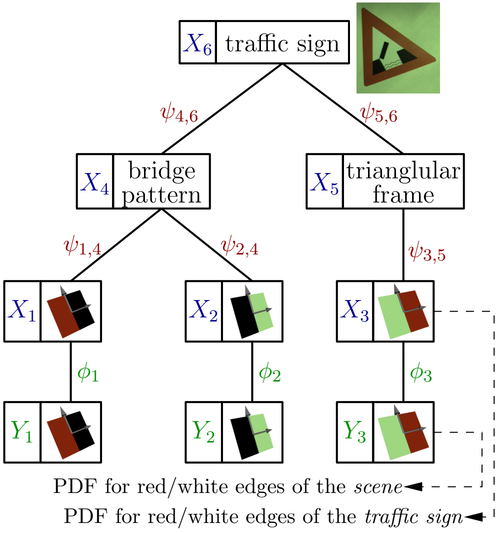
3D Probabilistic Representations for Vision and Action

Justus H. Piater and Renaud Detry
INTELSIG Group, Department of Electrical Engineering and Computer Science
University of Liège, Belgium

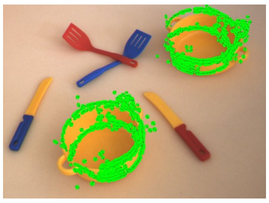
Autonomous robots must be able to construct their own representations that enable them to interact successfully with their environment. In less-than-tightly controlled environments, adequate management of (perceptual and action-related) uncertainty is crucial.

We present a framework for 3D visual representations that can be learned from visual training data without requiring external supervision. Once obtained, such representations can be used for fundamental interactive tasks such as object detection, recognition, and pose estimation. Moreover, they can serve as a basis for learning manipulative interaction.



Our framework represents an object by a Markov network, as illustrated on the left. Vertices are arranged in hierarchical layers. Each vertex is a random variable representing the spatial probability density of the presence of a *feature*. At the bottom level, a *primitive* feature (X_1, X_2 or X_3) represents the pose probability density of a given type of locally observable feature. It is inferred from a local observation Y_i via its observation potential ϕ_i. At higher levels, a *compound* feature (recursively) represents the presence of both of its children, and the compatibility potentials ψ represent pairwise relative 3D poses.

Primitive features are local, oriented 3D patches reconstructed by a biologically-inspired stereo vision system [3]. A set of such features constituting an object can be obtained e.g. via motion segmentation. Given one or more such scene reconstructions, we can construct our hierarchical representation, including its topology and compatibility potentials. Starting from the bottom layer, vertices at the next level are iteratively instantiated by combining pairs of features that reoccur at stable relative configurations. The compatibility potentials between the new feature and its two constituents is derived from the observed nonuniform relative-pose densities.



Representations learned in this way can be instantiated on a given stereo reconstruction by computing the observations Y_i from the reconstructed 3D patches, and then performing nonparametric belief propagation (NBP) throughout the network. Upon convergence, higher-level level vertices contain the probability density over object poses. Thanks to the probabilistic inference, the system exploits any available evidence without imposing arbitrary constraints, is robust to clutter, and can infer the poses of occluded parts. Except for degenerate input, the accuracy of the resulting pose estimate depends mostly on the number of particles used for NBP. For example, the incorrect result for the upper right dish is corrected by increasing the number of particles.

Importantly, nonvisual features can easily be incorporated into this framework. In this way, one may, for example, learn grasps by associating gripper positions to an object's representation.

Most of this work has already appeared elsewhere [1, 2].

[1] R. Detry and J. Piater. Hierarchical integration of local 3D features for probabilistic pose recovery. In *Robot Manipulation: Sensing and Adapting to the Real World (Workshop at RSS)*, 2007.

[2] R. Detry, N. Pugeault, and J. Piater. Probabilistic pose recovery using learned hierarchical object models. In *International Cognitive Vision Workshop (at ICVS)*, 2008.

[3] N. Krüger and F. Wörgötter. Multi-modal primitives as functional models of hyper-columns and their use for contextual integration. In *BVAI*, LNCS 3704, pp. 157–166, 2005.