Spatio-temporal Conditional Random Fields for recognition of sub-tropical crop types from multi-temporal images

Pedro Marco Achanccaray Diaz [1,3]
Raul Queiroz Feitosa [1,2]
Franz Rottensteiner[3]
Ieda Del'Arco Sanches[4]
Christian Heipke[3]

[1] Pontifical Catholic University of Rio de Janeiro – PUC-Rio
Rua Marquês de São Vicente, 225 – 22451-900 – Rio de Janeiro – RJ, Brazil
{pmad9589, raul}@ele.puc-rio.br

[2] Rio de Janeiro State University - UERJ
Rua São Francisco Xavier, 524 – 20550-900 – Rio de Janeiro – RJ, Brazil

[3] Institute of Photogrammetry and GeoInformation, Leibniz Universität Hannover – LUH
Nienburger Str.1, 30167 Hannover, Germany
{rottensteiner, heipke}@ipi.uni-hannover.de

[4] National Institute for Space Research – INPE
Av. dos Astronautas, 1.758 - Jardim da Granja, São José dos Campos - SP, Brazil
ieda.sanches@inpe.br

Abstract. Crop recognition from remote sensing images is a challenging task due to the dynamic behavior of different crops. The spectral appearance of a given crop changes over time because it is highly related to the phenological stage at each epoch or season, making it necessary to use sequences of images for a correct classification. Conditional Random Field (CRF) approaches have been increasingly applied for crop recognition due to their ability to consider contextual information in both, the spatial and the temporal domains. This work proposes a spatio-temporal CRF for modelling different crops and their respective phenological stages from a sequence of Landsat 5/7 images. The spatial context is introduced using a contrast-sensitive smooth labeling method. The interactions in the temporal domain are modeled based on the joint posterior probability of class relations between adjacent epochs given the observed data. These class relations are learnt using a Random Forest (RF) classifier. Comparisons between mono-temporal classification using RF, CRFs considering only spatial context information and the proposed model are presented. Furthermore, an analysis on how the sequence image length as well as the starting epoch affects the classification accuracy is carried out. Improvements in the overall accuracy of up to 12% and 6% over the RF and mono-temporal CRF approaches, respectively, are obtained using the proposed model considering sequences of up to 9 images.

Keywords: remote sensing, probabilistic graphical models, crop recognition, Landsat images.

1. Introduction

In recent years, agricultural monitoring has become more important for a wise management of natural resources due to constant population growth and urban expansion. Prediction of yields, estimation of food production and precise and accurate agricultural statistics are crucial in order to anticipate the market behavior. Remote sensing data provide a cost-effective tool for agricultural monitoring and management. The use of multi-temporal images sequences has shown significant improvement in classification of crops and vegetation (Lu & Weng, 2007), because image sequences can capture changes in spectral appearance over time which are related to different phenological stages of the plants.

Conditional Random Field (CRF) approaches are considered to be very suitable for crop recognition due to their capability to consider contextual information (in the spatial and the

temporal domain), which is particularly important for crop recognition because different crops have different phenological cycles based on their specific physiology, which cause changes in their spectral appearance over time. In spite of these benefits, just a few CRF-based approaches have been proposed. Hoberg & Müller (2011) used CRFs for spatio-temporal crop classification using a site wise feature differences in two epochs to model temporal dynamics and restrict class transitions over time. These restrictions, however have shown to be too restrictive for crop phenology changes. Later, Hoberg et al. (2015) modeled temporal interactions by a global transition matrix using expert knowledge, neglecting any dependency on the data. In Kenduiywo et al. (2016), a Dynamic Conditional Random Fields (DCRFs) approach is proposed to learn the phenological information from SAR images considering correlation between backscattering of crops in the same phenology stage.

In this work, a spatio-temporal CRF based approach for recognition of crop types in sub-tropical areas from a sequence of Landsat images is proposed. Our approach considers smooth labeling methods depending on the data for both spatial and temporal interactions. In our case, the spectral response of a crop varies significantly during the phenological cycle. For that reason, we consider every crop in a certain phenological stage as a unique class. In addition, we analyze the influence of the sequence length on the classification results.

2. Modeling spatial and temporal context with Conditional Random Fields

2.1. Conditional Random Fields

Conditional Random Fields (CRFs), firstly introduced by Lafferty et al. (2001) for one-dimensional text classification, belong to the family of undirected graphical models. Kumar & Hebert (2006) extended CRFs for two-dimensional image classification using discriminative models for class associations at individual sites as well as interactions for neighboring sites. Let $G = \{S, E\}$ be a graph with a set of nodes S and edges E and let the observed data from an input image be given by $\boldsymbol{x} = \{\boldsymbol{x}_i\}_{i \in S}$, where $\boldsymbol{x}_i$ is the data from i^{th} site. Their corresponding labels at the image sites are given by $\boldsymbol{y} = \{y_i\}_{i \in S}$, where $\boldsymbol{y}$ is indexed by the nodes of G and y_i belongs to a set of classes $L = [l_1, \dots, l_m]$. Then, CRF models the posterior probability $P(\boldsymbol{y}|\boldsymbol{x})$ of the labels given the data as follows:

$$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z}\left[\exp\left(\sum_{i \in S} A(y_i, \boldsymbol{x}) + \theta_{IS} \sum_{i \in S} \sum_{j \in N_i} IS(y_i, y_j, \boldsymbol{x})\right)\right] \quad (1)$$

where Z is a normalizing constant also called as partition function. $A(\cdot)$ and $IS(\cdot)$ are called unary or association potential and pairwise or interaction potential, respectively. The association potential measures how likely a site $i \in S$ will take a label y_i given the observed data $\boldsymbol{x}$, whereas the interaction potential determines how labels at neighboring sites i and j should interact given the data $\boldsymbol{x}$. N_i is the neighborhood of site i The parameter θ_{IS} expresses the weight of IS relative to A.

2.2. Multi-temporal CRFs

In a multi-temporal analysis, neighboring sites in adjacent epochs, which capture changes of a crop over time, are also taken into account. Let's consider a set of T coregistered images from different epochs, where a site i corresponds to the same geographical region in all epochs. The observed data of an image site i at epoch t is denoted as $\boldsymbol{x}_i^t$ and its corresponding label as y_i^t for $t = 1, \dots T$ being an epoch in the image sequence. Equation 1 is extended as follows:

$$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z}\left[exp\left(\sum_{t\in T}\sum_{i\in S} A^t(y_i^t, \boldsymbol{x}^t) + \theta_{IS}\sum_{t\in T}\sum_{i\in S}\sum_{j\in N_i} IS^t(y_i^t, y_j^t, \boldsymbol{x}^t) + \sum_{t\in T}\theta_{IT}^t\sum_{i\in S}\sum_{k\in C_i} IT^{tk}(y_i^t, y_i^k, \boldsymbol{x}^t, \boldsymbol{x}^k)\right)\right] \quad (2)$$

where A^t, IS^t and IT^{tk} are the association, spatial and temporal interaction potentials, resp. The temporal interaction potential IT^{tk} models the interaction at one site i in two adjacent epochs, namely t and k. C_i is the neighborhood of site i in adjacent epochs. $\theta_{IT}^t \in \boldsymbol{\theta}_{IT} = \{\theta_{IT}^1, \theta_{IT}^2, \dots, \theta_{IT}^{T-1}\}$ is the relative weight for the temporal interaction potential, assumed to depend on the epoch t and on the image sequence length.

The association potential $A^t(\cdot)$ measures how likely an image site i in epoch t will take a label y_i^t given its feature vector $\mathbf{f}_i(\boldsymbol{x}^t)$ that may depend on the entire image at epoch t. Thus, the association potential is given by $A^t(y_i^t, \boldsymbol{x}^t) = logP\left(y_i^t|\mathbf{f}_i(\boldsymbol{x}^t)\right)$, where $P\left(y_i^t|\mathbf{f}_i(\boldsymbol{x}^t)\right)$ is a local class conditional probability at image site i given $\mathbf{f}_i(\boldsymbol{x}^t)$. Any discriminative classifier with a probabilistic output can be used here. In this work, we adopted the Random Forest (*RF*) classifier (Breiman, 2001). RF generates an ensemble of randomized decision trees during training. For classification, each tree casts a vote for the most likely class based on its features. Then, the probability measure used to calculate the association potential is defined by the ratio of the sum of all votes for a class and the total number of trees.

The spatial interaction potential IS^t measures how labels at spatially neighboring sites i and j interact given the data $\boldsymbol{x}^t$ observed at time *t*. Contrast-sensitive smoothing labeling methods, which penalize label changes unless a significant data variation occurs in neighboring sites, have been successfully applied for this purpose (Schindler, 2012). Equation 3 presents the contrast-sensitive Potts model (Shotton et al., 2009) used in this work, which takes into account the similarity of adjacent site feature vectors by its Euclidian distance $d_{ij} = \left\|\mathbf{f}_i(\boldsymbol{x}^t) - \mathbf{f}_j(\boldsymbol{x}^t)\right\|$.

$$IS^t(y_i^t, y_j^t, \boldsymbol{x}^t) = \delta(y_i^t = y_j^t)\left[p + (1-p)e^{-\frac{d_{ij}^2}{2\sigma^2}}\right] \quad (3)$$

where σ^2 refers to the mean value of squared feature distances d_{ij}^2 and is computed during training and $\delta(\cdot)$ is a delta function returning 1 if its argument is true and 0 otherwise. The parameter $p \in [0,1]$ in Equation 3 controls the relative influence of the data-dependent and data-independent terms.

The temporal interaction potential IT^{tk} measures how labels y_i^t and y_i^k at the same site i in adjacent epochs t and k interact given their observed data $\boldsymbol{x}_i^t$ and $\boldsymbol{x}_i^k$. A generic model based on the joint posterior is used to design these potentials. At the cost of learning more parameters the model can express the fact that certain class relations may be more likely than others given the data (Niemeyer et al., 2016).

$$IT^{tk}(y_i^t, y_i^k, \boldsymbol{x}_i^t, \boldsymbol{x}_i^k) = logP\left(y_i^t, y_i^k|\mathbf{g}_i(\boldsymbol{x}_i^t, \boldsymbol{x}_i^k)\right) \quad (4)$$

where $P\left(y_i^t, y_i^k|\mathbf{g}_i(\boldsymbol{x}_i^t, \boldsymbol{x}_i^k)\right)$ represents the joint posterior probability of classes y_i^t and y_i^k at site i at adjacent epochs t and k given an interaction feature vector $\mathbf{g}_i(\boldsymbol{x}_i^t, \boldsymbol{x}_i^k)$, whose components are functions of the data observed at epochs t and k. Again, a Random Forest is used to determine these probabilities in the same way as for the association potential.

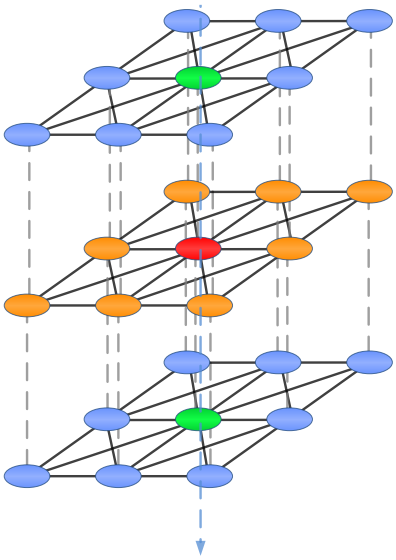


Figure 1. Multi-temporal graph structure corresponding to an image sequence of T epochs (here: 3), where an image site i (red) at epoch t has neighbors in the spatial domain N_i(orange) and the temporal domain C_i (green). The spatial and temporal interactions are represented by solid and dashed lines respectively.

In our application, the nodes correspond to the pixels of the georeferenced images. The neighborhoods considered to add contextual information in the spatial and temporal domains for both interaction potentials are illustrated in Figure 1. The red node represents an image site i in epoch t with spatial neighbors $j \in N_i$ (orange nodes) considering 8 neighbors and temporal neighbors $k \in C_i$ (green nodes) in adjacent epochs. The spatial and temporal interaction potential are represented as solid and dashed lines, respectively. The site-wise feature vectors $\mathbf{f}_\mathrm{i}(\boldsymbol{x}^t)$ are defined to consist of the spectral values directly observed at site i and the NDVI derived from the spectral values. The temporal interaction feature vectors $\mathbf{g}_\mathrm{i}(\boldsymbol{x}_i^t, \boldsymbol{x}_i^k)$ are obtained by concatenating the site-wise feature vectors $\mathbf{f}_\mathrm{i}(\boldsymbol{x}^t)$ and $\mathbf{f}_\mathrm{i}(\boldsymbol{x}^k)$.

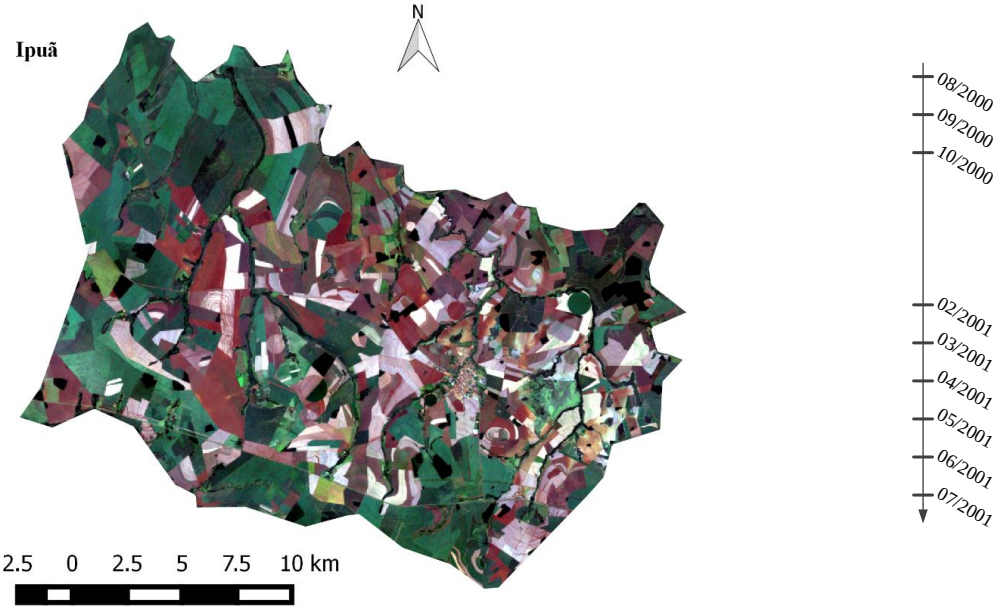
The classifiers for the association and temporal interaction potentials, respectively, were trained independently of each other. The weights for the spatial and temporal interaction potentials, θ_{IS} and $\boldsymbol{\theta}_{IT}$, respectively, are found using the Powell's search method (Kramer, 2010) setting all weights to one as starting solution, for a given image sequence length. The parameters of the spatial interaction potentials were determined empirically.

Exact inference, which is the task of finding the optimal label configuration $\boldsymbol{y}$ based on our model described by Equation 2, is computationally intractable for CRFs, except for special cases in binary classification (Kumar & Hebert, 2006). Thus, we used Loopy Belief Propagation (LBP) (Frey & MacKay, 1998), an approximation for graphs with cycles.

3. Experimental Analysis

3.1. Dataset

The study area has an extension of 46 km^2 and corresponds to the municipality of Ipuã, in the state of São Paulo, Brazil (see Figure 2a). A sequence of 9 Landsat scenes (see acquisition dates in Figure 2b) was taken, from either Landsat-5 (TM) or Landsat-7 (ETM+) with 30 m spatial resolution. The reference for each epoch was produced manually by a human expert.



(a) (b)

Figure 2. (a) Study area: Municipality of Ipuã in Sao Paulo, Brazil. (b) Acquisition dates corresponding to the image sequence in the dataset.

Agriculture is the main activity in this area; the most common crops are *sugarcane*, *soybeans* and *corn*. *Sugarcane* is a semi-perennial crop with cycles of 12 and 18 months. On the other hand, *soybeans* and *corn* are annual crops with cycles between 3 – 6 months (Schultz et al., 2015).

Based on the spectral differences observed in the images, the following phases were considered in our study: Initial phase of plant development (*IPPD*), Full Vegetative Vigor phase (*FVVP*) and Senescence phase (*SP*). During *IPPD*, the spectral response is dominated by the soil (when the soil is prepared before planting) or by the soil with straw (when planting is done over straw of the former crop) because the plants are very small and cannot be detected due to the coarse spatial resolution of orbital sensors such as Landsat. The next phase is *FVVP*, where the spectral response is dominated by photosynthetically active green vegetation (low reflectance in Red band and high reflectance in NIR band). Finally, *SP* comprises the phase where the plant leaves and grains, etc., begin to dry. Note that, depending on their physiology, not all crops pass all these three phases. In our study area, *corn* and *soybeans* pass by the three phases but *sugarcane* does not. However, there are some differences between the *SP* phase for *corn* and *soybeans* when they are ready for harvesting. In *corn* all components are dry (i.e. leaves, cobs, etc.), whereas in *soybeans*, only the pods are dry because leaves fall off after being dried. *Prepared Soil* and *Post-Harvest* stages, which were assigned to no crop, were also modeled. The *Prepared Soil* phase involves the ploughing and soil grooming processes, and in the *Post-Harvest* stage the vegetation residues still lie on the ground (i.e. straw). Even though *Pasture* and *Riparian Forest* are actually not crops, they were also treated as crop types in our model, and we assumed that both are permanently in one single stage (which means that there is one class for both (see Table 1). A class *Others* was considered, too, which represents other minor crops as well as urban areas and water bodies.

Due to the reduced availability of images with a low cloud coverage in our study area, there is a gap in the image sequence from November 2000 to January 2001. As a direct consequence of this gap, it was not possible to obtain images for all *soybeans*' phases. Table 1 summarizes all crops and phases considered. The first column represents the classes considered in this study.

Table 1. Classes considered in this study, which are formed by the combination of each crop with its corresponding phenological stage.

Class	Crop	Phase
PP	*No crop*	*Prepared Soil*
PH		*Post-Harvest*
SJ- FVVP	*Soybeans*	*Full Vegetative Vigor phase*
SJ-SP		*Senescence phase*
CR- IPPD	*Corn*	*Initial phase of plant development*
CR- FVVP		*Full Vegetative Vigor phase*
CR-SP		*Senescence phase*
SC- IPPD	*Sugarcane*	*Initial phase of plant development*
SC-FVVP		*Full Vegetative Vigor phase*
PS	*Pasture*	-
RF	*Riparian Forest*	-

3.2. Experimental Protocol

The proposed CRF model was tested for pixel-wise classification. A sequence of 9 images were used with references for all epochs; each image has approximately 500K pixels.

In our tests, we applied cross-validation taking approximately equal data partitions for training, determination of parameters (i.e., weights of the potentials) and testing and repeated the classification three times, varying the roles of the data subsets, so that each pixels was classified at least once in the test set. In each iteration of the cross-validation, the sequence length varied from 2 to T, where T is the number of epochs in the dataset. For each sequence length, the starting and end epochs were exhaustively moved across the whole sequence. For instance, for a sequence length of 2, sequences considering images from the 1st to 3rd, 2nd to 4th, 3rd to 5th and so on until 7th to 9th, were evaluated. Finally, the average of the overall accuracies of every sequence was computed.

The RF classifier used to compute the association and the temporal interaction potentials was applied with 250 trees (Hastie et al., 2009) and a tree depth of 25. These potentials and the parameter σ^2 of the contrast-sensitive Potts model for the spatial interaction potential were learned from training data. The parameter p of the spatial interaction potential, which represents the relative influence of the data-dependent and data-independent terms, was set to 0.5.

Our model, henceforth referred as CRF_{multi}, was compared with two mono-temporal approaches: the one that considers only the association potential and no spatial context, which leads us to a Random Forest classification (RF) and the second one that considers only the association and the spatial interaction (CRF_{mono}).

4. Results and Discussion

The results of our experiments are summarized in Table 2, where the average of the Overall Accuracy (OA) and Kappa Index ($Kappa$) are presented. A comparison between RF and CRF_{mono} shows significant performance improvements (from around 74% to nearly 80%) resulting from considering spatial context. Furthermore, with the addition of the temporal interaction potential as in the CRF_{multi} approach, even higher accuracies were obtained, attesting that the use of images of different epochs improves crops classification. The example shown in Figure 3 allows for a visual evaluation of the performance gains obtained with the incorporation of spatial and temporal interactions in the model.

Table 2. Average of Overall Accuracies and Kappa index using Random Forest (RF), a mono-temporal CRF (CRF_{mono}) and the multi-temporal CRF proposed in this work (CRF_{multi}).

Sequence Length	OA (%)			$Kappa$ (%)		
	RF	CRF_{mono}	CRF_{multi}	RF	CRF_{mono}	CRF_{multi}
2	73.9	79.5	**82.9**	65.7	72.2	**76.6**
3	73.8	79.4	**85.8**	65.6	72.1	**80.2**
4	73.9	79.6	**84.3**	65.8	72.5	**80.1**
5	74.1	79.8	**85.2**	66.1	72.8	**80.4**
6	73.9	79.6	**85.9**	65.8	72.5	**81.2**
7	73.8	79.4	**85.9**	65.6	72.1	**81.4**
8	73.9	79.5	**85.8**	65.7	72.2	**81.2**
9	73.8	79.3	**85.3**	65.5	71.9	**79.4**

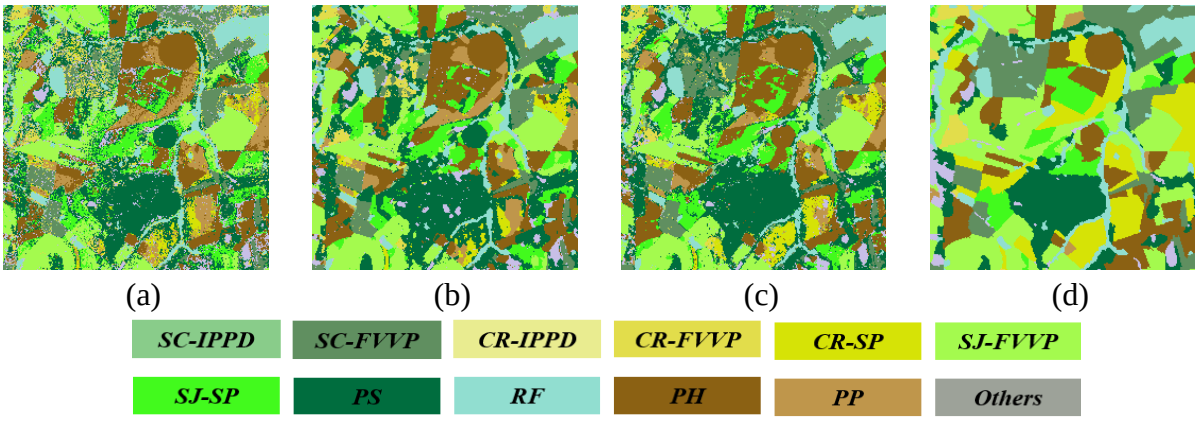


Figure 3. Snip of the maps obtained for an image taken in February 2001 using (a) Random Forest (RF), (b) a mono-temporal CRF (CRF_{mono}), (c) the multi-temporal proposed approach (CRF_{multi}) for a sequence length of 3 from the 2nd to 4th image in the sequence and (d) reference.

Clearly, the salt & pepper effect present in Figure 3a is attenuated in Figure 3b by the smoothing action of the contrast-sensitive Potts model. Similarly, the incorporation of the temporal interaction improved the classification in large areas, as shown in Figure 3c, although in some parts differences to the reference (Figure 3) remain.

In CRF_{multi}, there is a consistent improvement in the classification accuracy until a sequence length of 3, then, the accuracy decreases and increases again regularly with the sequence length. A possible explanation for this behavior is its relation to the phenological cycle and to the gap between the acquisition dates in the dataset, which introduces a high change, in crops and their phases, in the image sequence.

5. Conclusions

A spatio-temporal Conditional Random Field approach for crop recognition has been proposed and evaluated on a sequence of 9 Landsat images in this work. The new methodology lead to an increase of up to 6% in overall accuracies compared to mono-temporal context-based classification and of up to 12% compared to classification without considering context, demonstrating the importance of considering context sequences of images of different epochs.

The classification accuracy increases while the image sequence length increases until certain point, which seems to be related to the beginning of another phenological cycle.

Information about the duration of phenological cycles of each crop is thus believed to be very important in order to select a suitable sequence of image containing enough samples of each crop and its stage.

Extensions of this study will consider the usage of alternatives sensors such as SAR as well as the exploitation of expert knowledge to better model temporal context.

Acknowledgements

The authors acknowledge the support provided by CNPq (Conselho Nacional de Desenvolvimento e Pesquisa) and CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior).

References

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

Frey, B., MacKay, D. (1998). A Revolution: Belief propagation in Graphs with Cycles. A**dvances in neural information processing systems**, 10, 479-485.

Hastie, T., Friedman, J., Tibshirani, R. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics.

Hoberg, T., Müller, S. (2011). Multitemporal Crop Type Classification using Conditional Random FIelds and RapidEye data. **International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences**, XXXVIII-4(W19), 115-121.

Hoberg, T., Rottensteiner, F., Feitosa, R., Heipke, C. (2015). Conditional Random Fields for Multitemporal and Multiscale Classification of Optical Satellite Imagery. **IEEE Transaction on Geoscience and Remote Sensing**, 53(2), 659-673.

Kenduiywo, B., Bargiel, D., Soergel, U. (2016). Crop type mapping from a sequence of TerraSAR-X images with Dynamic Conditional Random Fields. **ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.**, III(7), 59-66.

Kramer, O. (2010). Iterated local search with Powell's method: a memetic algorithm for continuos global optimization. **Memetic Computing**, 2(1), 69-83.

Kumar, S., Hebert, M. (2006). Discriminative Random Fields. **International Journal of Computer Vision**, 68(2), 179-202.

Lafferty , J., McCallum, A., Pereira, F. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. 18th ICML (pp. 282-289). San Francisco, CA, USA: Morgan Kaufmann.

Lu, D., Weng, Q. (2007). A survey of image classification methods and techniques for improving classification performance. **International Journal of Remote Sensing**, 28(5), 823-870.

Niemeyer, J., Rottensteiner, F., Soergel, U., Heipke, C. (2016). Hierarchical higher order CRF for the classification of airborne Lidar point clouds in urban areas. **Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci**, XLI(B3), 655-662.

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

Schultz, B., Immitzer, M., Formaggio, A., Sanches, I., Barreto, A., Atzberger, C. (2015). Self-Guided Segmentation and Classification of Multi-Temporal Landsat 8 Images for Crop Type Mapping in Southeastern Brazil. **Remote Sensing**, 7(11), 14482-14508.

Shotton, J., Winn, J., Rother, C., 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), 2-23.