MULTIMODAL CLASSIFICATION WITH DEFORMABLE PART-BASED MODELS FOR URBAN CARTOGRAPHY

Hicham Randrianarivo, Bertrand Le Saux

Onera – The French Aerospace Lab
F-91761 Palaiseau, France
hicham.randrianarivo@onera.fr
bertrand.le_saux@onera.fr

Marin Ferecatu

CNAM – Laboratoire Cedric
292 rue St-Martin, 75141 Paris, France
marin.ferecatu@cnam.fr

ABSTRACT

Data from satellite and aerial images are now widely used by everyone. These images contain information from different frequency bands that help to characterize areas of interest. In this paper we study a framework for object detection in aerial image based on discriminatively-trained models trained on multimodal data. Specifically, we investigate a method to merge outputs of large margin classifiers trained on images from different sensors: we use the ranking ability of these classifiers to learn a probabilistic model.

1. INTRODUCTION

Nowadays state of the art detection methods in remote sensing are widely inspired by successful computer vision algorithms. Detectors such as Support Vector Machines (SVMs) (trained on Histogram of Oriented Gradients - HOGs - for example) have shown good results in object detection and are now extensively used in remote sensing. More recent methods like Discriminatively-trained Part Models (DPMs) [1] have also led to good results in remote sensing [2].

This paper investigates the various concepts behind the DPMs and propose an adapted DPM that handle multi-modal data. We show that DPMs can advantageously be used with descriptors that suit the sensor used to generate the data and propose an approach for combining DPMs trained over various sensors. It is organized as follows. In Sec. 2 we shed light on the components of the DPM framework and explain the multi-modal approach. Results are presented in Sec. 3 and analyzed in Sec. 4

2. DETECTION WITH DISCRIMINATIVELY-TRAINED PART MODELS

The available implementation of deformable part models was developed for the pascal voc challenge and includes specific tunings that led to successful results on this benchmark. However satellite imagery has peculiarities (such has image size and level of details) that has to be taken into account. In the following we identify the key parameters of DPMs and propose an adaptation of this algorithm for remote sensing.

2.1. Model

A DPM is a model which is composed of several components. The first component, the root filter F_0, captures the global appearance of the object and the set of N movable parts $\{P_i | i = 1, \ldots, n\}$, calculated at twice the resolution of the root filter, capture finer details of the object (Fig. 1). The model is composed by a set M of parts models, called a mixture of models, the role of which is to handle the variation of poses and orientations of the objects. Finally, this model is used to train a linear Support Vector Machine (SVM) and thus generating a discriminative classifier.

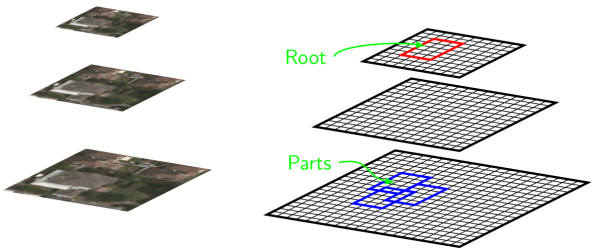


Fig. 1: Deformable Part Model synoptic: the Root filter captures the global appearance of the objects while Part filters get finer details at twice the resolution.

2.2. Key Mechanisms

As the name suggests, DPMs describe objects as a set of parts that can have various positions with respect to the object: it allows flexibility in the object representation, with benefit for dealing with non-rigid objects, occlusions or viewpoint changes. Keys mechanisms are the following

- Use of a mixture of models in order to model the variation in appearance and orientation of the samples. A recent paper investigates the importance the respective importance of deformations and mixture of models in DPMs [3].
- DPMs benefit from an appearance model based on a multiscale representation and pyramid matching as in [4].
- The HOG feature used in [1] is highly optimized in terms of image description and includes texture information.

2.3. Our Approach

2.3.1. Finding Subcategories

The issue of finding relevant subcategories in training data in order to learn a more accurate classifier is a subject of active research. There are methods based on latent membership to a cluster [1, 3], use of a novel measure of visual similarity [5] or decorrelated features [6]. Our approach is based on a 2-step process. First we use the aspect ratio of the positive training data to find a first set of clusters then we refine the clusters with appearance based clustering. The clusters based on the aspect ratio are found using the MeanShift algorithm [7] on the log of the aspect ratio. In remote sensing, the aspect ratio gives a good approximation of the object orientation. However to find relevant visual subcategories, a second step is needed. We use spectral clustering [8] on the decorrelated features of [6].

2.3.2. Ranking-Based Calibration

Our approach is based on the assumption that the outputs of the learning algorithm can be used to rank hypothesis. With this assumption the outputs are transformed into probabilities and theses probabilities are used compare mixture of classifiers learned on multimodal data. There are several methods to transform outputs of a large margin classifiers such as SVM into probabilities among which the most popular approach is Platt scaling [9] to estimate the probability.

The goal of this method is to map the scores of a classifier into probabilities by estimating the posterior probability $P(y = 1|w^T x)$. In [9] the author shows that the densities of the outputs of a large margin classifier can be modeled by a sigmoid function:

$$P(y = 1|w^T x) = \frac{1}{1 + \exp(\alpha^*(w^T x) + \beta^*)} \tag{1}$$

The parameters of the sigmoid are found by minimizing the negative log likelihood of the training data

$$\alpha^*, \beta^* = \arg\min_{\alpha,\beta} - \sum_i t_i \log(p_i) + (1 - t_i) \log(1 - p_i) \tag{2}$$

$$t_i = \frac{y_i + 1}{2} \tag{3}$$

$$p_i = \frac{1}{1 + \exp(\alpha w^T x_i + \beta)} \tag{4}$$

Once the classifiers are well calibrated they are applied on the corresponding modality (visible or LWIR). They return a set of hypothetical detections defined by a probability p and a position (x, y).

At this point there is a lot of redundant detection from the differents modality. We use the Non Maximum Suppression (NMS) to fuse multimodal detections and discards redundant detections. The NMS allows to keep the detection with the highest score and discard all detections that overlap this de-

Fig. 2: Illustration of the annotated dataset for object detection task. In green the tree category, in blue the car category and in maroon the house category.

tection. The overlap is computed as following for two detections W_i and W_j:

$$overlap(W_i, W_j) = \frac{W_i \bigcup W_j}{W_i \bigcap W_j} \quad (5)$$

3. EXPERIMENTS

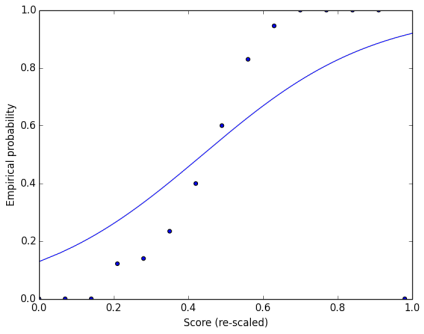
3.1. Dataset

The data we used are taken from the grss_dfc_2014 [10], it consists on two datasets acquired at different spectral ranges and spatial resolutions: a coarser-resolution long-wave infrared (LWIR, thermal infrared) hyperspectral data set and fine-resolution data acquired in the visible (VIS) wavelength range. The former is acquired by an 84-channel imager that covers the wavelengths between 7.8 to 11.5 μm with approximately 1-m spatial resolution. The latter is a series of color images acquired during separate flight-lines with approximately 20-cm spatial resolution. The two data sources cover an urban area near Thetford Mines in Qubec, Canada.

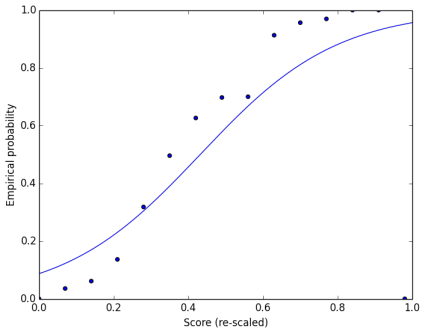
We extract from this dataset 273 cars, 235 trees and 226 houses for the training set. In the test image we extract 196 cars, 296 trees and 194 houses. Models are trained and calibrated on the training set then they are evaluated on the test image. An illustration of the annotated database is shown in Fig. 2

3.2. Calibration

The calibration of the classifier is evaluated with the reliability diagram of [11]. This diagram shows how well the empirical distribution of the probabilities fit the learned model. The more the empirical distribution is close to the curve the better the probabilities are estimated. Fig. 3 shows the reliability diagram for two differents classifiers learned separately.



(a)



(b)

Fig. 3: Diagram of reliability for classifiers calibrated using Platt calibration [9]. The more the dots fit the blue curve the better the probabilities are estimate. The model in Fig. 3b gives more reliable probabilities than the model evaluate in Fig. 3a.

3.3. Multimodal

In this section we evaluate the framework describe in 2.3. The detections from the optical and the LWIR images are fused based on the score of detection for each images. Precision-recall curves that evaluate the detectors are shown in Fig. 4.

4. CONCLUSION

We studied a framework that use classifiers learned on multimodal data to perform object detection. We show that a

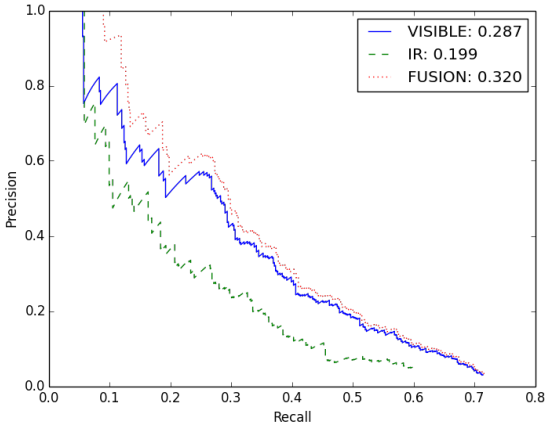


Fig. 4: Precision-recall curves for a tree detector. The solid blue curve evaluates the detector on the optical/visible image, the dash green curve evaluate the detector on the LWIR image and finally the dotted red curve evaluate the fusion of the previous classifier. The overall metric used to evaluate the detectors is the average precision.

proper calibration of the classifiers allows to compare outputs of these classifiers even if they are learned on heterogeneous data. We use this property to fuse detections hypothesis proposed by the mixture of classifiers and find out that the fusion increases the performances of the object detector.

ACKNOWLEDGEMENTS

”The authors would like to thank Telops Inc. (Quebec, Canada) for acquiring and providing the data used in this study, the IEEE GRSS Image Analysis and Data Fusion Technical Committee and Dr. Michal Shimoni (Signal and Image Centre, Royal Military Academy, Belgium) for organizing the 2014 Data Fusion Contest, the Centre de Recherche Public Gabriel Lippmann (CRPGL, Luxembourg) and Dr. Martin Schlerf (CRPGL) for their contribution of the Hyper-Cam LWIR sensor, and Dr. Michaela De Martino (University of Genoa, Italy) for her contribution to data preparation.

5. REFERENCES

[1] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan, “Object detection with discriminatively trained part-based models.,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–45, Sept. 2010. 1, 2

[2] Hicham Randrianarivo, Bertrand Le Saux, and Marin Ferecatu, “Urban structure detection with deformable part-based models,” in *IEEE International Geoscience and Remote Sensing Symposium*, 2013. 1

[3] SK Divvala, AA Efros, and Martial Hebert, “How important are Deformable Parts in the Deformable Parts Model?,” in *European Conference on Computer Vision*, 2012. 2

[4] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories,” in *IEEE Conference on Computer Vision and Pattern Recognition*. 2006, vol. 2, pp. 2169–2178, IEEE. 2

[5] Omid Aghazadeh and Hossein Azizpour, “Mixture component identification and learning for visual recognition,” in *European Conference on Computer Vision*, 2012, pp. 1–14. 2

[6] Bharath Hariharan, Jitendra Malik, and Deva Ramanan, “Discriminative decorrelation for clustering and classification,” *European Conference on Computer Vision*, vol. 1, pp. 1–14, 2012. 2

[7] Dorin Comaniciu and Peter Meer, “Mean shift: A robust approach toward feature space analysis,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 24, no. 5, pp. 603–619, 2002. 2

[8] Shi Jianbo and J. Malik, “Normalized cuts and image segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000. 2

[9] John Platt, “Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods,” *Advances in large margin classifiers*, 1999. 2, 3

[10] “2014 IEEE GRSS Data Fusion Contest,” *http://www.grss-ieee.org/community/technical-committees/data-fusion/*. 3

[11] Morris H. DeGroot and Stephen E. Fienberg, “The Comparison and Evaluation of Forecasters,” *The Statistician*, 1982. 3