Interactive Design of Object Classifiers in Remote Sensing

Bertrand Le Saux
ONERA The French Aerospace Lab
F-91761 Palaiseau, France
Email: bertrand.le_saux@onera.fr

***Abstract*—This paper deals with the interactive design of generic classifiers for aerial images. In many real-life cases, object detectors that work are not available, due to a new geographical context or a need for a type of object unseen before. We propose an approach for on-line learning of such detectors using user interactions. Variants of gradient boosting and support-vector machine classification are proposed to cope with the problems raised by interactivity: unbalanced and partially mislabeled training data. We assess our framework for various visual classes (buildings, vegetation, cars, visual changes) on challenging data corresponding to several applications (SAR or optical sensors at various resolutions). We show that our model and algorithms outperform several state-of-the-art baselines for feature extraction and learning in remote sensing.**

I. Introduction

Satellite and aerial images are now widely produced and (thanks to popular web applications) commonly used by everyone for exploration and searching places. Yet the information usually comes from existing maps and manually-added annotations, while today's high resolution would allow to extract lots of visual information. To that end, detection of visual patterns and classification in remote sensing has been an active field of research for many years. The global structure of the resulting algorithms consists in extracting relevant features that can be thresholded (for a few examples: the fractal error [1], the distribution of linear segments [2] or texture-based conditional random fields [3], [4]) or used to feed a machine learning algorithm that delivers the classification (like Support Vector Machines - SVMs - [5]).

But out of the lab, for practical situations, people often miss the right classifier for their purpose. For example, for disasters or crisis management, even with philanthropic procedures like the International Charter on Space and Major Disasters, the best image is the one that is available when the problem occurs. A solution to this problem is interactive learning: the user defines by himself the pattern of interest and learns it on the image to classify. The system in [6] keeps the complete image context visible, then learns the searched concept by using only a few selected pixels. Recently, pixel-based approaches have led to successful developments thanks to active learning [7], [8], especially with multi-spectral data. Several approaches that take their inspiration from content-based image retrieval have also been proposed: they segment images to small patches and display a ranked list of patches that users have to tag as good or bad. PicSOM [9] is based on self-organizing maps, VisiMine [10] on naive Bayes classifiers and Ikona [11] uses SVMs for relevance feedback.

Our approach takes the best of both worlds: it combines an intuitive selection of patches in their geographic context and a fast learning of classifiers of visual patterns. It allows to design generic detectors of objects or visual concepts that can be refined by relevance feedback, and thus extends the approach of [12]. This paper is organized as follows. In section II we define the principles of the approach and detail the learning algorithms. In section III we present results that assess the choices we made. In section IV we analyze the genericness of our approach and present some extensions that lead to practical applications in challenging conditions, before ending by concluding remarks in section V.

II. Interactive Learning

A. Training sample collection

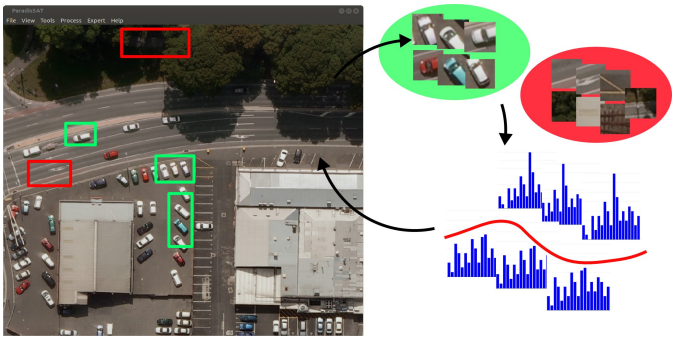


Fig. 1. On-line learning by analyst's selection of samples of what is looking for (green rectangles) and negative samples (red rectangles). These regions are segmented in small patches, from which meaningful features are extracted to constitute the training dataset. Detection results in Fig. 7.

Thanks to the development of web mapping applications like google maps and others, people are now used to geographic exploration of aerial image. Our interactive learning process follows this trend: the image analyst selects areas containing the object of interest and areas that do not contain it in a Geographic Information System (GIS) (cf. Fig. 1). First, these areas are segmented in small overlapping patches, thus allowing to harvest a large quantity of training samples. Second, patches are indexed by features that describe their content, typically d-dimensional vectors denoted by x_k. Along with their associated label, they constitute the training set $\{(x_k, y_k)_{1 \leq k \leq N}, x_k \in \mathbb{R}^d, y_k \in \{-1, 1\}\}$ of the learning algorithms.

B. Problems raised by interactivity

The aim of supervised learning is to build a function $f : \mathbb{R}^d \rightarrow \{-1, 1\}$ able to predict the label of an unknown descriptor x. Practically, this is performed by minimizing the misclassification risk:

$$R(f) = \frac{1}{N} \sum_k L(y_k, f(x_k)) \quad (1)$$

where $L(.,.)$ is a loss-function. The two major pitfalls of building interactively the training set are:

- Mislabeled data. If the user assigns a wrong label to a region or draws a region bigger than the target, some sample labels are false, so the learning algorithm should have good generalization properties.
- Unbalanced training sets. It is unlikely that the selection procedure yields in the same number of positive and negative samples. Typically, negative samples are much easier to find and should be over-numerous.

To deal with these problems, we present two approaches based on state-of-the-art learning algorithms.

C. On-line Gradient Boosting

In a nutshell, boosting is a machine learning approach which combines a set of weak classifiers f_m to build a good (*strong*) meta-classifier f:

$$f(x) = \sum_{m=1}^{M} f_m(x) \quad (2)$$

After the initial *Adaboost* algorithm [13], several variants have been proposed, including the *on-line boosting* used in [14] that offers an incremental mechanism. Boosting can be considered as an approximate gradient descent in the weak-classifier space [15], and this result yields in a more generic family of boosting methods named *on-line gradient-boost* [16]. They build the strong classifier f by minimizing the empirical risk of Eq. 1 with loss functions chosen among:

$$\begin{aligned} \text{exponential:} &\quad \exp(-yf(x)) \\ \text{logit:} &\quad \log(1 + \exp(-yf(x))) \\ \text{doomII:} &\quad 1 - \tanh(yf(x)) \\ \text{savage:} &\quad ((1 + \exp(2yf(x)))^2)^{-1} \\ \text{hinge:} &\quad \max(0, 1 - yf(x)) \end{aligned}$$

To deal with unbalanced data sets, we define a new set of loss functions that take into account the prior probabilities of the training sets:

$$L(x) \leftarrow \frac{L(x)}{p(y)} \quad (3)$$

where priors are estimated by counting the number of positive and negative samples in the training sets. This leads to weight the classification errors in the iterative minimization of risk according to the priors of each class, such giving more importance to under-represented samples.

Moreover, it has been shown [17] that non-convex loss functions (such that the DoomII function) that are less sensitive to mislabelings. Indeed, we show in Fig. 3 that such functions are more able to tolerate mislabelings for an image classification task.

D. Support Vector Machines

SVMs are a popular kernel method for minimizing risk. Even if incremental implementations of the SVM have been proposed [18], we chose to benefit from the fast computations of an implementation on Graphics Process Unit (GPU) of the SVM [19] to have tractable interaction times.

The soft margin principle allows some misclassifications due to mislabeling by setting an appropriate cost parameter. We handle unbalanced data in the same way as in boosting, by weighting different costs for each class according to their prior [20].

III. Experiments and Results

A. Man-made structure dataset

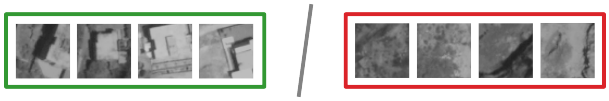


Fig. 2. Patch examples for the man-made structure dataset used for ground-truth: man-made structures (left) vs. clutter samples (right).

Man-made structure classification has many useful applications in the remote sensing domain, from urbanism (for urban development monitoring) to crisis management (for example refugee camp detection after a disaster). We build a ground-truth dataset by extracting 50x50 patches from a 2000x2000 QuickBird image (0.6m resolution) (cf. Fig. 2). It contains 615 positive samples (with houses and roads) and 1281 negative samples (woods and mountains).

In the following, this dataset is used to compute Receiver Operating Characteristic (ROC) curves for various classifiers: On-line Adaboost that is used as the baseline learning algorithm (Adaboost was used with Histograms Of Gradients - HOGs - for detection in remote sensing data in [14]), On-line Gradient-Boost with the prior-included DoomII loss-function and SVM with a Radial Basis Function (RBF) kernel. For each classification scheme, we average results on 5 runs of cross-validation using roughly 40% of the dataset for training and testing on the remaining samples. For each descriptor, the best parameters for each learning algorithm (i.e. number of selectors and number of weak learners by selectors for the on-line boosting, and kernel radius and cost for SVMs) are fine tuned by grid search.

B. Boosting classification

To test the capacity to handle mislabeled data of the different loss functions, we flipped in various proportions the class labels of the ground-truth data at training, and compared the classification rates. It appears on Fig. 3 that on-line gradient boosting with non-convex functions perform better than

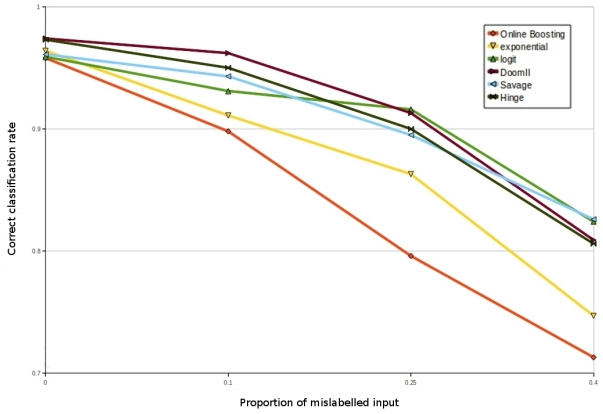


Fig. 3. Influence of training-data labeling errors on performances of on-line gradient-boost with various loss functions. Gradient-Boost learning is less sensitive to the mislabeling noise when non-convex loss functions are used.

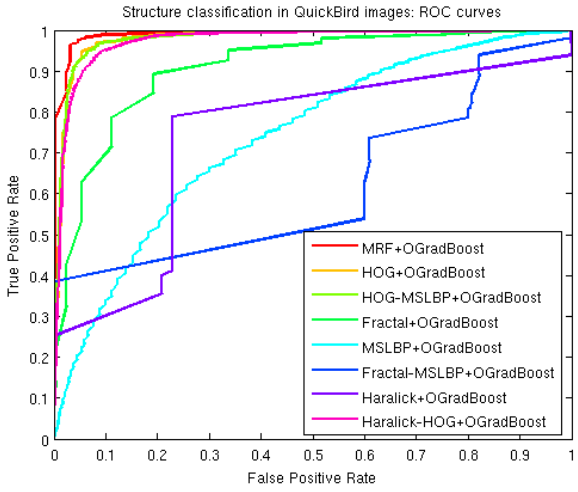


Fig. 4. ROC curves for man-made structure classification using On-line Gradient-Boost with the prior-included Doom-II loss-function. The comparison of various features shows that HOG-based features outperform the others.

others. DoomII has the highest performance with a limited amount of labeling noise, while with an increased mislabeling level ($> 20\%$ of mislabeled input) Savage performs better.

We then compared various image descriptors for a man-made structure classification task on the QuickBird dataset defined in III-A. We used image features commonly used in remote sensing for this task: Haralick features for texture description [21], multi-scale Linear Binary Patterns (MSLBP) [9], fractal error [1], HOGs [22], [9] and Markov Random Fields (MRF) [4] that are statistics computed on HOGs.

Fig. 4 shows ROC curves for On-line Gradient-Boost with the prior-included DoomII loss-function. It appears that all HOG-based features or combination of features outperform the other image descriptors, and that this learning scheme has the potential to discriminate man-made objects in the image (Area Under Curve - AUC - above 90%).

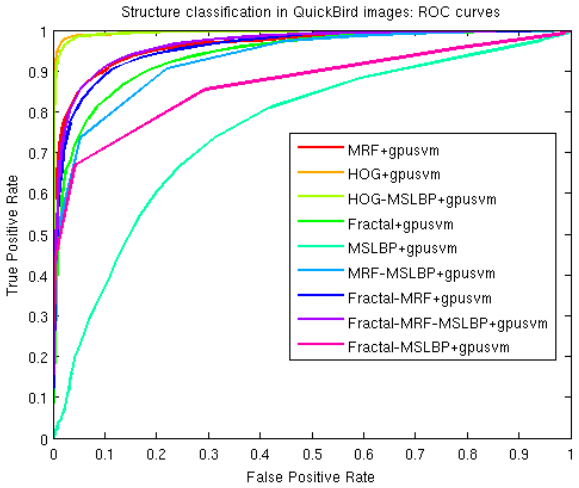


Fig. 5. ROC curves for man-made structure classification using SVM with a Radial Basis Function. The comparison of various features shows that HOGs outperform the others.

C. SVM classification

In Fig 5, the image descriptors of section III-B are compared using a SVM with a RBF kernel. HOGs are the descriptors that allow to obtain the best classification rates, slightly in front of MRFs. Both perform far better than other commonly used descriptors.

D. Overall classification

The best combinations of descriptors and learning algorithms are now compared. Table I compiles several performance measures for the various combinations: accuracy rates (both on the training and test data to emphasize overfitting if any), F1 score for and Area Under Curve (AUC). All algorithms obtain excellent results (AUC$> 90\%$) for HOG-

feature	classif.	train. accuracy	test accuracy	F1	AUC
MRF	Grad-Boost	0.975	0.970	0.970	0.993
HOG	Grad-Boost	0.956	0.950	0.915	0.980
HOG-MSLBP	Grad-Boost	0.957	0.945	0.921	0.980
Fractal	Grad-Boost	0.887	0.869	0.766	0.914
MSLBP	Grad-Boost	0.744	0.700	0.546	0.739
Haralick	Grad-Boost	0.797	0.784	0.661	0.856
Fractal-MSLBP	Grad-Boost	0.536	0.536	0.803	0.715
Haralick-HOG	Grad-Boost	0.794	0.784	0.661	0.860
MRF	AdaBoost	0.960	0.962	0.954	0.990
HOG	AdaBoost	0.914	0.910	0.886	0.962
HOG-MSLBP	AdaBoost	0.933	0.923	0.892	0.966
Fractal	AdaBoost	0.855	0.837	0.758	0.900
MSLBP	AdaBoost	0.662	0.651	0.457	0.688
Haralick	AdaBoost	0.754	0.745	0.599	0.590
MRF	SVM	0.915	0.924	0.880	0.965
HOG	SVM	0.990	0.977	0.965	0.997
HOG-MSLBP	SVM	0.989	0.976	0.965	0.996
MRF-MSLBP	SVM	0.934	0.927	0.886	0.961
Fractal	SVM	0.997	0.884	0.830	0.940
Fractal-MSLBP	SVM	0.998	0.885	0.824	0.925
MSLBP	SVM	0.721	0.725	0.589	0.764

TABLE I. COMPARISON OF COMBINATIONS (FEATURE + CLASSIFYING SCHEME) ACCORDING TO VARIOUS PERFORMANCE MEASURES: ACCURACIES ON BOTH THE TRAINING AND TEST SETS, F1-SCORE, AREA UNDER CURVE (AUC).

based features and (narrowly) the fractal error, thus pointing out obviously the feature of choice for describing the image content. Moreover the combination of different features does not usually improve the results, except for those which perform poorly.

Both SVM and Gradient Boost perform significantly better than the standard adaboost, thus showing that the mechanisms proposed in section II are efficient to control the unbalance of the training sets. SVM with HOGs is the best combination according to the AUC measure, while On-line Gradient Boost with (HOG-based) MRFs obtains the best F1 score. The significant difference between training and test accuracies for SVM with HOGs suggests that the SVM has slightly overfitted. These results hint that gradient orientation statistics are the most discriminant descriptor for discriminating structures in aerial images.

Fig. 6 shows ROC curves for these combinations of descriptors and learning algorithms, restrained to the upper-left domain of the ROC-space for better distinguishing between curves. When an operating point that tolerates a few false alarms is chosen, both On-line Gradient Boost and SVM are equivalent. However, SVMs allow to obtain a better precision at near-zero fall-out.

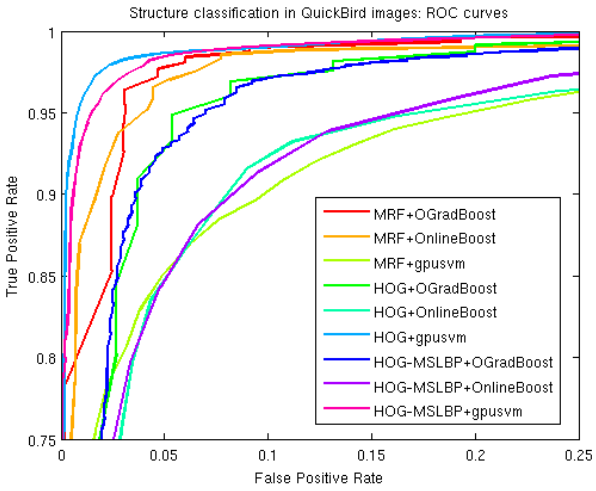


Fig. 6. ROC curves for classification using SVM and two flavors of Boosting on most-informative features (zoom on the $[0.75; 1] \times [0.75; 1]$ domain for better visualization).

IV. Analysis

A. What kind of visual objects can be learned ?

The experiments showed that both proposed algorithms are able to learn structures using HOGs, in conditions that are similar to interactive learning (especially unbalanced datasets). More precisely, regular man-made structures like buildings can be retrieved in images of resolution from 20m to 0.1m, as shown in [12]. Vegetation (such as tree foliage) can also be detected in the same wide range of resolutions, thanks to the isotropic nature of edge gradients in such patches. For example, Fig 9 shows detection results for both objects in 0.2m and 0.05m images.

With this framework, smaller objects like cars are detectable in high-resolution only. Two mechanisms are useful: more details to build the model, but also the fact that more sample patches are collected, thus preventing overfitting on singular data. Fig. 7 show car detections in an aerial image of resolution 0.1m in a challenging urban environment. Training areas were shown in Fig 1. The model was able to capture the appearance of cars at various orientations, even if a few false alarms appear on same-scale objects like air handler units.

In the following, we show that even more complex visual objects can be learned: visual changes between images.

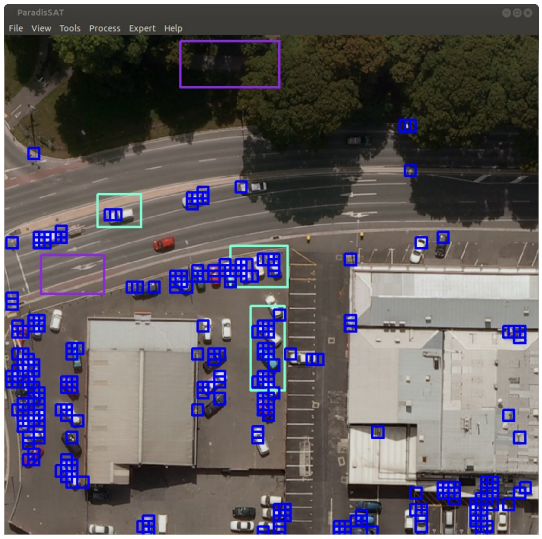


Fig. 7. Result of the interactive learning process of Fig. 1 with detections of cars (blue squares) on an orthorectified aerial image of resolution 0.1m.

B. Change detection

Change detection in SAR imagery consists in identifying new buildings or destruction between two images at two different dates [23]. Changes can be considered as visual patterns and learned by our approach with only minor adaptations [24]. Samples provided by the image analyst consist in modified areas (considered as positive samples) and areas that remain visually similar. A pre-processing step is required to generate a single image to compute features on: the Generalize-Likelihood Ratio Test (GLRT) [25] provides an estimate of the similarity of pixel distributions at pixel level. Then the learning process is the same as before, patches are extracted from the GLRT map and HOG features are computed to train the classifiers and thus estimate the spatial variations of changes.

This allows to distinguish between real changes of over-ground objects and some regular changes that are due to the geographic context (for instance the regular orientation of the streets) or to data capture (for example the registration error that always appears when viewing angles differ between the two image captures). Fig. 8 shows the method can retrieve various examples, such as new buildings or solar panels, even in dense urban areas.

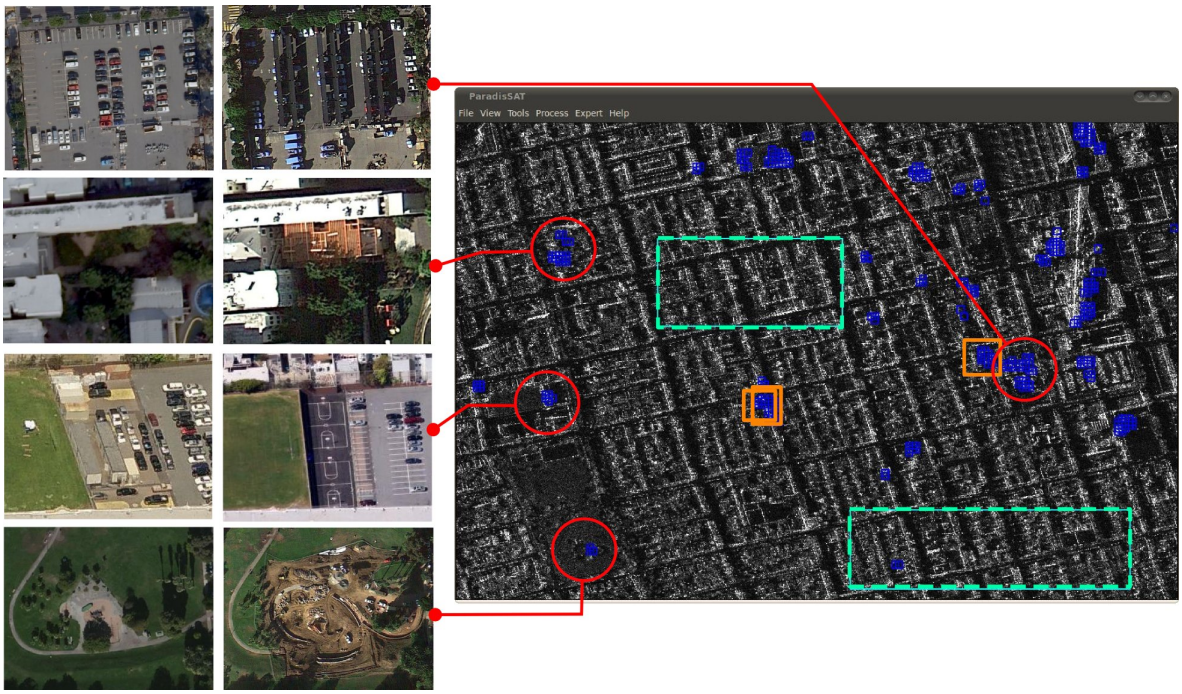


Fig. 8. Change detection in urban environment by comparison of TerraSAR-X images at two different dates. On right, SAR images: new building areas (orange rectangles) and unchanged areas (dashed green rectangles) are selected for training, while detection results are displayed using rectangle squares. Left column: checking (using more user-friendly optical images) of the detections (from top to bottom: solar panels, new buildings, playground construction, community park works).

C. A bit further: classifier adaptation for on-board camera

Fig. 9-a shows the interactive design of a detector in which, instead of a unique aerial or satellite image, image data are an ortho-rectified mosaïc built from the frames of a video captured by an Unmanned Aerial Vehicle (UAV) flying over the area to monitor. The objective is two-fold.

First, at mapping step, trees and vegetation are obstacles for the UAV landing, while houses and buildings can be considered either as targets or obstacles. On-site detectors allow to classify the area and plan safe paths and regions for subsequent flights (cf. Fig. 9-a).

Second, for navigation, the classifier is adapted to the video domain for detecting targets or obstacles as soon as they appear in the field of view [26]. This consists in projecting pixels from the video frames into the orthomosaïc geometry. Using projective coordinates, video pixels are denoted by $m'_k = (u', v', 1)^\top$ and projected points by $m_k = (u, v, 1)^\top$. Given the 3D location and the attitude of the UAV, the transform between the local coordinate systems of the camera and the world is defined by $H_{R,t} = R - \frac{tn^\top}{d}$, where R is the rotation matrix that depends on the UAV attitude; t the translation vector between the two origins; n the normal to the orthomosaïc plane and d the UAV altitude It is therefore possible to compute the homography that relates video pixels to projected points according to:

$$m_p = K_{\mathrm{OM}} \cdot H_{R,t} \cdot K_{\mathrm{Cam}}^{-1} \cdot m'_p \qquad (4)$$

where K_{Cam} is the intrinsic-parameter matrix of the on-board camera and K_{OM} is the transform matrix that encodes the change of origin and resolution between the world and the orthomosaïc image. The classifier is then applied on HOGs computed on patches interpolated from the rectified pixels.

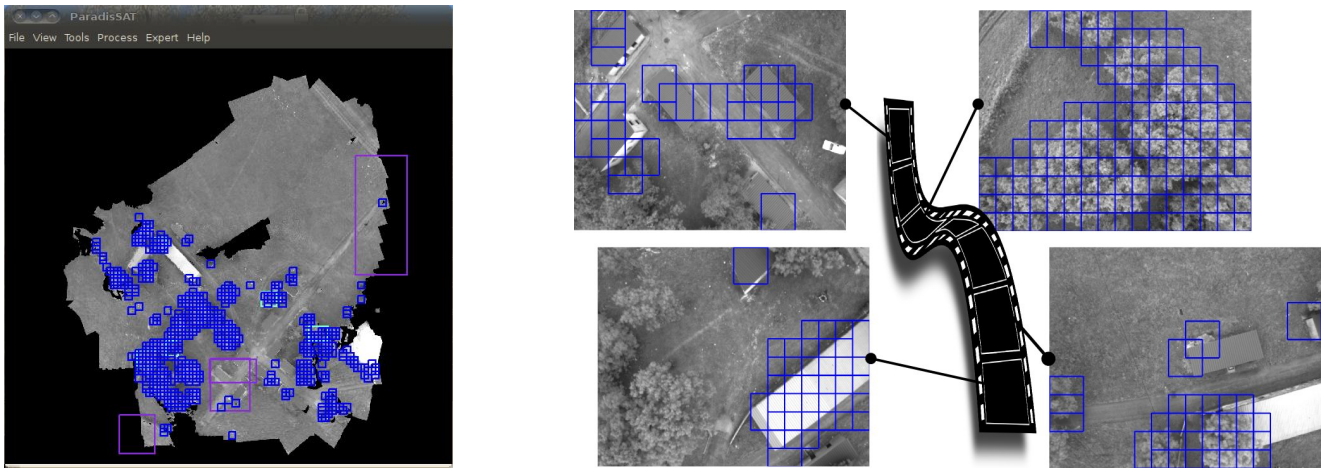
Fig. 9-b shows detection results in video-frames after such an adaptation. The two detectors (vegetation and building) were learned during a previous flight.

V. Conclusion

We presented an approach for interactive building of object detectors in aerial images, that combine user-friendly on-line collection of samples, HOG-based representation of objects and interactive-oriented learning methods based on gradient boosting or SVMs. The various use-cases explored in this paper (satellite or aerial images at various resolutions, video, change detection) assess the genericness of our approach to build detectors of visual patterns. It aims to transfer the design of these detectors from the lab to the end-user, who is the more able to define what is looked for. Further works will investigate how to distinguish between a larger variety of classes and concepts in the images, for example by using part-based models for more complexity and transfer learning for building detectors for classes with few samples from previously-made classifiers.

Acknowledgment

The authors would like to thank DigitalGlobe, Astrium Services, and USGS for providing TerraSAR-X images used in this study, and the IEEE GRSS Data Fusion Technical Committee for organizing the 2012 Data Fusion Contest. They would also like to thank New Zealand Aerial Mapping Limited for providing the orthorectified aerial image over Christchurch.



(a) (b)

Fig. 9. (a) Result of an interactive detection of vegetation (trees that are landing obstacles) in an orthomosaïc (resolution 0.2m) for mapping (blue squares) (b) after geometric adaptation of detectors learned interactively for on-board use in the video-domain, detection results (blue squares superimposed to the video frames) for buildings (on left) and vegetation (on right).

REFERENCES

[1] D. Chenoweth, B. Cooper, and J. Selvage, "Aerial image analysis using fractal-based models," in *Proc. of Aerospace Applications Conference*, 1995.

[2] C. Unsalan and K. L. Boyer, "Classifying land development in high-resolution satellite imagery using hybrid structuralmultispectral features," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 42, no. 12, 2004.

[3] A. Lorette, X. Descombes, and J. Zerubia, "Texture analysis through a markovian modelling and fuzzy classification: Application to urban area extraction from satellite images," *International Journal of Computer Vision*, vol. 36, no. 3, 2000.

[4] S. Kumar and M. Hebert, "Man-made structure detection in natural images using a causal multiscale random field," in *Proc. of Computer Vision and Pattern Recognition*, Madison, Wisconsin, 2003, pp. 119–126.

[5] G. Mountrakis, J. Im, and C. Ogole, "Support vector machines in remote sensing: A review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 66, pp. 247–259, 2011.

[6] M. Schröder, H. Rehrauer, K. Seidel, and M. Datcu, "Interactive learning and probabilistic retrieval in remote sensing image archives," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 38, no. 5, pp. 2288–2298, May 2000.

[7] L. Bruzzone and C. Persello, "Active learning for classification of remote sensing images," in *Proc. of International Geoscience and Remote Sensing Symposium*, Cape Town, South Africa, 2009.

[8] D. Tuia, F. Ratle, F. Pacifici, M. Kanevski, and W. Emery, "Active learning methods for remote sensing image classification," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 47, no. 7, pp. 2218–2232, 2009.

[9] M. Molinier, J. Laaksonen, and T. Häme, "Detecting man-made structures and changes in satellite images with a content-based information retrieval system built on self-organizing maps," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 45, no. 4, pp. 861–874, April 2007.

[10] K. Koperski, G. Marchisio, S. Aksoy, and S. Tusk, "Visimine: interactive mining in image databases," in *Proc. of International Geoscience And Remote Sensing Symposium*, vol. 3, Toronto, Canada, 2002, pp. 1810–1812.

[11] M. Ferecatu and N. Boujemaa, "Interactive remote-sensing image retrieval using active relevance feedback," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 45, no. 4, pp. 818–826, april 2007.

[12] N. Chauffert, J. Israël, and B. Le Saux, "Boosting for interactive man-made structure classification," in *Proc. of IEEE International Geoscience and Remote Sensing Symposium*, Munich, Germany, july 2012.

[13] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, 1997.

[14] T. T. Nguyen, H. Grabner, B. Gruber, and H. Bischof, "On-line boosting for car detection from aerial images," in *IEEE International Conference on Research, Inovation and Vision for the Future (RIVF'07)*, 2007, pp. 87–95.

[15] L. Mason, J. Baxter, P. Bartlett, and M. Frean, "Boosting algorithms as gradient descent," *Advances in Neural Information Processing Systems*, vol. 12, pp. 512–518, 2000.

[16] C. Leistner, A. Saffari, P. Roth, and H. Bischof, "On robustness of on-line boosting - a competitive study," in *Proceedings of ICCV Workshop on On-line Learning for Computer Vision*, Kyoto, Japan, 2009.

[17] P. Long and R. Servedio, "Random classification noise defeats all convex potential boosters," *Machine Learning*, vol. 78, no. 3, pp. 287–304, 2010.

[18] A. Bordes and L. Bottou, "The huller: a simple and efficient online svm," in *Machine Learning: ECML*, 2005, pp. 505–512.

[19] B. Catanzaro, N. Sundaram, and K. Keutzer, "Fast support vector machine training and classification on graphics processor," in *Proc. Int. Conf. Machine Learning*, Helsinki, Finland, 2008, pp. 104–111.

[20] K. Veropoulos, C. Campbell, and N. Christianini, "Controlling the sensitivity of support vector machines," in *Proc. of the International Joint Conference on AI*, 1999, pp. 55–60.

[21] X. Perrotton, M. Sturzel, and M. Roux, "Automatic object detection on aerial images using local descriptors and image synthesis," in *Proc. of International Conference on Vision Systems*, Santorini, Greece, 2008.

[22] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. of Computer Vision and Pattern Recognition*, Washington DC, USA, 2005, pp. 886–893.

[23] L. Bruzzone and D. Prieto, "Automatic analysis of the difference image for unsupervised change detection," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 38, no. 3, pp. 1171–1182, 2000.

[24] B. Le Saux and H. Randrianarivo, "Urban change detection in SAR images by interactive learning," in *Proc. of International Geoscience and Remote Sensing Symposium*, Melbourne, Australia, 2013.

[25] P. Lombardo and C. Oliver, "Maximum likelihood approach to the detection of changes between multitemporal SAR images," *IEE Proc. Radar, Sonar and Navig.*, vol. 148, no. 4, pp. 200–210, 2001.

[26] B. Le Saux and M. Sanfourche, "Rapid semantic mapping: Learn environment classifiers on the fly," in *Proc. of International Conference on Intelligent Robots and Systems*, Tokyo, 2013.