Object localization in ImageNet by looking out of the window

Alexander Vezhnevets
vezhnick@gmail.com
Vittorio Ferrari
vferrari@gmail.com

University of Edinburgh
Edinburgh, Scotland, UK

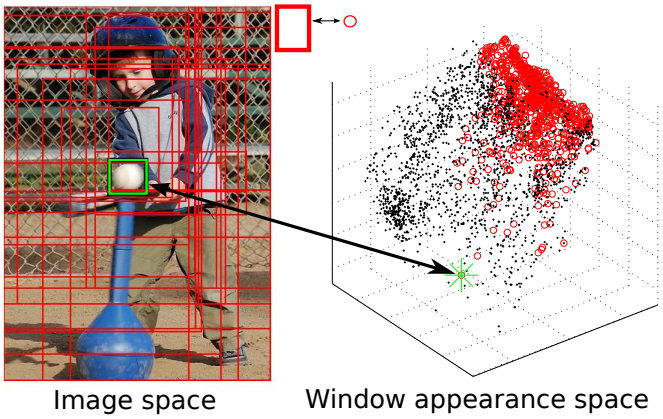

Figure 1: **Connecting the appearance and window position spaces.** *A window tight on the baseball (green star in the appearance space plot) and some larger windows containing it (red circles in the appearance space). Black points in appearance space represent all other candidate windows. The appearance space plots are actual datapoints, representing windows in 3-dimensional Associative Embedding of SURF bag-of-words.*

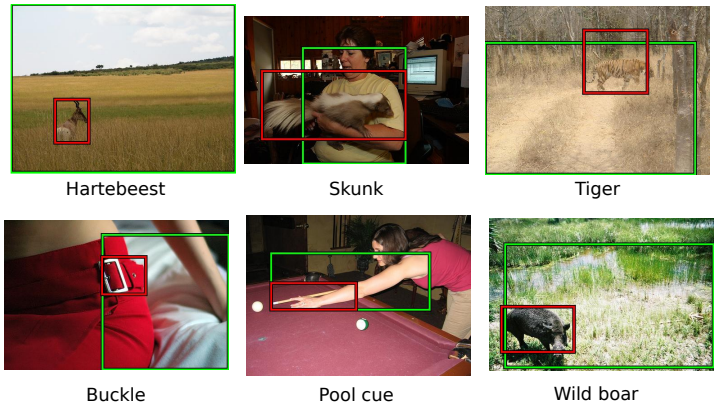

Figure 2: **Results of our method (red) vs a prior method [10] (green).** *Notice, how our method is able to detect small, off-center objects despite occlusion (pool cue) or the object blending with its surroundings (tiger).*

The ImageNet database [4] contains over 14 million images annotated by the class label of the main object they contain. However, only a fraction of them have bounding-box annotations (10%). Automatically annotating object locations in ImageNet is a challenging problem, which has recently drawn attention [6, 10]. These annotations could be used as training data for problems such as object class detection [3], tracking [7] and pose estimation [1]. Traditionally, object localization is cast as an image window scoring problem, where a scoring function is trained on images with bounding-boxes and applied to ones without. The image is first decomposed into candidate windows, typically by object proposal generation [8]. Each window is then scored by a classifier trained to discriminate instances of the class from other windows [3, 5, 8, 9] or a regressor trained to predict their overlap with the object [2, 10]. Highly scored windows are finally deemed to contain the object. In this paradigm, the classifier looks at one window at a time, making a decision based only on that window's appearance.

We believe there is more information in the collection of windows in an image. By taking into account the appearance of all windows *at the same time* and connecting it to their spatial relations in the image plane, we could go beyond what can be done by looking at one window at a time. Consider the baseball in fig. 1. For a traditional method to succeed, the appearance classifier needs to score the window on the baseball higher than the windows containing it. The container windows cannot help except by scoring lower and be discarded. By considering one window at a time with a classifier that only tries to predict whether it covers the object tightly, one cannot do much more than that. The first key element of our work is to predict richer spatial relations between each candidate window and the object to be detected, including part and container relations. The second key element is to employ these predictions to reason about relations between different windows. In this example, the container windows are predicted to contain a smaller target object somewhere inside them, and thereby actively help by *reinforcing* the score of the baseball window. By considering the configuration of all the windows in appearance space together we can reinforce its score.

In a nutshell, we propose to localize objects in ImageNet by scoring each candidate window in the context of all other windows in the image, taking into account their similarity in appearance space as well as their spatial relations in the image plane. To represent spatial relations of windows we propose a descriptor indicative of the part/container relationship of the two windows and of how well aligned they are. We learn a windows appearance similarity kernel using the recent Associative Embedding technique [10]. We describe each window with a set of hyper-features connecting the appearance similarity and spatial relations of that window to all other windows in the same image. These hyper-features are indicative of the object's presence when the appearance of a window alone is not enough (e.g. fig 1). These hyper-features are then linearly combined into an overall scoring function. We devise a fast and exact procedure to optimize our scoring function over all candidate windows in a test image, and we learn its parameters using structured output regression.

We evaluate our method on a subset of ImageNet containing 219 classes with more than 92000 images [6, 10]. The experiments show that our method outperforms a recent approach for this task [10], an MKL-SVM baseline [9] based on the same features, and the popular UVA object detector [8]. Figure 2 presents some qualitive results of our method compared to results of [10].

[1] M. Andriluka, S. Roth, and B. Schiele. Monocular 3d pose estimation and tracking by detection. In *CVPR*, 2010.

[2] M. B. Blaschko and C. H. Lampert. Learning to localize objects with structured output regression. In *ECCV*, 2008.

[3] N. Dalal and B. Triggs. Histogram of Oriented Gradients for human detection. In *CVPR*, 2005.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[6] M. Guillaumin, D. Küttel, and V Ferrari. ImageNet auto-annotation with segmentation propagation. *IJCV*, 2014.

[7] B. Leibe, K. Schindler, and L. Van Gool. Coupled detection and trajectory estimation for multi-object tracking. In *ICCV*, 2007.

[8] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. Selective search for object recognition. *IJCV*, 104(2): 154–171, 2013.

[9] Andrea Vedaldi, Varun Gulshan, Manik Varma, and Andrew Zisserman. Multiple kernels for object detection. In *ICCV*, pages 606–613, 2009.

[10] A. Vezhnevets and V. Ferrari. Associative embeddings for large-scale knowledge transfer with self-assessment. In *CVPR*, 2014.