MULTI-SOURCE MULTI-SCALE HIERARCHICAL CONDITIONAL RANDOM FIELD MODEL FOR REMOTE SENSING IMAGE CLASSIFICATION

Z. Zhang[a], M.Y. Yang[b], M. Zhou[a]

[a] Academy of OptoElectronics, Chinese Academy of Sciences, Beijing, China
[b] Institute for Information Processing (TNT), Leibniz University Hannover, Germany
yang@tnt.uni-hannovder.de

WG III/3

KEY WORDS: Classification, Fusion, Multisensor, LIDAR, Hierarchical, Vision, Performance

ABSTRACT:

Fusion of remote sensing images and LiDAR data provides complimentary information for the remote sensing applications, such as object classification and recognition. In this paper, we propose a novel multi-source multi-scale hierarchical conditional random field (MSMSH-CRF) model to integrate features extracted from remote sensing images and LiDAR point cloud data for image classification. MSMSH-CRF model is then constructed to exploit the features, category compatibility of multi-scale images and the category consistency of multi-source data based on the regions. The output of the model represents the optimal results of the image classification. We have evaluated the precision and robustness of the proposed method on airborne data, which shows that the proposed method outperforms standard CRF method.

1. INTRODUCTION

In the fields of photogrammetry and remote sensing, there exist many sources of earth observation data with the different characteristics of targets on the ground. For a long period, integration of the multi-source data reasonably and effectively has been an active topic. Fusion of remote sensing images and LiDAR data provides complimentary information for the remote sensing applications, such as object classification and recognition.

Many methods have been developed for the fusion of remote sensing images and LiDAR data. In general those methods are classified into three categories, namely image fusion (Parmehr et al., 2012), feature fusion (Dalponte et al., 2012, Deng and Su, 2012), and decision fusion (Huang et al., 2011, Shimoni et al., 2011). The methods for image fusion include different resolution data sampling and registration, so the processing is time-consuming, and the accuracy is affected by the accuracy of registration, which reduces the performance of the subsequent image classification. In the feature fusion methods, the features are usually extracted independently from different data source, and the fusion lacks consideration of correspondence of location and contextual information, by which the classification could be improved.

In order to overcome the limitations of the aforementioned methods, we present a novel multi-source multi-scale hierarchical conditional random field (MSMSH-CRF) model to fuse features extracted from remote sensing images and LiDAR point cloud data for image classification. In this paper, the major **contribution** is that both the category compatibility of the multi-scale image in a hierarchical structure and the category consistency of multi-source data are considered in the MSMSH-CRF model. The following sections are organized as follows. The related work is discussed in Section 2.. In Section 3., the MSMSH-CRF model is presented in detail. In Section 4., experimental results are presented. Finally, this contribution of this paper is concluded and the future work is discussed in Section 5..

2. RELATED WORK

In order to make full use of multi-source data for image classification and object recognition, many feature-based fusion methods have been proposed. One of the classic tools are graphical models (Bishop, 2006), i.e. probabilistic models defined on a graph describing the conditional dependence structure between random variables. As the one branch of the graphical model, Markov Random Fields (MRFs) have been used for image interpretation since 1986 (Besag, 1986), and their limiting factor only allowing for local image features has been overcome by Conditional Random Fields (CRFs) (Lafferty et al., 2001), where arbitrary features can be used for classification. CRFs have the ability to discriminatively model contextual dependencies, conditioned on observations, for capturing global as well as local image context, which makes them suitable for accurate labeling (Perez et al., 2012). Therefore, they have been receiving more and more attention in recent years (Yang and Förstner, 2011b, Zhang et al., 2012, Niemeyer et al., 2014).

(Schindler, 2012) gives a systematic overview of image classification methods, which impose a smoothness prior on the labels. Both local filtering-type approaches and global random field models developed in other fields of image processing are reviewed. He shows a detailed experimental comparison and analysis of the methods, using two different aerial data sets from urban areas with known ground-truth. Based on the standard CRF model (Shotton et al., 2009), (Yang and Förstner, 2011a) introduce a hierarchical conditional random field to deal with the problem of image classification by modeling spatial and hierarchical structures. (Perez et al., 2012) formulate a multi-scale CRF model to deal with the problem of region labeling in multispectral remote sensing images. (Zhang et al., 2013) propose the multi-source hierarchical conditional random field (MSHCRF) model to fuse features extracted from remote sensing images and LiDAR point cloud data for image classification. Hierarchical pairwise potentials are introduced to consider category consistency of multi-source data based on regions. (Niemeyer et al., 2014) integrate a random forest classifier into a CRF framework,

which is a flexible method for obtaining a reliable 3D classification in complex urban scenes. These methods exploit both spatial and hierarchical structures of objects in images. Considering the limitation of visual feature information from the images, the classification results could be potentially improved by incorporating information from different source data, such as the elevation information in LiDAR data and the spectral information in the hyperspectral images.

3. MSMSH-CRF MODEL FOR AUTOMATIC CLASSIFICATION

In this section, we start by presenting the graphical model to integrate an image and LiDAR data, so-called MSMSH-CRF model, with corresponding energy function. Then, we describe the model construction process. Afterward, we will derive the features from each region obtained from the unsupervised segmentation algorithm. Then, we will give particular formulations for each of the unary, pairwise, hierarchical potentials respectively. Finally, we will discuss the learning and inference of this graphical model.

3.1 MSMSH-CRF model

In the field of image analysis, the regions of interest are usually detected independently, but considering the relative position between regions in single source data and the correspondence between regions from multi-source data, the labeling of every region should not be independent. The CRF model is an effective way to solve the problem of prediction of the non-independent labeling for multiple outputs, and in this model, all the features can be normalized globally to obtain the global optimal solution.

Based on the standard CRF model, we propose the MSMSH-CRF model to learn the conditional distributions over the class labeling given an image and corresponding LiDAR data, and the model allows us to incorporate different features and correspondence information in a single unified model, as illustrated in Figure 3. The conditional probability of the class labels c given an image X and LiDAR data L, which has a distribution of the Gibbs form, is defined as follows

$$P(\mathbf{c}|\mathbf{X},\mathbf{L},\theta) = \frac{1}{Z(\theta,X,L)}\exp(-E(\mathbf{c}|X,L,\theta)) \quad (1)$$

And the energy function

$$\begin{aligned} E(\mathbf{c}|\mathbf{X},\mathbf{L},\theta) = &\sum_{i\in S} E_1(c_i,x_i,\boldsymbol{\theta_1}) \\ &+ \sum_{(i,j)\in N} E_2(c_i,c_j,x_i,x_j,\theta_2) \\ &+ \sum_{(i,k)\in M} E_3(c_i,c_k,x_i,x_k,\theta_3) \\ &+ \sum_{(i,t)\in H} E_4(c_i,c_t,x_i,l_t,\boldsymbol{\theta_4}) \end{aligned} \quad (2)$$

where $\theta = \{\boldsymbol{\theta_1},\boldsymbol{\theta_2},\theta_3,\boldsymbol{\theta_4}\}$ is the vector of model parameters, $Z(\theta,\mathbf{X},\mathbf{L})$ is the partition function, i,j and k respectively index regions x_i,x_j and x_k in the image, which correspond to nodes in the graph, and t index regions l_t in the LiDAR data, which also correspond to nodes in the graph. S is the set of all the nodes in image level of the graph, N is the set of corresponding pairs collecting neighborhood in both images and LiDAR data, M is the set of pairs collecting parent-child relations between regions with neighboring scales, and H is the set of corresponding pairs collecting neighborhood in both images and LiDAR data. E_1 is the unary potentials, which represent relationships between class labels and the observed data, E_2 is the pairwise potentials, representing relationships between class labels of neighboring regions within each scale. E_3 is the multi-scale hierarchical pairwise potential, which represents corresponding relationships between regions in neighboring scales of images. E_4 is the multi-source hierarchical pairwise potential, representing corresponding relationships between images and LiDAR data.

3.2 Model construction

In order to integrate features extracted from multi-source data for image classification, the MSMSH-CRF graphical model is consist of two levels: Image level and LiDAR level. In Image level, Texton is utilized to distinguish between different regions effectively and obtain the different segmented regions, which form all the nodes in image level of the graph. Meanwhile, we can change the amount of channels of the Texton filter (Shotton et al., 2009) to get different results which are similar to the multi-scale segmentation, and Figure 1 shows the example results of our algorithm. The neighborhood in Image level is defined as the relationship of two regions which have the common edge. In LiDAR level, the mean shift algorithm is used to get the flat regions corresponding to continuous planes of different targets in LiDAR data, which form all the nodes in LiDAR level of the graph.

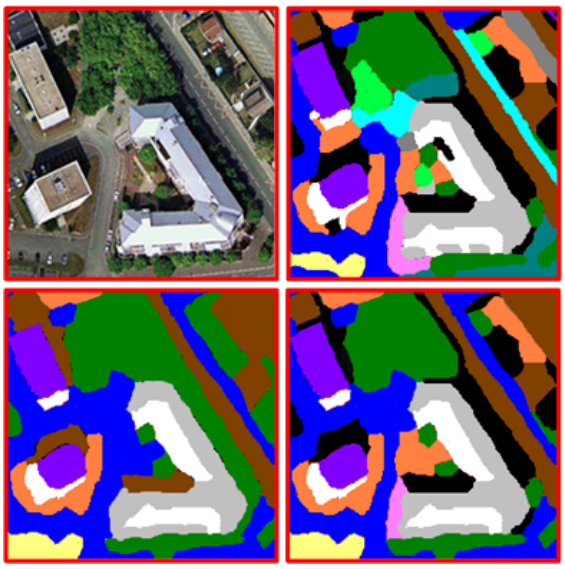

Figure 1: The example region images of Texton segmentation results at scale 1, 2, 3 respectively. The color of each region is assigned randomly that neighboring regions are likely to have different colors. *Top row left*: Original image, *Top row right*: segmentation result at scale 3. *Bottom row left*: segmentation result at scale 1, *Bottom row right*: segmentation result at scale 2.

For describing the consistency of multi-source data, we firstly choose the optimal scale of images to match with the LiDAR data. Assuming that there is a registration of multi-source data acquired on the same airborne platform, such as the algorithm introduced in literature (Mastin et al., 2009), and we calculate the center of each region (or line)RL_i in the depth image converted from LiDAR data, and the center should be inside the region (or line) and at the symmetric axis. Then based on the relative position of the centers, the corresponding regions (or lines) RL_{ia} in multi-scale images can be selected. The procedure of choosing optical scale images is illustrated in Figure 2. Therefore, for each pixel s in the region (or line) RL_i, we obtain the optimal scale of images

by

$$a^* = \arg\min_{a} \sum_{s \in \{RI_i \cup RI_{ia}\}} | RI_i(s) - RI_{ia}(s) | \quad (3)$$

and

$$RL_i(s) = \begin{cases} 1, & s \in RL_i, \\ 0, & s \notin RL_i \end{cases}, RL_{ia}(S) = \begin{cases} 1, & s \in RL_{ia}, \\ 0, & s \notin RL_{ia} \end{cases} \quad (4)$$

where i index the sequence number of all regions (or lines) in the depth image converted from the Mean Shift Feature (MSF) or Alpha Shape Feature (ASF) of LiDAR data.

Therefore, the MSMSH-CRF graphical model is constructed as follows, illustrated in Figure 3. Firstly, typical features are derived from the interest regions in multi-source data, where the regions are generated by an unsupervised segmentation algorithm. In the graphical model, the nodes correspond to regions. The blue edges represent the dependencies between neighboring regions, and the orange edges indicate the hierarchical relations between regions at different scales in a multi-scale segmentation. Purple edges indicate the hierarchical relations between regions from multi-source data, where the optimal scale of images is selected to match the LiDAR data. The MSMSH-CRF model is constructed to exploit the features and category compatibility of multi-scale images as well as the category consistency of multi-source data based on regions. The output of the model represents the optimal results of the image classification.

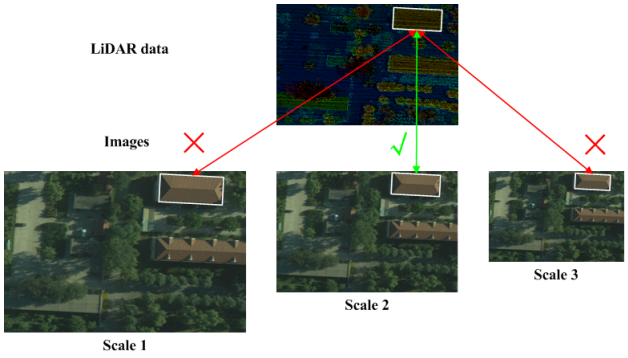

Figure 2: The example image of illustrating the procedure of choosing the optimal scale image to match the LiDAR data.

3.3 Features

Four types of features are extracted, namely the line features (LF), the texture features (TF), the mean shift features (MSF), and alpha shape features (ASF). The line features (LF) and the texture features (TF) are extracted from remote sensing images, whereas the mean shift features (MSF) and alpha shape features (ASF) are from LiDAR data.

Line Features (LF) Shape features, in particular line features, not only describe the structures of targets directly, but also are stable to light change, color change, etc. As a new and effective one of line features, the LSD (Line Segment Detector) (Grompone and Randall, 2010) can be used to give accurate results extracted, a controlled number of false detections, and requires no parameter tuning. In the method, the level-line orientation is defined and calculated by gradient magnitude, and then the pixels with the same level-line orientation are merged to cover the so-called line support regions, in which all the pixels are regarded as a long

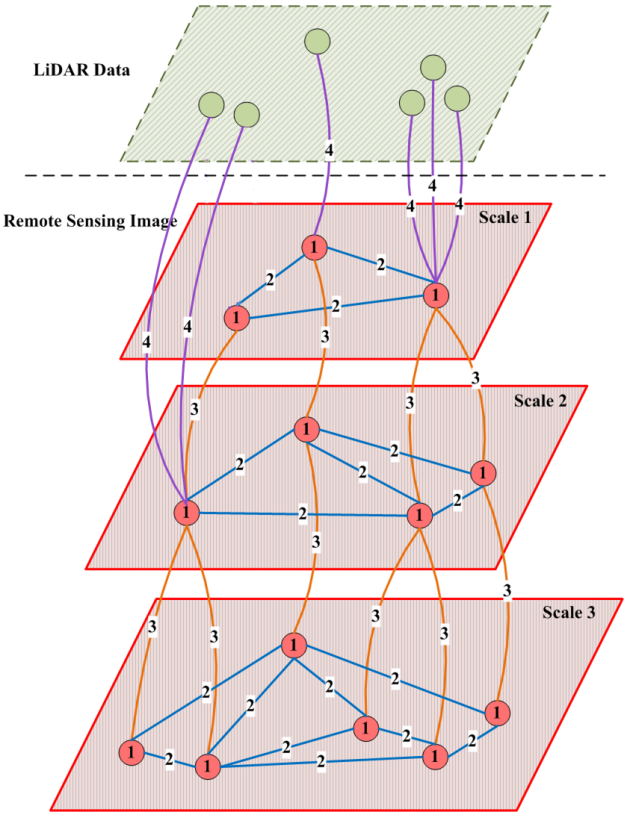

Figure 3: Illustration of the MSMSH-CRF model architecture. In Image level, red nodes (# 1) correspond to image regions, blue edges (# 2) linking red nodes represent the dependency between neighboring regions, and orange edges (# 3) linking red nodes in multi-scale indicate the hierarchical relation between regions at different scales corresponding to the multi-scale segmentation. In LiDAR level, green nodes represent the extracted regions. Purple edges (# 4) linking red and green nodes indicate the hierarchical relation between regions from multi-source data, where the optimal scale of images is selected to match the LiDAR data.

continuous segment. In accordance with the method introduced in (Grompone and Randall, 2010), we can calculate the response value of LSD at each pixel, denoted by $LF(s)$.

Texture Features (TF) Texture is one of the basic properties of objects, as well as the most direct and reliable way of characterization. The basic unit of texture is often referred to as Texton, and we can represent the texture most directly by describing the distribution of the components, namely Texton. In the process of textonization, images are convolved with a 17-dimensional filterbank. The 17D responses for all training pixels are then whitened (to give zero mean and unit covariance), and an unsupervised clustering is performed by the Euclidean-distance K-means clustering algorithm. Finally, each pixel in each image is assigned to the nearest cluster center, producing the texton map. Similar to the method in (Shotton et al., 2009), we can obtain the value of Texton classifier of each pixel in the image, denoted by $TF(s)$.

Mean Shift Features (MSF) The mean shift method (Comaniciu and Meer, 2002) is a robust clustering technique which does not require prior knowledge of the number of clusters, and does not constrain the shape of the clusters. The number of clusters is obtained automatically by finding the centers of the densest regions in the space, so this method is widely used for clustering of discrete points. In our model, the specific process of achieving

the MSF is introduced in (Georgescu et al., 2003), all the LiDAR points are clustered in different regions, and the elevation of all points in one region are assigned as the same value which is the mean of all the ones.

Alpha Shape Features (ASF) There are many methods for extracting the boundary of LiDAR data. Compared with other algorithms (Berger, 2012, Kong et al., 2012), Alpha Shapes algorithm works effectively in inner and outer boundaries extraction from LiDAR data with convex and concave polygon shape. Moreover, it can keep fine features of buildings adaptively and filter the footprints of non-building. Based on the MSF regions obtained, the alpha shape algorithm is used to extract the boundary contour of each region, and then the Delaunay triangulation is used to get the line feature. The extraction of the ASF refers to (Shen et al., 2011), similar to the MSF, all the points in one lines have the same elevation which is the mean of all the ones.

3.4 Unary potentials

The unary potentials consist of two element: LF and TF potentials, predict the label c_i of the region x_i based on the image $\mathbf{X}$

$$E_1(c_i, x_i, \theta_{\mathbf{1}}) = LF(c_i, x_i, \theta_{LF}) + TF(c_i, x_i, \theta_{TF}) \quad (5)$$

where $LF(c_i, x_i, \theta_{LF})$ is the LF potential and $TF(c_i, x_i, \theta_{TF})$ is the TF potential, and $\theta_{\mathbf{1}} = \{\theta_{LF}, \theta_{TF}\}$ is the vector of model parameters.

LF Potentials The LF potentials capture the (relatively weak) dependence of the class label and the boundaries of targets on the response value of LSD and absolute location of the pixel in the image. We can get the line segment image $LFI(s)$ by calculating the response value of LSD LF_s of each pixel s in the region x_i. The LF potentials take the form of a look-up table with an entry for each class ci and value of LSD LF_s and pixel location s

$$LF(c_i, x_i; \theta_{LF}) = -\log \sum_{s \in x_i} \theta_{LF}(c_i, LF_s, s) \quad (6)$$

where the parameter θ_{LF} represents the relationship among the value of each pixel LF_s, namely $LFI(s)$, the pixel location s and the label c_i.

TF Potentials Based on the Joint Boost algorithm, an adapted version of boosting learning algorithm, we can obtain the classifier of Texton, to which the responses are used directly as a potential in the MSMSH-CRF model, so that

$$TF(c_i, x_i; \theta_{TF}) = -\log \sum_{s \in x_i} P(c_i | TF_s) \quad (7)$$

where TF_s corresponds to the response of classifier at each pixel s, and $P(c_i|TF_s)$ is the normalized distribution given by the classifier using the learned parameters θ_{TF}.

3.5 Pairwise potentials

The pairwise potentials describe category compatibility between neighboring regions x_i and x_j obtained from the line segment image $LFI(s)$, and the responses of Texton classifier on the image $\mathbf{X}$.

$$\begin{aligned} E_2(c_i, c_j, x_i, x_j, \theta_{\mathbf{2}}) = {} & PLF(c_i, c_j, x_i, x_j, \theta_{PLF}) \\ & + PTF(c_i, c_j, x_i, x_j, \theta_{PTF}) \end{aligned} \quad (8)$$

where $PLF(c_i, c_j, x_i, x_j, \theta_{PLF})$ is the pairwise potentials of LF and $PTF(c_i, c_j, x_i, x_j, \theta_{PTF})$ is the pairwise potentials of TF, $\theta_{\mathbf{2}} = \{\theta_{PLF}, \theta_{PTF}\}$ is the vector of model parameters.

Pairwise Potentials of LF Based on the line segment image $LFI(s)$, we can calculate the pairwise potentials of LF as the form of the contrast-sensitive Potts model (Boykov and Jolly, 2001)

$$\begin{aligned} & PLF(c_i, c_j, x_i, x_j, \theta_{PLF}) = \\ & \theta_{PLF} \frac{1 + 6\exp(-2l(x_i, x_j))}{N_i + N_j} \sigma(c_i \neq c_j) \end{aligned} \quad (9)$$

where θ_{PLF} is the weight factor, $l(x_i, x_j)$ is the Euclidean metric of the pixel value between regions x_i and x_j in the LF images, N_i is the number of regions neighbored to region i, N_j is the number of regions neighbored to j, and $\sigma(\cdot)$ is a 0-1 indicator function, and the number 6 in Eq. (9) is set empirically. The pairwise potentials $PLF(c_i, c_j, x_i, x_j, \theta_{PLF})$ are scaled by N_i and N_j to compensate for the irregularity of the graph.

Pairwise Potentials of TF Similar to the pairwise potentials of LF, the pairwise potentials of TF take the form of the contrast-sensitive Potts model:

$$\begin{aligned} & PTF(c_i, c_j, x_i, x_j, \theta_{PTF}) = \\ & \theta_{PTF} \frac{1 + 4\exp(-2l(x_i, x_j))}{N_i + N_j} \sigma(c_i \neq c_j) \end{aligned} \quad (10)$$

where θ_{PTF} is the weight factor, $t(x_i, x_j)$ is the Euclidean metric of the value of Texton classifier at each pixel between regions x_i and x_j in the results of marked images, and the number 4 in Eq. (10) is set empirically. The pairwise potentials PTF are scaled by N_i and N_j to compensate for the irregularity of the graph.

3.6 Multi-scale hierarchical pairwise potentials

From the pairwise potentials in Section 3.5, there is a lack of longer range contextual relationship in the graphical modeling. To overcome those local restrictions, we analyze the image at multiple scales to enhance the model by evidence aggregation on a local to global level. Furthermore, we integrate multi-scale pairwise potentials to regard the hierarchical structure of the regions.

Based on results of multi-scale segmentation, the multi-scale hierarchical pairwise potentials describe category compatibility between hierarchically neighboring labels c_i and c_k given the image $\mathbf{X}$, which take the form of the contrast-sensitive Potts model:

$$\begin{aligned} & E_3(c_i, c_k, x_i, x_k, \theta_3) = \\ & \theta_3 \cdot [1 + 4\exp(-2m(x_i, x_j))]\, \sigma(c_i \neq c_k) \end{aligned} \quad (11)$$

where θ_3 is the weight factor, $m(x_i, x_j)$ is the Euclidean metric of the value of Texton classifier between regions x_i and x_j in the results of marked images, and the number 4 in Eq. (11) is set empirically. Multi-scale hierarchical pairwise potentials act as a link across scale, facilitating propagation of information in the model.

3.7 Multi-source hierarchical pairwise potentials

Compared to the remote sensing images, LiDAR data is sparse. The features extracted from multi-source data are different. In order to enhance the fusion performance, we introduce the hierarchical pairwise potentials, which represent correspondences between the data from different source in our MSMSH-CRF model. The hierarchical pairwise potentials describe category consistency between the corresponding regions in multi-source data, from which we can obtain the TF and MSF, which are named as planar features, and the LF and ASF, which are named as linear features. In order to enhance the fusion performance, we refer to the category consistency with the planar and linear features separately,

denoted as $HPP(c_i, c_t, x_i, l_t, \theta_p)$ and $HPL(c_i, c_t, x_i, l_t, \theta_l)$ respectively. So there is

$$E_4(c_i, c_t, x_i, l_t, \theta_4) = HPP(c_i, c_t, x_i, l_t, \theta_p) + HPL(c_i, c_t, x_i, l_t, \theta_l) \quad (12)$$

where $\theta_4 = \{\theta_p, \theta_l\}$ is the vector of model parameters.

Hierarchical pairwise potentials of planar features Based on the TF results of the optimal scale image, we firstly normalize the value $TF_s(x_i)$ of Texton classifier of each pixel s in the region x_i to get $NTF_s(x_i)$:

$$NTF_s(x_i) = TF_s(x_i)/TF_{max} \quad (13)$$

where TF_{max} is the maximum value of Texton classifier of each pixel in the image.

In the MSF results of LiDAR data, elevations of different regions are obtained, and the normalized elevation $NMSF(l_t)$ of all points in the regions l_t extracted is calculated:

$$NMSF(l_t) = MSF(l_t)/MSF_{max} \quad (14)$$

where $MSF(l_t)$ is the elevation of all points in the region l_t, and MSF_{max} is the maximum elevation of all flat regions in the LiDAR data.

So based on the normalized value $NTF_s(x_i)$ and $NMSF(l_t)$, the hierarchical pairwise potentials of planar features is defined by

$$HPP(c_i, c_t, x_i, l_t, \theta_p) = \theta_p \sum_{s \in x_i} \exp(-\epsilon_p |NTF_s(x_i) - NMSF(l_t)|^2) \sigma(c_i \neq c_t) \quad (15)$$

where $\epsilon_p = (2 < |NTF_s(x_i) - NMSF(l_t)|^2 >)^{-1}$ is the comparative item, $< \cdot >$ is the averaging operator, and θ_p is the weight.

Hierarchical pairwise potentials of linear features The hierarchical pairwise potentials of linear features take the form as

$$HPL(c_i, c_t, x_i, l_t, \theta_t) = \theta_t \sum_{s \in x_i} \exp(-\epsilon_t |NLF_s(x_i) - NASF(l_t)|^2) \sigma(c_i \neq c_t) \quad (16)$$

where $\epsilon_t = (< 2|NLF_s(x_i) - NASF(l_t)|^2 >)^{-1}$ is the comparative item, and θ_l is the weight. $NLF_s(x_i)$ is the normalized value from the LF results of the optimal scale image, and $NASF(l_t)$ is the normalized value from the ASF of LiDAR data.

3.8 Parameter Learning

In this paper, piecewise training method (Sutton and McCallum, 2005) is adopted for the learning of the parameters of MSMSH-CRF model. This method divides the MSMSH-CRF model into pieces corresponding to the different terms in Eq. (2). Each of these pieces is then trained independently, as if it were the only term in the model.

Parameters of LF Potentials The formula for calculating the parameters of LF Potentials respectively for each image is defined as

$$\theta_{LF}(c_i, LF_s, s) = 1 - \mid (\mid \sigma(c_i) - \sum_{s \in x_i} \sigma(LF_s) / \sum_{s \in x_i} 1 \mid) - w_{LF} \mid \quad (17)$$

where the small positive integer w_{LF} is set to 0.1 in practice.

Parameters of TF Potentials The learning of parameters of TF Potentials is based on Joint Boost algorithm, and an excellent detailed treatment of the learning process is given in literature (Shotton et al., 2009), but we briefly describe it here for completeness. Each training example s (a pixel in a training image) is paired with a target value $Z_s^c \in \{-1, +1\}$ ($+1$ if the example s has ground truth class c, -1 otherwise) and assigned a weight ω_s^c specifying its classification accuracy for class c after iteration of boosting. Each round of iteration chooses a new weak learner by minimizing an error function incorporating the weights. The training examples are then re-weighted ω_s^c to reflect the new classification accuracy. This procedure emphasizes poorly classified examples in subsequent rounds of iteration, and ensures that over many rounds, the classification for each training example approaches the target value and the parameters are optimal.

Parameters of other potentials The parameters of other potentials of MSMSH-CRF model,$\theta_{PLF}, \theta_{PTF}, \theta_3, \theta_p$ and θ_l, are selected manually such that the classification error is minimized on the training set.

3.9 Model Inference

Given a set of parameters learned for the MSMSH-CRF model, the optimal labeling $\mathbf{c}^*$, which minimizes the energy function in Eq. (2), is found by applying the alpha-expansion graph-cut algorithm (Boykov et al., 2001, Boykov and Jolly, 2001).

4. EXPERIMENTS

In this section, experiments are performed on the Beijing Airborne Data (Zhang et al., 2013), to evaluate the performance of the proposed method.

4.1 Dataset

We conduct experiments to evaluate the performance of the MSMSH-CRF model on the Beijing Airborne Data (Zhang et al., 2013), which include remote sensing images with a resolution of 0.12m and LiDAR data with a point density of 4 points/m^2, as illustrated in Figure 4. The objects in all images correspond to one of three classes: Building, Road and Vegetation. These classes are typical objects appearing in airborne images. In the experiments, we take the ground-truth label of a region to be the majority vote of the ground-truth pixel labels, and randomly divide the images into a training set with 50 images and a testing set with 50 images.

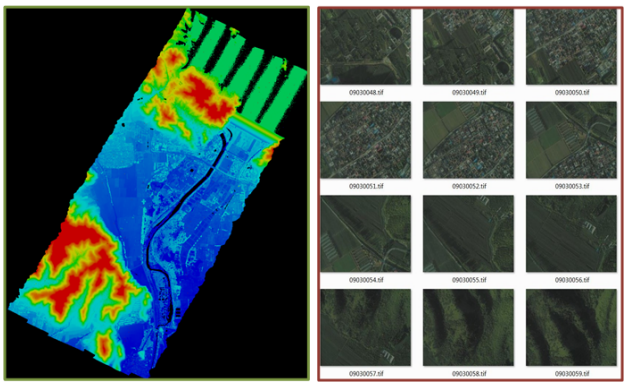

Figure 4: The example images of the Beijing Airborne Data. *Left*: LiDAR data, *Right*: remote sensing images of the surveying area.

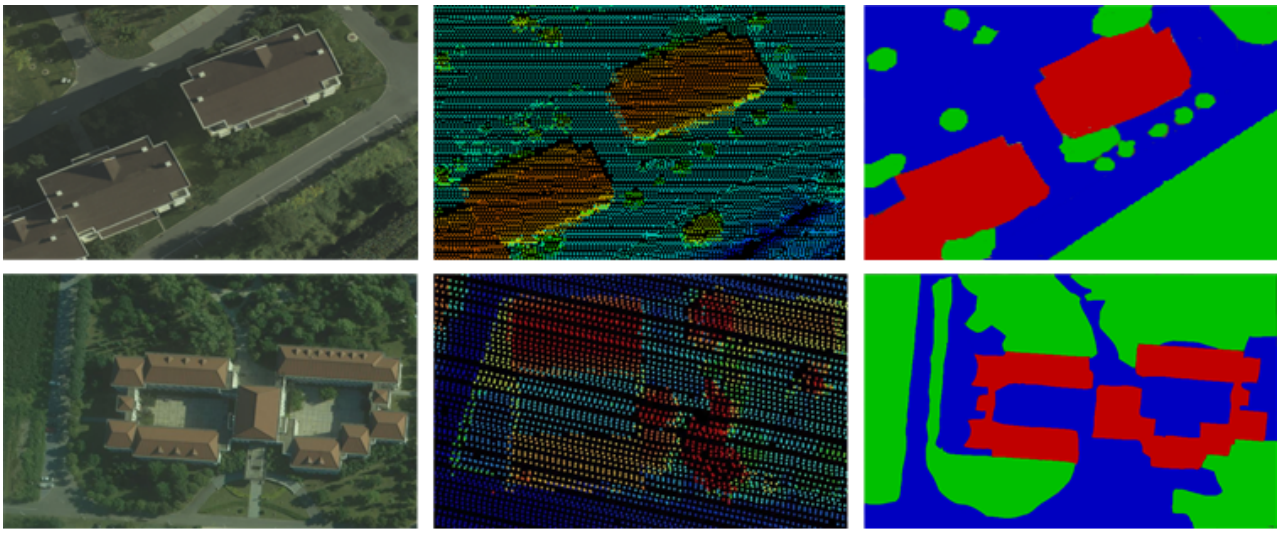

Figure 5: The classification result from the MSMSH-CRF model on the Beijing Airborne Data. *Left*: remote sensing image, *Middle*: LiDAR point cloud, *Right*: classification result (red - building, blue - road, green - vegetation).

Method	Accuracy (%)
(Shotton et al., 2009)	64.2
(Zhang et al., 2013)	73.6
Ours	83.7

Table 1: Average pixelwise accuracy of three methods on the Beijing Airborne Data.

4.2 Results

Figure 5 shows the example results of MSMSH-CRF classification method. The average pixelwise accuracy on the testing set is given in Table 1. The average classification accuracy of our method is 83.7%, which has 10.1% gain w.r.t. the accuracy of the MSHCRF model (Zhang et al., 2013) and 19.5% gain w.r.t. the accuracy of the standard CRF model (Shotton et al., 2009). The parameter, learned by cross validation on the training set, are $\theta_{PLF} = 0.22, \theta_{PTF} = 0.18, \theta_3 = 0.15, \theta_p = 0.2$, and $\theta_l = 0.25$. For the fairness of comparison, both the training set and the testing set are same for MSMSH-CRF, MSHCRF and standard CRF respectively. Figure 6 shows the classification accuracy with different parameters, with only one parameter is changing while the others are fixed.

Table 2 shows the confusion matrix obtained by applying standard MSMSH-CRF model to the whole test dataset. Accuracy values in the table are computed as the percentage of image pixels assigned to the correct class label, ignoring pixels labeled as void in the ground truth. Compared to the confusion matrices of standard CRF model and MSHCRF model in Table 3 and Table 4 respectively, the MSMSH-CRF model yields significant improvement on all three classes for integrating multi-scale hierarchical information of the regions in the images. Table 5 shows the performance comparison when dropping one types of potentials in the MSMSH-CRF model.

5. CONCLUSIONS

In conclusion, this paper presents a novel multi-source multi-scale hierarchical conditional random field model for automatic classification of remote sensing images. The main contributions of this work are summarized as follows: a novel CRF-based modeling scheme exploiting the complementarity of multi-source data

	building	road	vegetation
building	78.3	11.9	9.8
road	9.5	85.9	4.6
vegetation	9.7	8.7	81.6

Table 2: Pixelwise accuracy of the MSMSH-CRF classification on the Beijing Airborne Data. The confusion matrix shows classification accuracy for each class (rows) and is row-normalized to sum to 100%. Row labels indicate the true class, and column labels indicate the predicted class.

	building	road	vegetation
building	63.7	19.2	17.1
road	22.4	67.0	10.6
vegetation	11.3	15.2	73.5

Table 3: The confusion matrix: pixelwise accuracy of the standard CRF classification on the Beijing Airborne Data.

	building	road	vegetation
building	70.1	15.8	14.1
road	14.4	77.3	8.3
vegetation	12.3	13.8	73.9

Table 4: The confusion matrix: pixelwise accuracy of the MSHCRF classification on the Beijing Airborne Data.

Potentials	Accuracy (%)
With all potentials	83.7
Removing the pairwise potentials	63.9
Removing the multi-scale hierarchical pairwise potentials	73.6
Removing the Multi-source hierarchical pairwise potentials	70.1

Table 5: The performance comparison when dropping one types of potentials in the MSMSH-CRF model.

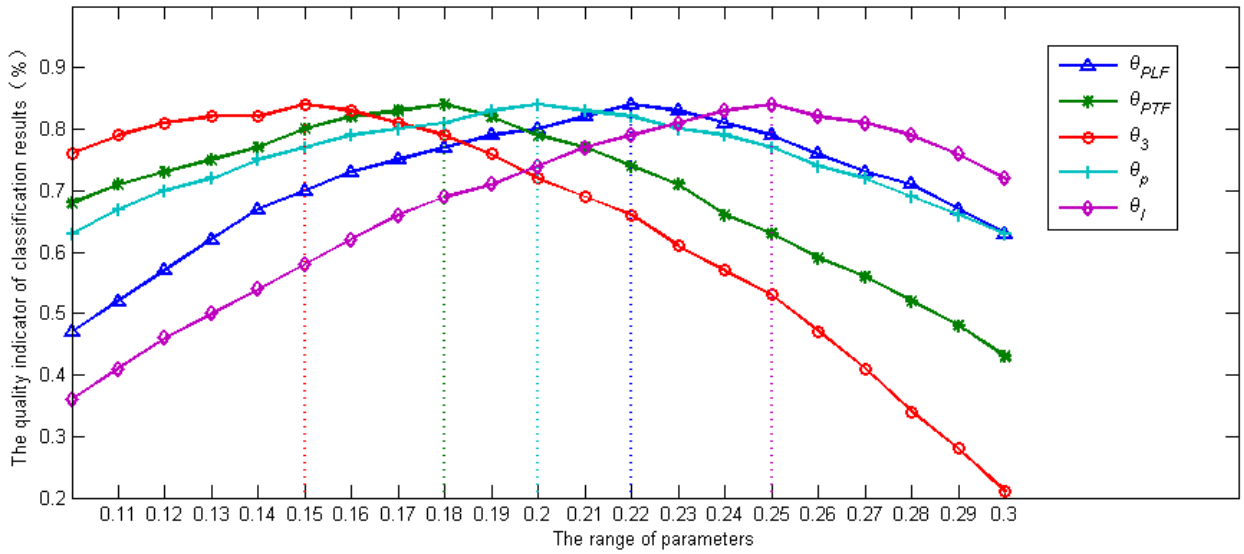

Figure 6: The classification accuracy with different parameters, with only one parameter is changing while the others are fixed.

such as the texture in remote sensing images and the elevation in LiDAR data. To exploit different levels of contextual information in images, the multi-scale hierarchical potentials are proposed in our model, which is then enhanced by evidence aggregation from a local to global level. Considering the interrelation of the same objects in remote sensing images and LiDAR data, multi-source hierarchical potentials are proposed in our model to make full use of the category consistency of multi-source data. We have evaluated the precision and robustness of the proposed approach on airborne data, which shows that the proposed method outperforms standard CRF method. However, feature extraction is crucial to the final classification accuracy. Feature selection is done in an ad-hoc fashion in the current stage. In our future work, we are interested in automatic feature selection that may further improve the classification performance.

ACKNOWLEDGEMENTS

The work is partially funded by National Natural Science Fund of China (Grant No. 40901177). The authors gratefully acknowledge these supports. The authors thank Dr. Franz Rottensteiner for his valuable assistance.

REFERENCES

Berger, C., 2012. Toward rich geometric map for slam: Online detection of planes in 2d lidar. In: Proceedings of the International Workshop on Perception for Mobile Robots Autonomy (PEMRA).

Besag, J., 1986. On the statistical analysis of dirty pictures. Journal of the Royal Statistical Society. Series B (Methodological) pp. 259–302.

Bishop, C., 2006. Pattern recognition and machine learning. Springer.

Boykov, Y., Veksler, O. and Zabih, R., 2001. Fast approximate energy minimization via graph cuts. IEEE Transactions on Pattern Analysis and Machine Intelligence 23, pp. 1222–1239.

Boykov, Y. Y. and Jolly, M.-P., 2001. Interactive graph cuts for optimal boundary & region segmentation of objects in nd images. In: IEEE International Conference on Computer Vision (ICCV), pp. 105–112.

Comaniciu, D. and Meer, P., 2002. Mean shift: A robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(5), pp. 603–619.

Dalponte, M., Bruzzone, L. and Gianelle, D., 2012. Tree species classification in the southern alps based on the fusion of very high geometrical resolution multispectral / hyperspectral images and lidar data. Remote Sensing of Environment 123, pp. 258–270.

Deng, F.and Li, S. and Su, G., 2012. Classification of remote sensing optical and lidar data using extended attribute profiles. IEEE Journal of Selected Topics in Signal Processing 6(7), pp. 856–865.

Georgescu, B., Shimshoni, I. and Meer, P., 2003. Mean shift based clustering in high dimensions: A texture classification example. In: IEEE International Conference on Computer Vision (ICCV), pp. 456–463.

Grompone, R., J. J. M. J. and Randall, G., 2010. Lsd: A fast line segment detector with a false detection control. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(4), pp. 722–732.

Huang, X., Zhang, L. and Gong, W., 2011. Information fusion of aerial images and lidar data in urban areas: vector-stacking, re-classification and post-processing approaches. International Journal of Remote Sensing 32(1), pp. 69–84.

Kong, D., Xu, L., Li, X. and Xing, W., 2012. Estimation of cluster centers on building roof from lidar footprints. In: IEEE International Conference on Imaging Systems and Techniques, pp. 254–258.

Lafferty, J., McCallum, A. and Pereira, F., 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: International Conference on Machine Learning (ICML), pp. 282–289.

Mastin, A., Kepner, J. and Fisher, J., 2009. Automatic registration of lidar and optical images of urban scenes. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2639–2646.

Niemeyer, J., Rottensteiner, F. and Soergel, U., 2014. Contextual classification of lidar data and building object detection in urban areas. ISPRS Journal of Photogrammetry and Remote Sensing 87, pp. 152–165.

Parmehr, E. G., Zhang, C. and Fraser, C. S., 2012. Automatic registration of multi-source data using mutual information. In: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, ISPRS Congress, pp. 301–308.

Perez, M., Chan, J. and Sahli, H., 2012. Multiscale conditional random fields for supervised region based labeling and classification. In: IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 1769–1772,.

Schindler, K., 2012. An overview and comparison of smooth labeling methods for land-cover classification. IEEE Transactions on Geoscience and Remote Sensing 50(11), pp. 4534–4545.

Shen, W., Zhang, J. and Yuan, F., 2011. A new algorithm of building boundary extraction based on lidar data. In: IEEE International Conference on Geoinformatics, pp. 1–4.

Shimoni, M., Tolt, G., Perneel, C. and Ahlberg, J., 2011. Detection of vehicles in shadow areas using combined hyperspectral and lidar data. In: IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 4427–4430.

Shotton, J., Winn, J., Rother, C. and Criminisi, A., 2009. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. International Journal of Computer Vision 81(1), pp. 2–23.

Sutton, C. and McCallum, A., 2005. Piecewise training for undirected models. In: Uncertainty in Artificial Intelligence (UAI), pp. 568–575.

Yang, M. Y. and Förstner, W., 2011a. A hierarchical conditional random field model for labeling and classifying images of man-made scenes. In: IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 196–203.

Yang, M. Y. and Förstner, W., 2011b. Regionwise classification of building facade images. In: Photogrammetric Image Analysis, Springer, pp. 209–220.

Zhang, Z., Yang, M. Y. and Zhou, M., 2013. Multi-source hierarchical conditional random field model for feature fusion of remote sensing images and lidar data. In: International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences; ISPRS Hannover Workshop, pp. 389–392.

Zhang, Z., Zhou, M., Tang, L. and Li, C., 2012. Automatic detection and mapping of urban buildings in high resolution remote sensing images. In: IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 5721–5724.