AGRICULTURAL FIELD DELIMITATION USING ACTIVE LEARNING AND RANDOM FORESTS MARGIN

Karim Ghariani (a), Nesrine Chehata(b,c), Arnaud Le Bris (d) and Philippe Lagacherie (e)

(a) Student at INSAT (Institut National des Sciences Appliquées et de Technologies), Tunis, Tunisia
(b) IRD/UMR LISAH El Menzah 4, Tunis, Tunisia
(c)EA 4592, Laboratoire G&E,Université de Bordeaux
(d)Université Paris-Est, IGN/SR, MATIS, Saint Mandé, France
(e)INRA Laboratoire LISAH, UMR 1221 Montpellier, France

ABSTRACT

Agricultural practices and spatial arrangements of fields have a strong impact on water flows in cultivated landscapes. In order to monitor landscapes at a large scale, there is a strong need for automatic or semi-automatic field delineation. Field measurements for delineating parcel network are not efficient, thus very high resolution satellite imagery should help delineating agricultural fields in a automatic way. This study focuses on agricultural field delineation based on the classification of very high resolution satellite imagery. A hybrid approach is proposed and combines a region-based approach and active learning (AL) techniques. Random forest (RF) classifier is used for classification and feature selection. The margin concept is used as uncertainty measure in active learning algorithm. Satisfying results are shown on a Geoeye image. AL RF model is compared to simple and global RF models that are built from adjacent and geographically distant fields respectively.

Keywords: Classification, active learning, segmentation, agricultural fields, very high resolution

1. INTRODUCTION

Agricultural practices are major drivers of water flows in cultivated landscapes. Especially, the spatial arrangements and connectivities of cultivated fields have a strong impact onto run off and soil erosion at the landscape and watershed scales.Thus, it is interesting to delineate automatically, at a large scale, cultivated fields in order to map the land use, to monitor crop rotations or parcel boundaries evolution or, to update an existing parcel database. Field methods are inadequate for characterizing the parcel network and detecting boundary changes at such scale. Very high resolution satellite images should help delineating the parcels. Our objective is to automatically detect parcel boundaries using very high resolution (VHR) images. The problem is challenging due to a high variability of agricultural field crops, and their boundaries. In practice, some boundaries may be delineated by embankments or various perennial vegetation. Another difficulty consists in a high intra-parcel variability that makes them difficult to segment as one entity.

In literature, few works exist in the remote sensing community, on agricultural field delineation, thus we searched for wider literature in the image processing community on edge and region extraction. Methods can be grouped into three approaches: 1) region-based approach 2) contour-based approach and 3) classification approach. Region-based approach is commonly used in the remote sensing community. Very high resolution satellites imagery has developed OBIA (Object based Image Analysis) approaches [1], based on segmentation algorithms. However, some regions semantically significant may appear at different scales, thus various works on agricultural or forest parcel segmentation are based on hierarchical segmentation [2, 3] and used Definiens Software [4]. These approaches are sensitive to intra-parcel variability. Contour-based approach can be based on edge detection using image gradient. However these methods tend to produce contours overdetection, due to tilled or cultivated fields. Another method, that was used for agricultural ditch drainage networks [5], is based on detecting lines using LSD algorithm and completing the parcel network. The third approach consists in detecting contours by supervised classification, learning a contour model. Luminance, color and texture features are used. It was applied successfully in [6] to natural images and could be interesting for parcel boundaries detection.

Our goal is to detect field boundaries using a binary classification point of view (field *boundary* Vs. *non-boundary*). The field boundaries may present a high variability which is often not well represented in learning datasets. Thus, active learning (AL) methods [7] appear to be an appropriate approach for solving this problem, progressively enriching the model to adapt it locally to various field boundaries. In this study, different issues are pointed out : 1) what are the most appropriate image descriptors for boundary delimitation ? 2) How to ensure a good classification at large scales? An hydrid approach is proposed, that combines a region-based ap-

proach and active learning techniques. The random Forest classifier is used for classification and feature selection. The unsupervised RF margin is used as uncertainty measure to select the most uncertain samples to add. Results are processed on VHR Geoeye images acquired on May 2009, during vegetation growth period.

2. METHODOLOGY

2.1. Random Forest classifier

Random Forests (RF) [8] is an ensemble of decision trees built from T multiple bootstrapped training samples. It does not require assumptions on the distribution of the data, which is interesting when different types or scales of input features are used. For each node of a tree, a subset of features is randomly selected. The best feature with regard to Gini impurity measure [8] is used for node splitting. For an input instance, each tree gives a unit vote for the most popular class. The final label is determined by a majority vote of all trees. Random Forests also provide a measure of feature importance that is processed on OOB data (Out-Of-Bag samples) and it is based on the permutation importance measure [8].

2.1.1. Unsupervised margin

For binary classification, an unsupervised margin is used which is the difference of base classifiers votes for each class. Suppose that the training samples consist of pairs (x_i, y_j) where x_i is an instance and $y_j \in \{1, -1\}$ its true label, 1 and -1 corresponding to *boundary* and *non-boundary* classes respectively. The margin m_i of instance x_i is computed as :

$$m_i = \text{margin}(x_i) = \frac{v_{(i,1)} - v_{(i,-1)}}{\sum_{y_j \in \{1,-1\}} v_{(i,y_j)}} \quad (1)$$

where $v_{(i,y_j)}$ is the number of votes for the class y_j. The margin ranges from -1 to +1. A high positive and negative sample margin values indicate a high confidence in classifying the sample as *boundary* and *non-boundary* respectively. A margin value near 0 indicates a high uncertainty of the classifier.

2.2. Input features

Three groups of image features were used: spectral, gradient-based and textural. The four initial bands and a derived vegetation index (NDVI) were used as spectral features. Two gradient-based features were used: gradient preceded by a Gaussian filtering and an anisotropic gradient. As texture features, first order mean and variance, Gabor filters, sift-based filters [9] were processed both on panchromatic and multispectral bands. Panchromatic texture features are then resampled at the multispectral resolution.

2.3. Learning strategies

Our goal is to provide a binary classification of cultivated fields into *boundary* and *non-boundary* classes. The proposed methodology should be applicable at a large scale . At such scale, data-shift may happen between different parts of the image due to clouds, local context or varying field boundaries. Different learning strategies have been tested; 1) a simple RF model over one set of adjacent fields 2) a global model constructed using various sets of fields geographically distant 3) an active learning RF model based on an automatic enrichment of the global RF model. For all strategies, due to a very highly imbalanced dataset, the training dataset is balanced using as many *boundary* samples as *non-boundary* ones.

2.3.1. Simple RF model

For the simple RF model, a set of adjacent fields is divided randomly into training and test datasets. 10% of boundary pixels are kept for training. The same number of non-boundary samples are then used to balance the training dataset. It is the most favorable case since all field boundaries are represented in the training set and are located in the same area.

2.3.2. Global RF model

The global RF model uses 50% of fields, that are geographically distant, as training dataset. The remaining 50% unknown fields are used as test dataset. This strategy is more appropriate for an operational use since a field boundary database can be produced and used for the learning step.

2.3.3. Active learning RF model

Generally, when previous models are applied to unknown fields, that are geographically distant, and not represented in the training dataset, they behave poorly. This is a critical issue for classification at a large scale. To overcome this issue, we propose to use active learning (AL) techniques [7]. Active learning is an iterative procedure of selecting the most informative unlabeled samples while allowing to preserve compact training sets. The algorithm starts with a small number of labeled instances. A first classifier is produced. Then, based on a ranking score computed on a model outcome, a few unlabeled samples are chosen and *queries* are presented to an expert to label them. The algorithm runs iteratively and improves its decision until a stopping criterion is met. The key issue is to select the most valuable samples and reduce the number of queries. The active learning strategy has been, at a vast majority, used with SVM classification. Two criteria are used and often coupled: uncertainty and diversity. The uncertainty criterion corresponds to the algorithm confidence in correctly classifying a set of samples.

In this study, an active learning (AL) of Random Forest model is proposed. It allows to define automatically the most relevant samples to enrich the global RF model. The most informative unlabeled samples correspond to the most uncertain for the initial classifier. In our method, RF unsupervised margin is used as uncertainty measure. The low margin samples may either correspond to noisy samples, class boundaries, new cultivated crop or to a new type of field boundaries that are not represented in the training set. In our method, only one iteration is made and a random low margin sampling among unlabeled samples is applied. Besides, an automatic labeling of selected unlabeled samples is proposed. It combines a region-based approach and the unsupervised margin values (*cf.* Figure 1). A multi-scale segmentation is processed using initial multispectral bands [10]. The region-based approach is complementary with the classification since they not use the same input features. Low positive margin samples that correspond to a region boundary are labeled as *boundary*. Low negative margin instances are labeled as *non-boundary*. Samples are then sorted by increasing margin. The first n samples are selected for each class, for active learning, in order to keep a balanced training set.

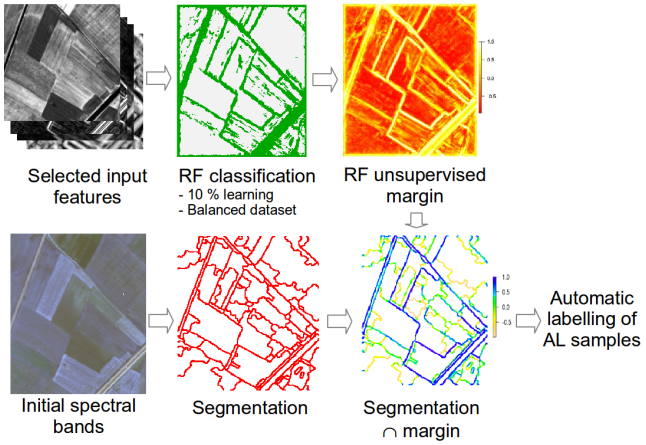


Fig. 1. Active learning using RF margin and region-based approach: automatic labelling of selected samples

3. RESULTS AND DISCUSSION

The study site is located in the North eastern Tunisia (Cap Bon region), over Lebna Catchment (≈ 210 km^2). 80 agricultural fields were studied. A VHR Geoeye image was acquired in May 2009 during the crop growth period. This period is not the most appropriate to distinguish the field boundaries, tillage period would be better for this purpose. The spatial resolution is 0.5m and 2m in Pan and MS bands(B,G,R,PIR) respectively. The ground truth has been processed by photo-interpretation on panchromatic image and dilated by a 3*3 structuring element.

3.1. Variable importance

The variable importance has been processed independently over 80 agricultural fields using simple models. Median, mean and standard deviation of variable importance have been calculated for more robustness. Results show that feature importance ranking may vary from one plot of fields to another. However, we can see that feature groups are always ranked similarly (*cf.* Figure 2). It appears that the more suitable image features for field boundary detection are Gabor filters processed on Panchromatic image, followed by gradient-based features over red and infrared channels and then 1^{st} order variance features processed on multispectral channels. SIFT-based features are the worst for this purpose since they tend to characterize the region content. These results were confirmed by applying the global RF model. For our experiments, the following input features were kept: variance, recursive gradient with Gaussian filtering and Gabor filters (Pan and MS).

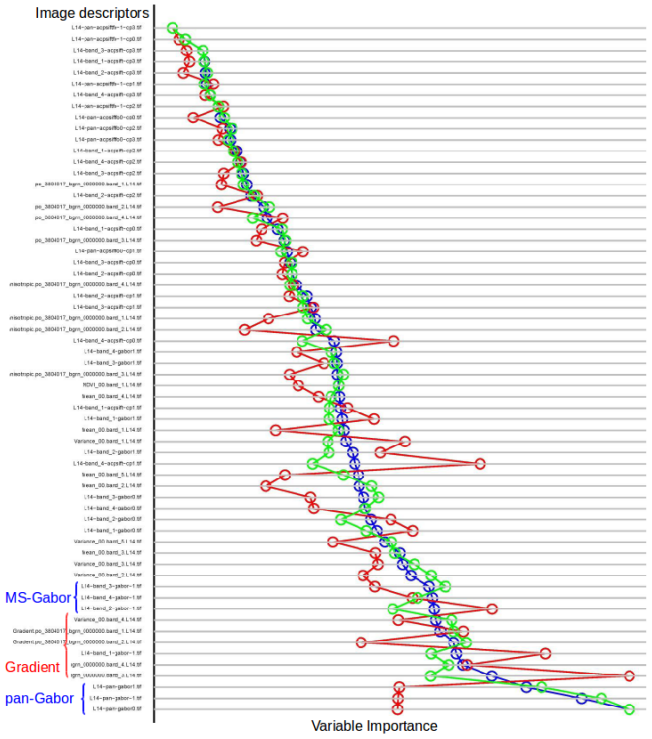


Fig. 2. variable importance for different input features. Median, mean and standard deviation values are shown in blue, green and red, respectively.

3.2. Classification results

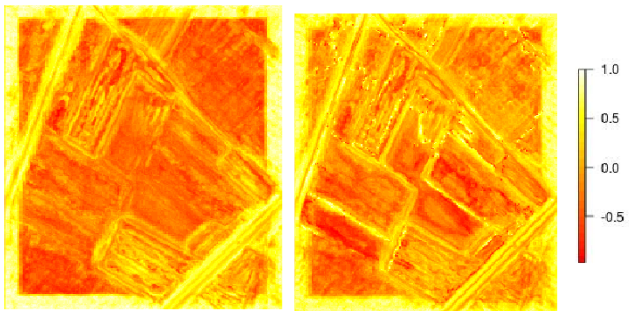
For AL RF global model, low margin threshold is fixed to 0.5. n=500 samples per class are selected for AL iteration. Results are shown on two agricultural sets of fields *L06 and L13* that are composed of 21 and 41 fields, respectively. Table 3.2 compares the accuracies (Average, produced and user

%	L06			L13		
	S	G	**AL**	S	G	**AL**
PA *"B"*	91.1	80.5	**82.1**	93.3	80.9	**83**
PA *"NB"*	95.7	72	**73.4**	93.2	61.2	**62.9**
PU *"B"*	98.7	90.9	**91.4**	94.7	73	**74.3**
PU *"NB"*	75.7	51.6	**54.2**	91.5	71.2	**74**
AA	93.4	76.3	**77.7**	93.3	71	**72.9**

Table 1. Comparison of RF global model and automatic RF AL model

accuracies) between RF simple, RF global and the enriched RF AL models. Classification accuracies are satisfying. *Non-boundary* class has lower accuracies due to the high variability of the *non-boundary* class. One can see that the simple model is the most favorable case. It is the most usual case in remote sensing literature but far from being operational to a large scale study. RF global model is more useful in practice since a field boundary database could be produced and used for the training step. As expected, RF AL model improves all the accuracies by 1.3 % to 2.8 %.

For visual clarity, AL model was processed to a simple RF model trained on a set of fields "A" and applied to a distant set of fields "B". One can see on Figure 3 that the margin values are higher with AL procedure, i.e. the classifier is more confident and undetected boundaries in the initial model are delineated with AL model.



(a) RF model A applied to B (b) RF AL model: enrichment from B

Fig. 3. Margin maps for a simple RF model (left) and an AL RF simple model (right

4. CONCLUSION

In this study we proposed an active learning procedure based on Random Forest model. The RF unsupervised margin is used as uncertainty measure. An automatic labeling of selected samples is proposed based on a combination between a region-based approach and the RF margin concept. First results are encouraging. The proposed method leads to a good localization of agricultural fields but presents some false positives. The obtained results are still not sufficient to directly derive a map of parcels but the obtained margin maps could be used in a global minimization framework as a data-term. The a priori term could integrate some criteria on the boundaries geometry. Further validation will be processed using VHR Pleiades images.

5. REFERENCES

[1] T. Blaschke, "Object based image analysis for remote sensing," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 65, no. 1, pp. 2 – 16, 2010.

[2] U.C. Benz, P. Hofmann, G. Willhauck, I. Lingenfelder, and M. Heynen, "Multiresolution, object-oriented fuzzy analysis of remote sensing data for gis-ready information," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 58, no. 3-4, pp. 239–258, 2004.

[3] R. Trias-Sanz, G. Stamon, and J. Louchet, "Using colour, texture, and hierarchial segmentation for high-resolution remote sensing," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 63, no. 2, pp. 156 – 168, 2008.

[4] M. Baatz and M. *Schäpe*, *Multiresolution segmentation - An optimization approach for high quality multi-scale image segmentation*, Strobl, J., Blaschke, T., Griesebner, G. (Eds) Angewandte Geographische Informations - Verarbeitung XII. Wichmann Verlag, Karlsruhe, 2000.

[5] J.-S. Bailly and F. Levavasseur, "Potential of linear features detection in a mediterranean landscape from 3d vhr optical data: Application to terrace walls," in *IEEE IGARSS*, 2012, pp. 7110–7113.

[6] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik, "Contour detection and hierarchical image segmentation," *IEEE PAMI*, vol. 33, no. 5, pp. 898–916, May 2011.

[7] D. Tuia, M. Volpi, L. Copa, M. Kanevski, and J. Muñoz-Marí, "A survey of active learning algorithms for supervised remote sensing image classification," *IEEE JSTARS*, vol. 5, no. 3, pp. 606–617, 2011.

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

[9] A. Le Bris, "Extraction of vineyards out of aerial orthoimage using texture information," in *22nd Congress of the ISPRS*, Melbourne, Australia, 2012, vol. I-3, pp. 383–388, ISPRS.

[10] L. Guigues, J.P. Cocquerez, and H. Le Men, "Scale-sets image analysis," *IJCV*, vol. 68, no. 3, pp. 289–317, 2006.