URBAN STRUCTURE DETECTION WITH DEFORMABLE PART-BASED MODELS

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

***Index Terms*—** image analysis, remote sensing, very high resolution, object recognition, object detection, deformable part-based models

ABSTRACT

In this paper we apply the deformable part model by Felzenszwalb et al., which is at this moment the state of the art in many computer vision related tasks, to detect different types of man made structures in very high resolution aerial images — a reputedly difficult problem in our field. We test the framework on a database of crops of aerial images at a definition of 10 cm/pixel and investigate how the model performs on several classes of objects. The results show that the model can achieve reasonable performance in this context. However, depending on the type of object, there are specific issues which will have to be taken into account to build an effective semi-supervised annotation tool based on this model.

1. INTRODUCTION

This paper focuses on object detection in aerial and satellite images. Our main objective is to be able to perform man-made structure detection. Today more and more images are produced at always increasing resolutions, which makes the task of the image analyst that has to annotate these images laborious. Several approaches have been proposed to cope with this problem. Markov fields where used to model the texture of images as the first step to urban area extraction [1]. More recently powerful machine learning algorithms like Supports Vectors Machines were used on sets of appearance features extracted from the image [2] . In [3] a graph of points of interest is built upon the image and graph-cut is performed to segment between urban and countryside areas.

Very High Resolution (VHR) images bring a lot of new information in remote sensing. It is possible to distinguish specific parts inside the buildings or the objects we want to model. This allows us to use powerful state of the art methods from computer vision: in this field the reference test for object detection is the Pascal VOC challenge [4] and the approach that has been successful for several years is based on deformable part models [5]. Modeling deformable shape for aerial image analysis is not a novel idea [6] but it has not been used on VHR images yet. Moreover appearance features for describing image parts can benefit from recent advances in feature extraction: Histograms of Oriented Gradients (HOG) [7] .

The paper is organized as follows: in the following section we describe the model and detail its implementation, in Sec. 3 we present results on two representative datasets, and we conclude in Sec. 4 with a discussion of possible improvements.

2. DEFORMABLE PART-BASED MODELS

This detection system uses a deformable part model framework developed in [5]. Briefly a model consists in a root filter that encodes the global appearance of the object we are looking for and several parts that encode the local appearance of the object.

2.1. MODEL

A model is defined by a root filter F_0 that captures the global appearance of the object and a set of n parts $\{P_i | i = 1, \ldots, n\}$, placed at twice the resolution of F_0, that capture finer details in higher resolution.

A part P_i of the model is composed by the filter F_i, the placement v_i of P_i with respect to the root filter and the deformation cost d_i. The filters F_i are HOG detectors [7] reshaped as 2D-filters, this feature is used because of good results in detection tasks [5] [8] but the model is feature agnostic and other features can be used, for example [9].

2.2. TRAINING

The training stage learns the model structure [10], the different filters, their location and the deformation costs using labeled bounding boxes from the dataset. To this end, the framework must in the first step learn the root filter without prior knowledge about the position of the parts and then, in a second step, find the optimal placement of the parts.

Fig. 1: The deformable part model superimposed over a building: the root filter (red bounding box) captures the global appearance of the image while the different parts (blue bounding boxes) capture the local information, thus increasing the precision of the model.

In the first step, to initialize the root filter, the framework uses a standard Support Vector Machine (SVM) trained on the positive bounding boxes, as in [7]. For the second step, [5] propose a modified SVM algorithm called latent SVM which used an extended search space. They define a score function f_β that test a model with a feature vector without prior knowledge of the positions of the parts:

$$f_\beta(x) = \max_{z \in Z(x)} \beta.\phi(x, z) \tag{1}$$

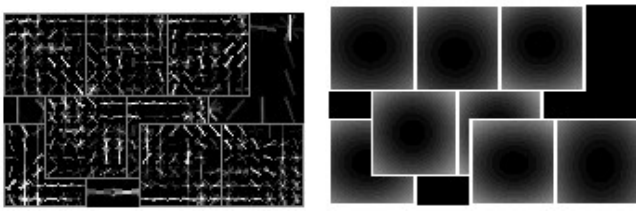
where β is the model, ϕ is the feature vector and $Z(x)$ are the possibles latent locations of the different parts of the model. This leads to the following SVM objective function:

$$L_D(\beta) = \frac{1}{2}||\beta||^2 + C\sum_{i=1}^{n} \max(0, 1 - y_i f_\beta(x_i)) \tag{2}$$

The trained root filter is then the root filter from the model β that minimizes the Eq. 2.

Fig. 2: Exemple of root filter learned after the latent SVM step for the building class boxes.

Third step: After the root filter is trained, the different parts of the model are successively placed in areas that have maximal energy according to an interpolated version of the root filter at twice the resolution. The set of parts constitute a star-graph of filters applied at smaller scale to describe details of the object.



(a) Part filters. (b) Cost of placing the parts.

Fig. 3: Part filters are placed over the root filter at twice the resolution.

Considering the variation of poses and orientations of an object, the framework provides a method to compute several models depending on the aspect of the objects from the training set. First, to increase the invariance to orientation, a mirrored version of the model is computed and the score is taken as the maximum value between the score of the model and the mirrored model. Second, in order to overcome the variability of appearance of an object the positives examples are clustered into n subsets and a model is trained over each subset. The clustering criteria used is called aspect-ratio, this is computed with the ratio between the height and the width of the bounding boxes. Each model represents an aspect of the object that we want to recognize (for example: front view vs. side view). Finally the score of the model is the maximum of all computed score for all model in the mixture.

2.3. MATCHING

During the matching process, first we compute the score of the root filter densely at several scales over the whole image. We keep as hypothesis for the detection the score of the root filter over a threshold computed during the learning phase. For each hypothesis the parts are placed in the same way as during the training phase and finally we compute the score of detection of the hypothesis with all the parameters. The score of the hypothesis is given by:

$$\begin{aligned} \text{score}(x_0, y_0, l_0) =& R_{0,l_0}(x_0, y_0) \\ &+ \sum_{i=1}^{n} D_{i,l_0-\lambda}(2(x_0, y_0) + v_i) + b \end{aligned} \tag{3}$$

- $R_{i,l}(x, y) = F_i \times \phi(x, y, l)$ is the score of the i-th filter
- ϕ is the feature vector extracted at position (x, y) and scale l
- $D_{i,l}(x, y) = \max_{dx,dy}(R_{i,l}(x+dx, y+dy)) - d_i.\phi_d(dx, dy)$ is the response of the part filter F_i,
- $(dx_i, dy_i) = (x_i, y_i) - 2(x_0, y_0) + v_i$
- $\phi_d(dx, dy) = (dx, dy, dx^2, dy^2)$

The detected model is the one with the maximum score over all possible models.

3. RESULTS

3.1. DATASETS

For testing we used two data sets at different resolutions:

- *Christchurch dataset*: 10cm/pixel orthonormal aerial images provided by New Zealand Aerial Mapping Limited[1] and captured after the earthquake that struck the town of Christchurch on 22 February 2011.
- *QuickBird images*: 2000×2000, 60 cm/pixel containing man-made structures previously evaluated in [8].

A training set of three image classes ('building', 'tree' and 'car') was manually built by annotating the content of the images.

3.2. EVALUATION

We want to predict the location (given by its bounding box) of each object from a specific class in a test image. We evaluate the algorithm by using the PASCAL VOC [4] procedure: a bounding box is associated with a score of confidence that is used to draw precision/recall curve. The overall score for a category is given is thus given by the average precision. To be considered as a correct detection, the area of overlap a_o between the predicted bounding box B_p and the ground truth bounding box B_{gt} must be superior to 50%.

$$a_o = \frac{\text{area}(B_p \cap B_{gt})}{\text{area}(B_p \cup B_{gt})} \tag{4}$$

However in the case of multiple detections of the same object only one detection is considered as a correct detection, the other ones are considered as false detections.

aspect-ratio	building	tree
1	0.3250	0.3040
2	**0.4301**	**0.3425**
3	0.2780	**0.3311**
4	0.2240	0.2752
5	0.1023	0.2831

Table 1: Average precision score for the detection on Christchurch dataset.

The table 1 shows that best performances are achieved for a mixture of two models. By viewing the models learned and with knowledge of the database we can explain this by the following observations:

- For the buildings: the database contains a large number of rectangular shaped buildings and small square houses, these respond very well to the model.
- For trees: 2 or 3 aspect-ratios both give good results. This class of objects has more isotropic features and models look strongly similar.

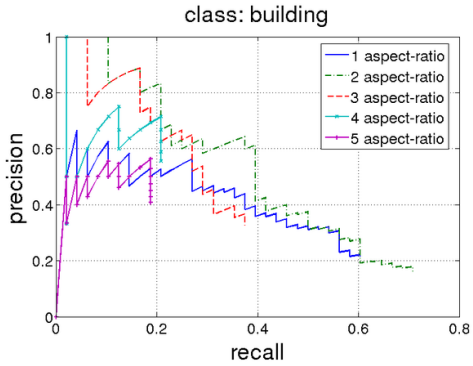


Fig. 4: Precision-recall curve for building on Christchurch.

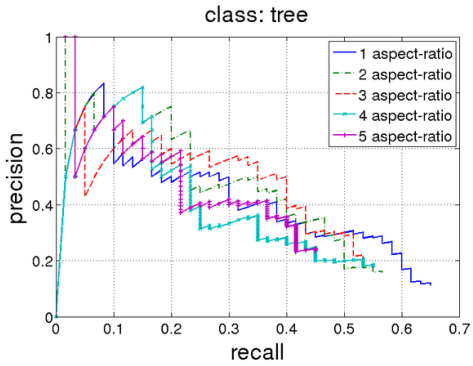


Fig. 5: Precision-recall curve for tree on Christchurch.

Fig. 6 and Fig. 7 show practical demonstration of the performance of the framework on one aerial image of Christchurch (that is not part of the training set). For buildings detection we notice that trees and roads are not flashed as false detections but due to the multiscale capacities of the model some groups of buildings are labeled as one building. For tree detection we notice that a lot of isolated trees in the city are well detected even if shrubs are not. The lot of false alarms can be interpreted as the response of the filters to homogeneous textures.

Fig. 8 shows a result of man-made structure detection in a QuickBird image. At this lower resolution, we can only look for coarser object categories. Urban areas as well as isolated houses were well retrieved in the image.

We don't present results on the other class we tested ('cars') because the model collapsed completely on these objects. This is likely due to several causes: absence of variation in the camera view, small size of the objects and the absence of invariance to rotation in the features. At a closer inspection, the clustering algorithm that produces the number

[1] http://nzam.com/

Fig. 6: Result for building detection.

Fig. 7: Result for tree detection.

of trained views of the model failed to work which hints at the absence of variability in the poses.

4. CONCLUSION

We present a deformable part-based framework for urban structure detection. The key contributions are efficient state of the art object detection framework transposed to aerial and satellite images. Especially, we show that computer vision methods, which are designed to work with very different types of content, can be specialized to be successfully used in very high resolution aerial and satellite images. In future we aim to bring orientation invariance to the features in order to increase the accuracy of the models and add context information during the training step of the classifier.

5. REFERENCES

[1] Lorette, A, Descombes, X and Zerubia, J (2000). Texture analysis through a markovian modelling and fuzzy classification: Application to urban area extraction from satellite images. *International Journal of Computer Vision*. **36** 221–36

[2] Inglada, J (2007). Automatic recognition of man-made objects in high resolution optical remote sensing images by SVM classification of geometric image features. *ISPRS journal of photogrammetry and remote sensing*. **62** 236–48

[3] Sirmacek, B and Unsalan, C (2009). Urban-area and building detection using SIFT keypoints and graph theory. *Geoscience and Remote Sensing, IEEE Transactions on*. **47** 1156–67

[4] Everingham, M, Zisserman, A, Williams, C and Van Gool, L (2006). The pascal visual object classes challenge 2006 (voc 2006) results

[5] Felzenszwalb, P F, Girshick, R B, McAllester, D and Ramanan, D (2010). Object Detection with Discriminatively Trained Part Based Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. **32** 1627–45

[6] Reno, A L, Gillies, D F and Booth, D M (1998). Deformable models for object recognition in aerial images. *Proceedings of the SPIE Conference Automatic Target Recognition VIII*. **3371** 323–33

[7] Dalal, N and Triggs, B (2005). Histograms of oriented gradients for human detection. *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. IEEE. **1** 886–93

[8] Chauffert, N, Israel, J and Le Saux, B (2012). Boosting for interactive man-made structure classification. *Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International*. IEEE. 6856–9

[9] Wang, X, Han, T X and Yan, S (2009). An HOG-LBP human detector with partial occlusion handling. *Computer Vision, 2009 IEEE 12th International Conference on*. 32–9

[10] Felzenszwalb, P F and Huttenlocher, D P (2003). Pictorial Structures for Object Recognition. *IJCV*. **61** 2005

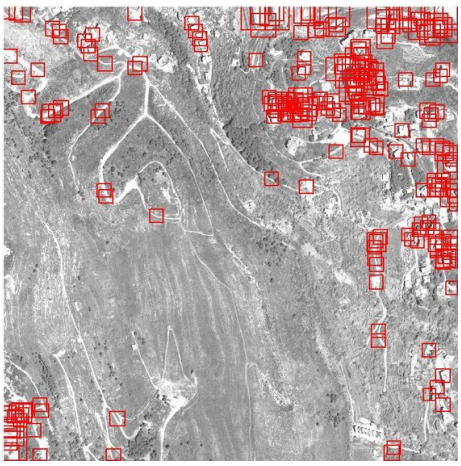


Fig. 8: Detection of buildings in a satellite test image from QuickBird satellite.