ADAPTIVE SPATIAL SAMPLING WITH ACTIVE RANDOM FOREST FOR OBJECT-ORIENTED LANDSLIDE MAPPING

A. Stumpf [a, b, c], *N. Lachiche* [d], *N. Kerle* [c], *Jean-Philippe Malet* [b], *A. Puissant* [a]

[a] Laboratoire Image, Ville, Environnement, CNRS ERL 7230, Université de Strasbourg, 3 rue de l'Argonne, F- 67083 Strasbourg, France

[b] Institut de Physique du Globe de Strasbourg, CNRS UMR 7516, Université de Strasbourg / EOST, 5 rue René Descartes, F-67084 Strasbourg, France

[c] ITC-Faculty of Geo-Information Science and Earth Observation, University of Twente, Department of Earth Systems Analysis, Hengelosestraat 99, P.O. Box 6, Enschede, 7500 AA, The Netherlands

[d] Image Sciences, Computer Sciences and Remote Sensing Laboratory, CNRS UMR 7005, Université de Strasbourg, Bd Sébastien Brant, BP 10413, F-67412 Illkirch, France

ABSTRACT

Active learning (AL) is a powerful framework to reduce labeling costs in supervised classification. However, spatial constraints on the sampling design have not yet received much attention and still pose problems for the application of AL on remote sensing data. In this study such issues are addressed in the context of landslide inventory mapping and it is shown that region-based query functions that focus the labeling efforts on compact spatial batches may provide several advantages over point-wise queries.

Index Terms—active learning, spatial sampling landslide inventory mapping, object-oriented image analysis

1. INTRODUCTION

Landslide inventory mapping is indispensable for landslide hazard and risk assessment, the quantification of erosion rates, and seismic hazard assessment. Very-High Resolution (VHR) remote sensing imagery to perform such tasks is now commonly available, whereas robust operational techniques to accelerate the mapping process are still lacking. A number of recent studies addressed this problem, developing object-oriented rule-based classifiers [1-3] that function without training data but often require adjustments of multiple thresholds when applied to new image data or geographic areas. Recently proposed supervised approaches fall into pixel-based studies using parametric classifiers [4] and approaches using object-oriented features and non-parametric learning algorithms [5]. While only object-oriented approaches can fully exploit the rich textural and spatial information content of VHR resolution images, both techniques still require an extensive amount of training data.

The acquisition of training data is typically associated with significant costs and an optimal training set should therefore be as small as possible, while still providing the representative characteristics for the target classes. In the domain of machine learning, AL has evolved as a key concept to reduce the labeling costs [6] and recently has seen successful applications in remote sensing [7]. The general underlying idea of AL is to initialize a machine learning model with a small training set, and to subsequently exploit the model state and the data structure to iteratively select the most valuable sample that should be labelled by the user and added in the training set. With relatively few queries and labelled samples, an AL strategy should ideally yield at least the same accuracy than an equivalent classifier trained with many randomly selected samples.

The iterative retraining of the classifier is typically a computational bottleneck of AL and it has been demonstrated that batch-mode query functions that consider the uncertainty and diversity of the samples [8, 9] are able to reduce the number of iterations significantly. Recent studies proposed AL strategies for semantic image segmentation by iteratively exploring hierarchical data representations [10, 11], and highlighted the value of integrating additional spatio-contextual features.

Most proposed approaches commonly query samples according to their position in the feature space and, assuming that the labeling costs of the queries are mutually independent, do not explicitly consider their position in geographic space. This typically yields a spatially dispersed distribution of samples and may incur the risk to revisit (during image interpretation or field work) the approximately same spatial location several times. Human scene interpretation generally involves the assessment of high-level contextual features [12], and this is in particular true for the identification of landslides in remote sensing images [13]. However, point-wise queries do not exploit the full interpreter knowledge of the spatial context around a particular point, suggesting region-based queries as a strategy that is more aligned with human perception. Fig. 1 summarizes the time expenditure of an image interpreter in a small experiment labeling (i) 20 queries selected by a region-based AL routine with a marker (Fig. 1a), and (ii) 20

queries selected by a point-wise AL routine individually (Fig. 1b). This experiment shows that the maker-based labeling of regions is about 15 times slower than the labeling of individual segments. On the other hand, maker-based labeling provides ~1900 labeled segments per iteration versus 1 for the segment-based labeling.

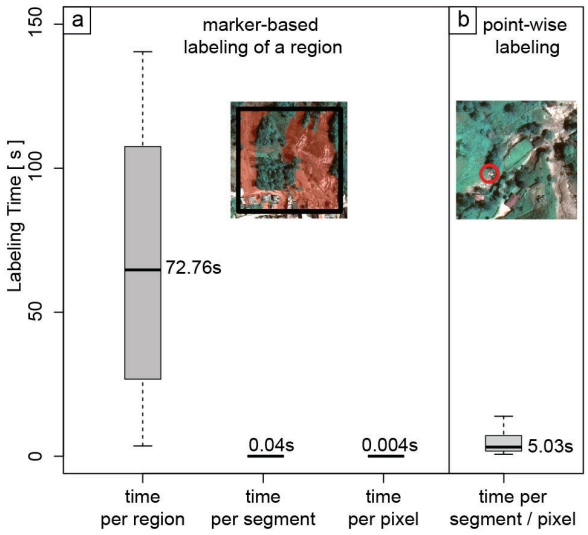

Fig. 1 Comparison of time expenditure for (a) marker-based (red marking on a 32400 m^2 region) and (b) point-wise (center of the red circle) labeling. Statistics were derived from different localities over 20 AL iterations, respectively.

The authors of [14] considered spatial distances among points to enhance margin-based sampling for point-wise queries, yielding an algorithm that encourages a dispersed spatial distribution of the training points. Liu et al. [15] formulated an AL heuristic as a traveling salesman problem in order to minimize, in each iteration, the travel distances to the most uncertain points.

To reduce the labeling costs in supervised landslide mapping, this study targets the development of a region-based AL approach that enables to limit the user's attention to a few interesting subsets of the study area, rather than querying individual points. It extends upon previous work on object-oriented landslide mapping from VHR remote sensing images [1, 5] and adopts the Random Forest (RF) framework [16], which already showed promising results as a base classifier for AL heuristics [17].

2. DATA AND METHODS

2.1. Test site and data

The study area is located in the Brazilian Serrana mountains, north of Nova Friburgo in the Rio de Janeiro state. On 11-12 January 2011 the area was affected by heavy rainfalls which triggered thousands of landslides and killed more than 1500 people [18]. Geoeye-1 images of the region were recorded on the 20 January 2011 and before the event on 26 May 2010. The bi-temporal dataset was used together with a global digital elevation model at 30 m resolution [19]. A reference landslide inventory was obtained through visual interpretation of both images. For all experiments, a subset area of 9 km^2 was selected.

2.2. Image segmentation and feature extraction

Image segmentation was performed on the post-landslide image with equal weights of the panchromatic and multispectral bands using the eCognition's multi-resolution segmentation [20] with a scale factor of 20. Regarding the targeted landslides this corresponds to a strong over-segmentation, which, compared to a coarser segmentation, reduces the impact of mixed segments on the accuracy of the final classified map [5]. For each of the resulting objects (~400.000), 106 features (e.g. intensity values, band ratios, texture, shape, neighborhood relationships, topographic location) were calculated and the class (landslide, non-landslide) was assigned considering the overlap (majority vote) with the reference inventory.

2.3. AL approaches

The adopted AL approach follows the query-by-committee (QBC) strategy where the next sample is chosen according to maximum disagreement of the ensemble [21]. A RF with 500 fully grown trees was adopted as a base classifier and the vote entropy (Eq. 1) was used as a measure of the classification uncertainty.

$$H = -\sum_{i \in (0,1)} p_i * \log(p_i) \qquad (1)$$

where p_0 and p_1 are the fractions of the trees that classify a sample as non-landslide and landslide, respectively. Based on this notion of uncertainty, three approaches differing mainly in their query function were tested.

The most basic approach is to start from only two samples (one per class) and choose in each iteration the segment with the highest entropy ($\arg\max H$) for labeling: this query per segment strategy is referred as QBC_{PS}.

Batch-mode AL can reduce computationally expensive classifier retraining [8, 9], and region-based batch queries enable the user to focus on one region and label hundreds of objects with a marker in relatively short time (Fig. 1). Therefore, a second approach is to query each time the region with the highest standard deviation of the entropy ($\arg\max \sigma_H$), thereby encouraging the presence of uncertain and diverse samples within the batch; this approach is referred as QBC_{PR}.

As a more explicit method to enforce uncertainty and diversity of the spatial batch, a third approach is to choose out of the m ($m > 1$) regions with the highest entropy the one with the highest diversity. To quantify diversity for each unlabeled point within the m pre-selected regions, the Euclidean distance (in feature space) to the nearest training point is computed. In each iteration the distance

computation considers only features with a RF variable importance greater than 1%. The variable importance is the mean decrease in accuracy if the variable is randomly permuted [22]. The final sampling region is chosen out of *m* candidates according to the maximum standard deviation of the Euclidian distances ($\arg\max \sigma_{DIST}$), to favor both diversity within the batch and distinctiveness from the already known training set. We refer to this heuristic further as QBC_{PR}+*D* and set *m*=3 for the experimental evaluation.

All approaches were implemented in `R` [23] and the diameter of the search window was kept at 180 m for both region-based approaches. Labels were queried from the reference landslide inventory.

3. RESULTS AND DISCUSSION

The three approaches were compared in terms of accuracy gains (F-measure) per iterations and algorithm runtime averaged over 10 randomly seeded runs. The runs were initialized with segments (or segment center-points for the region-based queries) sampled randomly with stratification to ensure the presence of at least one landslide example in the initial set.

Fig. 2 (a, b) indicate that all AL techniques clearly outperform random sampling. It also demonstrates that for the same accuracy level the region-based strategies require significantly less iterations as the point-wise approach, and especially the QBC_{PR}+*D* heuristic reduces the necessary number of iterations by a factor of 11-25 (depending on the accuracy level) compared to the point-wise QBC_{PS}.

Taking into account the estimated differences in labeling time (Fig. 1) QBC_{PR}+*D* provides slightly better accuracies in the same labeling time after 7 iterations (~500 s). A comparison of the results for the first 500 s remains rather inconclusive due to the high standard deviations (Fig. 2 c). Although the required labeling time is a clear indicator for the efficiency of AL algorithms it is important also to consider how the annotation time is distributed over individual queries to the user. For the current experiments labels were queried from a reference inventory and hence annotator accuracy as well as the annotator behavior could not be addressed. However, it seems evident that markings on 10 image patches place significantly less stress upon the user than 200 point wise yes-no decisions. Regarding algorithm runtime time the region-based approaches do not show significant differences but clearly outperform the per-segment query strategy (Fig. 2 d). This must be attributed to the strongly reduced number of iterations as a general feature of batch-mode AL.

Class-imbalance is an inherent issue in landslide mapping and may induce a bias toward the non-landslide class in the final map. Stratified bootstrap sampling can be used to adjust the class-ratio for the construction of each tree in the ensemble toward a ratio β (non-landslide/landslides), where user's and producer's accuracies balance.

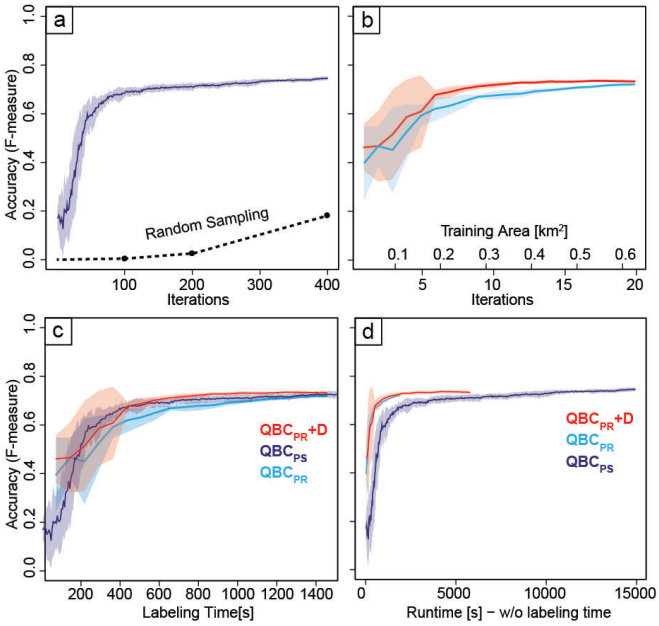

Fig. 2 (a,b) Comparison of the AL heuristics and random sampling in terms of accuracy gains (μ and σ over 10 runs) per query iterations (c) labeling time estimated according to the time expenditure per iteration as indicated in Fig. 1 and (d) the algorithm runtime without labeling time.

β is generally unknown but may be approximated on subsets of the training data. Each classification tree is built on a bootstrap sample that omits ~37% of the original training data. Those out-of-bag (OOB) samples can be used to assess the generalization error [16] and were adopted here to estimate β. For this purpose an iterative scheme that records the changes in user's and producer's according to the OOB samples, while systematically altering β, was implemented. A value of β=1.4 yielded a balance of both accuracy was observed and was used in the training of the final RF.

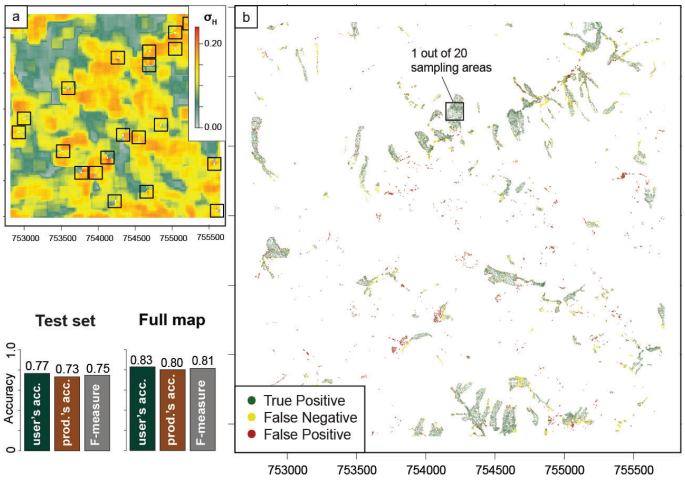

Fig. 3 (a) Queried regions after 20 AL iterations. (b) Results with QBC_{PR}+*D* after 20 iterations. Objects are represented by their center points.

Fig. 3 indicates that the estimated β yielded a balance of user's and producer's accuracy on the unlabeled test set with a F-measure of 0.75. Taking into account the segments that were labeled for training during 20 iterations the final map provided an accuracy of 0.81.

In summary, the proposed AL techniques allow to significantly reduce the amount of required training data. Region-based approaches justify an increased labeling time per iteration as they can reduce the number of required iterations by a factor of 11-25. They thereby reduce the overall labeling costs as well as the algorithm runtime, and provide significantly better map results when compared to point-wise queries. A heuristic that explicitly encourages diversity of the queried spatial batches (QBC_{PR}+D) provided the best results in the tested setting, but a more systematic assessment of parameters such as window size and variable importance are recommended. Further research should also address tests on different datasets and the integration of stratified bootstrapping into the AL routine.

Acknowledgements
The project SafeLand (Grant Agreement No. 226479) funded by the 7th Framework Programme of the European Commission and the project FOSTER funded by the French Research Agency (Contract ANR Cosinus, 2011–2013) supported this work and are thankfully acknowledged.

11. REFERENCES

[1] T. Martha, N. Kerle, C. J. van Westen, and K. Kumar, "Characterising spectral, spatial and morphometric properties of landslides for semi-automatic detection using object-oriented methods," *Geomorphology,* vol. 116, pp. 24-36 2010.

[2] P. Lu, A. Stumpf, N. Kerle, and N. Casagli, "Object-Oriented Change Detection for Landslide Rapid Mapping," *Geoscience and Remote Sensing Letters, IEEE,* vol. 8, pp. 701-705, 2011.

[3] T. Lahousse, K. T. Chang, and Y. H. Lin, "Landslide mapping with multi-scale object-based image analysis – a case study in the Baichi watershed, Taiwan," *Nat. Hazards Earth Syst. Sci.,* vol. 11, pp. 2715-2726, 2011.

[4] A. C. Mondini, F. Guzzetti, P. Reichenbach, M. Rossi, M. Cardinali, and F. Ardizzone, "Semi-automatic recognition and mapping of rainfall induced shallow landslides using optical satellite images," *Remote Sensing of Environment,* vol. 115, pp. 1743-1757, 2011.

[5] A. Stumpf and N. Kerle, "Object-oriented mapping of landslides using Random Forests," *Remote Sensing of Environment,* vol. 115, pp. 2564-2577, 2011.

[6] B. Settles, "Active Learning Literature Survey," *Computer Sciences Technical Report 1648,* p. 67, 2010.

[7] D. Tuia, M. Volpi, L. Copa, M. Kanevski, and J. Munoz-Mari, "A survey of active learning algorithms for supervised remote sensing image classification," *Selected Topics in Signal Processing, IEEE Journal of,* vol. 5, pp. 606-616, 2011.

[8] B. Demir, C. Persello, and L. Bruzzone, "Batch-Mode Active-Learning Methods for the Interactive Classification of Remote Sensing Images," *Geoscience and Remote Sensing, IEEE Transactions on,* vol. 49, pp. 1014-1031, 2011.

[9] M. Volpi, D. Tuia, and M. Kanevski, "Memory-Based Cluster Sampling for Remote Sensing Image Classification," *Geoscience and Remote Sensing, IEEE Transactions on,* vol. PP, pp. 1-11, 2012.

[10] J. Muñoz-Marí, D. Tuia, and G. Camps-Valls, "Semisupervised Classification of Remote Sensing Images With Active Queries," *Geoscience and Remote Sensing, IEEE Transactions on,* vol. PP, pp. 1-12, 2012.

[11] D. Tuia, J. Muñoz-Marí, and G. Camps-Valls, "Remote sensing image segmentation by active queries," *Pattern Recognition,* vol. 45, pp. 2180-2192, 2012.

[12] J. M. Henderson and A. Hollingworth, "High-Level Scene Perception," *Annual Review of Psychology,* vol. 50, pp. 243-271, 1999.

[13] R. Soeters and C. Van Westen, "Slope instability recognition, analysis and zonation," in *Landslides, investigation and mitigation.* vol. Special Report 247, A. K. Turner and R. L. Schuster, Eds., ed Washington, USA: Transportation Research Board, National Research Council, National Academy Press, 1996, pp. 129-177.

[14] E. Pasolli, F. Melgani, D. Tuia, F. Pacifici, and W. J. Emery, "Improving active learning methods using spatial information," in *Geoscience and Remote Sensing Symposium (IGARSS), 2011 IEEE International*, 2011, pp. 3923-3926.

[15] A. Liu, G. Jun, and J. Gho, "Spatially Cost-sensitive Active Learning," presented at the Ninth SIAM International Conference on Data Mining Sparks, Nevada, 2009.

[16] L. Breiman, "Random Forests," *Machine Learning,* vol. 45, pp. 5-32, 2001.

[17] A. Borisov, E. Tuv, and G. Runger, "Active Batch Learning with Stochastic Query-by-Forest (SQBF)," presented at the JMLR Workshop on Active Learning and Experimental Design, 2011.

[18] A. L. C. Netto, A. M. Sato, A. d. S. Avelar, L. G. G. Vianna, I. S. Araújo, D. L. C. Ferreira, P. H. Lima, A. P. A. Silva, and R. P. Silva, "January 2011: the extreme landslide disaster in Brazil," presented at the Second World Landslide Forum, Rome, Italy, 2011.

[19] ASTER-GDEM-VALIDATION-TEAM, "ASTER Global Digital Elevation Model Version 2 – Summary of Validation Results," METI/ERSDAC, NASA/LPDAAC, USGS/EROS, 2011.

[20] M. Baatz and A. Schäpe, "Multiresolution Segmentation – an optimization approach for high quality multi-scale image segmentation," presented at the Angewandte Geographische Informationsverarbeitung XII, Salzburg, 2000.

[21] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," presented at the Fifth Workshop on Computational Learning Theory, San Mateo, CA, 1992.

[22] K. Nicodemus, J. Malley, C. Strobl, and A. Ziegler, "The behaviour of random forest permutation-based variable importance measures under predictor correlation," *BMC Bioinformatics,* vol. 11, p. 110, 2010.

[23] R Development Core Team. R: A Language and Environment for Statistical Computing [Online]. Available: http://www.R-project.org