An Intellectual History of the Spatial Semantic Hierarchy

Benjamin Kuipers[1]

Computer Sciences Department, University of Texas at Austin, Austin, Texas 78712 USA
kuipers@cs.utexas.edu

Summary

The Spatial Semantic Hierarchy and its predecessor the TOUR model are theories of robot and human commonsense knowledge of large-scale space: the *cognitive map*. The focus of these theories is on how spatial knowledge is acquired from experience in the environment, and how it can be used effectively in spite of being incomplete and sometimes incorrect.

This essay is a personal reflection on the evolution of these ideas since their beginning early in 1973 while I was a graduate student at the MIT AI Lab. I attempt to describe how, and due to what influences, my understanding of commonsense knowledge of space has changed over the years since then.

1 Prehistory

I entered MIT intending to study pure mathematics. I was generally steeped in the ideology of pure mathematics, and I had every intention of staying completely away from practical applications in favor of abstract beauty and elegance. However, on a whim, in Spring of 1973 I took Minsky and Papert's graduate introduction to Artificial Intelligence. I was immediately hooked. I had always been fascinated by the idea of a science of the mind. But then in college I took a course in psychology, which was a crashing disappointment. The interesting parts weren't scientific, and the scientific parts weren't interesting. Now, in artificial intelligence, symbolic computation promised mathematical methods capable of rigorously modeling interesting aspects of the mind.

I spent that summer at the MIT AI Lab, reading papers and getting more and more excited. Marvin Minsky was circulating drafts of his "frames paper" [39], which advocated that research focus on representation and inference about complex symbolic descriptions of meaningful objects and situations, rather than on individual propositions and logical inference. Such a description was called a *frame*. It had a number of *slots*, which could contain *values*, and could be associated with symbol manipulation procedures for doing inference, including providing *default values* for empty slots. I recall telling Pat Winston once that I found the frames concept to be very compelling, but I wondered where the slots come from.

Minsky's classes introduced me to Piaget's theories of the development of children's knowledge of foundational domains, including space, time, causality, and so on. He, along with John McCarthy's writings, also convinced me that the nature and representation of commonsense knowledge was a bottleneck issue for artificial intelligence. This was the problem I wanted to work on.

Following up on an idea of Minsky's for model-based object recognition, and using the edge-and-vertex representation from Blocks World vision, I wrote a paper showing how a vision system could discriminate among a small set of block models, tracing a hypothesis from vertex to vertex along edges, and using contradictory evidence to force a jump to an alternate hypothesis when necessary.[1] This paper earned me an invitation to spend Summer 1974 at Xerox PARC as a summer student working with Danny Bobrow and Terry Winograd. I implemented and demonstrated my recognition system in Smalltalk on the Alto, alternately marveling at the wonderful new technology and taking it totally for granted. The revised paper was named "A frame for frames" [22] in conscious homage to Fillmore's far more influential "The case for case" [11].

As the end of the summer approached, before returning to MIT, I met with Danny Bobrow to ask his advice on research topics. I explained that I had enjoyed working on model-based object recognition, but I really wanted to work on the problem of commonsense knowledge, and I didn't know where to begin. Danny suggested that I look at some work being done by Joe Becker and Bill Merriam at BBN on a simulated robot learning the structure of a simulated city [3, 4].

I knew immediately that this was the right problem: *How can a robot learn a cognitive map from its own experience of the environment?* It focuses on spatial knowledge, which is not only important, but is arguably the foundation for most other kinds of commonsense knowledge [33]. It also looked like it would factor well, in the sense that I could define interesting sub-problems that were small enough to solve, but which could be assembled into solutions to larger problems as I made progress. It would make a great PhD thesis topic, and I went back to MIT happy.

2 Cognitive Background

Quite a bit was already known about how human knowledge of space is structured, and how people use spatial knowledge to solve problems. I immersed myself in that highly diverse literature, reading papers from cognitive and developmental psychology, urban planning, geography, linguistics, and the visual arts. Two books that par-

[1] Only with the benefit of much hindsight do I recognize the similarity with the process of building topological maps.

ticularly influenced me were *The Image of the City*[2] by Kevin Lynch [36] and *Image and Environment*, a new collection of papers edited by Downs and Stea [9]. Also, among the more cognitively oriented denizens of the MIT AI Lab, Piaget's "genetic epistemology" approach to developmental psychology (e.g., [43]) permeated the atmosphere.

What quickly emerged from all this reading was a view of spatial knowledge consisting of several quite different types of knowledge. Some was procedural, "how-to" knowledge about getting from one place to another. Some consisted of topological connections between places and travel paths. And some consisted of metrical layouts approximately analogous to the environment itself or to a printed map. But it was clear that accurate metrical layout descriptions came last, if at all, and depended on the earlier types of knowledge. Furthermore, spatial reasoning methods varied across individuals, with developmental stage, with experience in a particular environment, or simply with individual cognitive style. A year or so later, Siegel and White's masterful survey of the development of spatial knowledge [56] confirmed and deepened this view.

Since the differences between the representations for spatial knowledge are so central, I started collecting route directions and sketch maps from anyone available. These were informal probes, designed to elicit a wide range of behavior I could examine for qualitative features, not formal experiments designed to test or refute hypotheses. What I needed was to complement the literature review with an intimate sense of the phenomenon itself, as a basis for building a computational model.

One immediate conclusion was that there is a lot of individual variation in the amount, nature, and accuracy of spatial knowledge that different people have, and in how they express it. Another is that neither verbal directions nor sketch maps tend to be particularly accurate about absolute distances or directions. On the other hand, topological relations such as the order of places on a path, or the connections between paths at a place, tend to be represented accurately and when errors do creep in, they are usually detected.

A common style for drawing a map was to follow a mental route, drawing those places and paths needed for the route, and perhaps nearby structures. When the subject made an error in translating the route into the graphical map representation, the error was usually metrical, and could go unnoticed for quite some time as the map was elaborated in an incorrect direction. The error would be detected when it finally came time to close a loop, and two occurrences of the same place would be drawn far apart, sometimes separated by other structures. Detecting the problem was easy, but identifying the specific error or correcting it could be quite difficult.

Some subjects used a different style[3], sketching the overall structure of a region, such as the rectangular grid structure in Boston's Back Bay. Fortunately for my re-

[2] I later learned that both Lynch's *The Image of the City* [36] and Miller, Galanter, and Pribram's influential *Plans and the Structure of Behavior* [38] were inspired by Kenneth Boulding's seminal book, *The Image* [6].

[3] These two styles were also identified by Linde and Labov [35] in subjects' descriptions of their apartment layouts.

search, the geography of the Boston-Cambridge area abounds with interesting local structures that fail to generalize over larger regions, leading to easily detectable geographical fallacies and paradoxes in people's cognitive maps.

The overwhelming impression from both my own investigations and the published experimental studies is that human spatial knowledge consists of a number of distinct representations for different aspects of space. Some people have many of these cognitive modules, and they work together well, while others may have fewer of them, or they don't work together so well. As a working hypothesis, I took the position that there is a single "complete" structure for all of these modules, working well together, and that all the variants — with individual style, developmental stage, or amount of experience in a particular environment — are modified or restricted versions of the ideal. This is similar to both Piaget's "genetic epistemology" and to current notions of "ideal observer" models [12].

Since the target of my efforts was a structure of interacting modules, it was natural to do the research by identifying an interesting aspect of the phenomenon of the cognitive map, constructing and testing individual modules to explain that aspect, and then looking for further parts of the natural phenomenon not adequately explained by existing modules.

3 The TOUR Model

My doctoral thesis described the representation of knowledge of large-scale space — the *cognitive map* [23, 24]. Space is considered *large-scale* if its relevant structure is at a scale larger than the sensory horizon, so knowledge of the structure must be acquired from exploration within it. The focus on large-scale space allowed me to avoid the difficult problems of computer vision and scene understanding. I focused my attention on spatial representation and inference, and specifically, on the problem of how global spatial structure can be inferred from local sensory experience. The *TOUR model* is a computational model of this kind of knowledge, including in most cases how that knowledge is learned from experience.

The TOUR model describes an agent[4] that receives a sequence of experiences as it travels through the environment, and builds its own cognitive map of that environment. The cognitive map is a symbolic representation, consisting of a set of frames for describing different types of objects such as places, paths, and regions; each type with its own collection of attributes; each instance with values for some or all of those attributes.[5] A place includes an attribute for the set of paths it is on, and a path includes an attribute for the partially-ordered set of places on it. An agent on a path faces in one of two directions: up or down the place-ordering on that path.

[4] The TOUR model and the Spatial Semantic Hierarchy are intended to describe both human and robotic agents.

[5] The equivalence between frames and first-order predicate logic is now well understood [15]. Jimi Crawford and I later formalized the intuitions behind this version of frames as "Access-Limited Logic" and its implementation, Algernon [7, 8].

As the agent receives experiences, it draws only those conclusions that can be inferred efficiently with information available at the time. This kind of "opportunistic" inference puts a premium on representations capable of expressing incomplete knowledge, so the results of small inference steps can be represented and stored, rather than being lost if attention moves elsewhere. Because of this strategy, inference is very efficient, but several travels along a particular route may be necessary for the TOUR model to infer all of the conclusions that follow logically from the experience.

The TOUR model divides spatial representation into three levels: procedural, topological, and metrical.[6] At the procedural level, experience is modeled as a sequence of GO-TO and TURN actions, with associated distance or angular magnitudes, respectively. The action description can be augmented with descriptions of the states before and after the action, each modeled as place, path, and direction along the path. When not provided explicitly, these may be inferred from context.

The inferential heart of the TOUR model is the "TOUR machine", a finite-state, rule-driven automaton. It has a set of registers called the "You-Are-Here pointer" describing the current place, path, direction, etc. Instead of an infinite tape, its memory is a potentially infinite set of frames reachable through the attributes of existing frames. Knowledge of the current state fills in the initial-state description in the current action. If the current place or path description can predict the final-state of the current action, it does; if not, new descriptions are created. In either case, the results update the You-Are-Here pointer, and they are stored as part of the action, place, and path descriptions, extending or confirming what was previously stored. Since the world itself is assumed to have a single consistent structure, and since the representation is supposed to be sufficiently expressive of incomplete knowledge for the results of opportunistic inference, contradictions between stored and newly-inferred information should be rare. The problem of more extensive reorganization and correction of the map when such an error is detected was beyond the scope of this research.

The sequence of GO-TO and TURN actions representing the agent's experience is provided by a simple natural language interface. The interface is based on Vaughan Pratt's elegant LINGOL parser [49], which allows context-free grammar rules to be annotated with semantic interpretation routines. The grammar makes it easy to describe the agent's experiences in natural-sounding route instructions, such as:

> Start on Broadway, at the intersection of Broadway and Prospect Street, facing Kendall Square.
> Turn right onto Prospect Street.
> Take Prospect Street to Central Square.
> Turn right onto Mass Ave.
> Take Mass Ave to Putnam Circle.

The topological level of the TOUR model is based on the connectivity of places and paths, the circular order of directed paths at each place, and the partial ordering

[6] This division into levels is updated to reflect the later perspective of the Spatial Semantic Hierarchy [32, 18].

of places on each path. It also includes boundary relations, whereby places can be described as “to the right” or “to the left” of a path. Boundary relations can be used to define regions in terms of bounding paths. All of these are learned by the TOUR model through opportunistic inference from experience in the form of GO-TO and TURN actions. Another form of topological knowledge is a region hierarchy, which allows the environment to be described, and route-planning problems to be solved, at many different levels of abstraction. For the region hierarchy, the TOUR model describes the representation and use of the knowledge, but provides no learning theory.

The metrical level of the TOUR model consists of attributes and relations with continuous values, like distance and direction. Analog spatial representations such as 2D occupancy grids [42] were still far in the future. Every GO-TO action includes a description of the magnitude of travel from one place to another along a given path. This provides a constraint on the relative location of the two places in the 1D frame of reference of that path. Enough observations of distances between pairs of places on the same path determines the layout of places within the path. Similarly, observations of TURN magnitudes at a given place provides a radial layout of the directed paths at that place. These radial layouts can be interpreted as defining the heading of an agent at that place, path, and direction, but only in a frame of reference local to the place, so headings cannot be compared from place to place. However, if the GO-TO action magnitude is extended to include a “net angular displacement” attribute $\Delta\theta$, then a single frame of reference can propagate along GO-TO actions to include multiple places. For places within a single frame of reference, GO-TO and TURN actions provide relative distance and direction measurements, from which a 2D layout of places can be inferred.

The TOUR model [23, 24] was the first computational model of the cognitive map that explicitly addressed the multiple types of spatial knowledge that must be represented. It specifically focused on the topological representations whose importance was well-understood by researchers deeply familiar with human cognitive mapping, but which was widely overlooked by many others in psychology, geography, and robotics. The major limitations of the TOUR model were the oversimplified interface to the agent’s actual sensorimotor experience in the world, and the inadequate treatment of analog metrical representations.

4 Explicit Representation of Sensory Views

One problem with the original TOUR model is that the procedural level too thoroughly abstracts away the agent’s sensory input from the environment. The route-direction-like input representation was unable to express either gaps in the sequence of experience or perceptual aliasing (different places that look the same). Part of solving this was to provide an explicit representation for sensory experience [25]. A *view* is an abstracted description of the sensory image experienced by the agent at a particular state (i.e., place, path, and direction). The TOUR model avoids the problem of interpreting input from any particular sensor (e.g., vision, sonar, laser) by treating views as atomic symbols that can only be used as retrieval keys or matched for identity. The specific representation or implementation of views is outside the scope of the theory (until later; see Sect. 7).

Given the concept of view we can define a more natural interface, representing the agent's experience as an alternating sequence of views and actions:

$$v_0, a_0, v_1, a_1, v_2, \cdots v_{n-1}, a_{n-1}, v_n.$$

An action a_i can have type Turn or Travel, with an associated magnitude.

We can now replace the procedural description of travel experience with a collection of causal schemas $\langle v, a, v' \rangle$, where the view v describes the context when action a is initiated, and v' describes the result after a has completed [25]. A schema $\langle v, a, v' \rangle$ has the declarative interpretation that in context v, after performing action a, one can expect resulting view v', and the imperative interpretation that if the agent experiences the context view v, it should do action a.

Knowledge of an experienced route is represented as a collection of schemas, indexed by their context views. This representation can express several very plausible states of incomplete knowledge. A gap in the route, perhaps due to inattention during exploration, corresponds to omitted schemas in the route description. If all the schemas $\langle v, a, v' \rangle$ in a route description are complete, they form a linked list, as the result v' of each schema allows retrieval based on the context v of the next schema along the route. However, incomplete schemas $\langle v, a, \_ \rangle$ can be constructed if working memory is disrupted during the possibly-extended time while a is taking place, before the result v' becomes available. Incomplete schemas still have their imperative meanings, and can still be used to traverse the route physically in the environment, since the environment will provide the result of each action. What is lost is the ability to review the route in the absence of the environment.

In these ways and others, the schema representation is very expressive of states of incomplete knowledge of a route. Variations may depend on developmental stage, amount of experience with this route, amount of computational resources available, and frequency of disruptions. We extended this concept to describe one aspect of individual variation in cognitive style, corresponding to the set of rules available for constructing partial schemas [26]. [7]

As it happens, it took a while to recognize that a good formal structure for representing route experience is the familiar finite-state automaton, or more generally, the partially-observable Markov decision process (POMDP) [10, 1, 2]. We require a set of underlying states x, that are themselves unobservable, but which map to observable views v. The set of schemas $\langle x, a, x' \rangle$ represents the transition function for the automaton, and the relation $view(x, v)$ represents the mapping from unobservable state to observable view. In full generality, POMDP learning of automata with stochastic transition and observation functions is intractable. However, this direction of investigation takes us farther away from an understanding of human commonsense spatial knowledge.

[7] Starting around 1978-79, I decided to change research direction for a variety of reasons [19]. This led to a productive line of work on medical reasoning and qualitative simulation [31, 16, 17, 28]. Spatial knowledge became a secondary concern until the mid-1990s.

In the Spatial Semantic Hierarchy [18, 50], we assume that transitions $\langle x, a, x' \rangle$ among states are deterministic (reflecting the error-correcting capabilities of feedback control laws), and that the relation $view(x, v)$ is a function, though not necessarily one-to-one. With these assumptions, and when exploring physical space, learning a minimal underlying automaton from observational experience is generally feasible in practice.

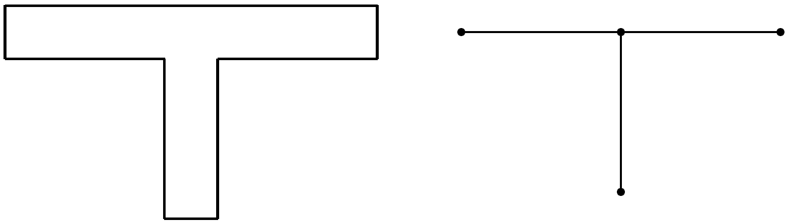

Fig. 1. A T-shaped space, and its topological model.

5 Abstracting Continuous Experience to Discrete States

A second problem with the original TOUR model is that it presupposes that the continuous experience of the agent has already been abstracted to a discrete sequence of states and transitions. This was justified by Kevin Lynch's observation that humans tend to represent knowledge about decision points, with much less about the spaces between them [36]. Nonetheless, this unexplained abstraction remained a gaping hole in the theory, and it was a barrier to robot implementation.

My cognitive mapping research had been on hiatus for several years, with QSIM receiving all of my attention, when a new grad student named Yung-Tai Byun approached me in 1986, wanting to do research on robot exploration and mapping. In the course of our discussions, we ran directly into the problem of relating the robot's continuous behavior to the kind of discrete topological map that the TOUR model creates. When we contemplated the simplest non-trivial environment I could think of — two corridors joined to form a T (Fig. 1) — the concept of *distinctive place* became clear. If we overlay the obvious T-shaped topological map onto the continuous polygonal environment, the natural locations for the four topological places are at the dead-ends and the intersection, at locations equidistant from the nearest obstacles. The segments connecting places are clearly corridor midlines. These loci corresponding to topological places and topological paths naturally suggest the attractors of hill-climbing and trajectory-following control laws, respectively. This basic idea, of letting the attractors of continuous control laws define the topological features of large-scale space, led to several influential papers, including [29, 30]. Fig. 2 demonstrates this approach to the exploration of a simulated environment.

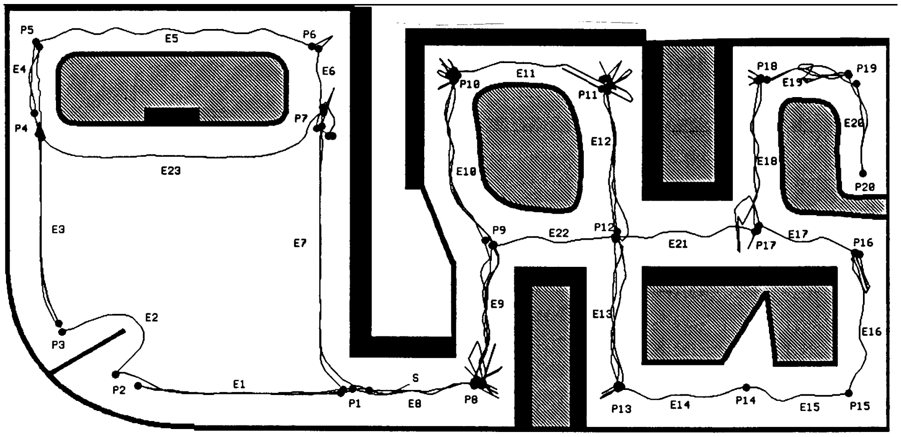

Fig. 2. A simulated robot applies the SSH exploration and mapping strategy. It identifies a topological graph of distinctive places and connecting path segments according to the behavior of control laws in the environment.

The selection of a control law couples the robot and its environment into a continuous dynamical system, which moves through its state space toward an attractor. The selection, execution, and termination of these control laws can be defined based entirely on sensory features available "from the inside" of the agent, without any appeal to the external semantics of the sensors or of the features. (It wasn't until later that we actually tried to *learn* the sensors, features, and control laws without appeal to external semantics [47]. See Sect. 8.) This method for defining symbolic entities referring to topological places and path segments in terms of the behaviors of control laws is a concrete example of a solution to the Symbol Grounding Problem [14].

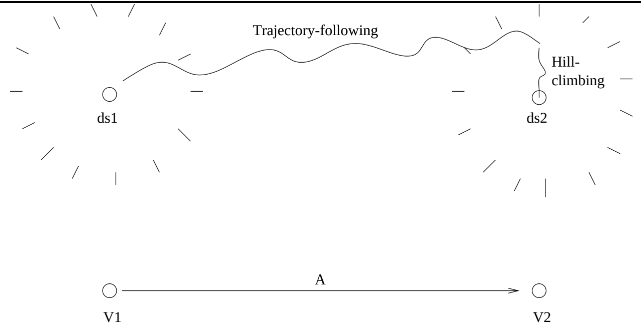

Fig. 3. Motion from one distinctive state to another via trajectory-following and hill-climbing control laws eliminates cumulative error. Reliable behavior can be abstracted to the causal schema $\langle V_1, A, V_2 \rangle$.

By physically hill-climbing to the local optimum of a “distinctiveness measure” defined over the local neighborhood, the robot localizes itself within that neighborhood with minimal assumptions about the nature of its sensors (Fig. 3). Because the dynamical system defines motion over the robot’s state space (location plus orientation), rather than over the work space (location alone), we came to realize that what is distinctive is the state, rather than the place, so we began to refer to *distinctive states* rather than *distinctive places*. For example, the single topological place at a T-intersection corresponds to four distinctive states, with the same location and different orientations. The Turn actions that link them correspond to trajectory-following control laws that change only orientation, followed by hill-climbing control laws to align with the walls of the corridors. (Later, in sect. 7, we will see a new conception of places and place neighborhoods.)

Motion among distinctive states avoids the problem of cumulative error that typically plagues robot mapping. There is no attempt to maintain an accurate location in a single global frame of reference. Rather, the purpose of an action is to move reliably from one distinctive state to another one. Any error that accumulates during trajectory-following is eliminated by the hill-climbing step, as long as the error is not so large as to miss entirely the basin of attraction of the destination distinctive state.

6 The Spatial Semantic Hierarchy

We started with the idea that the cognitive map consists of different representations for knowledge of space. As we come to understand spatial knowledge more deeply, the actual representations have evolved. We can best organize these different representations by grouping them according to *ontology*: the types of objects that can be described and the relations that can hold among them.

The *Spatial Semantic Hierarchy* (SSH) describes the cognitive map as consisting of four different levels, each with its own ontology, and each level grounded in the ones below [32, 30, 18, 50].

- At the *control level*, the agent and its environment are described as parts of a continuous dynamical system. The agent acts by selecting trajectory-following and hill-climbing *control laws*, subject to their applicability and termination conditions, so the agent-environment system moves toward an attractor. The stable attractor of a hill-climbing control law is called a *distinctive state*.
- At the *causal level*, the agent and its environment are described as a partially known finite-state automaton, whose *states* correspond to the distinctive states identified at the control level, and whose *actions* correspond to sequences of control laws. *Views* are the observable properties of states. A discrete state transition at the causal level corresponds to the extended evolution of dynamical systems at the control level.
- At the *topological level*, the environment is described in terms of *places*, *paths*, and *regions*, with relations such as connectivity, order, and containment. A state of the agent, described at the causal level, corresponds to being at a place, on a

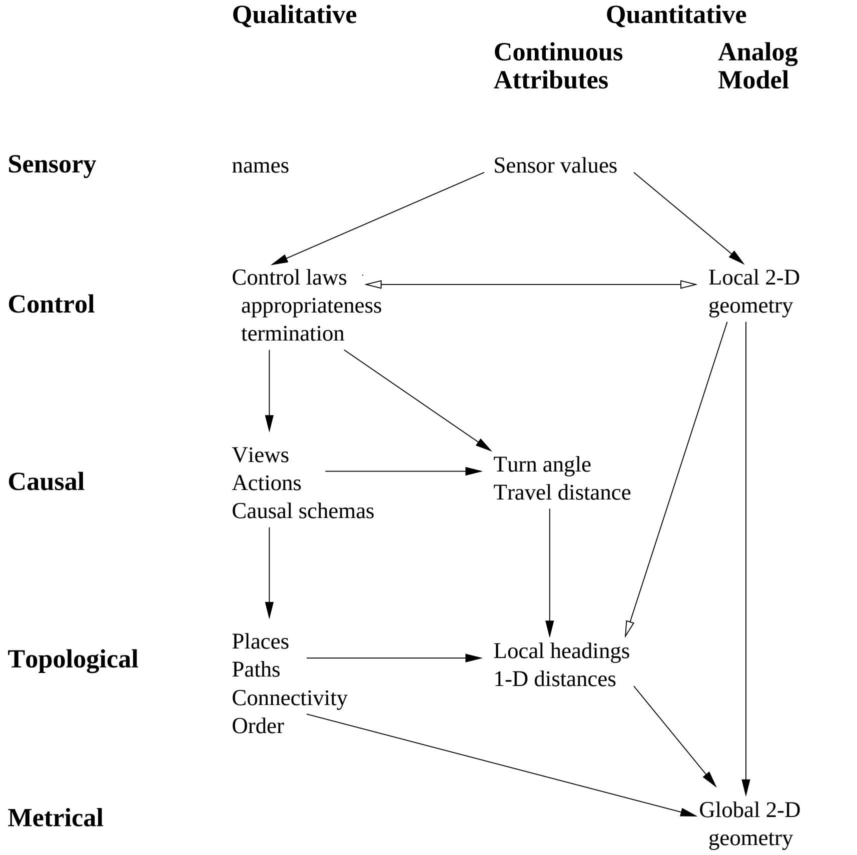

Fig. 4. The Spatial Semantic Hierarchy. Closed-headed arrows represent dependencies; open-headed arrows represent potential information flow without dependency.

path, and facing along the path in one of two directions. The topological map is created by a process of *abduction*, to explain the sequence of views and actions that represent the agent's experience at the interface between the control and causal levels [50].

- The *metrical level* has several different aspects. The causal and topological levels may include attributes with quantitative values, such as the magnitudes of actions, distances between places along paths, and angles between paths at places. A local place neighborhood can be described by a two-dimensional spatial analog such as an occupancy grid, with a single frame of reference. A spatial analog model of the large-scale environment can be created, based on the skeleton provided by the topological map.

There are logical dependencies (Fig. 4) among the levels, which constrain the combinations of representations that can occur. Different parts of the cognitive map may represent knowledge at different SSH levels, but each part of the map must respect the dependency structure. The agent's cognitive map may have a global metrical map of one portion of the environment, a topological map of another, simply causal knowledge of the sequence of actions to take in a third, and then use the control level to explore unknown territory. Or, when pressed for time or preoccupied with other concerns, the agent may access only causal knowledge to follow a familiar route even though topological and metrical knowledge may be available.

Emilio Remolina's doctoral work [50] provided a major step forward in the clarity of the SSH. He provided a formal axiomatization for the SSH causal and topological levels, plus the quantitative attribute portion of the metrical level. Since the topological map is the result of an abduction process, finding the best consistent explanation of the available observations, the formalization required a non-monotonic logic, in this case circumscription as embodied in Vladimir Lifschitz' nested abnormality theories [34]. The axioms express the consistency requirements for topological maps, and the nesting structure and the prioritized circumscription policy express the preference ordering on consistent maps. If a new observation should refute the current most preferred consistent map, then the preference ordering can be used to help select a preferred map from those still considered consistent.

This non-monotonic logical inference is implemented as an algorithm that creates a tree of all possible topological maps and imposes a preference order on the leaves.[8] At any point in time, the leaves of the tree represent the topological maps consistent with experience so far. After a travel action reaches and describes a new place neighborhood, some maps at the leaves of the tree are refuted as inconsistent, some are confirmed as consistent, and others branch on all consistent extensions. Branches only take place when there is perceptual aliasing; that is, when different places can have the same appearance. Then if a travel action reaches a place that appears the same as a previously-known place, two hypotheses must be created: one that the new place really is the same as the old one, and a second that the new place is genuinely new, but has the same appearance as the old one.

By initially creating all possible consistent successors, and refuting only the inconsistent ones, we maintain the guarantee that the correct topological map is present in the tree [50, 21]. In subsequent work, Francesco Savelli augmented the existing topological axioms with a test for the planarity of the topological map, which could be applied either as a consistency requirement or as a preference criterion [55]. It will also be important to use probability as well as prioritized circumscription policies to order the consistent maps [13].

The SSH treats observations gathered during exploration as the fundamental source of experience for building a cognitive map of large-scale space. However, there are other ways to obtain information about the structure of the environment. Verbal route directions translate naturally into sequences of actions (and minimal descriptions of views) at the SSH causal level [37]. Informal sketch maps translate

[8] Strictly speaking, the abduction searches for the best set of equality and inequality axioms over the symbols representing distinctive states. The algorithm creates models of those sets of axioms, and tests them for consistency.

naturally into subgraphs at the SSH topological level. And precise graphical maps provide information at the SSH metrical level. These and other forms of spatial communication are a topic for active research in psychology, linguistics, and cognitive science. One role for the SSH is to provide a useful description of the target representation for such communication.

7 The Hybrid Spatial Semantic Hierarchy

The four levels of the basic SSH framework start to look pretty satisfactory. This lets us turn our attention to certain assumptions and issues whose resolution will help us broaden and improve the Spatial Semantic Hierarchy.

First, the basic SSH treats perception as a black-box process that returns "view" symbols, abstractions of the full sensory image, capable only of being matched for equality or used as retrieval keys. We are ready to break down the hard separation between large-scale space and small-scale perceptual space. A more realistic theory of perception of the local environment, with both laser range-finders and computer vision, needs to be integrated with the cognitive mapping process.

Second, the basic SSH assumes that distinctive states are identified through the agent's physical motion, hill-climbing to the location in the environment that maximizes the current distinctiveness measure. This physical motion seems awkward and unnecessary.

Third, there has been an explosion of successful work on the SLAM (simultaneous localization and mapping) problem, building metrical maps of increasing size directly from sensory input within a single global frame of reference [57]. This approach differs significantly from the human cognitive map and from the multi-representation approach of the SSH. Do the two approaches compete? Are they complementary? Is one suitable for modeling humans while the other is for building robots? We need to understand the relationship between these two approaches.

Fortunately, there is a synergy between these three concerns that leads to their resolution [21]. Having defined *large-scale space* as space whose structure is larger than the sensory horizon, it is natural to define *small-scale space* as space whose structure is within the sensory horizon. Small-scale space is described by a *local perceptual map* that is metrically accurate and is constructed directly from sensory input. Recently developed SLAM methods are well suited for creating such a local perceptual map. We avoid the problem of closing large loops by confining the map to the agent's local perceptual surround, where we can apply the strengths of existing SLAM methods. When reasoning about small-scale space, we are concerned only with the frame of reference of the local perceptual map, and not with its inevitable drift with respect to the world frame of reference. We call the resulting combined model of large-scale and small-scale space, the *hybrid SSH*.

Local SLAM methods continually maintain the agent's localization in the frame of reference of the local map. Accurate incremental localization supports accurate incorporation of observations into the local map, and accurate local motion planning. In the basic SSH, hill-climbing provides the same benefit of accurate localization under weaker assumptions about sensors and effectors, but at the cost of physical motion to the distinctive state. In the hybrid SSH, when the agent has enough knowledge about its sensors and effectors to maintain its localization within the local perceptual map, it no longer requires physical hill-climbing.

Where the basic SSH treats views as atomic symbols, matched only for equality, the hybrid SSH treats the local perceptual map as the observable manifestation of a topological place [21]. The local perceptual map of a place neighborhood is parsed to define a local topology that describes how directed path segments join at that place. Distinctive states in the basic SSH causal level correspond to *gateways* within the local perceptual map of the place. Two local perceptual maps are matched by first matching their local topology descriptions, and then matching their perceptual maps to give a probability that they correspond to the same state. The local perceptual map with its local topology description bind together the small-scale-space and large-scale-space descriptions of the same place neighborhood, and thus bind together the continuous sensorimotor ontology and the discrete topological ontology.

The agent's experience in the environment is an alternating sequence of views and actions. However, in the hybrid SSH, a view corresponds to a pose within the local perceptual map, a turn action corresponds to motion within the local perceptual map of the current place neighborhood, while a travel action moves from one place neighborhood with its local perceptual map, to another place neighborhood. In addition to fixed local perceptual maps of place neighborhoods, a scrolling local perceptual map is used by trajectory-following control laws as an "observer" process to model obstacles in the agent's immediate surround. A topological place is detected at a change in the qualitative properties of the local topology of the scrolling local perceptual map during execution of a trajectory-following control law [5]. The topological map is built by abduction to explain this sequence of experiences. Where it is possible to have *perceptual aliasing* (two different places look the same), we build a tree of topological maps consistent with the same sequence of experiences. After sufficient exploration, inconsistent maps are refuted, and a single simplest or most probable map can be identified.

At this point, we can combine the global topological map with local perceptual maps of place neighborhoods to build a global metrical map of the large-scale environment in a single frame of reference [40]. Each local perceptual map defines a local frame of reference for accurate metrical knowledge at a place neighborhood, but the frame of reference will drift enough during travel to make it unusable globally. A consistent topological map hypothesis embodies a decision about which experiences of perceptually similar places were actually visits to the same place. Travel along each path segment between places can be used to estimate the displacement of each place in the local frame of reference of its predecessor. These local displacements between adjacent places can then be merged into a layout of the local place frames within a single global frame of reference, typically by applying a relaxation algorithm to the displacements. (The resulting probability of the global layout given the topological map and the displacements can be used as part of the preference ordering of topological maps in the tree of consistent maps.) The entire trajectory of robot poses can now be described in the global frame of reference, anchored by the poses

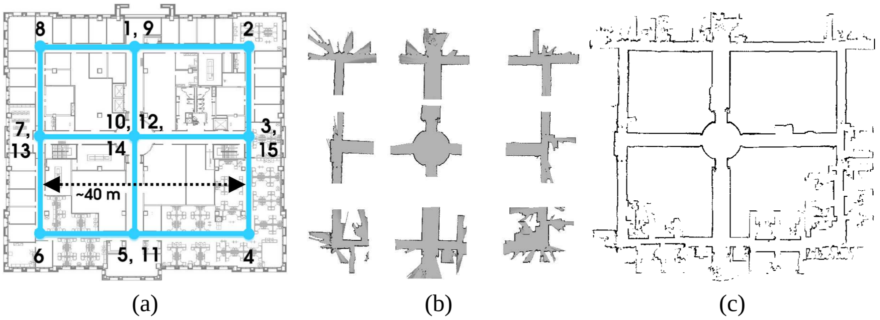

Fig. 5. The Hybrid SSH builds a global metrical map: (a) The robot explores an office environment with multiple nested large loops, identifying places in the sequence shown. (b) After inferring the correct topological map, the layout of local place maps in the global frame of reference. (c) The global map is created by localizing the trajectory poses in the global frame of reference, anchored by the poses in the local place maps, then creating the global map from the laser range-finder observations.

at both ends of each path segment, which already have accurate localization within the local frames of reference. Finally, an accurate global metrical map can be constructed, given the accurately localized trajectory of poses. This factors the problem of global metrical mapping into three tractable steps.

Part of the original motivation for the TOUR model of the cognitive map was the observation that humans do *not* typically create an accurate global metrical map from observations during travel. However, with increasing experience in the environment, they can learn a cognitive map that is increasingly faithful to the correct Euclidean model of the world [43]. Furthermore, accurate global metrical maps are valuable engineering and scientific tools, so it is useful for a robot to be able to build them. We demonstrate the value of combining different representations of space by showing how to build a correct global metrical map on the skeleton provided by an accurate global topological map, using observations from experience in the local perceptual map.

8 Foundational Learning

We have jumped over a research thread that has important implications for the future. The Spatial Semantic Hierarchy, both basic and hybrid, presumes that the agent has a collection of control laws for coupling its sensors, effectors, and environment together. This, in turn, presumes that the agent possesses (or embodies) knowledge of which sensory features are useful, and how its effectors change those features. In an artificially constructed robot, much of this knowledge is built in by the designer. In a biological creature, some of this knowledge is innate. We ask, how can this knowledge be learned? Biologically, some of the learning is done by the species over evolutionary time, while the rest is done by the individual.

This question was inspired by a challenge problem proposed by Ron Rivest at MIT in 1984 [27]. Suppose an agent wakes up in an unknown world, with a sense vector and a motor vector, but with no knowledge of how they are related to its world. How can such an agent learn to predict the results of future actions? This challenge led Rivest, Sloan, and Schapire to a series of results about learning finite automata from observations [51, 54, 52, 53]. My own approach was to try to learn the sensorimotor foundation for the TOUR model from exploration experience [27].

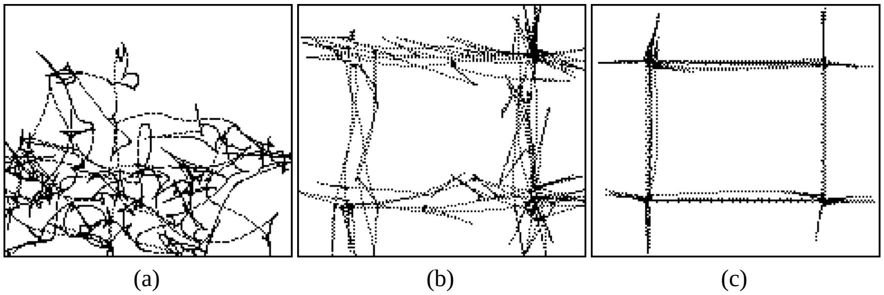

Fig. 6. Exploring a simple world at three levels of competence. (a) The robot wanders randomly while learning a model of its sensorimotor apparatus. (b) The robot explores by randomly choosing applicable homing and open-loop path-following behaviors based on the static action model while learning the dynamic action model. (c) The robot explores by randomly choosing applicable homing and closed-loop path-following behaviors based on the dynamic action model.

Around 1988, David Pierce and I began to investigate this question for an agent with continuous experiences in a continuous world. After developing some preliminary pieces of the puzzle [45, 48, 46], we demonstrated a learning agent that started with an uninterpreted sensorimotor system in an unknown world, and learned: (a) to separate the sense vector into distinct sensory modalities; (b) to learn a low-dimensional spatial structure for the sense elements (“pixels”) in a particular modality; (c) to identify primitive actions from the sensory flow fields induced on this spatial structure; (d) to identify a set of stable sensory features that can be extracted and tracked in the sensory image; (e) to learn which actions cause reliable changes to which perceptual features in which contexts; (f) to construct useful homing (i.e., hill-climbing) and trajectory-following control laws from those actions; and (g) to define distinctive states and actions linking them [44, 47]. Thus, by bootstrapping through a number of intermediate representations, the agent learned a sufficient foundation to reach the “bottom rung” of the SSH ladder. While there were a number of assumptions and limitations in this work, it genuinely demonstrated that a computational agent could learn its own sensorimotor grounding from its own interaction with the environment (Fig. 6).

This research thread returned to the back burner for several years, until Patrick Beeson and I started looking at the problem of place recognition [20]. A realistic robot receives a high-dimensional sensory image at any given moment. For the basic SSH causal level, that image must be abstracted to one of a discrete set of views. Our goal was to learn a view representation such that each view correctly determines a unique distinctive state. We build on the fact that perceptual aliasing of distinctive states can be overcome by continued exploration, proposing candidate topological maps and refuting the incorrect ones when predictions are violated.

We gave the name *bootstrap learning* to the learning method we developed.[9] Start by creating an over-abstract but usable view representation: cluster sensory images aggressively enough that each distinctive state corresponds to only one view, even at the cost of multiple states having the same view (perceptual aliasing). Then the standard SSH exploration and mapping methods can converge to the correct topological map after enough exploration. The correct topological map provides a correct association between distinctive states and the high-dimensional sensory images, even if the views are aliased. So now we can use supervised learning (more powerful than unsupervised clustering), to learn correct associations between sensory images and distinctive states. In two experiments with rich sensors and real environments, the learning agents rapidly reached 100% accurate place recognition.

The generic structure of this bootstrap learning scenario is: (1) approximately abstract the problem using an unsupervised method; (2) use a much more expensive inference method to find the correct answer; (3) use supervised learning to find the correct level of abstraction. We believe that this pattern can be applied to other abstraction-learning problems.

More recently, Joseph Modayil and I have been considering the problem of how a higher-level ontology of objects and actions can be learned from experience with a lower-level ontology of individual sense elements (“pixels”) and motor signals [41]. This, too, requires a multi-stage learning process. It was developed and demonstrated using the range-sensor-based local perceptual map (implemented as an occupancy grid) used by our exploring robots. First, we identify those sensor returns in the current sensor image that are explained by static features of the environment, represented by cells in the occupancy grid that have high confidence of being occupied, and have never had high confidence of being free space. The remaining sensor returns are explained by cells whose occupancy has changed at some time in the past. Second, we cluster these “dynamic” sensor returns in the current sensory image frame; and third, we attempt to track these clusters from frame to frame over time. These trackable clusters are hypothesized to be explainable as images of objects. The fourth step is to collect a sequence of images of an object from different perspectives to describe its shape; and the fifth is to create a classification hierarchy of object types based on this described shape. Ongoing work considers the abstraction of actions applied to these learned objects.

[9] We have since extended the term “bootstrap learning” to apply to this general approach to foundational learning.

9 Conclusions

I began studying the cognitive map as a manageable subset of commonsense knowledge. I hoped that this problem would *not* be “AI Complete” — that is, it could be sufficiently separated from other major issues in AI and cognitive science that it would be possible to make useful progress without simultaneously solving every other major problem in AI. At the same time, knowledge of space is clearly a fundamental part of commonsense knowledge [43, 33], so progress in understanding the cognitive map contributes to the overall enterprise of understanding commonsense knowledge, and hence the nature of mind.

It seems to me that these hopes were well justified, and the research efforts have paid off. Boundaries separating one scientific problem from another are always artificial scaffolding, used to make a problem tractable for human minds. Once enough progress has been made on one formulation of a problem, it becomes time to move the scaffolding so progress can be made on a larger formulation. The progress from the TOUR model to the Basic SSH and then to the Hybrid SSH seems to me to have exactly this character. Each problem definition served its purpose, led to an improved understand of the nature of spatial knowledge, and was replaced by a new, larger, problem definition. The focus of the TOUR model was primarily on the role of topological knowledge of space. The focus of the Basic SSH was on the role of control laws and dynamical systems. The focus of the Hybrid SSH is on the role of metrical knowledge and perception.

When I first learned about Minsky’s frames for knowledge representation, I wondered where the slots come from. The multiple representations of the TOUR model and the Spatial Semantic Hierarchy are clearly distinct theories with distinct ontologies. The flexibility and robustness of commonsense knowledge depends on having multiple ontologies for the same domain of knowledge. The question of where the slots come from has been transformed into the question, *How can an agent learn, not just new knowledge within an existing ontology, but a new ontology it does not already possess?*

The foundational learning problem is not simply an enlarged version of the cognitive mapping problem. Rather, now that we have a reasonably solid theory of spatial knowledge in the cognitive map, we can ask questions about its foundation with a degree of specificity that was not possible before. We can also evaluate foundational learning methods according to their ability to support higher-level theories that we already understand. In my own case, the theory of the cognitive map serves this role. However, the learning methods we seek will serve as foundations for a much larger body of commonsense knowledge.

References

1. K. Basye, T. Dean, and L. P. Kaelbling. Learning dynamics: system identification for perceptually challenged agents. *Artificial Intelligence*, 72:139–171, 1995.
2. K. Basye, T. Dean, and J. S. Vitter. Coping with uncertainty in map learning. *Machine Learning*, 29(1):65–88, 1997.

3. J. D. Becker. “Robot” computer problem solving system. Technical Report 2316, Bolt Beranek and Newman, September 1972.
4. J. D. Becker and E. W. Merriam. “Robot” computer problem solving system. Technical Report 2792, Bolt Beranek and Newman, April 1974.
5. P. Beeson, N. Jong, and B. Kuipers. Towards autonomous topological place detection using the extended Voronoi graph. In *IEEE International Conference on Robotics and Automation*, 2005.
6. Kenneth Boulding. *The Image*. University of Michigan Press, Ann Arbor, 1956.
7. J. M. Crawford and B. J. Kuipers. Toward a theory of access-limited logic for knowledge representation. In *Proc. 1st Int. Conf. on Principles of Knowledge Representation and Reasoning*, San Mateo, CA, 1989. Morgan Kaufmann.
8. J. M. Crawford and B. J. Kuipers. Negation and proof by contradiction in access-limited logic. In *Proc. 9th National Conf. on Artificial Intelligence (AAAI-91)*. AAAI/MIT Press, 1991.
9. R. M. Downs and D. Stea. *Image and Environment*. Aldine Publishing Company, Chicago, 1973.
10. G. Dudek, M. Jenkin, E. Milios, and D. Wilkes. Robotic exploration as graph construction. *IEEE Trans. on Robotics and Automation*, 7(6):859–865, 1991.
11. C. Fillmore. The case for case. In E. Bach and R. T. Harms, editors, *Universals in Linguistic Theory*. Holt, Rinehart and Winston, Chicago, 1968.
12. W. S. Geisler. Sequential ideal-observer analysis of visual discriminations. *Psychological Review*, 96:267–314, 1989.
13. D. Hähnel, S. Thrun, B. Wegbreit, and W. Burgard. Towards lazy data association in SLAM. In *Proc. Int. Symp. on Robotics Research (ISRR-03)*, 2003.
14. Stevan Harnad. The symbol grounding problem. *Physica D*, 42:335–346, 1990.
15. P. J. Hayes. In defence of logic. In *Proc. 5th Int. Joint Conf. on Artificial Intelligence (IJCAI-77)*, pages 559–565, 1977.
16. B. Kuipers. Commonsense reasoning about causality: Deriving behavior from structure. *Artificial Intelligence*, 24:169–204, 1984.
17. B. Kuipers. Qualitative simulation. *Artificial Intelligence*, 29:289–338, 1986.
18. B. Kuipers. The Spatial Semantic Hierarchy. *Artificial Intelligence*, 119:191–233, 2000.
19. B. Kuipers. Why don’t I take military funding? http://www.cs.utexas.edu/users/kuipers/opinions/no-military-funding.html, 2003.
20. B. Kuipers and P. Beeson. Bootstrap learning for place recognition. In *Proc. 18th National Conf. on Artificial Intelligence (AAAI-2002)*, pages 174–180. AAAI/MIT Press, 2002.
21. B. Kuipers, J. Modayil, P. Beeson, M. MacMahon, and F. Savelli. Local metrical and global topological maps in the hybrid spatial semantic hierarchy. In *IEEE Int. Conf. on Robotics & Automation (ICRA-04)*, 2004.
22. B. J. Kuipers. A frame for frames: representing knowledge for recognition. In D. G. Bobrow and A. Collins, editors, *Representation and Understanding*, pages 151–184. Academic Press, New York, 1975.
23. B. J. Kuipers. *Representing Knowledge of Large-Scale Space*. PhD thesis, Mathematics Department, Massachusetts Institute of Technology, Cambridge, MA, 1977. http://www.cs.utexas.edu/users/qr/papers/Kuipers-PhD-77.html.
24. B. J. Kuipers. Modeling spatial knowledge. *Cognitive Science*, 2:129–153, 1978.
25. B. J. Kuipers. Commonsense knowledge of space: learning from experience. In *Proc. 6th Int. Joint Conf. on Artificial Intelligence (IJCAI-79)*, pages 499–501, Tokyo, Japan, August 1979.

26. B. J. Kuipers. Modeling human knowledge of routes: Partial knowledge and individual variation. In *Proc. 3rd National Conf. on Artificial Intelligence (AAAI-83)*, Los Altos, CA, 1983. Morgan Kaufmann.
27. B. J. Kuipers. The map-learning critter. Technical Report AI TR 85-17, University of Texas at Austin, Artificial Intelligence Laboratory, Austin, TX, 1985.
28. B. J. Kuipers. *Qualitative Reasoning: Modeling and Simulation with Incomplete Knowledge*. MIT Press, Cambridge, MA, 1994.
29. B. J. Kuipers and Y. T. Byun. A robust qualitative method for spatial learning in unknown environments. In *Proc. 7th National Conf. on Artificial Intelligence (AAAI-88)*, pages 774–779, Los Altos, CA, 1988. Morgan Kaufmann.
30. B. J. Kuipers and Y.-T. Byun. A robot exploration and mapping strategy based on a semantic hierarchy of spatial representations. *Journal of Robotics and Autonomous Systems*, 8:47–63, 1991.
31. B. J. Kuipers and J. P. Kassirer. Causal reasoning in medicine: analysis of a protocol. *Cognitive Science*, 8:363–385, 1984.
32. B. J. Kuipers and Tod Levitt. Navigation and mapping in large scale space. *AI Magazine*, 9(2):25–43, 1988.
33. G. Lakoff and M. Johnson. *Metaphors We Live By*. The University of Chicago Press, Chicago, 1980.
34. V. Lifschitz. Nested abnormality theories. *Artificial Intelligence*, 74:351–365, 1995.
35. C. Linde and W. Labov. Spatial networks as a site for the study of language and thought. *Language*, 51:924–939, 1975.
36. Kevin Lynch. *The Image of the City*. MIT Press, Cambridge, MA, 1960.
37. M. MacMahon. A framework for understanding verbal route instructions. In A. Schultz, editor, *The Intersection of Cognitive Science and Robotics: From Interfaces to Intelligence*, Papers from the AAAI Fall Symposium, pages 97–102, 2004.
38. G. A. Miller, E. Galanter, and K. H. Pribram. *Plans and the Structure of Behavior*. Holt, Rinehart and Winston, 1960.
39. M. Minsky. A framework for representing knowledge. In P. H. Winston, editor, *The Psychology of Computer Vision*. McGraw-Hill, NY, 1975.
40. J. Modayil, P. Beeson, and B. Kuipers. Using the topological skeleton for scalable, global, metrical map-building. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2004.
41. J. Modayil and B. Kuipers. Bootstrap learning for object discovery. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2004.
42. H. Moravec and A. Elfes. High resolution maps from wide angle sonar. In *IEEE International Conference on Robotics and Automation*, pages 116–121, 1985.
43. Jean Piaget and Baerbel Inhelder. *The Child's Conception of Space*. Norton, New York, 1967. First published in French, 1948.
44. D. Pierce and B. Kuipers. Learning to explore and build maps. In *Proc. 12th National Conf. on Artificial Intelligence (AAAI-94)*. AAAI/MIT Press, 1994.
45. D. M. Pierce. Learning turn and travel actions with an uninterpreted sensorimotor apparatus. In *IEEE International Conference on Robotics and Automation*, Sacramento, CA, 1991.
46. D. M. Pierce and B. J. Kuipers. Learning hill-climbing functions as a strategy for generating behaviors in a mobile robot. In J.-A. Meyer and S. W. Wilson, editors, *Proceedings of the International Conference on Simulation of Adaptive Behavior: From Animals to Animats*, Cambridge, MA, 1991. MIT Press/Bradford Books.
47. D. M. Pierce and B. J. Kuipers. Map learning with uninterpreted sensors and effectors. *Artificial Intelligence*, 92:169–227, 1997.

48. D.M. Pierce. Learning a set of primitive actions with an uninterpreted sensorimotor apparatus. In L. Birnbaum and G. Collins, editors, *Machine Learning: Proceedings of the Eighth International Workshop*, San Mateo, CA, 1991. Morgan Kaufmann.
49. V. Pratt. A linguistics oriented language. In *Proc. 3rd Int. Joint Conf. on Artificial Intelligence (IJCAI-73)*, 1973.
50. E. Remolina and B. Kuipers. Towards a general theory of topological maps. *Artificial Intelligence*, 152:47–104, 2004.
51. R. L. Rivest and R. E. Schapire. A new approach to unsupervised learning in deterministic environments. In *Proceedings of the Fourth International Workshop on Machine Learning*, 1987.
52. R. L. Rivest and R. E. Schapire. Inference of finite automata using homing sequences. In *Proc. 21st ACM Symposium on Theory of Computing*, pages 411–420. ACM, 1989.
53. R. L. Rivest and R. E. Schapire. Diversity-based inference of finite automata. *Journal of the ACM*, 41(3):555–589, May 1994.
54. R. L. Rivest and R. Sloan. Learning complicated concepts reliably and usefully. In *Proc. 7th National Conf. on Artificial Intelligence (AAAI-88)*, pages 635–640. AAAI Press/The MIT Press, 1988.
55. F. Savelli and B. Kuipers. Loop-closing and planarity in topological map-building. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS-04)*, 2004.
56. A. W. Siegel and S. H. White. The development of spatial representations of large-scale environments. In H. W. Reese, editor, *Advances in Child Development and Behavior*, volume 10, pages 9–55. Academic Press, New York, 1975.
57. S. Thrun. Robotic mapping: A survey. In G. Lakemeyer and B. Nebel, editors, *Exploring Artificial Intelligence in the New Millenium*. Morgan Kaufmann, 2002.