Towards Bootstrap Learning for Object Discovery *

Joseph Modayil and Benjamin Kuipers
Computer Science Department
University of Texas at Austin
Austin, Texas 78712 USA
{modayil,kuipers}@cs.utexas.edu

Abstract

We show how a robot can autonomously learn an ontology of *objects* to explain aspects of its sensor input from an unknown dynamic world. Unsupervised learning about objects is an important conceptual step in developmental learning, whereby the agent clusters observations across space and time to construct stable perceptual representations of objects. Our proposed unsupervised learning method uses the properties of allocentric occupancy grids to classify individual sensor readings as static or dynamic. Dynamic readings are clustered and the clusters are tracked over time to identify objects, separating them both from the background of the environment and from the noise of unexplainable sensor readings. Once trackable clusters of sensor readings (i.e., objects) have been identified, we build shape models where they are stable and consistent properties of these objects. However, the representation can tolerate, represent, and track amorphous objects as well as those that have well-defined shape. In the end, the learned ontology makes it possible for the robot to describe a cluttered dynamic world with symbolic object descriptions along with a static environment model, both models grounded in sensory experience, and learned without external supervision.

Introduction

Part of the symbol anchoring problem is understanding how an agent *learns* this skill. It is a daunting task to anchor symbols to the environment in a large intelligent system. We intend to make the problem of anchoring symbols more tractable by examining how an agent can construct its own symbol representations directly from sensation.

For a robot to learn about an unknown world, it must learn to identify the objects in it, what their properties are, how they are classified, and how to recognize them. The robot's sensorimotor system provides a "pixel-level" ontology of time-varying sensor inputs and motor outputs. Even after a substantial learning process (Pierce & Kuipers 1997) provides the organization on the sensors along with the ability to follow control laws and defines distinctive states to describe the large-scale structure of the environment, the robot's ontology still does not include *objects*. In this paper, starting from a lower-level ontologies that includes egocentric range sensors and incremental motion, and an occupancy grid model of the local environment, we show how an ontology of objects can be learned without external supervision. These generated representations facilitate the creation of controls, the recognition of objects, and the development of object based rules. The method is designed to work for a mobile robot; it works in unstructured environments, it uses online algorithms, and it is computationally efficient.

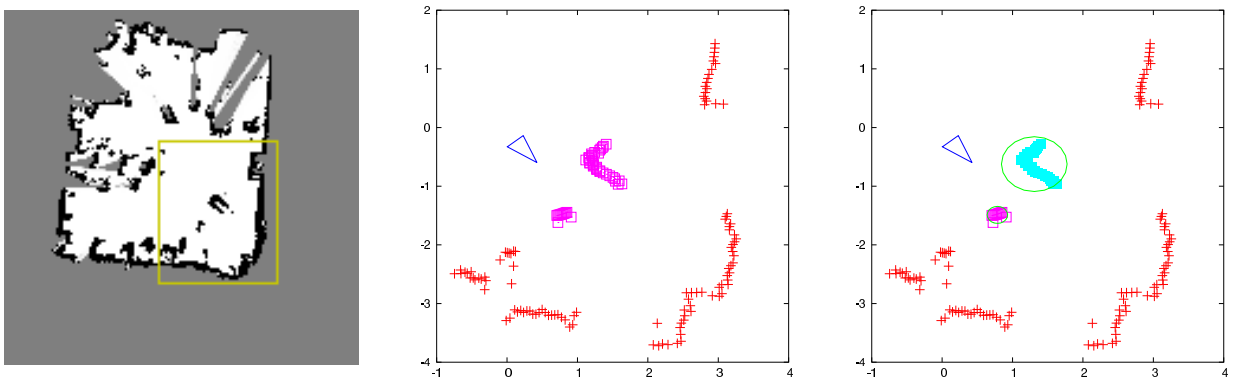
Learning about Objects

We claim that a robot can learn a working knowledge of *objects* from unsupervised sensorimotor experience by representing moveable objects in four steps: Individuation, Tracking, Image Description, and Categorization. We demonstrate this learning process using a mobile robot equipped with a laser range sensor, experiencing an indoor environment with significant amounts of dynamic change.

This is a kind of "bootstrap learning" (Kuipers & Beeson 2002) since we combine multiple learning stages, each stage learning the prerequisites for subsequent stages. In particular, recognition depends on image description that relies on tracking that in turn relies on individuation. The described sequence of stages provides an initial pathway for developing object representations before rich prior knowledge is available. In future work, we intend to show how this initial sequence can be used to gather informative prior knowledge that is required for other algorithms (Schulz & Burgard 2001).

A major motivation for this work is to understand how complex cognitive structures can autonomously develop in a learning agent. We know that tremendous leaps in cognitive complexity occur through evolution and during infant development, using high dimensional sensory experience acquired in unconstrained environments. Computational learning theory tells us that learning is exponentially hard in the dimensionality of the representation space (Hastie, Tibshirani, & Friedman 2001). Learning in a high dimensional representation space (such as an observation stream) should be vastly harder than learning in a low dimensional (sym-

*This work has taken place in the Intelligent Robotics Lab at the Artificial Intelligence Laboratory, The University of Texas at Austin. Research of the Intelligent Robotics lab is supported in part by the Texas Higher Education Coordinating Board, Advanced Technology Program (grant 003658-0656-2001), and by an IBM Faculty Research Award.
Copyright © 2004, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.



(a) (b) (c)

Figure 1: Object Individuation. (a) The occupancy grid representation of the environment generated online by a SLAM algorithm up to the current time t. The boxed region is shown in the following plots. (b) Sensor readings at time t classified as static ($+$) or dynamic ($\square$) according to the occupancy grid cells they fall on. The robot ($\triangleright$) is in the upper-left portion of the plot, so nearby dynamic objects occlude parts of the static environment. (c) Dynamic readings are clustered and hence individuated into objects. Each of the two clusters is assigned to a tracker (circles). [*All of these figures are clearer in the color PDF than in grayscale prints.*]

bolic) representation. The premise of bootstrap learning is that an agent can apply a variety of high bias, but unsupervised learning algorithms to simple tasks (recognizing movable objects) to transform a high dimensional representation (an observation stream) into one with significantly lower dimension (a symbolic representation).

Individuation

The process of individuation starts by using an occupancy grid to classify sensor readings. The occupancy grid representation for local space does not include the concept of "object." It assumes that the robot's environment is static, that it can be divided into locations that are empty and those that are occupied. A cell of an occupancy grid holds the probability that the corresponding region of the environment is occupied. Simultaneous localization and mapping (SLAM) algorithms can efficiently construct an occupancy grid map and maintain accurate localization of a mobile robot within it using range sensor data (Moravec 1988; Thrun, Fox, & Burgard 2000; Eliazar & Parr 2003).

The occupancy grid representation embodies a static world assumption. Sense data reflecting dynamic change in the environment are treated as noise. Fortunately, occupancy grid algorithms are quite robust to failures of the static world assumption. If changes in the environment are slow relative to repeated observation (12 Hz for the laser range-finder), changes in occupancy are quickly washed out by new observations, restoring the grid to a reasonably accurate description of the current state of the environment. We exploit this property and add a new attribute to the occupancy grid. A grid cell is labeled *transient* if it has ever been unoccupied (i.e., the probability of occupancy falls below a threshold), and *permanent* if it has never been unoccupied.[1]

The low-resolution occupancy grid cell labeling is used to classify individual high-resolution range sensor readings. Each individual range sensor reading is labeled as *static* or *dynamic*, depending on whether the endpoint of the reading falls in a cell labeled as permanent or transient, respectively. Permanent grid cells and static sensor readings represent the static background environment, and the learning algorithm restricts its attention to the dynamic range sensor readings. Note that a non-moving object such as a trash bin would be labeled with dynamic sensor readings if the robot had *ever* observed the space the readings are located in as unoccupied.

Next, the learning algorithm clusters the endpoints of the dynamic range sensor readings.[2] The coordinates of the endpoints x_i are represented in the fixed local frame of reference of the occupancy grid. Two endpoints are considered close if their distance is less than the threshold value δ_I:

$$close(x_i, x_j) \equiv \|x_i - x_j\| < \delta_I.$$

The individual clusters are the connected components of the $close$ relation: i.e., the equivalence classes of its transitive closure. Within a single observation frame at time t, these clusters $\{S_{i,t}\}$ are called *object snapshots*. They are the initial representation for individual objects. The process of individuation is shown in Figure 1.

Tracking

An object snapshot $S_{i,t}$ at time t has a spatial location and extent $<\mu_i, r_i>$: its center of mass μ_i and the distance r_i

[1] To account for small localization errors, a transient cell may also require that all of its neighbors cells are unoccupied, which leaves permanent cells surrounded by a thin rim of unlabeled cells.

[2] Recall that the endpoints of range sensor readings, like the localization of the robot, are not limited to the resolution of the occupancy grid, but have real-valued coordinates, albeit with limited precision and accuracy.

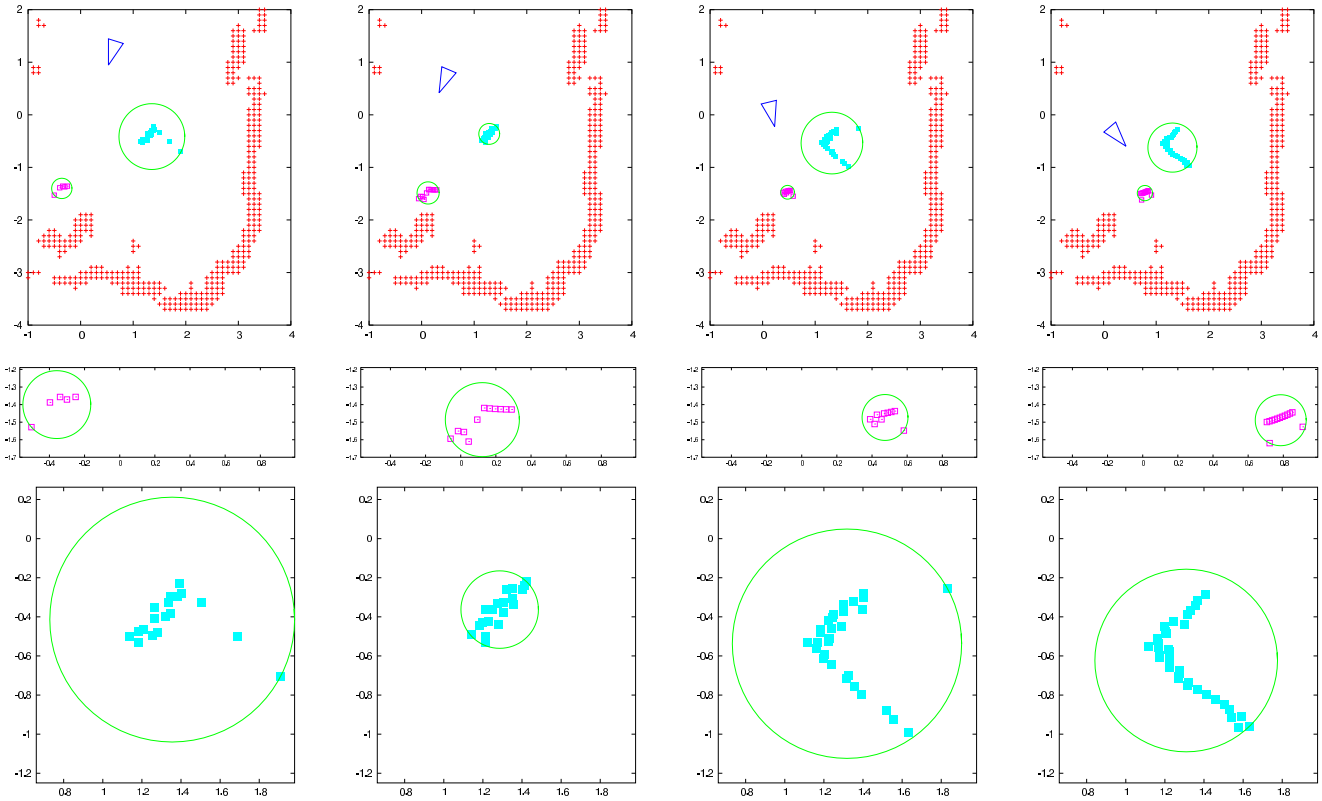


Figure 2: Object Tracking. The shape of an object can vary greatly during tracking whether it has a rigid body or not. This figure shows a sequence of time steps prior to the scene in Figure 1. The actual trackers use data at much finer temporal granularity than the time-points (columns) shown. Note that the robot is moving while tracking. **Top**: The tracked dynamic objects, superimposed for reference on a low-intensity display of the permanent cells in the occupancy grid. **Middle**: A tracked pedestrian object, showing its irregular shape over time. **Bottom**: Tracked snapshots of a non-moving object (an ATRV-Jr).

from its center of mass to its farthest reading. The dissimilarity between two snapshots S_i and S_j is

$$d_S(S_i, S_j) = \|\mu_i - \mu_j\| + |r_i - r_j|.$$

This function is robust to random noise and incorporates both the observed center and radius since the snapshots of a moving, dynamic object (such as a person) will vary in both dimensions. Where the successor to time t is t', we say that object snapshot S_t has *unique clear successor* $S'_{t'}$ if

$$d_S(S_t, S'_{t'}) < \delta_T \text{ and}$$

$$\forall S''_{t'} \neq S'_{t'} \quad d_S(S_t, S''_{t'}) > d_S(S_t, S'_{t'}) + \delta_R.$$

An *object tracker* is a function $T_k(t)$ whose value is an object snapshot $S_{i,t}$ at time t, such that for successive time-points t and t', $T_k(t')$ is the unique clear successor of $T_k(t)$. An object tracker T_k thus defines a collection of corresponding object snapshots extending from frame to frame in the observation stream, with at most one snapshot in each frame. The process of object tracking is depicted in Figure 2.

The domain of a particular object tracker ends at the time-points where the *unique clear successor* relation cannot be extended. "Object permanence", the ability of an object tracker to tolerate breaks in the sequence of frames, is clearly a learned ability in young children (Spelke 1990). Our current implementation includes the ability to tolerate two missing frames in a sequence. Three missing frames terminates a tracker. New trackers are generated for large unexplained snapshots. Small snapshots without trackers are treated as noise and ignored.

Dynamic objects being tracked will converge and diverge, for example pedestrians in a crowded hallway. We have not incorporated velocity estimation into the tracker since it increases the complexity of state estimation. Object trackers will successfully track individuals over segments of their behavior, losing them when they get too close together and their readings are merged into a single snapshot. When they separate again, new trackers will be created to track the different individuals. More sophisticated methods for "object permanence" will be required to infer the identity of object trackers across such merges and splits. Following our bootstrap learning approach, we learn properties of objects during the periods of time when tracking is unambiguous and learning is easy. We expect those properties will make it possible to track objects under more difficult circumstances.

We define these trackable clusters of dynamic sensor readings to be objects. Each tracker represents a distinct sym-

bolic identity which is assumed to be the cause of the readings associated with it. At this point, objects have only two properties: spatial location and temporal extent. These properties are sufficient for the trackers to guide the robot's actions to acquire additional information about the object. For example, control laws for following, circling and avoidance are easily specified using trackers to specify the desired goals. The next step will be to acquire properties of the object instances that are stable across changes in space and time. This makes it possible to categorize them into object classes.

Image Description

We have defined the *object snapshot* to be the set of sensor readings associated with an object at a particular time. The *shape model* for an object is a subset of the object snapshots collected over the time that the object is tracked.

The problem is how (and whether) the snapshots can be aggregated into a consistent, object-centered frame of reference. We consider it important to describe both objects with stable shapes that can be learned, and objects that are *amorphous* in the sense that they can be individuated and tracked, but their shape is beyond the capacity of the agent to describe and predict. For our robot learning agent, at its current level of sophistication, *pedestrians* are good examples of amorphous objects. At a later stage, the learning agent may be able to model a pedestrian as two alternately-moving legs (observed as 2D blob shapes), but for now, object snapshots of pedestrians change too much to form stable shape models.

Consider a temporarily non-moving object such as an ATRV-Jr (a mobile robot). To be individuated and tracked as an object, it must be located at a position that was unoccupied at some time, so its sensor readings are considered dynamic. Since the object doesn't move in the environment, tracking is quite simple. However, as the robot moves around it, the object snapshot still changes slowly (Figure 2).

The agent creates a shape model by accumulating distinctive snapshots while the object appears to be non-moving (Figure 3). Both tasks, detecting the lack of object motion and determining distinctiveness, are accomplished by a non-symmetric dissimilarity function d_D that compares snapshots.

$$d_D(S_{new}, S_{old}) = \frac{1}{|S_{new}|} \sum_{s \in S_{new}} \min(1, \frac{1}{\epsilon} \min_{t \in S_{old}} \|s - t\|)$$

When successive snapshots differ by a large amount, δ_M, the agent assumes the object has moved, and discards the current shape model. Otherwise, if the current snapshot is sufficiently distinct, δ_N, from snapshots currently in the shape model, the new snapshot is added to the shape model. Finally, snapshots in the shape model are discarded if they are incompatible with the full set of current sensor readings.

The shape model also records the directions from which the snapshots have been observed, and is considered *complete* when the full 360° surround has been sufficiently densely sampled.[3]

[3]In the current implementation, this means at least one snapshot exists in each of six 60° pose buckets around the object.

While the shape model is incomplete, it is considered "amorphous". When the shape model is complete, the agent creates a *standard shape image* for the object by placing the snapshots of the shape model into a canonical frame of reference. The snapshots are first rotated so that the primary axis of the readings is aligned with the y-axis. This is accomplished by rotating the shape model to minimize the entropy of the projection onto the x-axis. Next, the shape model is translated to minimize the distance of the farthest points from the origin. (See Figure 4.)

Categorization

Once an individual object has a standard shape image, the agent must categorize it. Note that the learning agent is responsible for building its own classes. Moreover, since the object observations come in incrementally, the agent must add new classes incrementally. The task of adding new classes incrementally is known as online clustering, and several algorithms exist (Duda, Hart, & Stork 2001). For simplicity however, we solve this clustering task with a distance function.

We define the asymmetric dissimilarity function between two aligned shape images V and W by comparing their component snapshots

$$d'(V, W) = \frac{1}{|V|} \sum_{v \in V} \min_{w \in W} d_D(v, w).$$

We use this to define the symmetric distance measure

$$d_C(V, W) = \max(d'(V, W), d'(W, V)).$$

If the image of an instance is less than a threshold distance, δ_C, from multiple known types, then its classification is uncertain. If there is only one known type within δ_C, then it is classified as that type. If it is more than δ_C from any known type, then a new category is formed. For example, when the shape model in Figure 3 is converted into a standard shape image and compared to the known categories in Figure 4, it is recognized as an instance of the ATRV-Jr category. It is then displayed as a known type in Figure 5(d).

The robot does not learn a shape model by observing a continuously moving object, but it can learn a shape model if the object stops for a short period. Once an object has been classified, the tracker retains this classification and the corresponding shape model even when perception is difficult. Furthermore, the robot can obtain individual snapshots of a moving object, and we predict that those snapshots will be useful as evidence toward the classification of a moving object within an existing class hierarchy.

Even without a complete shape model, the robot can still generate a standard shape image for an object. For an incomplete image, the dissimilarity function is useful because it has the property that if $V \subset W$, then $d'(V, W) = 0$. This makes it suitable for comparing an incomplete model of an instance V with complete models that are already known. Also, this can be used to guide active perception by defining the observations that are most informative for classification.

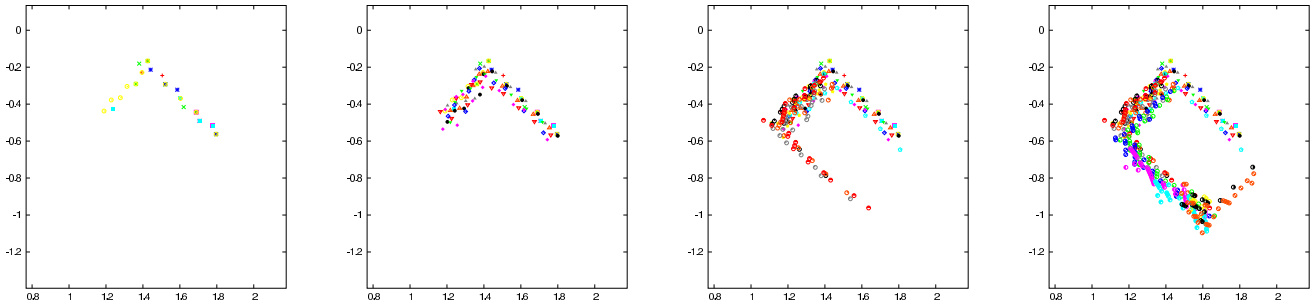
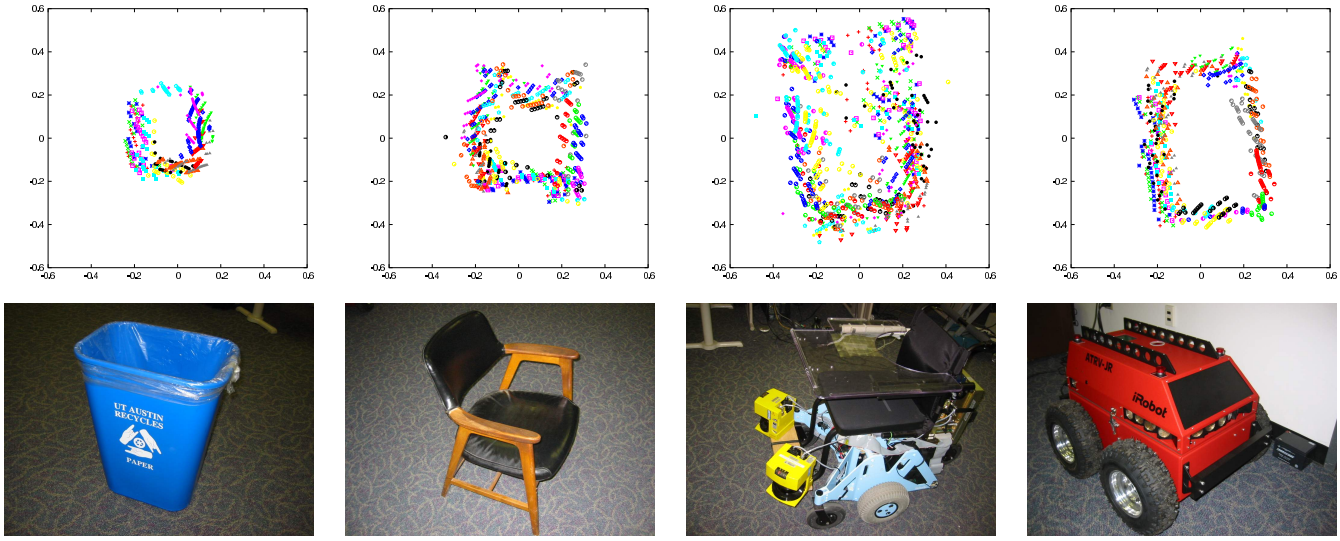


Figure 3: Object Shape Model. This shows the incremental shape model creation for the ATRV-Jr observed in Figure 2. The range sensor endpoints in each snapshot are shown with different symbols. Selected snapshots combine to form a shape model.



(a) (b) (c) (d)

Figure 4: Categorization entails both clustering and classification. Standard shape images and photographs for four learned object classes: (a) recycling bin, (b) chair, (c) robot wheelchair, and (d) an ATRV-Jr robot.

Experimental Results

The above system was implemented on a RWI Magellan Pro robot equipped with a SICK PLS laser rangefinder. The parameters mentioned in the paper had the following values: $\delta_I = 0.5m$, $\delta_T = 1.0m$, $\delta_R = 0.01m$, $\delta_M = 0.5$, $\delta_N = 0.1$, and $\delta_C = 0.33$. A set of learned shape images is shown in Figure 4.

An occupancy grid representation of the environment (shown in Figure 1(a)) is generated online in the presence of object motions. The process of individuation is displayed in the subsequent two images, first showing the classification of laser scans as static or dynamic, and then clustering the dynamic readings to form snapshots. The snapshots are associated with trackers in Figure 2, providing temporal extent to the object representation. The ATRV-Jr robot is not moving during this time, so an image description is incrementally accumulated, as shown in Figure 3. When the description is sufficiently complete, the agent compares it to the set of objects, shown in Figure 4. The agent discovers that the image description best matches that of the ATRV-Jr robot.

The agent's world description is graphically represented in Figure 5 along with a photo of the same scene. The result is a discretization of a natural environment into several entities which are useful for later reasoning: a coarsely represented fixed environment (walls+furniture), a localized agent (the Magellan Pro robot), an amorphous moving object (a pedestrian), and a classified known object (the ATRV-Jr). Moreover, since the agent can autonomously generate new categories online, its ability to succinctly and accurately describe nearby objects should improve with experience.

These results demonstrate several qualitative goals. The system is able to go from an ontology of sensation to an ontology of objects. It can learn quickly without requiring complex *a priori* models. However, it gathers statistics of objects, and these statistics will be useful for the principled construction of prior models in future work. The system creates symbols that are ground in sensation. We have not performed a quantitative evaluation because it is not yet clear

which quantities accurately capture our goal of understanding how an agent can learn to create objects from sensations. While the components of the system could be compared to state of the art algorithms, the comparison is not meaningful. Hence, we are presenting purely qualitative results in this paper.

Related Work

There are multiple earlier articles that motivate both the need for the gradual construction of an object ontology, and the techniques we employed.

There is a large body of literature on individuation in both psychology and computer vision. Work in developmental psychology (Spelke 1990) suggests that infants learn Gestalt principles of perception. Work in perceptual psychology (Geisler & Diehl 2003) demonstrates that the natural statistics of the environment can provide sufficient training data for acquiring grouping mechanisms. Individuation in vision has been achieved by a variety of criteria using the normalized cut algorithm (Shi & Malik 2000).

Our approach to tracking unknown objects provides the learning agent with symbols (trackers) that are ground in the sensory experience (snapshots). Issues related to anchoring symbols have been explored in (Coradeschi & Saffiotti 2001). This work describes how anchored symbols can be used for planning and reasoning in a dynamic world.

The ARGUS vision system (Gribble 1995) used local windows on visual features to track moving objects through the visual field. The tracker made it possible to collect and analyze visual shape data in a stable object-centered frame of reference, even while the object itself was moving too quickly for traditional methods to recognize it. We use a similar idea for describing shapes of moving objects.

There is extensive work on view based categorization, particularly in the vision community. The "Chorus of Prototypes" (Edelman 1999) uses prototype silhouettes to efficiently index a space of object views. Using distances between views provides an efficient mechanism for determining when a view is sufficiently distinct from previously known models. We intend to incorporate these ideas when our system learns sufficiently many objects.

Multiple researchers have examined how prior knowledge can be used to improve performance on various state estimation tasks. With prior shape models, it becomes possible to estimate pose of an object in the environment (Schulz & Burgard 2001). Articulated shape models can be created from three dimensional range snapshots, provided the snapshots correspond to the same object (Anguelov *et al.* 2004).

State estimation techniques can also be applied to create shape models from occupancy grids (Biswas *et al.* 2002; Anguelov *et al.* 2002), thereby generating new object types. They assume that the world is static during observation, which permits the use of a standard SLAM algorithm to capture the shape of the objects in a grid representation. The assumption that the entire environment stays static is fairly restrictive, since many environments and objects of interest move regularly. Moreover, their algorithm uses an offline learning process. This makes the online incremental acquisition of new object types difficult.

Finally multiple studies describe how models of objects can be acquired in an unsupervised manner.

The construction of shape models of non-rigid objects has been explored in (Hähnel, Thrun, & Burgard 2003). Using a variant of the iterative closest point algorithm, they are able to merge dense three-dimensional range scans into a single coherent shape model even when the object undergoes small motions. This algorithm creates a qualitatively consistent model when an person moves their arms or head between successive scans. Because it relies on having significant amounts of data to align the scans, it is unclear that this method can be extended to handle non-rigid motion as observed by a two-dimensional range scanner.

Recent work on the Navlab project (Wang, Thorpe, & Thrun 2003) has demonstrated the feasibility and value of tracking unknown objects in the environment. This work describes how a truck equipped with multiple range sensors is able to detect and track moving objects while driving down a road. The ability to track unknown moving objects is required for their goal of safe autonomous control at high speeds on urban streets. They are also able to recognize instances of a few object classes. A significant difference from the work in this paper is their inability to generate new object types.

Work on the VSAM project (Collins *et al.* 2000) demonstrated visual detection and tracking of objects using multiple cameras at fixed locations. Objects are detected and tracked using frame differencing and background subtraction. These objects were later classified using silhouette models of the object shapes. This work does not address the needs of mobile robotics very well, since their vision algorithms rely heavily on fixed locations for the cameras.

Object discovery has been convincingly demonstrated in vision. Work on simultaneous language and object learning has shown impressive results (Yu, Ballard, & Aslin 2003). Using a vector to describe image features, they are able to bind phoneme sequences (words) to objects. A flexible alternative image description is provided by constellations of features (Li, Fergus, & Perona 2003). Segmenting objects in static scenes using the statistics from dynamic scenes has been demonstrated in vision (Ross & Kaelbling 2003). It is difficult to make direct comparisons with these works since vision and range sensors have very different characteristics, though it would be very valuable to integrate the two sensors.

Conclusions and Future Work

We have described and implemented a method for an agent to autonomously learn properties of novel dynamic objects in a natural environment without complex prior knowledge. This paper demonstrates how a learning agent can efficiently build an ontology of objects as part of a bootstrap learning process. Using this autonomously acquired ontology, a robot can categorize the dynamic objects it encounters in the world. This system demonstrates the feasibility of learning to ground object symbols to sensation without supervision.

This work may be incrementally improved in multiple ways. Small errors in localization cause the shape models

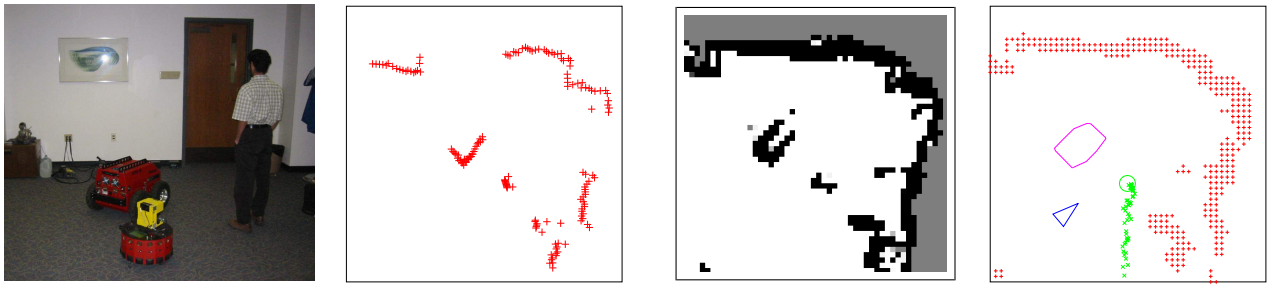


Figure 5: Multiple representations of the scene in Figure 1. The robot observer is the small round robot in the foreground. The larger ATRV-Jr is used as a non-moving object. **(a)**: A photograph of the scene. **(b)**: A range scan representation of the scene. **(c)**: An occupancy grid representation of the scene. **(d)**: An iconic representation of the scene. This is a symbolic description of the robot's environment enabled by the learned object ontology. The location of the observing robot is indicated by a small triangle (▷). A moving object (pedestrian) of amorphous shape is shown with its trajectory. A non-moving object (ATRV-Jr) has been classified (as an instance of Figure 4(d)), and is shown by the convex hull of its shape model. The permanent cells in the occupancy grid are shown for reference, representing the static environment.

to become noisy, a problem that may be alleviated by better snapshot alignment. Also, the method is specified for a range sensor, so testing it with stereo vision is desirable.

An important part of bootstrap learning has not yet been explored here, namely utilizing acquired knowledge to construct informed priors to improve competence in harder tasks. This leads to several directions for future work: examining how class knowledge can aid in image description (by selecting discriminating observation angles), examining how image description can aid in tracking (by providing feedback on the plausible motion of the object), and using tracking to aid in individuation (by providing feedback for separating objects). Finally, we would like to examine how the learned object ontology can be used to speed up further learning tasks.

References

Anguelov, D.; Biswas, R.; Koller, D.; Limketkai, B.; and Thrun, S. 2002. Learning hierarchical object maps of non-stationary environments with mobile robots. In *Proc. Uncertainty in Artificial Intelligence (UAI)*.

Anguelov, D.; Koller, D.; Pang, H.; Srinivasan, P.; and Thrun, S. 2004. Recovering articulated object models from 3D range data. In *Proceedings of the Uncertainty in Artificial Intelligence Conference(UAI)*.

Biswas, R.; Limketkai, B.; Sanner, S.; and Thrun, S. 2002. Towards object mapping in non-stationary environments with mobile robots. In *IROS*, 1014–1019.

Collins, R. T.; Lipton, A. J.; Kanade, T.; Fujiyoshi, H.; Duggins, D.; Tsin, Y.; Tolliver, D.; Enomoto, N.; Hasegawa, O.; Burt, P.; and Wixson, L. 2000. A system for video surveillance and monitoring. Technical Report CMU-RI-TR-00-12, The Robotics Institute, Carnegie Mellon University.

Coradeschi, S., and Saffiotti, A. 2001. Perceptual anchoring of symbols for action. In *Proc. 17th Int. Joint Conf. on Artificial Intelligence (IJCAI-01)*, 407–412.

Duda, R. O.; Hart, P. E.; and Stork, D. G. 2001. *Pattern Classification*. New York: John Wiley & Sons, Inc., Second edition.

Edelman, S. 1999. *Representation and recognition in vision*. Cambridge, MA: MIT Press.

Eliazar, A., and Parr, R. 2003. DP-SLAM: Fast, robust simultaneous localization and mapping without predetermined landmarks. In *Proc. 18th Int. Joint Conf. on Artificial Intelligence (IJCAI-03)*, 1135–1142. Morgan Kaufmann.

Geisler, W. S., and Diehl, R. L. 2003. A Bayesian approach to the evolution of perceptual and cognitive systems. *Cognitive Science* 27(3):379–402.

Gribble, W. S. 1995. Slow visual search in a fast-changing world. In *Proceedings of the 1995 IEEE Symposium on Computer Vision (ISCV-95)*.

Hähnel, D.; Thrun, S.; and Burgard, W. 2003. An extension of the ICP algorithm for modeling nonrigid objects with mobile robots. In *Proc. 18th Int. Joint Conf. on Artificial Intelligence (IJCAI-03)*, 915–920.

Hastie, T.; Tibshirani, R.; and Friedman, J. 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.

Kuipers, B., and Beeson, P. 2002. Bootstrap learning for place recognition. In *Proc. 18th National Conf. on Artificial Intelligence (AAAI-2002)*, 174–180. AAAI/MIT Press.

Li, F.-F.; Fergus, R.; and Perona, P. 2003. A Bayesian approach to unsupervised one-shot learning of object categories. In *Proc. Ninth IEEE ICCV*, 1134–1141.

Moravec, H. P. 1988. Sensor fusion in certainty grids for mobile robots. *AI Magazine* 61–74.

Pierce, D. M., and Kuipers, B. J. 1997. Map learning with uninterpreted sensors and effectors. *Artificial Intelligence* 92:169–227.

Ross, M. G., and Kaelbling, L. P. 2003. Learning ob-

ject segmentation from video data. Technical Report AIM-2003-022, MIT Artificial Intelligence Lab.

Schulz, D., and Burgard, W. 2001. Probabilistic state estimation of dynamic objects with a moving mobile robot. *Robotics and Autonomous Systems* 34(2-3):107–115.

Shi, J., and Malik, J. 2000. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(8):888–905.

Spelke, E. S. 1990. Principles of object perception. *Cognitive Science* 14:29–56.

Thrun, S.; Fox, D.; and Burgard, W. 2000. Monte Carlo localization with mixture proposal distribution. In *Proc. 17th National Conf. on Artificial Intelligence (AAAI-2000)*, 859–865. AAAI Press/The MIT Press.

Wang, C.-C.; Thorpe, C.; and Thrun, S. 2003. Online simultaneous localization and mapping with detection and tracking of moving objects: theory and results from a ground vehicle in crowded urban areas. In *IEEE International Conference on Robotics and Automation*, 842–849.

Yu, C.; Ballard, D. H.; and Aslin, R. N. 2003. The role of embodied intention in early lexical acquisition. In *25th Annual Meeting of Cognitive Science Society (CogSci 2003)*.