Bootstrap learning of foundational representations*

B. Kuipers, P. Beeson, J. Modayil and J. Provost
Computer Science Department
University of Texas at Austin
Austin, Texas 78712 USA

{kuipers,pbeeson,modayil,jp}@cs.utexas.edu
Tel: 1-512-471-9561
Fax: 1-512-471-8885

January 27, 2006

Abstract

To be autonomous, intelligent robots must learn the foundations of commonsense knowledge from their own sensorimotor experience in the world. We describe four recent research results that contribute to a theory of how a robot learning agent can bootstrap from the 'blooming buzzing confusion' of the pixel level to a higher-level ontology including distinctive states, places, objects, and actions. This is not a single learning problem, but a lattice of related learning tasks, each providing prerequisites for tasks to come later. Starting with completely uninterpreted sense and motor vectors, as well as an unknown environment, we show how a learning agent can separate the sense vector into modalities, learn the structure of individual modalities, learn natural primitives for the motor system, identify reliable relations between primitive actions and created sensory features, and can define useful control laws for homing and path-following.

Building on this framework, we show how an agent can use self-organizing maps to identify useful sensory featurs in the environment, and can learn effective hill-climbing control laws to define distinctive states in terms of those features, and trajectory-following control laws to move from one distinctive state to another. Moving on to place recognition, we show how an agent can combine unsupervised learning, map-learning, and supervised learning to achieve high-performance recognition of places from rich sensory input. And finally, we take the first steps toward learning an ontology of objects, showing that a bootstrap learning robot can learn to individuate objects through motion, separating them from the static environment and from each other, and can learn properties useful for classification. These are four key steps in a larger research enterprise on the foundations of human and robot commonsense knowledge.

Keywords: bootstrap learning; ontology learning; spatial learning; learning places, objects, actions.

*To appear in *Connection Science* **18**.2, 2006, special issue on Developmental Robotics.

1 Introduction

Commonsense knowledge is a bottleneck problem on the way to artificial intelligence [McCarthy, 1968]. Common sense, and hence most other human knowledge, is built on knowledge of a few foundational domains, such as space, time, objects, action, causality, and so on [Piaget and Inhelder, 1967, Minsky, 1975]. Spatial knowledge is arguably the most fundamental of these foundational domains [Lakoff and Johnson, 1980]. We are investigating how the foundations of spatial knowledge can be learned from unsupervised sensorimotor experience.

We have done extensive work on human and robot knowledge of large-scale space (the cognitive map), leading to the Spatial Semantic Hierarchy [Kuipers, 2000, Remolina and Kuipers, 2004, Kuipers et al., 2004]. The multiple levels of the SSH demonstrate how higher levels of representation can be based on lower, simpler levels. The SSH Control level, the lowest, makes a good target for bootstrap learning.

The basic idea behind bootstrap learning is to reach a learning goal by composing multiple simple machine learning methods, using weak but general learning methods to create the prerequisites for applying stronger but more specific learning methods. The result is a lattice of learning methods that collectively learn the desired knowledge.

We assume that a learning agent[1], human or robot, starts with a low-level ontology for describing its sensorimotor interaction with the world. William James called this the "blooming buzzing confusion" that confronts the infant from its unfamiliar senses. From a robotics perspective, we call it the "pixel level", referring to the individual pixels of a camera image, the individual measurements in a laser range scan, the incremental motions caused by motor signals, the individual cells of an occupancy grid map, and so on. The learning task is to create useful higher-level representations for space, time, objects, actions, etc, to support effective planning and action in the world, bootstrapping up from experience at the pixel level.

In the remainder of this paper, after discussing the methodological framework for learning without prior domain-specific knowledge, we describe four recent research results that carry us significantly further toward autonomous learning of the foundational representations for commonsense knowledge.

Section 3 describes a method for starting with a completely uninterpreted sensorimotor system, applying a hierarchy of learning methods to define sensor modalities and their structures, primitive actions, sensory features and how they are affected by actions, control laws and distinctive states. Section 4 re-examines how sensory features and control laws are learned, providing an unsupervised method based on self-organizing maps. This method, SODA, uses self-organizing maps to learn the abstraction from "pixel-level" sensor inputs and motor outputs to perceptual features, distinctive states, and hill-climbing and trajectory-following control laws. Section 5 shows how highly reliable place recognition can be learned through a bootstrap learning process, combinging unsupervised learning, map-learning, and supervised learning. Section 6 describes how an ontology of *objects* can be learned from pixel-level experience. These are initial steps toward a foundational theory of how commonsense knowledge is possible.

2 A Methodological Framework

There are some serious questions about how one even begins to investigate the problem of learning foundational representations from uninterpreted sensors and effectors.

There are four different nested learning problems, where each defines the learning target for its predecessor.

[1] We use the term "robot" to refer to the physical system and its sensors and effectors. The "learning agent" is the computational process observing and learning to control the robot. Body and mind, if you wish.

1. As human researchers with limited resources of various kinds, we need to develop a suitable research strategy and research methodology to allow this overall research enterprise to be broken down into projects of manageable size. This research enterprise is a collective and extended search for ...

2. ...a learning algorithm that takes as input a set of domain-independent statistical learning methods and an uninterpreted set of sensors and effectors allowing exploratory behavior in the environment, and eventually learns ...

3. ...a higher-level ontology for commonsense knowledge, including such foundational concepts as space, time, motion, objects, actions, causality, etc, along with inference methods for abduction and planning, in order to ...

4. ...build higher-level models ("maps") that describe the environment in terms of the concepts in the learned high-level ontologies, to explain the observations the agent obtains during exploratory behavior. The quality of such a "map" is determined by its utility for generating effective predictions and plans.

Our own extensive work on learning cognitive maps of unknown environments through exploration is a contribution to problem 3, devising an appropriate ontology for knowledge of large-scale space, and problem 4, the task of building the maps given that ontology. The result of this work is the Spatial Semantic Hierarchy (SSH), which we use to define the target ontology for bootstrap learning of spatial knowledge as part of problem 2.

Our own work on bootstrap learning (summarized in this paper) contributes in various ways to problem 2. Continuing progress in our map-learning research makes the SSH a moving target. Nonetheless, we believe that the progress we have made so far on bootstrap learning will have sufficient generality and robustness to survive changes in the target ontologies it is intended to reach.

Problem 2 is itself an enormous problem, with many diverse aspects, so it must be solved by the research community over an extended period of time, which raises the strategic questions of problem 1.

Problem 1 is how to break the overall research enterprise of problem 2 into manageable pieces. Each piece necessarily makes assumptions about what other research results, some to be created in the future, will provide. Ideally, as the individual pieces are created, they will fit together into a larger intellectual structure, the assumptions of each piece being satisfied by the results from some pieces it rests on. In reality, we don't always guess right in making those assumptions or deciding how to break the large problem into manageable pieces, so some work will inevitably need to be modified or redone.

One research strategy that we have adopted is to avoid placing a prior constraint on the set of domain-independent statistical learning methods to be used in solving problem 2. We are pragmatic in our choice of methods, driven by the needs of the research problem at hand, while attempting to ensure that the statistical learning methods chosen are as general as possible, without domain-specific assumptions. After we have discovered sets of learning methods sufficient to provide solutions for problem 2, then we can begin to identify minimal subsets of those methods compatible with particular implementation technologies, including biological ones.

Another research strategy we have adopted is to place secondary importance on the question of whether a particular learning problem is solved by the species (over evolutionary time) or by the individual (over developmental time). In our work, we frame the learning problem as a problem for the individual, but this is a *gedankenexperiment* or "intuition pump" to help us develop useful insights at this early stage of our research enterprise. Eventually, we will need to consider whether particular kinds of learning take place during evolution, embryogeny, development, or mature behavior. In the biological world, the answers clearly vary from species to species.

The underlying hypothesis behind problem 2 is that the sophisticated higher-level ontology of human commonsense knowledge can arise from an undirected bootstrap-learning search through a space of representations. In the end, a solution to problem 2 in its entirety must be evaluated according to the ability of the learning process to construct a useful higher-level ontology for a complex world, without external direction. Our hypothesis is that the learning process is directed by the structure of the knowledge that is being learned, not by external supervision based on explicit goals.

However, Problem 1 recognizes that human research efforts cannot be undirected. A particular research project, intended to be accomplished by a few researchers within a year or two or three, must have an explicit target for the learning process to be developed. Is this inconsistent with the overall goal of undirected learning? No.

We attempt to define manageable projects directed as particular aspects of successful human commonsense knowledge. If we are successful in defining a project with an initial set of assumptions and a learning target along the path that human learning followed, then the target should be reachable by some undirected learning method. If the project succeeds, our confidence in the selection of the initial assumptions and the learning target, not to mention the learning method that was discovered, are all increased. If the project fails, it could be due to insufficient cleverness in finding a learning method, but it also could be due to poor choice of initial assumptions or learning target.

Considered within the context of a larger research enterprise, an individual research project can succeed, not only by solving a stated problem, but by modifying the given problem into one that has a good solution. Such a problem-solution pair must interact well with its neighbors, in the sense that the initial assumptions are satisfied by some prior process, and the target of this learning process is useful to other processes.

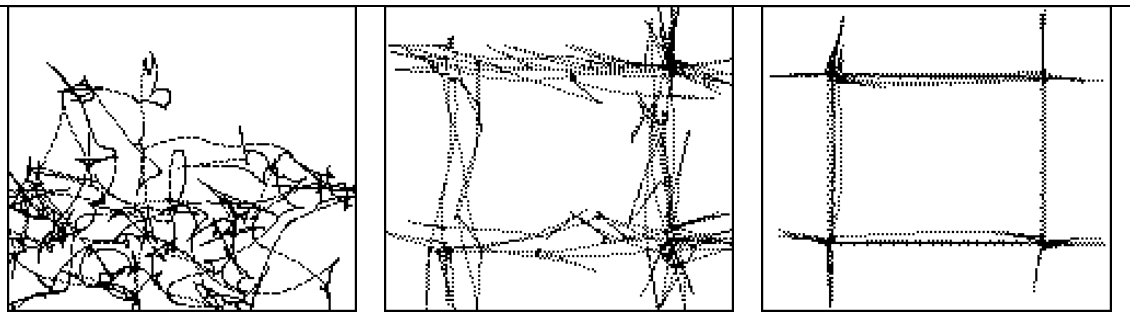
Thus, we can resolve the apparent paradox of a directed research project to develop an undirected learning method. Each individual research project is a part of the larger research enterprise. Each explicit learning target is a working hypothesis about how the solution to the overall research enterprise will look. The methods used in each project must be undirected, encoding no domain-specific knowledge about the learning target. The solution to the overall enterprise is obtained if the choice of intermediate learning targets leads to a set of compatible undirected learning methods that, together, reach the overall goal. (Figure 2 is a preliminary example of this.)

3 Learning from Uninterpreted Sensors and Effectors

The lowest level problem is faced by a learning agent in an unknown environment with unknown sensors and effectors. Our goal is to learn the foundation for the Spatial Semantic Hierarchy [Kuipers, 2000]. The SSH rests on a set of hill-climbing and trajectory-following control laws and the knowledge of the sensorimotor interface to support them. How can this knowledge be learned from unsupervised experience?

Pierce and Kuipers [Pierce and Kuipers, 1997] answered this question in the context of a simulated mobile robot with unknown sensors and effectors. The learning agent conducted a variety of experiments and analyzed the data, building a hierarchy of representations of both the sensory and motor systems, and eventually creating control laws that could define distinctive states (Figure 1). The experiment had the following steps.

1. Gather observations during random sequences of actions. First, coarsely cluster the sensors according to the qualitative properties of a histogram of values returned by each sensor. Then, within appropriate clusters, compute pairwise correlations among sensor values and interpret them as similarity measures.



(a) (b) (c)

Figure 1: Exploring a simple world at three levels of competence. (a) The robot wanders randomly while learning a model of its sensorimotor apparatus. (b) The robot explores by randomly choosing applicable homing and open-loop path-following behaviors based on the static action model while learning the dynamic action model (see text). (c) The robot explores by randomly choosing applicable homing and closed-loop path-following behaviors based on the dynamic action model.

2. Assign the sensors in a cluster to positions in a high-dimensional space reflecting their pairwise similarities. Project to a low-dimensional subspace (2D in our examples) that best accounts for most of the variance in the cluster. Once sensor values have a spatial as well as temporal dependence, we can calculate spatial as well as temporal derivatives, and thus define motion fields.

3. The motion fields corresponding to different motor signals are analyzed using principal component analysis to determine the most significant motion effects and the motor signals that correspond to them. These signals are used as the natural primitives for the motor space.

4. Higher-level sensory features are proposed, based on the spatial and temporal attributes of the field of primitive sensory values. These include features such as discontinuities, local minimum and local maximum, with magnitude, position, and scope. Proposed features are evaluated according to stability, predictive power, and extensibility.

5. Evidence is collected of the effects of primitive motor commands on higher-level features, searching for motor commands that change features in predictable ways. “Local state variables” are defined for particular neighborhoods in the environment. Trajectory-following and hill-climbing control laws are defined according to which local state variables correspond to features that are both observable and controllable.

6. Open-loop control laws are defined by identifying commands that reliably change one feature while keeping another one relatively constant. Closed-loop control laws are defined by searching for and identifying commands that can reduce deviations in the relatively constant feature, actively keeping it close to a desired setpoint. (Think of moving along a wall, turning slightly to maintain a desired distance from it. Compare figures 1(b,c).)

Figure 1 shows exploration traces at three stages of the learning process. The analysis uses a variety of mathematical methods, but only ones that can be applied to weakly interpreted data, using local computations such as neural networks. The sequencing of the learning steps arises because later learning methods depend

on prerequisites learned by earlier ones. Figure 2 shows the lattice of learning methods that supported these conclusions.

One lesson from this work is that learning even an apparently simple sensorimotor skill such as wall-following, starting from a pixel-level ontology, requires a large number of distinct learning algorithms, constructing a lattice of different representations of the sensory and motor capabilities of the robot.

4 Learning Distinctive States

The learning of high-level sensory features and hill-climbing and trajectory-following control laws in Pierce and Kuipers [Pierce and Kuipers, 1997] made use of certain background knowledge about sensors and control, albeit of an abstract and domain-independent kind. In order to eliminate this use of background knowledge, Provost, et al [Provost et al., 2006] use more generic learning methods such as self-organizing maps and reinforcement learning to achieve the same goals.

Modern robots are endowed with rich sensory systems, in which a high-dimensional sense vector provides a high-bandwidth stream of information about a continuous environment. In addition, many important real-world robotic tasks have *high diameter*, that is, their solutions require a large number of primitive actions by the robot, for example, navigating to distant locations using primitive motor control commands. Reinforcement learning (RL) methods show promise for automatic learning of robot behavior, but extending these methods to high-dimensional, continuous, high-diameter problems remains a major challenge. Thus, the success of RL on real-world tasks still depends on human analysis of the robot, environment, and task to provide a useful set of perceptual features and an appropriate decomposition of the task into subtasks. Our goal is to create autonomous learning agents, relying on few assumptions about the nature of the robot and its world.

Self-Organizing Distinctive-state Abstraction (SODA) is a new method for automatic discovery of high-level perceptual features and large-scale actions for reinforcement learning in continuous environments [Provost et al., 2006]. A *distinctive state* is the isolated local maximum of a selected measure defined over the local neighborhood, so that a hill-climbing control law can bring the robot to the distinctive state from any point in its neighborhood.

In SODA, we use a version of self-organizing maps (SOMs) [Kohonen, 1995] called the Growing Neural Gas (GNG) [Fritzke, 1995] to learn a small and general set of prototype units to represent the sensory experience available in the domain. Unlike the original SOM, the GNG allows the number of units and the topology of the mesh to adapt to the properties of the domain. To define distinctive states, we define the activation level of the leading GNG unit to be the target value for hill-climbing. The activation levels of the GNG units therefore serve as the perceptual features for this agent.

As the agent moves around the environment, different GNG units will have the leading activation level, and will thus define distinctive states based on different hill-climbers. Motion from one distinctive state neighborhood to another is done using trajectory-following control laws. In this preliminary version of SODA, these are simply repetitions of particular primitive actions until the dominant GNG unit changes, so they correspond to open-loop path-following control loops in [Pierce and Kuipers, 1997] (Figure 1(b)).

Thus, without prior knowledge of the robot's sensorimotor system or its environment, SODA does several distinct types of abstraction. It does *perceptual abstraction* by abstracting a high-bandwidth sense vector to a small set of GNG units, which serve as perceptual features. It does *state abstraction* by defining locally distinctive states to represent large portions of the continuous state space. And it does *temporal abstraction* by defining higher-level actions that take the agent from one distinctive state to the next, combining the effect of a trajectory-following control law to take the agent to a new neighborhood, and a hill-climbing control law to

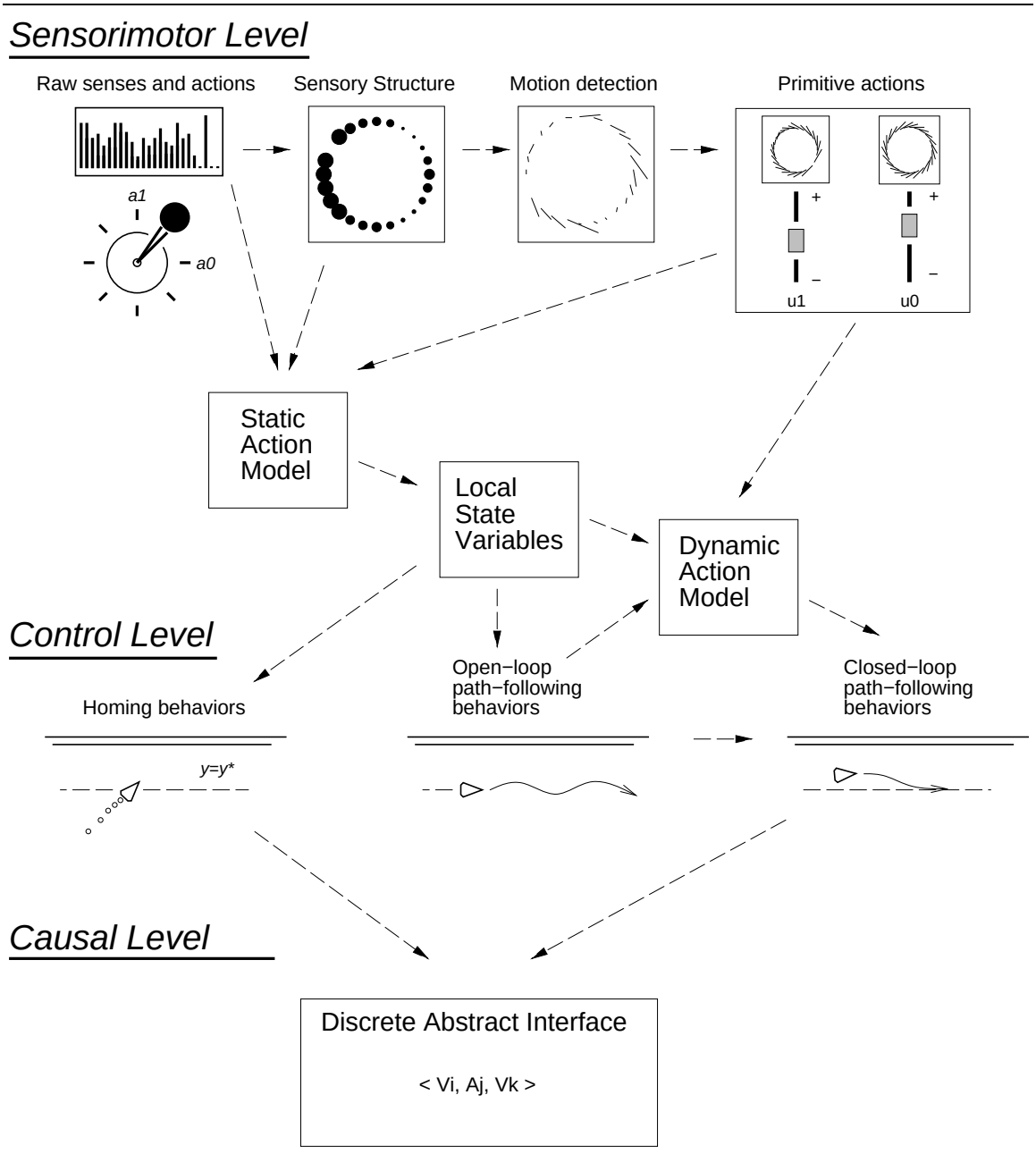


Figure 2: The lattice of learning methods and their results, from Pierce & Kuipers [1997].

reach the distinctive state itself within that neighborhood. These abstractions reduce both the dimensionality and the diameter of the robot's tasks.

Given high-dimensional, continuous-valued sensory input and continuous motor output, SODA works as follows.

1. Explore the environment with primitive ($\mathcal{A}^0$) actions, using a Growing Neural Gas [Fritzke, 1995] self-organizing feature map to learn a set of high-level perceptual features that define distinctive states in the environment. Figure 3(bottom) shows examples of the learned perceptual features.

2. Learn a set of high-level ($\mathcal{A}^1$) actions in the form of control laws that carry the robot from one distinctive state to another. Each action consists of a trajectory-following control law that repeats a primitive action until a new perceptual feature becomes dominant, followed by a hill-climbing control law that maximizes the new dominant feature.

3. Use reinforcement learning in the abstracted space of high-level distinctive states and actions to learn a policy for the same high-diameter task, which now has much lower diameter with respect to the $\mathcal{A}^1$ actions.

Each distinctive state created by SODA is characterized by the GNG unit whose activation is maximized at that state. Different underlying states may be aliased according to this criterion, leading to an action model that is non-deterministic due to state aliasing. In future work, we plan to draw on our methods for learning causal and topological maps [Remolina and Kuipers, 2004], to learn when distinctive states are and are not aliased, and to create a deterministic map. Nonetheless, even with a non-deterministic state-action map, reinforcement learning on the abstracted states and actions is very successful.

An experiment on a simulated robot navigation task (Figure 3) shows that the agent using SODA can learn to perform a task requiring 300 small-scale, local actions using as few as 9 autonomously-learned, temporally-extended, abstract actions. The learning time is substantially improved (Figure 4).

The methods discussed so far can learn the properties of the pixel-level sensorimotor system well enough to support autonomous learning of control laws and distinctive states suited to the environment the robot is embedded in. These distinctive states and the actions connecting them are the first steps toward a higher-level ontology for describing the robot's world. We now turn to two learning scenarios that build further on this higher-level ontology. First we look at the problem of *place recognition*: overcoming the variability of the pixel-level sensory image to recognize the current distinctive state directly and correctly from sensory input. And second, we take an important step toward learning the concept of *object*, a higher-level explanatory concept that makes it possible to learn useful causal regularities about the world.

5 Bootstrap Learning for Place Recognition

It is valuable for a robot to know its position and orientation with respect to its map of the environment. This allows it to plan actions and predict their results, using its map. Kuipers and Beeson [Kuipers and Beeson, 2002] applied the bootstrap learning approach to the problem of learning to recognize places that may have originally been perceptually aliased.

We define *place recognition* as identifying the current position and orientation in a large-scale space, a task sometimes called "global localization" [Thrun et al., 2001]. However, not every location in the environment is a "place", deserving of independent recognition. Humans tend to remember places which are distinctive, for example by serving as decision points, better than intermediate points during travel [Lynch, 1960].

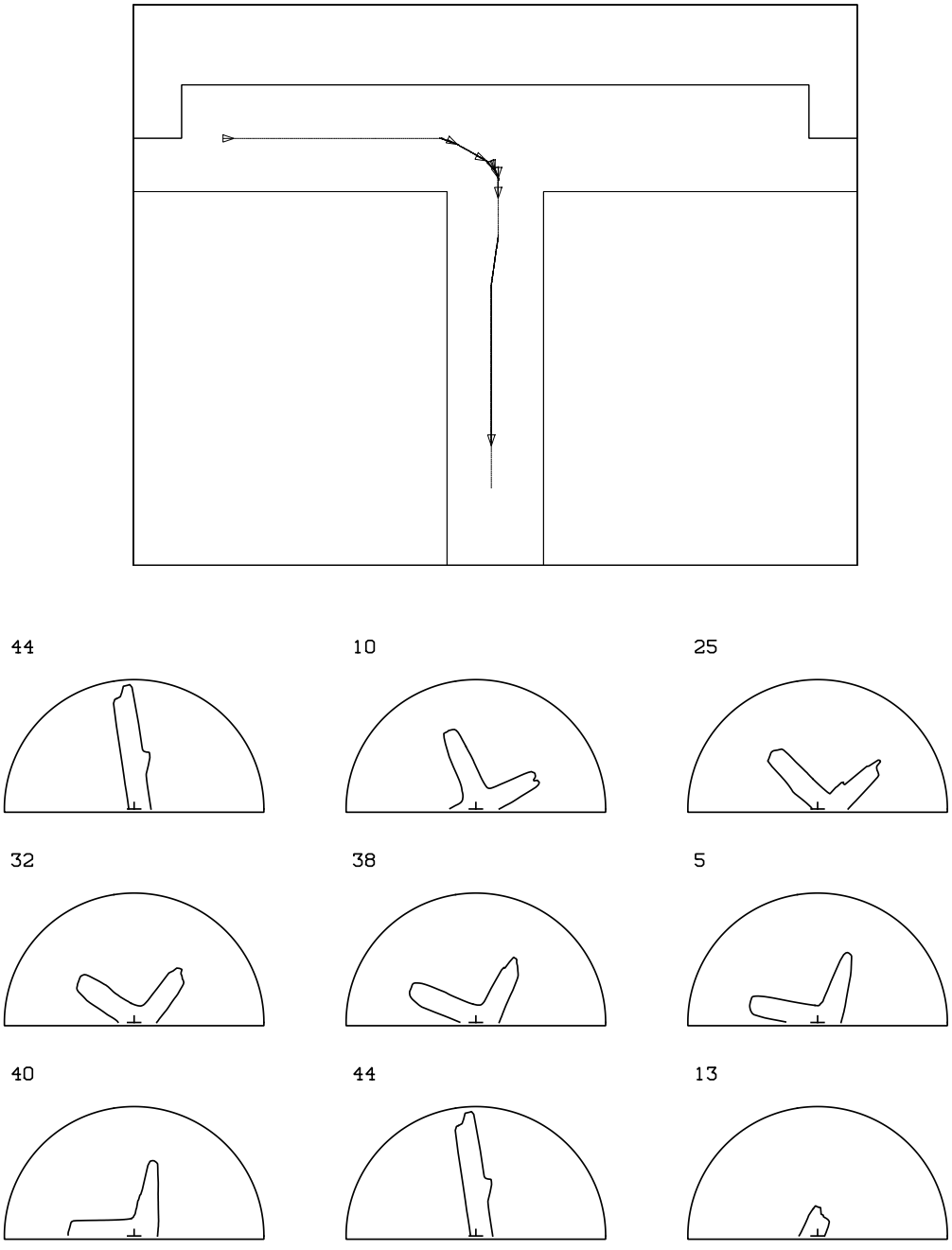


Figure 3: **Navigation using Learned Abstraction**. The upper diagram shows the robot's environment and an example episode after the agent has learned the task using the $\mathcal{A}^1$ actions. The triangles indicate the distinctive states the robot is in at the start of each $\mathcal{A}^1$ action. The bottom part of the figure shows the sequence of perceptual features corresponding to these distinctive states. The narrow line indicates the sequence of $\mathcal{A}^0$ actions used by the $\mathcal{A}^1$ actions. Navigating to the goal requires only 9 $\mathcal{A}^1$ actions, instead of hundreds of $\mathcal{A}^0$ actions – task diameter is vastly reduced. (Figure from [Provost et al., 2006].)

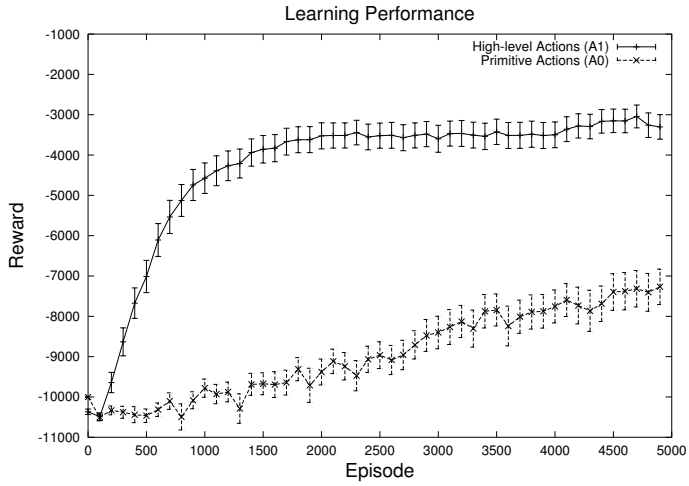


Figure 4: **Learning Performance** Comparison of the reward earned per episode using $\mathcal{A}^0$ actions vs. using $\mathcal{A}^1$ actions. Each curve is an average of 12 runs using each of 10 different learned feature sets. Error bars indicate +/- one standard error.

We assume that the world and the agent's sensors are both very rich, so distinguishing information exists, but is hard to find. Real sensors are imperfect, so important but subtle image features may be buried in sensor noise. Two complementary problems stand in the way of reliable place recognition.

- *Perceptual aliasing*: different places may have similar or identical sensory images.
- *Perceptual variability*: the same position and orientation may have different sensory images on different occasions, for example at different times of day.

These two problems trade off against each other. With relatively impoverished sensors (e.g., a sonar ring) many places have similar images, so the dominant problem is perceptual aliasing. With richer sensors such as vision or laser range-finders, discriminating features are more likely to be present in the image, but so are noise and dynamic changes, so the dominant problem for recognition becomes image variability. For this research, we use only real sensors in real environments, in order to avoid assumptions that restrict us to certain types of sensors or make it difficult to scale up to large environments.

When unique place recognition cannot be done using the current sensory image alone, active exploration will provide history information that can localize the robot and determine the correct place. However, when subtle features, adequate for discriminating between different places, are buried in the noise due to image variability, we want to recover those features.

We build on the abstraction of the continuous environment to a discrete set of *distinctive states* (dstates), provided by the *Spatial Semantic Hierarchy* (SSH) [Kuipers, 2000]. We assume that the agent has previously learned a set of features and control laws adequate to provide reliable transitions among a set of distinctive states in the environment [Pierce and Kuipers, 1997, Provost et al., 2006]. The steps in our solution to the place recognition problem apply several different learning methods (Figure 5).

1. Restrict attention to recognizing *distinctive states* (dstates). Distinctive states are well-separated in the robot's state space.

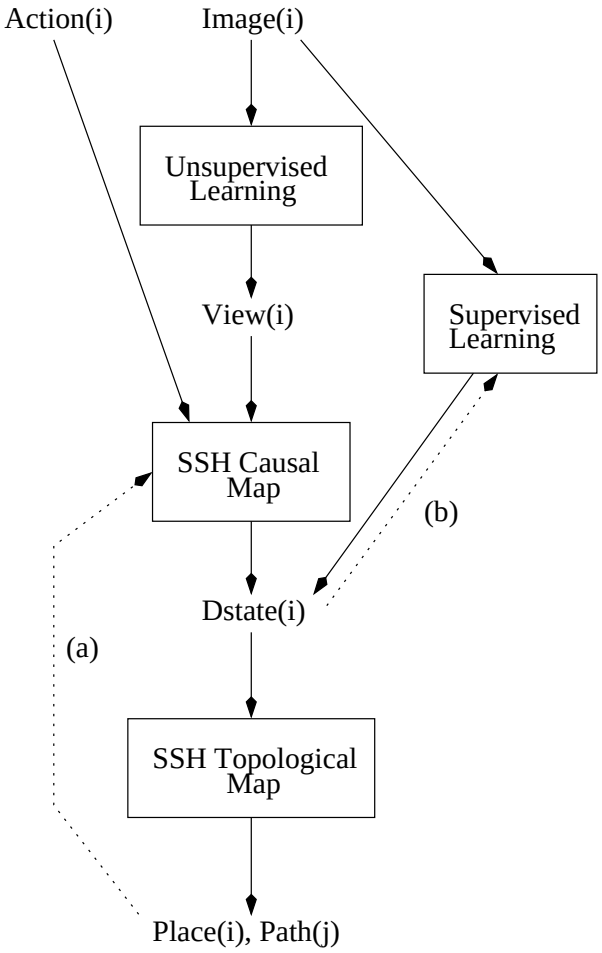


Figure 5: Bootstrap learning of place recognition. Solid arrows represent the major inference paths, while dotted arrows represent feedback.

2. Apply an unsupervised clustering algorithm to the sensory images obtained at the dstates in the environment. This reduces perceptual variability by mapping different images of the same dstate into the same cluster, even at the cost of increasing perceptual aliasing by mapping images of different states into the same cluster. We define each cluster to be a *view*, in the sense of the SSH [Kuipers, 2000].

3. Build the SSH causal and topological maps — symbolic descriptions made up of dstates, views, places, and paths — by exploration and abduction from the observed sequence of views and actions [Kuipers, 2000, Remolina and Kuipers, 2004]. This provides an unambiguous assignment of the correct dstate in the map to each experienced image, which is feedback path (a) in Figure 5.

4. The correct causal/topological map labels each image with the correct dstate. Apply a supervised learning algorithm to learn a direct association from sensory image to dstate. The added information in supervised learning makes it possible to identify subtle discriminating features that were not distinguishable from noise by the unsupervised clustering algorithm. This is feedback path (b) in Figure 5.

We evaluated this method in experiments in two different real-world environments, one constructed to have a subtle distinguishing feature in an otherwise simple and symmetrical environment, and the other the main corridor in an office building. In both cases, unsupervised clustering produced significant amounts of perceptual aliasing, but with the help of the learned topological map, supervised learning was able to converge rapidly to 100% accurate place recognition.

This is a paradigm example of *bootstrap learning*. A weak learning method (k-means clustering) provides the prerequisites for an abductive method (topological map-building), which in turn provides the labels required by a stronger supervised learning method (nearest neighbor), which finally achieves high performance.

6 Bootstrap Learning of Object Representations

The blooming buzzing confusion of the pixel-level world is too variable to contain meaningful causal regularities useful for prediction and planning. Among the many important achievements in early childhood development is learning the higher-level concept of *object*, which along with the higher-level concept of *action* is capable of supporting learning of causal regularities useful for understanding and manipulating the world [Spelke, 1990].

In recent work toward this goal [Modayil and Kuipers, 2004], we have shown how an agent can autonomously learn an ontology of *objects* to explain many aspects of its sensor input from an unknown dynamic world. For an agent 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 time-varying sensor inputs and motor outputs. From this, we assume that it can construct a description of the local environment in the "pixel-level" ontology of occupancy grid models.[2] The learning scenario described here takes place in "small-scale space", the space within the immediate sensory surround of the agent where it can reliably localize itself [Kuipers et al., 2004].

The occupancy grid representation for local space does not include the concept of *object*. The occupancy grid representation assumes that the robot's environment can be divided into cells that are empty and those that are occupied. Evidence provided by range sensors is used to update the probability of occupancy of each

[2]The learning methods in Pierce and Kuipers [Pierce and Kuipers, 1997] can learn the properties of sensors and effectors from experience. We assume that the occupancy grid representation and inference method can be learned in a similar way. We have a sketch of such a learning scenario, but it is outside the scope of this research on objects.

cell. 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 et al., 2000].

In this bootstrap learning scenario, the learning agent acquires a working knowledge of *objects* from unsupervised sensorimotor experience. We begin by using the properties of occupancy grids to classify individual sensor readings as static or dynamic. A cell in the occupancy grid is considered *static* if it is labeled occupied with high confidence, and has never been labeled free with high confidence. A cell is considered *dynamic* if it has ever been labeled free with high-confidence, even if it later becomes occupied. An individual sensor reading is labeled static or dynamic according to the label of the cell it falls in. Static readings are considered to be explained by the structure of the fixed environment, and are not considered parts of objects.

The representation of objects is constructed from dynamic sensor readings in four steps: Individuation, Tracking, Image Description, and Categorization. Dynamic readings are clustered and the clusters are tracked over time to identify objects, separating them both from the background of the static 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 when shape is a stable and consistent property of these objects. However, the representation can tolerate, represent, and track amorphous objects as well as those that have well-defined shape. The shape models are classified, so that instances of the same type of object can be categorized together.

In Modayil and Kuipers [Modayil and Kuipers, 2004], 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. The agent learned to individuate and track dynamic objects in the scene, acquired shape models where the shape was stable, and created a categorization of shape models. The scene could then be described in terms of the static environment (grounded to the static portions of the occupancy grid), and the dynamic objects (whose identities and trajectories could be described symbolically, grounded to the tracked objects in the scene). Figure 6 shows selected steps leading to this result.

By this process, the agent has learned substantial portions of the concept of *object*. It has learned to separate objects from the background environment, describing it as an individual that has a spatial extent and persists over time. Some individual objects have a consistent shape, which can be used to categorize individual objects into object types. A straight-forward consequence of this, but one not explored yet in this paper, is the ability to identify new individual elements of the object type, which might not have been identified using previous methods, perhaps because it has never moved, or perhaps because its image has always been entangled with other objects or the environment. We are also investigating the learning of appropriate actions for interacting with object individuals and types.

7 Conclusions

To be autonomous, a robot must be able to learn its own ontology of higher-level concepts from its own pixel-level experience with the world, rather than obtaining it from an external programmer. We have described recent research that shows how the structure of unknown sensors and effectors can be learned [Pierce and Kuipers, 1997]; how high-level perceptual features and actions can be learned and used to define distinctive states [Provost et al., 2006]; how high performance place recognition can be learned by bootstrapping unsupervised learning, map-building, and supervised learning [Kuipers and Beeson, 2002]; and how an ontology of objects can be learned from low-level experience with a dynamic world [Modayil and Kuipers, 2004].

There are many other aspects of commonsense knowledge of the physical world still to be learned. We

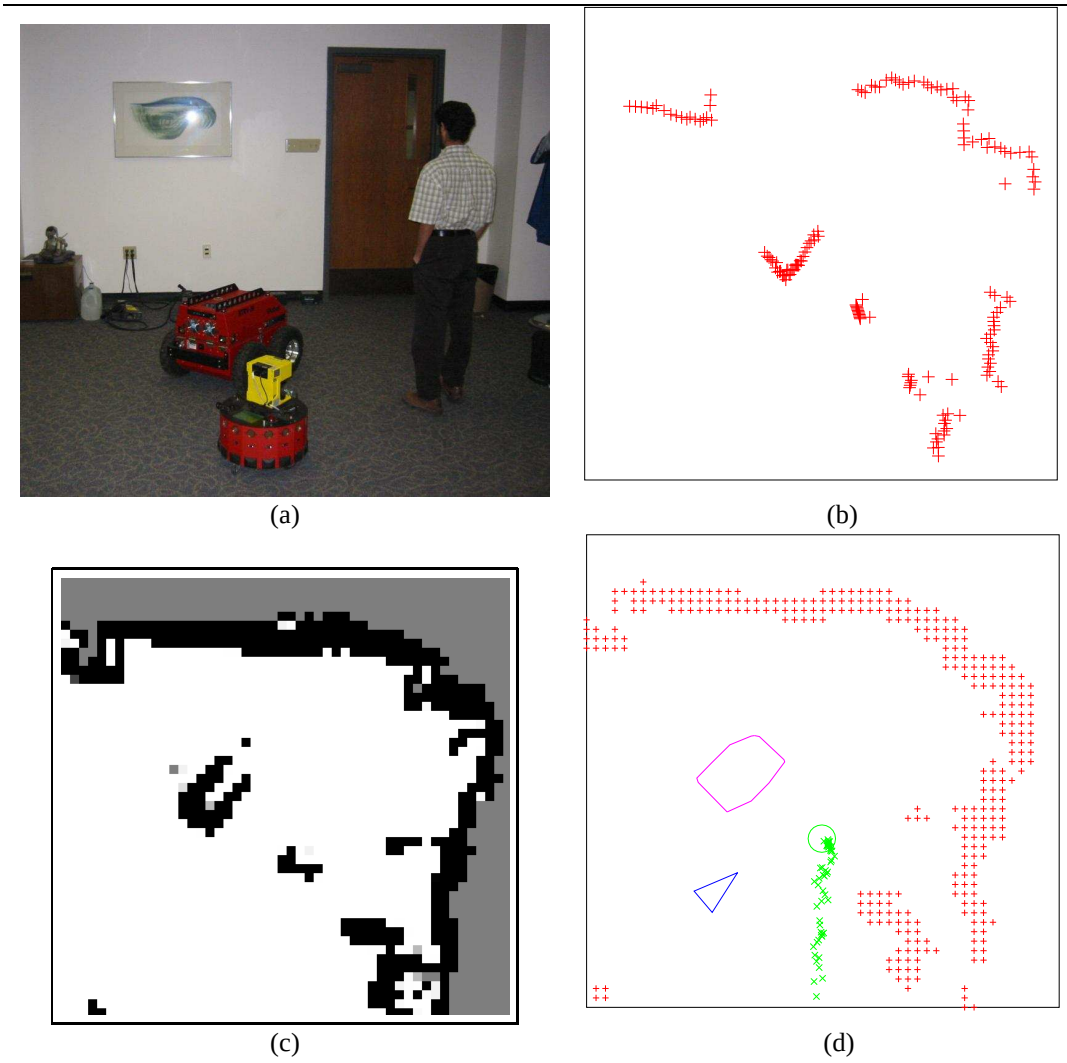


Figure 6: Multiple representations of a scene. The robot observer is the small round robot in the foreground. The larger ATRV-Jr is used as a non-moving object. **(a)**: A photograph of the scene. **(b)**: A range image of the scene at approximately the same time. **(c)**: An occupancy grid representation 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, and is shown by the convex hull of its shape model. The permanently occupied cells in the occupancy grid represent the static environment.

have already mentioned the need to learn the occupancy grid representation, or more generally, a local perceptual map representation of the immediate sensory surround [Kuipers et al., 2004]. We are also extending the learned theory of objects with the actions that affect those objects, along with their preconditions and postconditions [Modayil and Kuipers, 2004]. Another important research direction will be learning to use vision as a sensory modality. Naturally, this kind of learning will straddle the evolutionary/developmental boundary.

Bootstrap learning of foundational representations may also be an important part of developing a scientific theory of consciousness [Kuipers, 2005]. One of several aspects of consciousness is a property that philosophers call *intentionality*: the referential connection from concepts in the mind to objects in the external environment [Searle, 2004]. Critics of "strong AI" claim that robots can never have "original" intentionality (instrinsic to itself), but can only have "derived" intentionality (from the mind of a human author or programmer). However we have seen, albeit in very simple forms, bootstrap learning of concepts of *place* and *object*, complete with referential connections to individual places and objects in the world through the causal properties of the sensorimotor system. The ability of a robot to learn *its own* higher-level concepts from *its own* low-level experience is the foundation for having *its own* original intentionality.

Acknowledgements

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 grants from the National Science Foundation (IIS-0413257 and IIS-0538927), from the National Institutes of Health (EY016089), and by an IBM Faculty Research Award. A preliminary version of this paper appeared in [Kuipers et al., 2005].

References

[Fritzke, 1995] Fritzke, B. (1995). A growing neural gas network learns topologies. In Tesauro, G., Touretzky, D. S., and Leen, T. K., editors, *Advances in Neural Information Processing Systems 7*, pages 625–632. MIT Press.

[Kohonen, 1995] Kohonen, T. (1995). *Self-Organizing Maps*. Springer, Berlin; New York.

[Kuipers, 2000] Kuipers, B. (2000). The Spatial Semantic Hierarchy. *Artificial Intelligence*, 119:191–233.

[Kuipers, 2005] Kuipers, B. (2005). Consciousness: drinking from the firehose of experience. In *Proc. 20th National Conf. on Artificial Intelligence (AAAI-05)*, pages 1298–1305. AAAI.

[Kuipers and Beeson, 2002] Kuipers, B. and Beeson, P. (2002). Bootstrap learning for place recognition. In *Proc. 18th National Conf. on Artificial Intelligence (AAAI-02)*, pages 174–180. AAAI/MIT Press.

[Kuipers et al., 2005] Kuipers, B., Beeson, P., Modayil, J., and Provost, J. (2005). Bootstrap learning of foundational representations. In *Developmental Robotics*, AAAI Spring Symposium Series, Stanford, CA.

[Kuipers et al., 2004] Kuipers, B., Modayil, J., Beeson, P., MacMahon, M., and Savelli, F. (2004). Local metrical and global topological maps in the hybrid spatial semantic hierarchy. In *IEEE Int. Conf. on Robotics & Automation (ICRA-04)*.

[Lakoff and Johnson, 1980] Lakoff, G. and Johnson, M. (1980). *Metaphors We Live By*. The University of Chicago Press, Chicago.

[Lynch, 1960] Lynch, K. (1960). *The Image of the City*. MIT Press, Cambridge, MA.

[McCarthy, 1968] McCarthy, J. (1968). Programs with common sense. In Minsky, M. L., editor, *Semantic Information Processing*, pages 403–418. MIT Press, Cambridge, MA.

[Minsky, 1975] Minsky, M. (1975). A framework for representing knowledge. In Winston, P. H., editor, *The Psychology of Computer Vision*. McGraw-Hill, NY.

[Modayil and Kuipers, 2004] Modayil, J. and Kuipers, B. (2004). Bootstrap learning for object discovery. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*.

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

[Piaget and Inhelder, 1967] Piaget, J. and Inhelder, B. (1967). *The Child's Conception of Space*. Norton, New York. First published in French, 1948.

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

[Provost et al., 2006] Provost, J., Kuipers, B. J., and Miikkulainen, R. (2006). Developing navigation behavior through self-organizing distinctive-state abstraction. *Connection Science*, 18.2. In press.

[Remolina and Kuipers, 2004] Remolina, E. and Kuipers, B. (2004). Towards a general theory of topological maps. *Artificial Intelligence*, 152:47–104.

[Searle, 2004] Searle, J. R. (2004). *Mind: A Brief Introduction*. Oxford University Press.

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

[Thrun et al., 2000] 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)*, pages 859–865. AAAI Press/The MIT Press.

[Thrun et al., 2001] Thrun, S., Fox, D., Burgard, W., and Dellaert, F. (2001). Robust Monte Carlo localization for mobile robots. *Artificial Intelligence*, 128:99–141.