Context Forest for Object Class Detection

Davide Modolo
davide.modolo@gmail.com
Alexander Vezhnevets
vezhnick@gmail.com
Vittorio Ferrari
vittoferrari@gmail.com

University of Edinburgh
Edinburgh, UK

Global image appearance carries information about properties of objects in the image. For instance, a picture of a highway taken from a car is more likely to contain cars from the back viewpoint than from the side (fig. 1). This shows how the global image appearance of images can help understanding what objects are present and what they look like. Moreover, another property that can be inferred from global image appearance is the rough location of object instances [7]. For instance, an urban scene with cars parked in front of a building, shows cars in the bottom half of the image.

In this paper we exploit this observation for object class detection. We propose *Context Forest (ConF)*: a technique for learning the relation between the global appearance of an image and the properties of the objects it contains. Given only the global appearance of a test image, ConF retrieves a subset of training images that contain objects with similar properties. ConF is based on the Random Forest [2] framework, which provides high computational efficiency and the ability to learn complex, non-linear relations between global image appearance and objects properties. It is very flexible and only requires these properties to be defined through a distance function between two object instances, e.g. their appearance similarity or difference in location. We demonstrate ConF by learning to predict three properties: aspects of appearance, location in the image, and class membership.

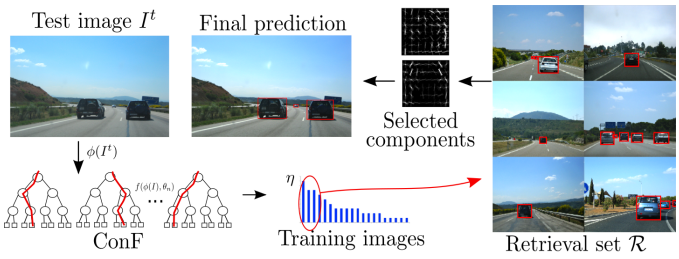

Figure 1: Illustration of ConF selecting components for a test image.

Aspects of appearance. Multi-component detectors [3, 5] model an object class as a mixture of components, each trained to recognize a different aspect of appearance. For example, different viewpoints (e.g. front and back view of a car) or articulation states (e.g. a person sitting vs standing). When trained on a large set, such a detector has many components [9], which all need to be evaluated on a test image, making it slow. Instead, we use ConF to select a subset of model components which is most relevant to a particular test image. We then run only those components, obtaining a speed-up.

We present experiments on two detectors: DPM [3] and EE-SVM [5] on a large 2-class dataset we call *BigCH*. This combines 6 existing datasets and, in total, the dataset has 15766 images containing 28548 car instances and 10107 images containing 13071 horse instances. We compare to building retrieval sets by kNN, and to a baseline which randomly selects components without looking at the test image. ConF outperforms the baseline and kNN for both object classes, for both detectors, and over the whole range of experiments. By employing ConF, we closely match the performance of the full DPM model by running roughly half of the components. We match the performance of a full EE-SVM when running less than 10% of the components. Even in the extreme case of running just *one* EE-SVM component, the AP is about 90% of that of the full model. Interestingly, for EE-SVM on the horse class, ConF *improves AP by 3%* over the full ensemble using all components, when running *10× fewer components*.

Location. At test time, a typical detector scores windows in a test image, based on their appearance only. We propose here to augment the detector's scores by adding knowledge about likely positions and scales

Figure 2: Detection obtained before (top row) and after (bottom row) applying ConF as location model. Green bounding-boxes highlight correct detections, while red ones show false positives.

of the object class, derived purely from the global appearance of the image. We train a second ConF to predict at which positions and scales objects are likely to appear in a given test image, analogue to [4, 7, 8]. By incorporating this information in the detector score at test time, we reduce the false positive rate by removing detections at unlikely locations.

We experiment with DPM and EE-SVM on BigCH, as for the aspect of appearance property. Results show that ConF improves AP for both classes and both detectors (+2% for cars and +1% for horses). Instead, kNN does not bring any improvement.

Class membership. In multi-class problems, a typical system would run detectors for all classes on all test images [6]. Instead, here we use ConF to predict what classes are present in each image, and run only the corresponding detectors. This greatly reduces the number of detectors run, and removes some false-positives.

We experiment with EE-SVM on the ILSVRC2014 dataset [1], which has a large number of classes (200). Without any context, EE-SVM achieves an mAP of 16.3%. Selecting classes based on kNN retrieval sets improves performance by +3.3% (mAP 19.6%), while ConF delivers a larger improvement of +4.8% (mAP 21.1%). The improvements are due to removing false positives produced by detectors of classes unlikely to be present in the image. Interestingly, ConF selects less than 10 classes per image on average, and therefore runs 20× fewer EE-SVM detectors than the context-free baseline.

To conclude, all these experiments demonstrate that ConF is a general technique that can predict various kinds of object properties. An extensive comparison to standard nearest-neighbour techniques for such context-based predictions also shows that ConF predicts object properties from global image appearance more accurately, while being 60× more memory efficient and much faster: kNN requires a number of distances computations linear in the number of training images, whereas ConF requires only a logarithmic number of thresholding operations.

[1] Imagenet large scale visual recognition challenge (ILSVRC). `http://www.image-net.org/challenges/LSVRC/2014/index`, 2014.

[2] L. Breiman. Random forests. *ML Journal*, 45(1):5–32, 2001.

[3] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Trans. on PAMI*, 32(9), 2010.

[4] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing: Label transfer via dense scene alignment. In *CVPR*, 2009.

[5] T. Malisiewicz, A. Gupta, and A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, 2011.

[6] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *IJCV*, 2015. doi: 10.1007/s11263-015-0816-y.

[7] B. Russell, A. Torralba, C. Liu, R. Ferugs, and W. Freeman. Object recognition by scene alignment. In *NIPS*, 2007.

[8] A. Torralba. Contextual priming for object detection. *IJCV*, 53(2):153–167, 2003.

[9] X. Zhu, C. Vondrick, D. Ramanan, and C. Fowlkes. Do we need more training data or better models for object detection? In *BMVC*, 2012.