Learning a Hierarchical Compositional Representation of Multiple Object Classes

Aleš Leonardis
University of Ljubljana
ales.leonardis@fri.uni-lj.si

Visual categorization, recognition, and detection of objects has been an area of active research in the vision community for decades. Ultimately, the goal is to recognize and detect a large number of object classes in images within an acceptable time frame. This problem entangles three highly interconnected issues: the internal object representation which should expand sublinearly with the number of classes, means to learn the representation from a set of images, and an effective inference algorithm that matches the object representation against the representation produced from the scene.

Through the decades of research the main focus has been on the issue of representation. A number of interesting and diverse ideas have been proposed ranging from full 3D geometric models, completely appearance-based approaches to bags-of-features which render the structure irrelevant. An excellent overview of the field can be found in [1].

The ideas on hierarchies, compositionality, and image grammars have been an important stream of research in the very beginning of computer vision. Later, however, modern hardware gave way to structurally much simpler approaches, which have, despite the inherent limitations with respect to the general vision problem, managed to overshadow the high level concepts with tangible recognition results. However, the complexity of the goals set by the modern vision which entails recognizing objects on a larger scale, have brought back grammar-based approaches into the community.

In this talk I will focus on hierarchical compositional representations. The general idea of these frameworks is to take a set of elementary features and build increasingly larger structures by virtue of simple production rules. Their recursive architecture enables them to capture exponential structural variability with far less stored internal information than that coded in flat representations [2]. One of the main challenges in the large-scale hierarchy design is *learning* with little or no explicit supervision.

In the main part of the talk I will present our framework for learning a hierarchical compositional representation of multiple object classes. Learning is unsupervised, *statistical*, and is performed *bottom-up*. The approach takes simple contour fragments and learns their frequent spatial configurations which recursively combine into increasingly more complex and class-specific contour compositions. The top-level compositions pertain to the whole objects, each exerting a high degree of class invariance. The learned representation is highly compact, compositional all the way through to the objects, enables fast learning of multiple object classes and exerts a high degree of feature sharing between them. I will show experimental results on several important issues: applied to a large collection of natural images, the approach learns hierarchical detectors for various curvatures, corners, and junctions as predicted by the Gestalt school. For multi-class object detection the approach achieves (i) a highly sub-linear growth in the size of the hierarchy and, consequently, a sub-linear inference complexity as the number of modeled classes increases; (ii) a decrease in time needed to train a class as more classes are modeled; (iii) competitive detection performance with respect to the current state-of-the-art. These results provide an important showcase that highlights compositional hierarchies as a suitable form of representing a higher number of object classes.

References

[1] S. Dickinson. The evolution of object categorization and the challenge of image abstraction. In S. Dickinson, A. Leonardis, B. Schiele, and M. J. Tarr, editors, *Object Categorization: Computer and Human Vision Perspectives*. Cambridge University Press, 2009.

[2] S. Zhu and D. Mumford. A stochastic grammar of images. *Found. and Trends in Comp. Graphics and Vision*, 2(4):259–362, 2006.