Learning to Sense Robustly and Act Effectively

Benjamin Kuipers, P.I. and Silvio Savarese, co-P.I.

EECS, University of Michigan,
Ann Arbor, Michigan

Our project explores the hypothesis that, to understand its dynamic environment, an intelligent agent must learn foundational concepts such as Space, Object, and Action from its own interactions with its world. Our Spatial Semantic Hierarchy (SSH) [Kuipers, 2000; Beeson, et al, 2010] represents knowledge of large-scale space. We have developed the Object Semantic Hierarchy (OSH) [Xu & Kuipers, 2010] to represent knowledge of objects, and QLAP [Mugan & Kuipers, 2009] to represent knowledge of actions. Our collaboration integrates these concepts with state-of-the-art work in computer vision.

Building on the OSH, and treating the surrounding environment as an "object", Tsai, Xu, Liu & Kuipers [ICCV, 2011] present a new method whereby an embodied agent using visual perception can efficiently create a model of a local indoor environment from its experience moving within it.
Our method uses a single-image analysis, not to attempt to identify a single accurate model, but to propose a set of plausible hypotheses about the structure of the environment from an initial frame. We then use data from subsequent frames to update a Bayesian posterior probability distribution over the set of hypotheses. The likelihood function is efficiently computable by comparing the predicted location of point features on the environment model to their actual tracked locations in the image stream. Our method runs in real time, and avoids the need for extensive prior training and the Manhattan-world assumption, which makes it more practical and efficient for an intelligent robot to understand its surroundings compared to most previous scene understanding methods. Experimental results on a collection of indoor videos suggest that our method is capable of an unprecedented combination of accuracy and efficiency.

Another major contribution is a new Bayesian probabilistic framework for joint object detection and scene 3D layout estimation from a single image. Co-PI Savarese and his students explore novel ways of encoding the geometrical relationship among objects, the physical space and the observer [Bao, Sun, Savarese CVPR 2010]. In this framework, object detection becomes more and more accurate as additional evidence about a specific scene becomes available. In turn, improved detection results enable more stable and accurate estimates of the scene layout and object supporting surfaces. Extensive quantitative and qualitative experimental analysis on existing and newly proposed datasets validates the theoretical claims.

A final major effort is dedicated to the problem of activity recognition from videos. Liu, Kuipers & Savarese [CVPR, 2011] explore the idea of using high-level semantic concepts (attributes) to represent human actions from videos and to enable the construction of more descriptive models for human action recognition. A unified framework is proposed wherein attributes are: i) selected in a discriminative fashion so as to account for intra-class variability; ii) integrated coherently with data-driven attributes to make the attribute set more descriptive; iii) effectively used for classifying novel action classes for which no training samples are available (zero-shot learning).