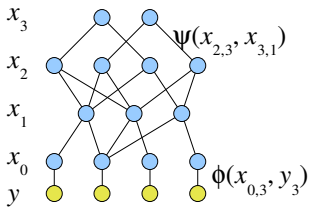
Vision as Inference in a Hierarchical Markov Network

Justus H. Piater, Fabien Scalzo, and Renaud Detry
University of Liège, INTELSIG Group
Grande Traverse 10; 4000 Liège – Sart Tilman, Belgium
Justus.Piater@ULg.ac.be, FScalzo@ULg.ac.be, Renaud.Detry@student.ULg.ac.be

Cortical visual processing involves both bottom-up propagation of perceptual stimuli and modulation by top-down signals. Lee and Mumford (2003) suggested that the visual processing stream from the LGN via V1, V2 and V4 to IT might perform Bayesian inference within an undirected Markov chain. A cortical layer x_i (say, V1) computes its activity $P(x_i \mid x_{k \neq i}) = P(x_i \mid x_{i-1}, x_{i+1}) = P(x_{i-1} \mid x_i)\ P(x_i \mid x_{i+1}) / Z_i$, that is, a posterior probability distribution given bottom-up input from x_{i-1} (LGN), under top-down priors x_{i+1} (V2). The parameters of the Markov network are specified via pairwise compatibility potentials ψ, where $P(x_i \mid x_{i-1}, x_{i+1}) = \psi(x_{i-1}, x_i)\ \psi(x_i, x_{i+1}) / Z_i$. These potentials must be learned from experience with the world; Lee and Mumford (2003) do not comment on how this might be done. A crucial aspect of this model is that ambiguities at low levels should persist and propagate upwards until they can be resolved by integrating larger-scale evidence or top-down expectations. As a biologically plausible implementation of inference with arbitrary, possibly multi-modal probability densities, Lee and Mumford (2003) suggest belief propagation using particle representations. Moreover, they provide a wealth of neurophysiological and psychophysical evidence for such a computational model.

We are currently developing representations and methods for visual inference that constitute, at least at the vague level of detail given above, a working computer implementation of central aspects of Lee and Mumford's model. Without making explicit reference to specific cortical layers, our approach is based on a Markov network such as the didactic example shown on the right, with vertices arranged in layers corresponding to those of Lee and Mumford's. Each vertex is a random variable representing the spatial probability density of the presence of a *feature*. At level 0, a *primitive* feature $x_{0,j}$ is the spatial probability density of a given type of locally observable feature. It is inferred from local image appearance y_j via its observation potential $\phi(x_{0,j}, y_j)$. At higher levels, a *compound* feature (recursively) represents the presence of both of its children, and the compatibility potentials ψ represent pairwise spatial relationships. For example, in the figure, feature $x_{3,1}$ represents the spatial probability density of features $x_{2,1}$ and $x_{2,3}$ occurring in the relative configuration encoded by $\psi(x_{2,1}, x_{3,1})$ and $\psi(x_{2,3}, x_{3,1})$.



We construct such networks, including their topology and compatibility potentials, using unsupervised learning. The input layer y, fixed at the outset, is successively exposed to visual stimuli. The system records the occurrences of known features (primitive or compound), as well as the spatial relations between them. When reoccurring constellations of features $x_{i,a}$ and $x_{i,b}$ are detected, the observed non-uniform spatial co-occurrence probability densities are turned into the compatibility potentials $\psi(x_{i,a}, x_{i+1,c})$ and $\psi(x_{i,b}, x_{i+1,c})$ with respect to a newly-instantiated feature $x_{i+1,c}$, located between them.

In practice, we arrive at networks of on the order of 10 layers and on the order of between 10 and 100 vertices per layer. What they represent depends on the training data. For example, we have trained networks that represent individual objects, from a fixed viewpoint or from a variety of viewpoints. If a network is trained on images of several distinct objects, higher-level vertices will specialize to become view-tuned cells of specific objects. If objects share common parts, these are likely to be represented by the same lower-level subgraphs.

Networks learned in this way can be instantiated on a given input stimulus by computing the observations y, optionally instantiating higher-level vertices according to prior expectations, and performing nonparametric belief propagation using particle representations (similar to Sudderth et al., 2003) throughout the network until convergence. Thanks to bidirectional propagation, the network converges to a globally coherent interpretation of the scene, where each vertex $x_{i,j}$ contains its best possible interpretation of its children, under the priors provided by the parents. We have used this procedure successfully for object detection, recognition, and pose estimation, in 2D and in 3D, as well as for inference of occluded object parts.

Lee, T. S. and Mumford, D., 2003. Hierarchical Bayesian inference in the visual cortex, J. Opt. Soc. Am. 20(7), 1434-1448.

Sudderth, E. B., Ihler, A. T., Freeman, W. T., and Willsky, A. S., 2003. Nonparametric Belief Propagation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (1), 605-612.

12th International Conference on Cognitive and Neural Systems (ICCNS), May 14-17, 2008
This work was supported by the Belgian National Fund for Scientific Research (FNRS)
and by the EU Cognitive Systems project IST-FP6-IP-027657 PACO-PLUS.