SEMANTIC SEGMENTATION OF FOREST STANDS OF PURE SPECIES AS A GLOBAL OPTIMIZATION PROBLEM

Clément Dechesne, Clément Mallet, Arnaud Le Bris, Valérie Gouet-Brunet

Univ. Paris-Est, LASTIG MATIS, IGN, ENSG, F-94160 Saint-Mande, France
firstname.lastname@ign.fr

Commission II, WG II/6

KEY WORDS: Forest stands, classification, regularization, semantic segmentation, graph-cut, graphical models, MRF, CRF.

ABSTRACT:

Forest stand delineation is a fundamental task for forest management purposes, that is still mainly manually performed through visual inspection of geospatial (very) high spatial resolution images. Stand detection has been barely addressed in the literature which has mainly focused, in forested environments, on individual tree extraction and tree species classification. From a methodological point of view, stand detection can be considered as a semantic segmentation problem. It offers two advantages. First, one can retrieve the dominant tree species per segment. Secondly, one can benefit from existing low-level tree species label maps from the literature as a basis for high-level object extraction. Thus, the semantic segmentation issue becomes a regularization issue in a weakly structured environment and can be formulated in an energetical framework. This papers aims at investigating which regularization strategies of the literature are the most adapted to delineate and classify forest stands of pure species. Both airborne lidar point clouds and multispectral very high spatial resolution images are integrated for that purpose. The local methods (such as filtering and probabilistic relaxation) are not adapted for such problem since the increase of the classification accuracy is below 5%. The global methods, based on an energy model, tend to be more efficient with an accuracy gain up to 15%. The segmentation results using such models have an accuracy ranging from 96% to 99%.

1. INTRODUCTION

The analysis of forested areas from a remote sensing point of view can be performed at three different levels: pixel, object (mainly trees) or stand. When a joint mapping and statistical reasoning is required (e.g., land-cover (LC) mapping and forest inventory), forest stands remain the prevailing scale of analysis (Means et al., 2000, White et al., 2016). A stand can be defined in many different ways in terms of homogeneity: tree specie, age, height, maturity, and its definition varies according to the countries. Most of the time in national forest inventories, for reliability purposes, each area is manually interpreted by human operators using very high resolution (VHR) geospatial images with a near infra-red channel (Kangas and Maltamo, 2006).
Among the large body of available remote sensing data today, airborne laser scanning (ALS) and Very High spatial Resolution (VHR) hyper/multispectral images are both well adapted and complementary inputs for stand segmentation (Dalponte et al., 2012, Dalponte et al., 2015, Lee et al., 2016). ALS provides a direct access to the vertical distribution of the trees and to the ground underneath. Hyperspectral and multispectral images are particularly relevant for tree species classification: spectral and textural information from VHR images can allow a fine discrimination of many species, respectively. Multispectral images are often preferred due to their higher availability, and higher spatial resolution. One should note that the literature remains focused on individual tree extraction and tree species classification, developing site-specific workflows with similar advantages, drawbacks and classification performance. Consequently, no operational framework embedding the automatic analysis of remote sensing data has been yet proposed in the literature for forest stand segmentation (Dechesne et al., 2017). More surprisingly, only few methods have addressed such an issue from a research perspective. More authors have focused on forest delineation (Eysn et al., 2012), that do not convey information about the tree species and their spatial distribution.

The analysis of the lidar and multispectral data is performed at three levels in (Tiede et al., 2004), following the hierarchy of the nomenclature of forest LC species database. The multi-scale analysis offers the advantage of alleviating the standard limitations of individual tree crown detection, and of retrieving labels related here to forest development stage. Nevertheless, the pipeline is highly heuristic and under-exploits lidar data. Besides, significant confusions between classes are reported.
The automatic segmentation of forests in (Diedershagen et al., 2004) is also performed with lidar and VHR multispectral images. The idea is to divide the forests into higher and lower sections according to the height provided by the lidar sensor. An unsupervised classification process is applied and pre-defined thresholds enable to obtain the desired delineation of stands. This method is efficient if the canopy structure is homogeneous and requires a strong knowledge on the area of interest. Based on height information only, it cannot differentiate two stands of similar height but different species.
In (Leppänen et al., 2008), a stand segmentation technique for a forest composed of *Scots Pine*, *Norway Spruce* and *Hardwood* is defined. A hierarchical segmentation on the Crown Height Model followed by a region growing procedure is performed on images composed of rasterized lidar data and Colored Infra-Red images. The process was only applied on a limited area of Finland, preventing from drawing broad conclusions. However, the quantitative analysis enhances again that lidar data can help to define statistically meaningful stands and that multispectral images are inevitable inputs for tree species discrimination.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
doi:10.5194/isprs-annals-IV-1-W1-141-2017

Eventually, in (Dechesne et al., 2017), forest stand segmentation is considered from semantic segmentation point of view. Forest areas are first classified according to the tree species at the pixel and tree levels using lidar and multispectral airborne images. Then, the label map is smoothed using an energetic framework that integrates both lidar and optical features.

In this paper, we specifically focus on semantic segmentation of forest stands through the regularization/smoothing process of an existing label map of pure species, following the strategy proposed in (Dechesne et al., 2017). Therefore, we build upon the vast amount of literature dealing with tree species classification at the tree level and investigate how the combined use of airborne lidar and VHR multispectral images can provide more accurate label maps. Simple smoothing methods are first investigated as well as more complex energy formulations. We aim to determine whether a complex formulation of the problem helps to obtain better results in such non-structured environments.

2. RELATED WORKS

2.1 How to smooth a label map?

Pixel-wise classification is not sufficient for both accurate and smooth land-cover mapping with VHR remote sensing data. This is particularly true in forested areas: the large intra-class and low inter-class variabilities of classes result in noisy label maps at pixel or tree levels. This is why various regularization solutions can be adopted from the literature (from simple smoothing to probabilistic graphical models, see Section 2.1).

According to (Schindler, 2012), both local and global methods can provide a regularization framework, with their own advantages and drawbacks. In local methods, the neighborhood of each element is analyzed by a filtering technique. The labels of the neighboring pixels (or the posterior class probabilities) are combined so as to derive a new label for the central pixel. Majority voting, Gaussian and bilateral filtering can be employed if it is not targeted to smooth class edges.

Global methods consider the full area of interest at the same time. They are based on Markov Random Fields (MRF), the labels at different locations are not considered to be independent. The optimal configuration of labels is retrieved when finding the Maximum A Posteriori over the entire field (Moser et al., 2013). The problem is therefore considered as the minimization procedure of an energy E over the full image I. Despite a simple neighborhood encoding (pairwise relations are often preferred), the optimization procedure propagates over large distances. Depending on the formulation of the energy, the global minimum may be reachable. However, a large range of optimization techniques allow to reach local minima close to the real solution, in particular for random fields with pairwise terms (Kolmogorov and Zabih, 2004). For genuine structured predictions, in the family of graphical probabilistic models, Conditional Random Fields (CRF) have been massively adopted during the last decade. Interactions between neighboring objects, and subsequently the local context can be modelled and learned. In particular, Discriminative Random Fields (DRF, (Kumar and Hebert, 2006)) are CRF defined over 2D regular grids, and both unary/association and binary/interaction potentials are based on labelling procedure outputs. Many techniques extending this concept or focusing on the learning or inference steps have been proposed in the literature (Kohli et al., 2009, Ladický et al., 2012). A very recent trend even consists in jointly considering CRF and deep-learning techniques for the labelling task (Kirillov et al., 2015).

In standard LC classification tasks, global methods are known to provide significantly more accurate results (Schindler, 2012) since contextual knowledge is integrated. This is all the more true for VHR remote sensing data, especially in case of a large number of classes (e.g., 10, (Albert et al., 2016)), but presents two disadvantages. For large datasets, their learning and inference steps are expensive to compute. Furthermore, parameters should often be carefully chosen for optimal performance, and authors that managed to alleviate the latter problem still report a significant computation cost (Lucchi et al., 2011).

2.2 Semantic segmentation is a suitable solution

The classification process can be eased with segmentation techniques. Such algorithms provide strong local spatial supports (namely superpixels), sometimes at various scales (Lucchi et al., 2011). This is the so-called Object-Based Image Analysis (OBIA) framework. A pure bottom-up process is however not sufficient in our case. Alternatively, it can be achieved with more sophisticated top-down processes, e.g., based on pattern recognition methods but, emphasis is then put on localization of the objects of interest instead of sharp boundary retrieval (i.e., the reverse advantage of per-pixel classification). The best of both worlds is obtained with semantic segmentation, which aims to solve the interleaved issue of classification and segmentation by combining top-down and bottom-up cues. It defines the task of partitioning an image into regions that delineate meaningful objects and labelling those regions with an object label. While it is very popular in computer vision (Ladický et al., 2010, Boix et al., 2011, Arbeláez et al., 2012, Chen et al., 2015), it has been barely addressed in the remote sensing community (Montoya-Zegarra et al., 2015, Zheng and Wang, 2015). Segmentation segmentation frameworks have demonstrated their usefulness in particular in structured environments such as urban areas. Emphasis is put on context learning in (Volpi and Ferrari, 2015) and on the design of robust yet locally discriminative modelling strategy for urban area classification of VHR multispectral images. It is based on a flexible energetical framework, namely a CRF. The adoption of fully-connected CRF can even allow to learn longer class interactions such as shown in (Li and Yang, 2016). Finally, semantic segmentation can be achieved using Deep Neural Networks, assuming the standard procedure is accompanied with proper deconvolution steps or with a Fully Connected Network such as in (Marmanis et al., 2016).

In forested areas, the combined use of airborne lidar (for height structure) and VHR multispectral images (for species composition) into such a smoothing process would allow (i) to retrieve homogeneous patches, (ii) to define the homogeneity criterion/criteria and (iii) to control the level of generalization of the final label map.

3. METHODS

3.1 General strategy

The proposed method assumes that a label map is provided for the areas of interest, and is accompanied with a class membership probability map, which provides, for each pixel of the image, the posterior class membership for all classes of interest. These are the necessary inputs for all methods described below. In practice, the strategy proposed in (Dechesne et al., 2017) is as followed: a supervised classification is performed on a selection of features

extracted both from 3D lidar point clouds and aerial multispectral images. The training pixels are selected according to an existing forest LC geodatabase. The used classifier is the Random Forest (RF) classifier. This is an efficient classifier, that directly handles multiple classes, and provides posterior probabilities for each class.
Here, both local and global methods are tested. For local techniques, majority voting and probabilistic relaxation are selected (Sections 3.2 and 3.3). For global methods, various energy formulations based on a feature-sensitive Potts model are proposed (Section 3.4).

3.2 Filtering

An easy way to smooth a probability map is to filter it. All the pixels in a $r \times r$ pixels moving window $\mathcal{W}$ are combined in order to generate an output label of the central pixel. The most popular filter is the majority filter. Firstly, the class probabilities are converted into labels, assuming that the label of pixel $\mathbf{x}$ is the label of the most probable class.

$$C(\mathbf{x}) = [c_i | P(\mathbf{x}, c_i) \geq P(\mathbf{x}, c_j) \forall j], \tag{1}$$

with $i, j \in [1, n_c]$, where n_c is the number of classes. From this label image, the final smoothed result is obtained by taking the majority vote in a local neighborhood.

$$C_{smooth}(\mathbf{x}) = \arg\max_i \left[\sum_{\mathbf{u} \in \mathcal{W}} [C(\mathbf{u}) = c_i] \right]. \tag{2}$$

Many other filters have been developed but are not investigated in this paper.

3.3 Probabilistic relaxation

The probabilistic relaxation aims at homogenizing probabilities of a pixel according to its neighboring pixels. The relaxation is an iterative algorithm in which the probability at each pixel is updated at each iteration in order to have it closer to the probabilities of its neighbors (Gong and Howarth, 1989). It was adopted for simplicity reasons. First, good accuracies are reported with decent computing time, which is beneficial over large scales. Secondly, it offers an alternative to edge aware/gradient-based techniques that may not be adapted in semantically unstructured environments like forests. The probability $P_k^t(\mathbf{u})$ of class k at a pixel $\mathbf{u}$ at the iteration t is defined by $\delta P_k^t(\mathbf{u})$ which depends on:

- The distance $d_{\mathbf{u},\mathbf{v}}$ between the pixel $\mathbf{u}$ and its neighbors $\mathbf{v}$ (the pixels that are distant of less than r pixels from $\mathbf{u}$).
- A co-occurrence matrix $T_{k,l}$ defining a priori correlation between the probabilities of neighboring pixels. The local co-occurrence matrix has been tuned arbitrarily, but can also be estimated using training pixels (Volpi and Ferrari, 2015). The matrix is expressed as follow:

$$T_{k,l} = \begin{bmatrix} 0.8 & p & \cdots & p \\ p & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & p \\ p & \cdots & p & 0.8 \end{bmatrix}, \text{with } p = \frac{0.2}{n_c - 1}.$$

The update factor is then defined as:

$$\delta P_k^t(\mathbf{u}) = \sum_{\mathbf{v} \in \mathcal{N}_\mathbf{u}} d_{\mathbf{u},\mathbf{v}} \sum_{l=1}^{n_c} T_{k,l}(\mathbf{u}, \mathbf{v}) \times P_l^t(\mathbf{v}). \tag{3}$$

In order to keep the probabilities normalized, the update is performed in two steps using the unnormalized probability $Q_k^{t+1}(\mathbf{u})$ of class k at a pixel $\mathbf{u}$ at the iteration $t+1$:

$$Q_k^{t+1}(\mathbf{u}) = P_k^t(\mathbf{u}) \times \left(1 + \delta P_k^t(\mathbf{u})\right), \tag{4}$$

$$P_k^{t+1}(\mathbf{u}) = \frac{Q_k^{t+1}(\mathbf{u})}{\sum_{l=1}^{n_c} Q_l^{t+1}(\mathbf{u})}. \tag{5}$$

3.4 Global smoothing

The global smoothing method uses only a small number of pairwise cliques between neighboring pixels (4-neighbors or 8-neighbors) to describe the smoothness. Over the entire resulting first order random fields, the maximization of the posterior probability leads to a smoothed results. This can be done by finding the minimum of the negative log-likelihood, $\arg\min\limits_C E(I, C, A)$ with

$$\begin{aligned} E(I, C, A) = \sum_{\mathbf{u} \in I} E_{\text{data}}(\mathbf{u}, P(\mathbf{u})) + \\ \gamma \sum_{\mathbf{u} \in I, \mathbf{v} \in \mathcal{N}_\mathbf{u}} E_{\text{pairwise}}(\mathbf{u}, \mathbf{v}, C(\mathbf{u}), C(\mathbf{v}), A(\mathbf{u}), A(\mathbf{v})), \end{aligned} \tag{6}$$

where $P(\mathbf{u}) = [P(\mathbf{u}, c_i) | P(\mathbf{u}, c_i) \geq P(\mathbf{u}, c_j) \forall j]$, $A(\mathbf{u})$ are the values of the features at pixel $\mathbf{u}$ (such as height, reflectance...) and $\mathcal{N}_\mathbf{u}$ is the 8-connected neighborhood of the pixel $\mathbf{u}$ (only the 8-connected neighborhood is investigated in this paper). When $\gamma = 0$, the pairwise term has no effect in the energy formulation; the most probable class is attributed to the pixel, leading to the same result as the classification output. When $\gamma \neq 0$, the resulting label map becomes more homogeneous, and the borders of the segments/stands are smoother. However, if γ is too high, the small areas are bound to be merged into larger areas, removing a part of the useful information provided by the classification step. The automatic tuning of the parameter γ has been addressed in (Moser et al., 2013) but is not used here.
In this paper, two formulations of E_{data} (unary term) and four formulations of E_{pairwise} (prior) are investigated.

3.4.1 Unary term

A widely used formulation for the unary term is the log-inverse formulation using the natural logarithm. It corresponds to the information content in information theory and is formulated as follow:

$$E_{\text{data}} = -\log(P(\mathbf{u})). \tag{7}$$

It highly penalizes the low-probability classes but increase the complexity with potential infinite values.
An other simple formulation for the unary term is the linear formulation,

$$E_{\text{data}} = 1 - P(\mathbf{u}). \tag{8}$$

It penalizes less than the log-inverse formulation but has the advantage of having values lying in $[0, 1]$.

3.4.2 Prior

In this work, the prior has a value depending on the class of neighboring pixels. In the four formulations, two neighboring pixels pay no penalty if they are assigned to the same class. Two basic and popular priors, the *Potts model* and the *contrast-sensitive Potts model* (called here *z-Potts model*), are investigated. In the *Potts model*, two neighboring pixels pay the same penalty if they are assigned to different labels, the prior for the *Potts model* is:

$$\begin{aligned} E_{\text{pairwise}}(C(\mathbf{u}) = C(\mathbf{v})) &= 0, \\ E_{\text{pairwise}}(C(\mathbf{u}) \neq C(\mathbf{v})) &= 1. \end{aligned} \tag{9}$$

In the *z-Potts model*, the penalty for a change of label depends on the gradient of height between two neighboring pixels. The *z-Potts model* is a standard *contrast-sensitive Potts model* applied to the height obtained from the point clouds. Here, since we are dealing with forest stands that are likely to exhibit distinct heights, the gradient of the height map (given with the 3D lidar point cloud) is computed for each of the four directions separately. The maximum M_g over the whole image in the four directions is used to compute the final pairwise energy. A linear function has been used: the penalty is highest when the gradient is 0, and decreases until the gradient reaches its maximum value. The prior of the *z-Potts model* is therefore:

$$\begin{aligned} E_{\text{pairwise}}(C(\mathbf{u}) = C(\mathbf{v})) &= 0, \\ E_{\text{pairwise}}(C(\mathbf{u}) \neq C(\mathbf{v})) &= 1 - \frac{g_{\mathbf{u}\to\mathbf{v}}}{M_g}, \end{aligned} \tag{10}$$

where $g_{\mathbf{u}\to\mathbf{v}}$ is the gradient between pixel $\mathbf{u}$ and pixel $\mathbf{v}$, i.e., the absolute value of the height difference of the two pixels.

An other pairwise energy investigated is a global feature sensitive energy (called here *Exponential-features model*). The pairwise energy is computed with respect to a pool of n features. When the features have close values, the penalty is high and decreases when the features tends to be very different. The pairwise energy in this case is expressed as follows:

$$\begin{aligned} E_{\text{pairwise}}(C(\mathbf{u}) = C(\mathbf{v})) &= 0, \\ E_{\text{pairwise}}(C(\mathbf{u}) \neq C(\mathbf{v})) &= \frac{1}{n}\sum_{i=1}^{n} \exp(-|A_i(\mathbf{u}) - A_i(\mathbf{v})|), \end{aligned} \tag{11}$$

where $A_i(\mathbf{u})$ is the value of the i^{th} feature of the pixel $\mathbf{u}$. To compute such energy, the features need to be first normalized (i.e., zero mean, unit standard deviation) in order ensure that they all have the same dynamic.

The last formulation investigated is also a global feature sensitive energy (called here *Distance-features model*). The pairwise energy is still computed with respect to a pool of n features. In this case, the energy is computed according to the distance between the two neighboring pixels in the feature space, the penalty is high when the pixels are close in the feature space and decrease when they get distant. The pairwise energy in this case is expressed as follow:

$$\begin{aligned} E_{\text{pairwise}}(C(\mathbf{u}) = C(\mathbf{v})) &= 0, \\ E_{\text{pairwise}}(C(\mathbf{u}) \neq C(\mathbf{v})) &= 1 - ||A(\mathbf{u}); A(\mathbf{v})||_{n,2}, \end{aligned} \tag{12}$$

with

$$||A(\mathbf{u}); A(\mathbf{v})||_{n,2} = \frac{1}{\sqrt{n}} \sqrt{\sum_{i=1}^{n} \left(A_i(\mathbf{u}) - A_i(\mathbf{v})\right)^2}. \tag{13}$$

To compute such energy, the features need to be first normalized (i.e., zero mean, unit standard deviation) in order ensure that they all have the same dynamic. They are then rescaled between 0 and 1 to ensure that $||A(\mathbf{u}); A(\mathbf{v})||_{n,2}$ lies in $[0; 1]$ $\forall(\mathbf{u}, \mathbf{v})$.

In (Dechesne et al., 2017), a high number of features was extracted from available lidar and optical images ($\sim$ 100) but can be selected. They can also be weighted according to their importance, computed through the Random Forest classification process. Since the most important features (20) are almost all equally weighted, it does not bring additional discriminative information for the global feature sensitive energy.

3.4.3 Energy minimization The energy minimization is performed using graph-cut methods. The graph-cut algorithm employed is the quadratic pseudo-boolean optimization (QPBO). The QPBO is a popular and efficient graph-cut method as it efficiently solves energy minimization problems (such as the proposed ones) by constructing a graph and computing the min-cut (Kolmogorov and Rother, 2007). α-expansion moves are used, as they are an efficient way to deal with the multi-class problems (Kolmogorov and Zabih, 2004).

4. DATA AND RESULTS

4.1 Data

The different smoothing methods have been conducted on 4 mountainous test areas. Each area has a surface of 1 km^2. The IRC ortho-images of the test areas are presented in Figure 1. The proposed areas exhibit a large range of species (>4). The airborne multispectral images were captured by the IGN digital cameras (Souchon et al., 2012). They have 4 bands: 430-550 nm (blue), 490-610 nm (green), 600-720 nm (red) and 750-950 nm (near infra-red) at 0.5 m ground sample distance (spatial resolution).

The airborne lidar data were collected using an Optech 3100EA device. The footprint was 0.8 m in order to increase the probability to reach the ground. The point density for all echoes ranges from 2 to 4 points/m^2. Our multispectral and lidar data fit with the standards used in many countries for large-scale operational forest mapping purposes (White et al., 2016). The multispectral images and the lidar data were acquired simultaneously.

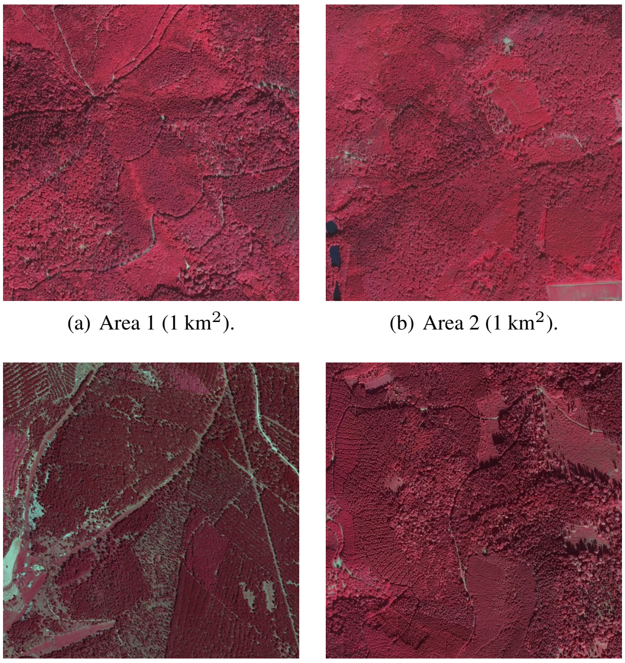

(a) Area 1 (1 km^2).
(b) Area 2 (1 km^2).
(c) Area 3 (1 km^2).
(d) Area 4 (1 km^2).

Figure 1. Orthoimages of the areas of interest.

4.2 Results

The results for all methods are presented for Area 2 in Table 1. The overall accuracy is computed by comparing the labelled pixels in the forest LC, to the pixels of the output images. The filtering method performs the worse with a gain of less than 1% compared to the classification, even with large filters. Furthermore, the larger the filter is, the longer are the computation times. The probabilistic relaxation has also poor results (+5% than the classification) and has also important computation times, since the iterative process runs until the convergence has been reached. The global smoothing methods have great results, increasing the accuracy up to 15%. The *z-Potts model* tends to have slightly worse results than the other methods. The *Potts model* and the *Distance-features model* have very close results regardless of the unary term. The *Exponential-features model* have the greatest results with the linear unary, but have poor results with the log-inverse unary. It appears that it is the only energy that is sensitive to the unary term, indeed, for the *Potts model*, the *z-Potts model* and the *Distance-features model*, the difference between the linear unary and the log-inverse unary is less than 0.2%.

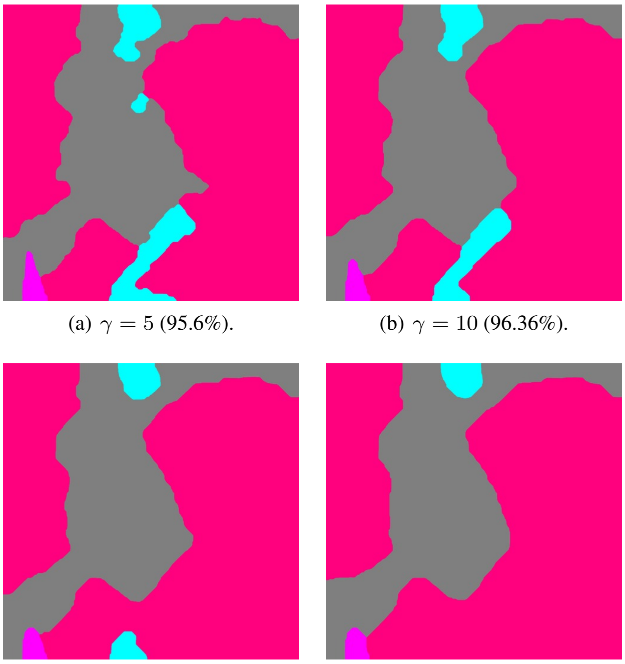
Methods		Smoothing overall accuracy	Parameter
Filtering		82.33%	$r = 5$
Filtering		82.41%	$r = 25$
Probabilistic relaxation		86.89%	$r = 5$
Potts	log-inverse unary	93.34%	$\gamma = 5$
		95.24%	$\gamma = 10$
		95.61%	$\gamma = 15$
		96.03%	$\gamma = 20$
	linear unary	95.96%	$\gamma = 5$
		96.24%	$\gamma = 10$
		94.08%	$\gamma = 15$
		92.32%	$\gamma = 20$
z-Potts	log-inverse unary	93.02%	$\gamma = 5$
		95.09%	$\gamma = 10$
		95.53%	$\gamma = 15$
		95.96%	$\gamma = 20$
	linear unary	95.52%	$\gamma = 5$
		96.00%	$\gamma = 10$
		94.04%	$\gamma = 15$
		93.23%	$\gamma = 20$
Exponential-features	log-inverse unary	92.73%	$\gamma = 5$
		95.13%	$\gamma = 10$
		95.54%	$\gamma = 15$
		95.78%	$\gamma = 20$
	linear unary	95.6%	$\gamma = 5$
		96.36%	$\gamma = 10$
		95.27%	$\gamma = 15$
		94.09%	$\gamma = 20$
Distance-features	log-inverse unary	93.12%	$\gamma = 5$
		95.24%	$\gamma = 10$
		95.61%	$\gamma = 15$
		96.05%	$\gamma = 20$
	linear unary	95.63%	$\gamma = 5$
		96.23%	$\gamma = 10$
		94.12%	$\gamma = 15$
		92.34%	$\gamma = 20$

Table 1. Results for the proposed methods for Area 2. The classification has an overall accuracy of 81.41%.

The best results for the four areas are presented in Figure 2. It shows that the proposed formulation is very efficient to retrieve forest patches of pure species with smooth borders. The accuracy ranges from 96% to 99%. Furthermore, the borders between adjacent classes fit well with the ones from the forest LC borders, which validates the relevance of our model. However, in areas where no data is available, it is hard to ensure that our model has relevant results, but, from a visual point of view, the results seem good.

The results for all the proposed models using the log-inverse unary are presented for Area 1 in Figure 3. It appears clearly that basic methods (such as filtering or probabilistic relaxation) are not adapted to our problem since the results remain very noisy. However, having a too binding unary term in the model also leads to noisy results. Even if the accuracy is higher than the accuracy using the linear unary, the small patches produced with the log-inverse unary are not acceptable for a forest LC.

The effect of the parameter γ is presented in Figure 4. When γ is low, the borders are rough and small regions might appear (Figure 4(a)). Increasing γ smooth the borders, however, a too high value have a negative impact on the results, reducing the size of meaningful segments (Figure 4(c)) or even removing them (Figure 4(d)). The tuning of the parameter γ is an important issue, since different values of γ might be acceptable depending on the level of detail expected for the segmentation. In forest inventory, having small regions of pure species is interesting for the understanding of the behavior of the forest. For generalization purposes (such as forest LC), the segments must have a decent size and may exhibit variability.



(a) $\gamma = 5$ (95.6%).

(b) $\gamma = 10$ (96.36%).

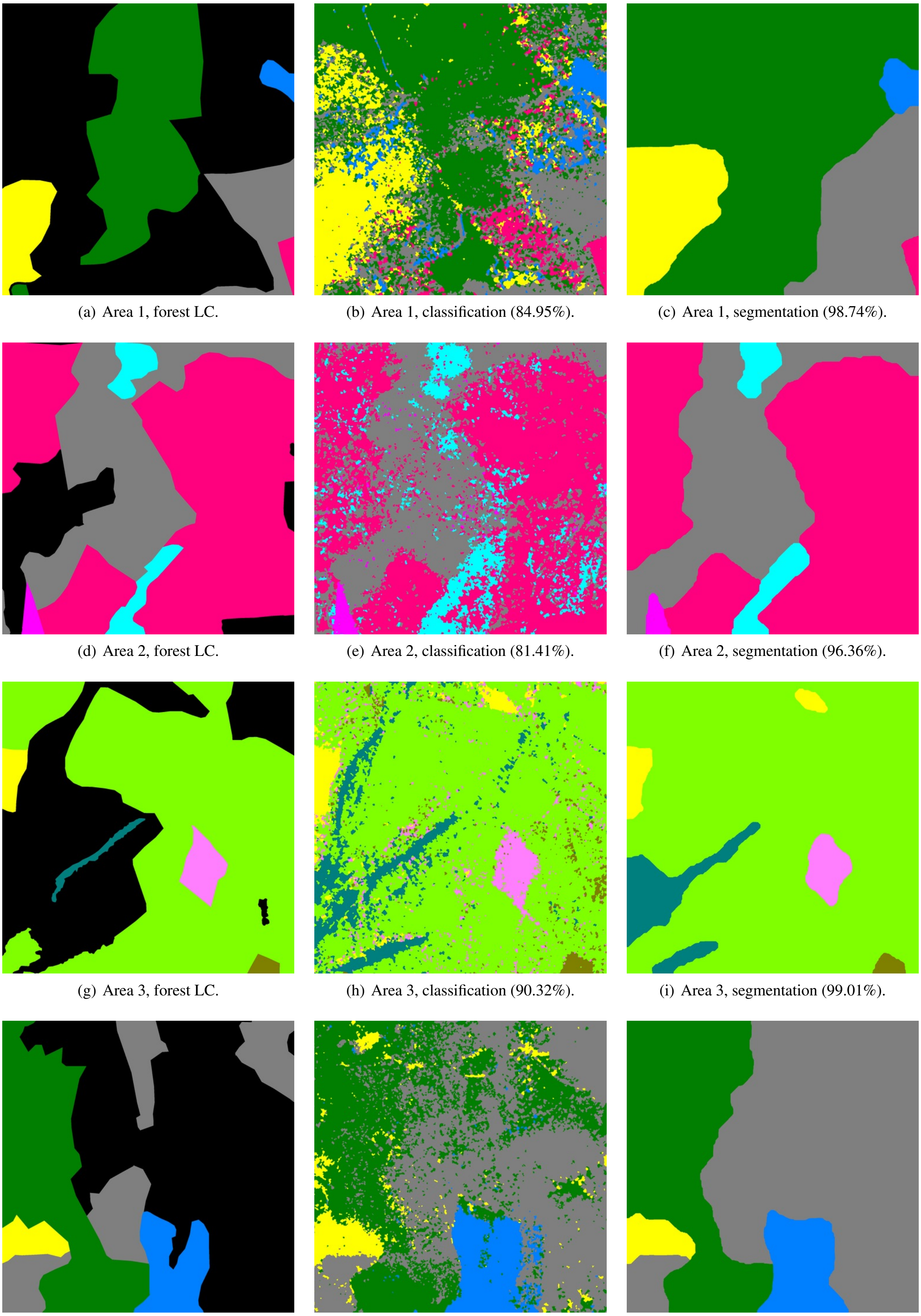
(c) $\gamma = 15$ (95.27%).

(d) $\gamma = 20$ (94.09%).

Figure 4. Results of the *Exponential-features model* with linear unary for different values of γ for Area 2, the overall accuracy is specified in brackets. Color codes: ●*deciduous oaks*, ●*fir or spruce*, ●*chestnut*, ●*robinia*.

5. CONCLUSION

The semantic segmentation of forest stands can be achieved by the fusion of ALS data and multispectral images. These two remote sensing modalities produce very satisfactory results since they both provide complementary observations. Good discrimination scores are already obtained with standard features and



(a) Area 1, forest LC.

(b) Area 1, classification (84.95%).

(c) Area 1, segmentation (98.74%).

(d) Area 2, forest LC.

(e) Area 2, classification (81.41%).

(f) Area 2, segmentation (96.36%).

(g) Area 3, forest LC.

(h) Area 3, classification (90.32%).

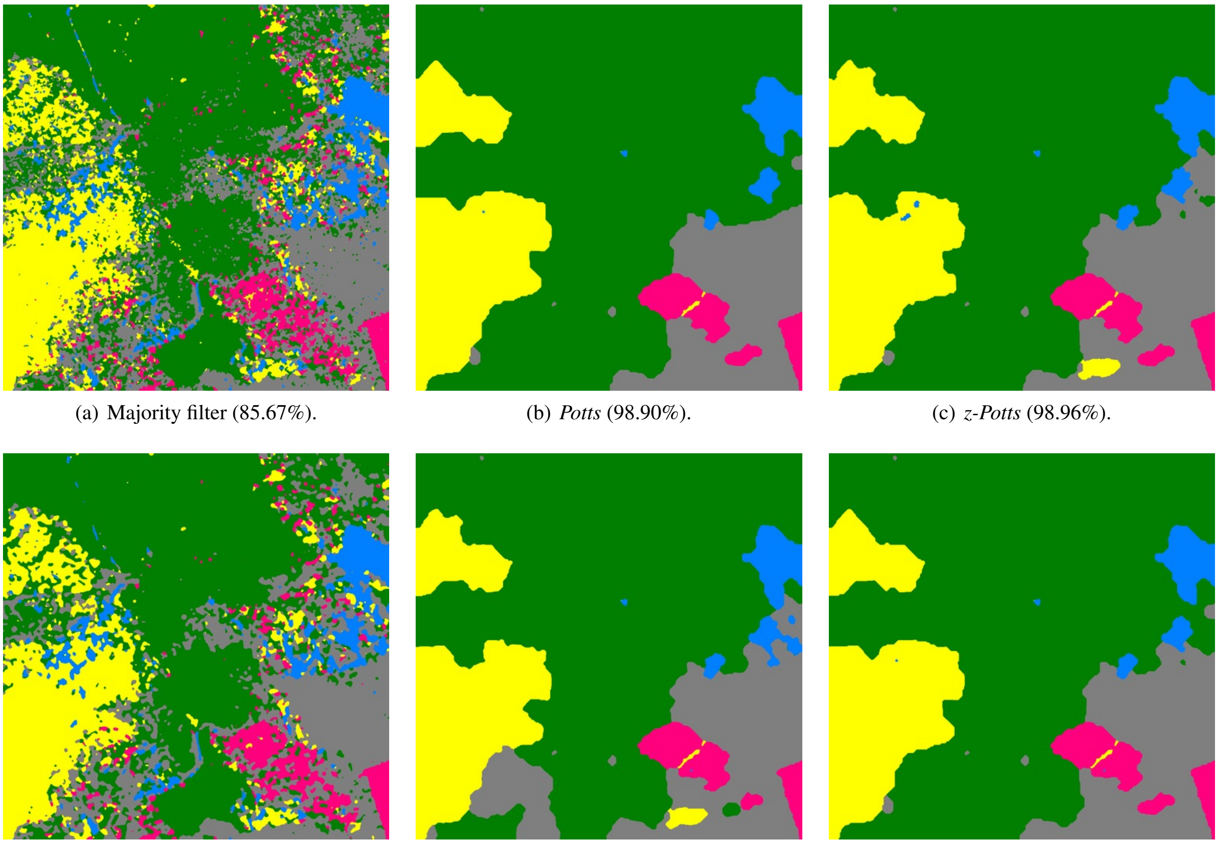
(i) Area 3, segmentation (99.01%).

(j) Area 4, forest LC.

(k) Area 4, classification (86.69%).

(l) Area 4, segmentation (97.39%).

Figure 2. Results for all the 4 areas, the overall accuracy is specified in brackets. The smoothing is performed using the *Exponential-features model* with linear unary ($\gamma = 10$). Color codes: ●*beech*, ●*deciduous oaks*, ●*Scots pine*, ●*Douglas fir*, ●*fir or spruce*, ●*chestnut*, ●*robinia*, ●*larch*, ●*non-pectinated fir*, ●*black pine*, ●*herbaceous formation*, ●no data.



(a) Majority filter (85.67%).

(b) *Potts* (98.90%).

(c) *z-Potts* (98.96%).

(d) Probabilistic relaxation (88.89%).

(e) *Exponential-features* (97.63%).

(f) *Distance-features* (98.89%).

Figure 3. Results of the different proposed models for Area 1, the overall accuracy is specified in brackets. Color codes: ●*beech*, ●*deciduous oaks*, ●*Scots pine*, ●*Douglas fir*, ●*fir or spruce*.

classifier, which is a strong basis for an even more accurate delineation. This delineation can then achieved using several smoothing methods. It appears that a too simple smoothing model (such as filtering or probabilistic relaxation) is not sufficient in order to obtain consistent segments. A global smoothing method based on an energy model tends to be well adapted to the problem. A simple *Potts model* with a linear unary term provides excellent results. The model can be improved using the features used for the classification. Such model produces slightly better results, but also increases the complexity. However, having a too complex model (such as *Exponential-features model* with log-inverse unary) decreases the performance of the segmentation. In order to obtain homogeneous areas in terms of species with smooth borders, a global method based on a simple energy model is sufficient.

REFERENCES

Albert, L., Rottensteiner, F. and Heipke, C., 2016. Contextual land use classication: How detailed can the class structure be? *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLI-B4, pp. 11–18.

Arbeláez, P., Hariharan, B., Gu, C., Gupta, S., Bourdev, L. and Malik, J., 2012. Semantic segmentation using regions and parts. In: *Proc. of CVPR*, pp. 3378–3385.

Boix, X., Gonfaus, J. M., Weijer, J., Bagdanov, A. D., Serrat, J. and Gonzàlez, J., 2011. Harmony potentials. *International Journal of Computer Vision* 96(1), pp. 83–102.

Chen, L., Papandreou, G., Kokkinos, I., Murphy, K. and Yuille, A., 2015. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. In: *Proc. of ICLR*.

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.

Dalponte, M., Reyes, F., Kandare, K. and Gianelle, D., 2015. Delineation of individual tree crowns from ALS and hyperspectral data: a comparison among four methods. *European Journal of Remote Sensing* 48, pp. 365–382.

Dechesne, C., Mallet, C., Le Bris, A. and Gouet-Brunet, V., 2017. Semantic segmentation of forest stands of pure species combining airborne lidar data and very high resolution multispectral imagery. *ISPRS Journal of Photogrammetry and Remote Sensing* 126, pp. 129–145.

Diedershagen, O., Koch, B. and Weinacker, H., 2004. Automatic segmentation and characterisation of forest stand parameters using airborne lidar data, multispectral and fogis data. *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences* 36(8/W2), pp. 208–212.

Eysn, L., Hollaus, M., Schadauer, K. and Pfeifer, N., 2012. Forest delineation based on airborne lidar data. *Remote Sensing* 4(3), pp. 762–783.

Gong, P. and Howarth, P., 1989. Performance analyses of probabilistic relaxation methods for land-cover classification. *Remote Sensing of Environment* 30(1), pp. 33–42.

Kangas, A. and Maltamo, M., 2006. *Forest inventory: methodology and applications*. Vol. 10, Springer Science & Business Media.

Kirillov, A., Schlesinger, D., Forkel, W., Zelenin, A., Zheng, S., Torr, P. and Rother, C., 2015. A generic CNN-CRF model for semantic segmentation. *arxiv:1511.05067*.

Kohli, P., Ladický, Ľ. and Torr, P., 2009. Robust higher order potentials for enforcing label consistency. *International Journal of Computer Vision* 82(3), pp. 302–324.

Kolmogorov, V. and Rother, C., 2007. Minimizing nonsubmodular functions with graph cuts-a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29(7), pp. 1274–1279.

Kolmogorov, V. and Zabih, R., 2004. What energy functions can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26(2), pp. 147–159.

Kumar, S. and Hebert, M., 2006. Discriminative random elds. *International Journal of Computer Vision* 68(2), pp. 179–201.

Ladický, Ľ., Russell, C., Kohli, P. and Torr, P., 2012. Inference methods for CRFs with co-occurrence statistics. *International Journal of Computer Vision* 103(2), pp. 213–225.

Ladický, Ľ., Sturgess, P., Alahari, K., Russell, C. and Torr, P., 2010. What, where and how many? combining object detectors and CRFs. In: *Proc. of ECCV*.

Lee, J., Cai, X., Lellmann, J., Dalponte, M., Malhi, Y., Butt, N., Morecroft, M., Schnlieb, C. B. and Coomes, D. A., 2016. Individual tree species classification from airborne multisensor imagery using robust pca. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 9(6), pp. 2554–2567.

Leppänen, V., Tokola, T., Maltamo, M., Mehtätalo, L., Pusa, T. and Mustonen, J., 2008. Automatic delineation of forest stands from lidar data. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 38(4/C1)* pp. 5–8.

Li, W. and Yang, M. Y., 2016. Efficient semantic segmentation of man-made scenes using fully-connected Conditional Random Field. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLI-B3, pp. 633–640.

Lucchi, A., Li, Y., Boix, X., K.Smith and Fua, P., 2011. Are spatial and global constraints really necessary for segmentation? In: *Proc. of ICCV*, pp. 9–16.

Marmanis, D., Wegner, J. D., Galliani, S., Schindler, K., Datcu, M. and Stilla, U., 2016. Semantic segmentation of aerial images with an ensemble of CNNs. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences* III-3, pp. 473–480.

Means, J. E., Acker, S. A., Fitt, B. J., Renslow, M., Emerson, L. and Hendrix, C. J., 2000. Predicting forest stand characteristics with airborne scanning lidar. *Photogrammetric Engineering & Remote Sensing* 66(11), pp. 1367–1372.

Montoya-Zegarra, J. A., Wegner, J. D., Ladick, L. and Schindler, K., 2015. Semantic segmentation of aerial images in urban areas with class-specific higher-order cliques. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences* II-3/W4, pp. 127–133.

Moser, G., Serpico, S. and Benediktsson, J., 2013. Land-cover mapping by Markov modeling of spatial contextual information in very-high-resolution remote sensing images. *Proceedings of the IEEE* 101(3), pp. 631–651.

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.

Souchon, J.-P., Thom, C., Meynard, C. and Martin, O., 2012. A large format camera system for national mapping purposes. *Revue Française de Photogrammétrie et de Télédétection* (200), pp. 48–53.

Tiede, D., Blaschke, T. and Heurich, M., 2004. Object-based semi automatic mapping of forest stands with laser scanner and multi-spectral data. *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences* 36(8/W2), pp. 328–333.

Volpi, M. and Ferrari, V., 2015. Semantic segmentation of urban scenes by learning local class interactions. In: *Proc. of CVPR Workshops*, pp. 1–9.

White, J., Coops, N., Wulder, M., Vastaranta, M., Hilker, T. and Tompalski, P., 2016. Remote sensing technologies for enhancing forest inventories: A review. *Canadian Journal of Remote Sensing,* 42(5), pp. 619–641.

Zheng, C. and Wang, L., 2015. Semantic segmentation of remote sensing imagery using object-based markov random field model with regional penalties. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 8(5), pp. 1924–1935.