# Optimization framework for learning a hierarchical shape vocabulary for object class detection

Sanja Fidler
http://vicos.fri.uni-lj.si/sanja
Marko Boben
http://vicos.fri.uni-lj.si/markobob
Aleš Leonardis
http://vicos.fri.uni-lj.si/alesl

University of Ljubljana
Faculty of Computer and Information Science
Slovenia

Approaches that learn visual codebooks of appearance and/or shape, and combine them with simple object geometry have been shown to give the most successful performance for object class detection to date. Most of these works, however, use flat visual vocabularies where each object is represented as an immediate aggregate of codebook features. Recently, hierarchical approaches have demonstrated appealing computational and qualitative advantages [2, 4, 7]. Hierarchical vocabularies incorporate structural dependencies among the codebook entries at multiple levels: objects are defined in terms of a collection of parts, which are further composed from a set of simpler constituents, etc. This 1.) increases the reliability of detections, and 2.) reduces inference time because the features are not only shared among the distinct classes, but are also shared among the features themselves — at multiple layers of the representation.

This paper proposes a stochastic optimization framework for unsupervised learning of a *hierarchical vocabulary of object shape* intended for object class detection. We build upon the approach by [2]. The original idea of [2] is to find a vocabulary of shape models that well represent the distribution of the spatial layouts of contour fragments inside local neighborhoods (*receptive fields* or RFs). The sizes of the RFs are increased with each layer and part configurations are learned to explain larger and larger image areas. Performing spatial contraction at each layer successfully deals with the problem of increasingly larger RFs, while the principle of hierarchical compositionality successfully handles their increasing complexity. In the top layer the RF covers the whole image (object).

In [2] the RFs were assumed independent, i.e. for each local neighborhood the frequency of the composition that best explained its contour content was updated and the method finally selected the set of most frequently occurring compositions. Since there is a huge variety of local shape configurations, the number of compositions quickly increases to a large number which makes further combination-learning difficult (as also seen in [5, 6]). The selection based only on frequency of appearance also makes the method prone to overfitting of certain parts of object shape while losing the more discriminative shape information. Here we exploit the fact that the top, object-layer models will have a tree structure, meaning they are composed of disjunct parts at each layer. This means that we do not need our vocabulary shapes to explain each of the RFs in an image, but must only be able to explain all the contour fragments in an image in a global sense — as a union of a few matched compositions.

The goal of learning is to find an "optimal" hierarchical shape vocabulary for object class representation, where we define the optimal vocabulary to be the one that, among all the vocabularies that well represent the shape distribution of object classes, attains the lowest complexity of inference. The idea of this paper is to cast the vocabulary learning into an optimization framework that iteratively improves the hierarchy as a whole. Optimization is two-fold: one that learns and selects the vocabulary of shapes at each layer in a bottom-up phase and the other that extends/improves it by top-down feedback from the higher layers. The algorithm then loops between the two learning stages several times.

We have evaluated the proposed learning approach for object class detection on 11 diverse object classes taken from the standard recognition data sets. Examples of learned shapes in the vocabulary are depicted in Fig. 1. Compared to the original approach [2], we obtain a 3 times more compact vocabulary, a 2.5 times faster inference, and a 10% higher detection performance at the expense of 5 times longer training time (25min vs 5min). The comparison is presented in Fig. 2. Specifically, it takes 100–200Kb to store a hierarchy for one object class on a hard disk. Inference for an image of size roughly $700 \times 500$ takes on average between 2–4 seconds. The detection rates are given in Table 1. The approach attains a competitive detection performance with respect to the current state-of-the-art at both, faster inference as well as training times.

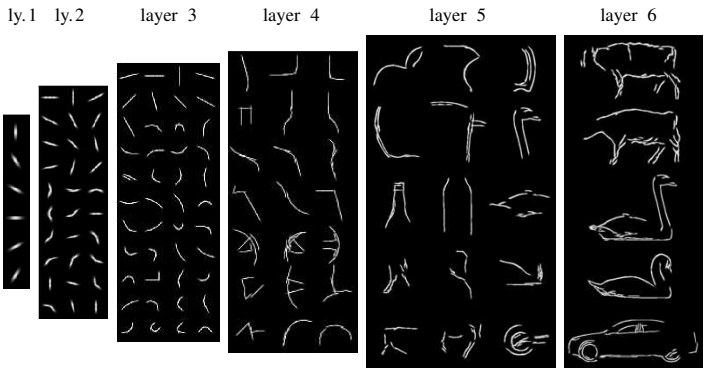
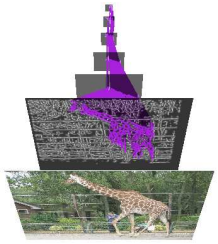



Figure 1: Example shapes in the learned hierarchical vocabulary. Each shape in the hierarchy is a composition of shapes from the layer below. Each shape also models spatial relations between its constituent parts, which are not shown — only the mean of each shape is depicted.

Table 1: Detection results. *Left:* An example detection of a giraffe. *Right:* Average detection-rate (in %) at 0.4 FPPI for ETH and INRIA datasets.



| | | [1] | [3] | **our appr.** |
|---|---|---|---|---|
| ETH shape | apple | 83.2 (1.7) | 89.9 (4.5) | 87.3 (2.6) |
| | bottle | 83.2 (7.5) | 76.8 (6.1) | **86.2** (**2.8**) |
| | giraffe | 58.6 (14.6) | 90.5 (5.4) | 83.3 (4.3) |
| | mug | 83.6 (8.6) | 82.7 (5.1) | **84.6** (**2.3**) |
| | swan | 75.4 (13.4) | 84.0 (8.4) | 78.2 (5.4) |
| | **avg.** | 76.8 | 84.8 | 83.7 |
| INRIA | horse | 84.8(2.6) | / | **85.1**(**2.2**) |

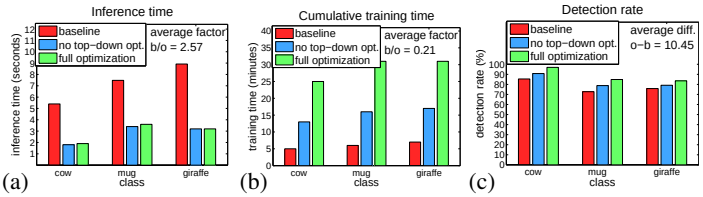



Figure 2: A comparison of (a) inference, and (b) training times, and (c) detection rates for the *baseline*, *no top-down optimization*, and our proposed *full optimization* approach.

[1] V. Ferrari, L. Fevrier, F. Jurie, and C. Schmid. Accurate object detection with deformable shape models learnt from images. In *IEEE CVPR*, 2007.

[2] S. Fidler and A. Leonardis. Towards scalable representations of visual categories: Learning a hierarchy of parts. In *IEEE CVPR*, 2007.

[3] M. Fritz and B. Schiele. Decomposition, discovery and detection of visual categories using topic models. In *IEEE CVPR*, 2008.

[4] B. Ommer and J. M. Buhmann. Learning the compositional nature of visual objects. In *IEEE CVPR*, 2007.

[5] M. A. Ranzato, F.-J. Huang, Y-L. Boureau, and Y. LeCun. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *IEEE CVPR*, 2007.

[6] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Object recognition with cortex-like mechanisms. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 29(3):411–426, 2007.

[7] L. Zhu, C. Lin, H. Huang, Y. Chen, and A. Yuille. Unsupervised structure learning: Hierarchical recursive composition, suspicious coincidence and competitive exclusion. In *ECCV*, volume 2, pages 759–773, 2008.