples. In this case, different trees have different learning experiences, and make decisions almost independently from each other. Even though each decision tree may not be accurate, the aggregated decision is quite accurate [4]. In literature, the majority vote rule is often applied.

Let $d_k(\mathbf{y}) \in \Lambda$ be a decision from the i^{th} decision tree. For a given pixel s, the number of possible outcomes from the random forest is L^K. Thus, we can define

$$w(\mathbf{y}) = \sum_{k=1}^{K} d_k(\mathbf{y}) L^{k-1} \tag{8}$$

Hence, the number of possible values of w is L^K. Of course, the possible values of w increase rapidly as the number of decision trees increases, and it can become more than the possible values of the observed vectors, $\mathbf{y}(s)$. Due to the near-independence nature of each decision tree in an RF, the conditional PDF of w, given the underlying land cover classes, can

be approximated as the product of condition PDFs of each decision tree, i.e.,

$$\Pr\left(w\,|x\right) \approx \prod_{k=1}^{k} \Pr\left(d_k\,|x\right) \tag{9}$$

such that $w = \sum_{k=1}^{K} d_k L^{k-1}$. Here and throughout the rest of the paper, we omit s for the sake of abbreviation. The only remaining part is to estimate $\Pr\left(d_k\,|x\right)$. Luckily, in the construction of random forest (RF), a portion of training samples is reserved as the out-of-bag [4] samples. These out-of-bag samples are used to examine the performance of the RF and determine the number of decision trees, K. Hence, we can use these out-of-bag samples to estimate $\Pr\left(d_k\,|x\right)$, and we have

$$\Pr\left(d_k\,|x\right) \approx \frac{N_{x,d_k}}{N_x} \tag{10}$$

where N_x is the number of out-of-bag samples in Class x, and N_{x,d_k} is the number of out-of-bag samples in Class x that is classified into Class d_k by the k^{th} decision tree.

In general, $E\left(X, W\right)$ is a non-convex function and, therefore, conventional optimization algorithms may not be applied for the solution of Eq.(6). Furthermore, the number of possible LCMs is also very large. Therefore, to solve Eq.(6) within a reasonable computational time, the mean field (MF) [5] approximation method is applied. The MF approximation replaces the complex interaction between the neighboring pixels with the expected values. As a result, the first term on the right hand side in Eq. (7) can be approximated as

$$\sum_{C\in\mathcal{S}} V_C(X) \approx \sum_{s\in\mathcal{S}} \sum_{r\in Ns} E_{x(r)}\left[V_{\{s,r\}}\left(x(s), x(r)\right)\right] \tag{11}$$

where $E_{x(r)}\left[V_{\{s,r\}}\left(x(s), x(r)\right)\right]$ is the expected value of the Gibbs potential function for a given configuration at a pixel s, over the configuration of its neighboring pixel r. In this paper, we define the Gibbs potential function as

$$V_{\{s,r\}}\left(x(s), x(r)\right) = \begin{cases} -\beta; & \text{if } x(s) = x(r) \\ 0; & \text{if } x(s) \neq x(r) \end{cases} \tag{12}$$

where β is the Gibbs parameter. It is obvious that

$$E_{x(r)}\left[V_{\{s,r\}}(x(s), x(r))\right] = -\beta \Pr\left(x(r) = x(s)\right) \tag{13}$$

Hence, the energy can be rewritten as

$$E\left(X, W\right) \approx \sum_{s\in\mathcal{S}} E\left(x(s), w(s)\right) \tag{14}$$

where

$$\begin{aligned} E\left(x(s), w(s)\right) &= -\sum_{k=1}^{K} \log\left(\frac{N_{x,d_k}}{N_x}\right) \\ &\quad -\beta \sum_{r\in Ns} \Pr\left(x(r) = x(s)\right) \end{aligned} \tag{15}$$

Our algorithm begins by having an analyst select a sufficient number of training pixels from the observed image. The majority of training pixels are used to train the RF, and the remainder (out-of-bag samples) are used to estimate $\Pr\left(d_k\,|x\right)$ in Eq.(10). The trained RF are the inputs to the iteration phase that employs the MF approximation to find the optimum LCM. The iteration phase is given as follows.

1. Use the trained RF to classify the observed image. Let X_0 be an initial LCM. Set the number of iterations to be one ($h = 1$) and assign the value to the Gibbs parameter, β.
2. Assign $\Pr\left(x(r) = l\right)$ to one if $x_0(r) = l$ and to 0 otherwise.
3. For all pixels $s \in S$, compute $E(x, w)$ from Eq.(15) for all $x \in \Lambda$. Recompute $\Pr\left(x(s)\right) = \frac{e^{-E(x,w)}}{z_s}$ where z_s is a normalizing constant.
4. Let $h = h + 1$ go to 3 until $h > h_{max}$ or a termination criteria is reached.
5. Create the LCM by assigning each pixel $s \in \mathcal{S}$ to

$$x(s) = \arg\max_{l\in\Lambda} \Pr\left(x(r) = l\right). \tag{16}$$

5. EXPERIMENTAL RESULTS

In this section, we examined the performance of the proposed algorithm on a small dataset of one QuickBird multispectral image of size 150×300 pixels at 2.4m resolution. The multispectral image consists of four bands: blue, green, red and near infrared. There are six land cover classes ($L = 6$) in the image: red, white ,blue, green, black and dark green for Building1, Building2, water, grass, shadow and tree, respectively. Fig. 1 and Fig. 2 display a multispectral and the corresponding ground truth images, respectively. Here, the ground truth image was manually labeled.

Fig. 1. True color composite of QuickBird image of a part of Kasetsart University

Next, we randomly selected 10% of the image as a training set, and applied this training set to the traditional RF algorithm where $1/3$ of samples were kept as the out-of-bag samples. Here, the maximum number of decision trees with a maximum depth of 15 was set to be 350. The resulting LCM

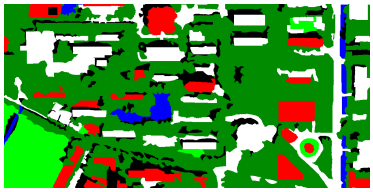

Fig. 2. Ground truth image of Fig. 1

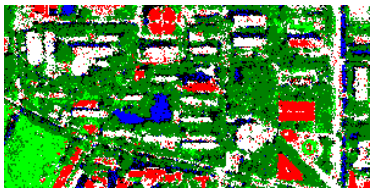

Fig. 3. The LCM using the traditional RF algorithm [4]

from the traditional RF was shown in Fig. 3. The out-of-bag samples together with the initial LCM from the traditional RF algorithm ware submitted to our iteration phase, and the resulting LCM from our proposed algorithm with $\beta = 57.7$ was shown in Fig. 4. By visual inspection, it is clear that our proposed algorithm outperforms the traditional RF algorithm.

To further demonstrate the superiority of our proposed algorithm against traditional approaches, we repeated the experiment 50 times where new sets of training samples were randomly chosen at each run. These training samples were used to train six image classification algorithms, namely, the traditional RF (TRF) [4], decision tree (DT) [3], maximum likelihood (ML) classifier, MRF classifier (MRF) [2], RF with unequal weight (RFW), and our proposed algorithm (MRFRF). For both ML and MRF, $\Pr(\mathbf{y}|x)$ was assumed to be Gaussian. The RFW is our proposed algorithm with $\beta = 0$. The averaged classification accuracies over the 50 runs are provided in Tab. 1 for different sizes of training samples (TS) ranging from 10% to 100% of the observed image. In all scenarios, our proposed algorithm outperforms all traditional approaches, even with $\beta = 0$ (RFW). This outstanding performance is due to the ability of our proposed algorithm to combine the strengths of the RF in describing the observed

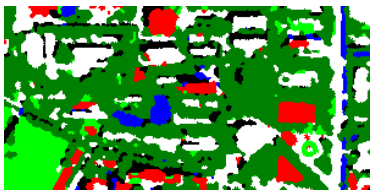

Fig. 4. The LCM using the proposed algorithm

Table 1. The percentage of correctly classified pixels as a function of the size of training samples for different image classification algorithms

TS	TRF	DT	ML	MRF	RFW	MRFRF
10%	67.0	61.7	69.2	69.3	68.2	**80.0**
30%	73.9	70.3	69.0	69.0	75.3	**85.6**
50%	78.0	74.0	69.0	69.0	79.2	**89.4**
70%	82.4	77.8	69.1	69.1	83.5	**92.8**
100%	88.5	83.8	69.1	69.1	90.1	**95.9**

data at a given pixel, and the MRF in characterizing the class dependency among neighboring pixels in the LCM.

6. SUMMARY

We have proposed a new land cover mapping algorithm that combines the strength of a random forest with a Markov random field model. The idea is to map a remote-sensing image into a new domain such that the observation probability can be estimated without using any probabilistic models. Since the random forest classifier contains many weak decision trees, the new domain used in this paper is a set of decisions derived from all the trees. Furthermore, due to how the decision trees are trained, the joint probability of these decisions can be approximated as the multiplication of the marginal probabilities of each individual one. We examined the performance of our proposed algorithm using a real remote-sensing image and found that the performance of our algorithm is superior to the traditional random forest classifier, MRF-approach with Gaussian model, maximum likelihood classifier under Gaussian model, and decision tree.

7. REFERENCES

[1] T. Kasetkasem, M.K. Arora, and P.K. Varshney, "Super-resolution land cover mapping using a markov random field-based approach," *Remote Sens. Environ*, vol. 96, pp. 302314, 2005.

[2] G. Winkle, *Image Analysis Random Fields and Dynamic Monte Carlo Methods*, Springer-Verlag, New York, NY USA, 1995.

[3] J.R. Quinlan, *Induction of Decision Tree*, Kluwer Academic Publishers, Boston, 1986.

[4] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[5] F. Forbes and N. Peyrard, "Hidden markov random field model selection criteria based on mean field-like approximations," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 25, no. 19, pp. 1089–1101, 2003.