Recognition and Pose Estimation for Manipulating Native Objects

Author Names Omitted for Anonymous Review. Paper-ID [add your ID here]

***Abstract*—Robots need to pick stuff up. We feel that perception and planning will both benefit from eachother's feedback. So we contribute a *framework* which includes the structure of a *memory network* and the *algorithms* (methods) that we use to maintain the network. Perception can structure data, structured data facilitates planning, planning allows more principled perception. We form a closed loop through perception and planning, allowing them to interact by means of structured data. Our system is supported by online, human-in-the-loop algorithms. Non-technical human participants can easily teach and collaborate with the system.**

I. Problem Setting

II. Some Standard Tools

Here we review the standard AI, vision, and ML algorithms we employ.

III. Structured Memory Facilitates Planning

Here we symbolically describe our memory network, which is the first contribution of the paper and the first half of our framework.

Working Memory What object tokens exist in the world? What are their categories and their names? (names map particular object instances I see now to instances that persist over time, that is, for which I have a history) Where are they in the scene?

Long Term Memory Intrinsic (Autobiographical) Episodic Memory This is the sensory data that we are collecting. This information cannot be provided by external sources, it comes from the perceptual capabilities of the system. Traditionally this would be thought of as Training Data X. Extrinsic Episodic Memory Labels and names for categories and peristant instances. This information can be provided by another source and will eventually be provided by the entity itself. Traditionally this would be thought of as Training Labels Y. Implicit Memory These are the models, such as SVM or kNN, that are involved in mapping intrinsic memory to the extrinsic memory. Semantic Memory Facts about stuff that can be used to reason. Affordances Procedural Memory Grasps and collision free planning are procedural memory.

Short Term Persistent Memory Structurally identical to long term memory but restricted to the current context. This is much like a cache for long term memory. Short term memories are incorporated into long term memory either online through the day or offline during a sleep phase.

IV. Online Learning With Humans in the Loop

Here we describe the algorithms which we use to build and grow the memory network, which are the second contribution and second half of our framework. Note that performing inference with the framework is actually the same as growing the memory, since the results of inference are recorded in working and short term memory directly and strongly contribute to long term memory.

V. Evaluation

Our system can be evaluated in two important ways. Firstly, how effective is the system at facilitating other actions? Secondly, how accessible to users is the system?

Because this system is designed as a component in a complex system, raw recognition rates and pose estimation accuracy would not be very informative even if it were not impractical to obtain them. Besides, we are not trying to extend the state-of-the-art on those tasks. Rather, we are trying to provide an interactive framework which will raise the maximum automated vision available to the average user.

We therefore establish a baseline for performance by training the system in a representative domain specific setting, which tells us how well it can perform on laboratory objects when trained by an expert. This represents the best that the system could be expected to perform.

We then go on to compare the performance of the system when trained to various degrees by naive and technical non-expert users.

A. Baseline Domain Specific Pick and Place

We qualitatively evaluate the performance by picking and placing a respresentative set of objects and reporting the final success rate for each object separately.

We report the performance of the system as a function of user interactions.

We report the performance of the system as a function of program lifetime.

Our representative set consists of a block, a spoon, a bowl, a diaper, and a sippy cup. A *single cut video* showing multiple grasps of all objects is available here.

B. Usability Experiments

We repeat the DSPP experiments with naive users in two settings. In the first, they train laboratory objects. In the second, they provide their own objects.

[1]

VI. Extensions

References

[1] S. Tellex, T. Kollar, S. Dickerson, M.R. Walter, A. Banerjee, S. Teller, and N. Roy. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proc. AAAI*, 2011.