Inference and Learning with Hierarchical Compositional Models

Iasonas Kokkinos
Laboratoire MAS, Ecole Centrale de Paris
Equipe GALEN, INRIA Saclay - Ile-de-France, Orsay

Alan Yuille
University of California at Los Angeles

In this work we consider the problem of object parsing, namely detecting an object and its components by composing them from image observations. We build on [1] where we address the computational complexity of the inference problem. For this we exploit our hierarchical object representation to efficiently compute a coarse solution to the problem, which we then use to guide search at a finer level. Starting from our adaptation of the A^* parsing algorithm of [3] to the problem of object parsing, we then propose a coarse-to-fine approach that is capable of detecting multiple objects simultaneously. Details can be found in [1].

We extend this work to automatically learn a hierarchical model for a category from a set of training images for which only the bounding box is available. Our approach consists in (a) automatically registering a set of training images and constructing an object template (b) recovering object contours (c) finding object parts based on contour affinities and (d) discriminatively learning a parsing cost function.

Initially we use the method of [2] to automatically register the images in our training set. Finding object boundaries and symmetry axes is then feasible as we gather information from the whole training set: edge and ridge features are unreliable on individual images, but become robust when *averaged* over *multiple, registered* images.

We build an hierarchical representation of the object that can model these contours. Initially we break the edge and ridge maps into a set of straight line segments, which are then grouped into parts using a bottom-up, pairwise clustering method. We thereby bypass the difficulties of structure learning for graphical models, by exploiting the geometrical information in our problem.

In specific, we treat each line segment as a node on a graph; the edges in this graph identify line segments that could be grouped into a larger structure based on perceptual grouping cues: edge-to-edge and ridge-to-ridge connections are established based on continuity, while an edge is connected to a ridge based on parallelism. The weight for each edge is computed based on deformation statistics and co-occurrence probabilities, and is larger for pairs of lines that move and appear together.

Spectral clustering [5] is then used to find the object parts: this turns a pairwise relation among graph nodes into a segmentation of the whole graph, such that within each segment the nodes are well connected, while only weak connections exist between the separated parts. Visually appealing parts are extracted in this way for several categories, which closely match our own perception of object parts.

Finally, we extend the work in [1] that was developed from a purely generative viewpoint by adopting a discriminative training approach. The parsing criterion optimized in [1] is a sum of the log-likelihoods of the production rules used to generate the image conditioned on the object. Here we consider treating the individual log-likelihood terms as features, and learning a weighted combination of them that can better discriminate between objects and background. The detection algorithm of [1] can then be used as is, as long as the weights are positive.

For training all we know is the label of each training image, while its correct parse is unknown. For this we follow an approach based on *Multiple-Instance Learning (MIL)* [4] which addresses the problem of having limited labeling information. For each training image we use our algorithm to provide us with several local minima of our cost function, which are then grouped in a 'bag' of 'instances' for each training image. We consider that a training sample should be labelled positive if *at least* one of the elements in its bag corresponds to an object, and negative if *none* of its elements does.

These constraints can be incorporated with logistic regression as in [4]. The training criterion is differentiable in the classifier parameters, and is optimized using BFGS. Further, we adopt an iterative algorithm: at each iteration, after estimating the parameters we re-parse the training set, and augment the instances corresponding to each training image with the new parses. We thus deal with the potentially infinite number of possible parses.

References

[1] I. Kokkinos and A. Yuile. HOP: Hierarchical Object Parsing. In *CVPR*, 2009.

[2] I. Kokkinos and A. Yuille. Unsupervised Learning of Object Deformation Models. In *ICCV*, 2007.

[3] P. Felzenszwalb and A. McAllester. The generalized A* Architecture. *Journal of Artificial Intelligence Research*, 2007.

[4] S. Ray and M. Craven. Supervised versus multiple instance learning: an empirical comparison. In *ICML*, 2005.

[5] U. von Luxburg. A Tutorial on Spectral Clustering. Technical report, MPI, 2006.

978-1-4244-3993-5/09/$25.00 ©2009 IEEE