Report on the Hybrid Spatial Semantic Hierarchy

Tony Pan
tonypan@umich.edu

September, 2019

Introduction

At the forefront of robotics and AI research, one may be tempted with techniques such as machine learning and computer vision. Mobile robots, however, are more fundamentally dependent on navigation maps, because a robust and accurate representation of an exploration environment dictates the range of motion of the robots. There are two main parts to the robot navigation challenge: localization and selection of control laws. For localization, robots often rely on odometry, the estimation of current pose based on sensorimotor data. Small inconsistencies such as wheel diameters, encoder accuracy, and travel surfaces can introduce error in odometry, leading to cumulative drift in pose estimation. Although methods such as Simultaneous Localization and Mapping (SLAM) and the use of landmarks can re-calibrate and correct for odometry errors, they also require more computing power and longer processing time. To make matters more complicated, metrical errors are difficult to identify and hence correct for. While there are many algorithms for selecting where a robot should travel to and what control laws should be applied, they are limited when there are incomplete or incorrect representations in the navigation maps. Furthermore, it is inefficient to search for paths and conduct motion planning from global metrical maps due to the complexity and size of such computation.

Instead of complex mathematics models and expensive equipment, the most optimal solution to the robot exploration and navigation problems may lie within the common sense human spatial knowledge. The human spatial knowledge is "arguably the foundation for most other kinds of commonsense knowledge" (Kuipers, 2008). The human spatial knowledge consists of several distinct representations for different aspects of space, including procedures for getting from place to another (local metrical map), topological network maps of an environment (global topological layout), and geometrical models of the environment (global metrical layout). The human spatial knowledge is ideal also because it is grounded in sensorimotor experience. Humans cannot see the entire Earth, so we build representation of space in our neighborhood based on travel experience, and piece each segment together to represent environments larger than our eyes can see. Most importantly, humans can work with incorrect and incomplete spatial knowledge and deduce a general sense of direction. Children can learn symbolic representations of space and develop skills to construct various forms of spatial knowledge, and the exact symbolic representation varies from person to person. Although the human spatial knowledge is imperfect, modeling after the human cognitive map allows for representation of the environment using human-like concepts, such as place and path, human-robot interaction, and logical reasoning.

Over the years, Professor Benjamin Kuipers has developed three models: TOUR, SSH, and HSSH. Each model uses properties in the human cognitive map to build maps describing an environment through an agent's exploration experience. The TOUR model first solved the problem of inferring local sensory experience to build a global spatial structure. The SSH models large-scale space through robot exploration in several levels that merge into a global map. The HSSH further improves the map building of globally consistent metrical maps from a topological skeleton from robot travel experience and connecting local frames of reference.

TOUR

The TOUR model builds a computational model of global spatial structure from local sensory experience. It draws conclusions locally on small steps and requires several travel actions on the same tour to build the global map. The model abstracts spatial knowledge into procedural (discrete travel actions), topological (connectivity of places and paths), and metrical levels (continuous attributes and relations). The TOUR machine is a finite-state, rule-driven automaton with states describing the place, path, and direction. Although the TOUR model can represent space from experience like the human cognitive map, it has many drawbacks. The procedural level abstracts away sensory inputs from the environment too thoroughly such that the model becomes unable to express gaps in travel action or perceptual aliasing. It can only identify but not interpret symbols that the agent senses. The travel gaps lead to an incomplete schema, which the agent can still use for navigation, but cannot review the route outside of the environment. Furthermore, TOUR model's presumption that the continuous experience of a traveling "agent has already been abstracted to a discrete sequence of states and transitions" makes it difficult for implementation on physical robots (Kuipers, 2008).

SSH

The Spatial Semantic Hierarchy (SSH) model represents large-scale space both qualitatively and quantitatively in the sensory, control, causal, topological, and metrical levels, each with its own ontology. Qualitative spatial knowledge can be matched for equality or be compared but cannot necessarily support a difference operation. At the sensory level, the agent continuously senses its local surroundings' environment in the form of continuous sensor values and produces behavior of the robot.

At the SSH control level, continuous control laws bind the agent and the environment into a dynamical system including the appropriateness and termination of each selected control law. A control law specifies the relation between sensory input and motor output. The two main types of control laws are hill climbing control laws that moves agent from one distinctive state to the next by seeking the isolated local maximum of a distinctiveness measure, and trajectory following control laws that moves the agent from one state to a place neighborhood without entering another distinctive state. A dynamical system, modeled by a differential equation, determines an agent's behavior, which depends on the sensory data, the control law selected, and the time derivative of the state and control input (Kuipers, 2004). The dynamical system modeled by SSH takes in sensory input and finds the observed current local state from the observer's perspective and the desired local state from the gradient of distinctiveness measure based on sensory input. The dynamical system then takes both the current local state and the desired local state as input and calculates the motor output. The dynamical system in SSH makes it feasible to train agents motion control in unknown environments. The transition between control laws is based on weighted appropriateness measures. The control level builds a local two-dimensional geometry broken into individual neighborhoods based on the dynamical system and local sensory data. A locally distinctive state can be found by alternating the said control laws. Abstraction from continuous behavior to symbolic map is achieved by first applying trajectory-following control law to the neighborhood of a distinctive state and then applying hill-climbing control law to reduce the cumulative error during travel actions. The closure criteria on the control laws leaves each distinctive state with no dead ends, and that each trajectory terminates at a distinctive state. If the choice of hill-climbing control law is unique, the causal level will be deterministic, and if closely-competing hill-climbing control laws take the agent to different distinctive states, then causal level will be non-deterministic.

At the SSH causal level, the continuous world and agent behavior are abstracted into discrete sensory views, actions, and their causal relations. A view is a description of the sensory input vector at a locally distinctive state and can help uniquely identify current state. An action is a sequence of applications of one or more control laws which can start at a locally distinctive state and terminates at another. A schema is a tuple representing the starting view, the travel action, and the terminating view. A declarative schema is observed before and after the action. A procedural action implies the initiation of an action if a view is observed at the current locally distinctive state. Partially filled schema without the view after the action needs to be stored in the working memory for the creation of a complete schema. A routine is a set of schemas indexed by initial view. A routine is complete if it contains the views after action for each schema and can thus support cognitive operations in the absence of the explored environment. An adequate routine contains partially filled schemas without the terminating view and can only be used within the environment. The causal level also includes quantitative attributes of turn and travel actions such as turn angle, travel distance, and agent orientation, which can be translated back to the control level. A turn action leaves the agent at the same place while a travel action takes the agent from one place to another. The view-action interface among the control, causal, and topological level is consistent with the human spatial knowledge.

At the SSH topological level, the ontology of places, paths, regions, and their connectivity, order and containment relations are abducted from the views, actions, and causal schemas from the causal level. Regions are sets of places on the same side of a boundary. The hypothesized places, paths, regional structure, local headings, and one-dimensional distances can augment the topological map to improve motion planning. For route planning, one can use the left and right containment boundary or a topological grid. An abstraction region represents the set of places in a large granularity map. The topological level is beneficial for reducing cumulative error and for path finding.

Finally, at the SSH metrical level, a patchwork of local frames builds a global metrical representation of the large-scale space from the agent's travel experience. The local geometric maps are merged by topological descriptions into a global two-dimensional geometry.

The SSH functions well in "environments where locally distinctive states can be defined, and where motion in qualitatively uniform regions is sufficiently reliable between locally distinctive states" (Kuipers, 2000). However, if the environment is so uniform that there are no distinctive states, then modeling with SSH becomes impossible. On the other hand, the views, symbols, and abstractions of the full sensory image in SSH can only be matched for equality, perception of the local environment can be improved by computer vision. Additionally, the physical motion of hill-climbing to location maximizing the current distinctiveness measure becomes obsolete with SLAM. Most importantly, the SSH model depends on the assumption that the environment naturally decomposes into place neighborhoods connected by path segments, which can then be abstracted to a topological map. With the improvement of modern sensor technology, stronger assumptions can be made about sensory input available to the agent.

HSSH

The Hybrid Spatial Semantic Hierarchy (HSSH) takes advantage of those sensors by extending the metrical mapping techniques to create precise observational models of the local surroundings. HSSH abstracts spatial knowledge into several levels of representations: the local metrical level, the local topological level, the global topological level, and the global metrical level. The biggest improvement of HSSH over SSH is the more detailed representation of the local environment with the use of a local perceptual map (LPM), a metrically accurate local map within an agent's sensory horizon. Views and symbols that can only be matched for equality in the SSH are replaced by local topology in the HSSH. HSSH also uses gateways for motion control instead of a dynamic system model.

At the HSSH Local Metrical Level, the agent builds and localizes itself in the LPM. The LPM is beneficial for both motion planning and hazard avoidance at the local level. It describes a place and identifies local obstacles in one frame of reference and scrolls to update. SLAM makes it unnecessary for the agent to physically hill-climb to the distinctive states. The HSSH model utilizes the occupancy grid and the particle filter Markov localization to reduce odometry error. Nevertheless, the Local Metrical Representation resembles the SSH control level because it takes in sensory input from an observer's perspective and sends motion commands to the hardware level.

At the HSSH Local Topological Level, the agent identifies discrete places and symbolically describes the configuration of the paths through the place. A place has a circular order of directed paths radiating from it. The local topology partitions direction concisely into left and right. An agent is on path if it is traveling via control laws and is not on-path if it is in a place neighborhood intersection or dead end. Neighborhoods are connected nodes of places in the topological map. Grounding local paths in the LPM is achieved with gateways, which are boundaries between local places. A gateway is also an interface between the two types of travel, motion along paths and motion within place neighborhoods, which are required for abstraction from small to large scale space. The HSSH model finds gateways by first pruning a Voronoi skeleton, followed by defining the core of a local region, and finally looking for constrictions in the frontiers. Pruning reduces noise and removes islands, occupied or unknown cells surrounding by free cells. A spanning tree only connecting exits, terminal points that reach edges of the LPM or unexplored cells, is the best approach for pruning. Place detection can be achieved by recalculating gateways and the local topology defined by gateways at each time step. Gateways also provide a geometric method for motion control.

At the HSSH Global Topological Level, the agent resolves structural ambiguities and determines how to describe the global environment as a graph of places, paths, and regions. The global topological map needs to be consistent with exploration experience, but closing large loops makes it difficult. A single topological map is generated for each qualitatively distinct alternative to a loop when structural ambiguity arises. A small-scale star describes the circular order and the correspondence between directed local-paths and oriented gateways. In the large-scale space, when the directed local-path passes through a place, the distinctive state corresponds with two differently oriented gateways, one entering and the other departing. To simplify the construction of the global topological map, exploration experience alternates between travel actions and place neighborhoods. The topological map maintains a tree whose nodes are pairs of topological map and distinctive state representing the agent's current position. The leaves represent all possible topological maps consistent with travel experience. If a new territory is discovered by an action, then it is either a new distinctive state or a loop is closed. Because only complete local topologies are matched, the tree of maps only branches on travel actions. Both the SSH Causal Level and the HSSH Global Topological Level describes a distinctive state as being at a place along a directed path. In SSH, distinctive states are grounded by isolated distinctive states where hill-climbing control laws terminate; in HSSH, distinctive states are grounded by a directed local-path from the LPM of a place neighborhood.

At the HSSH Global Metrical Level, the layout of places, paths, and obstacles are specified in one global frame of reference. The metrical inference comes mainly from defining an appropriate set of reference frames and estimating the values of metrical quantities. Pose estimation is the most important step when searching for the maximum-likelihood path the agent traveled. When there is only one global topology, the probability of the pose given sensorimotor data is straightforward. When multiple topologies exist, the tree of topological maps is used to construct metrical map for each potential topological map. The pose estimation process depends on the place-to-place displacements derived from local metrical maps, the metrical layout of places in the global topological map, and the global metrical layout of the agent's trajectory. Lastly, the global metrical map is constructed by projecting the recorded range measurements from poses in the new global coordinates (Beeson, Modayil, & Kuipers, 2010).

Discussion

The HSSH is modeled after the human cognitive map, so its representation of space is compatible with human interaction. It can guide agents to navigate local environments with incomplete or incorrect knowledge and build global metrical representations of unknown environments through travel experience. While many questions have been answered throughout the years of development, there are still many interesting areas for research.

A physical implementation of HSSH on the Anki Cozmo robots is feasible. Metrical values of travel actions can be obtained from the Cozmo odometry to establish the basis for the LPM. While Cozmo does not have SLAM features, it localizes itself with respect to landmarks (cubes, charger, and other objects with fiducials on them). It uses an occupancy grid and annotates the cells with not only the probability of occupancy but also the object type and ID if known. Cozmo's memory navigation map stores the cells in a tree structure with each leaf nodes being the occupied content and the other nodes splitting into children nodes. A local topological structure and a tree of topological maps can be built from the Cozmo navigation memory map with some modifications. While Cozmo itself does not use gateways for motion control or building topological maps, such method can be implemented with the images streamed from the Cozmo camera. The gateway can be used to re-calibrate the Cozmo pose and reduce cumulative odometry error during the travel experience. The motion planning and control can be implemented and executed onboard Cozmo, and the more computationally heavy global maps can be generated on a computer running on a separate Python thread.

One of the drawbacks of the HSSH and SSH is that they do not work in uniform environments where it is difficult to differentiate places. Humans have been able to navigate the Micronesian archipelago without modern instruments and islands in sight for long periods of time. The use of HSSH can be greatly extended if it can also explore places like the ocean, surface of another planet, and similar environments with few human-made objects and differentiating views. An anchor based gateway for coastal navigation was proposed in which an agent starts at a known edge; and then defines control laws into the unknown space; during the exploration process, the agent stays localized using SLAM; and finally estimates pose when it arrives at an "island" in the unknown open space. Building an implementation of the anchor based gateway in the HSSH in either simulation or with physical robots can be an interesting project.

In the SSH, the dynamical system and control laws imply an agent can learn which sensorimotor features are useful through travel experience. A bootstrap learning process was developed to enable an agent to learn a view representation that correctly determines a unique distinctive state by continued exploration and supervised learning. However, humans rely not only on vision to navigate and explore, but also on verbal directions. An agent could perhaps learn a topological layout from verbal instructions that describe paths, places, regions, and boundaries. What if a similar bootstrap learning method is implemented such that high-level natural language verbal descriptions are given to an agent, and then supervised learning is used to teach the exact association between the verbal description and the topology? Can the HSSH model build a representation of the environment from both travel experience and verbal instructions?

The LPM is a representation of local space built from within an agent's sensory horizon, and information scrolling off the LPM is discarded. While this keeps the LPM a fixed time and size, it also maybe assuming the explored environment is static. If the scrolled off environment is dynamic, can the agent still identify the place being the same place, or will it see it as a new place and cause loop closing issues? For example, if a car drives into an empty parking space already explored by the agent, and the parking space scrolls off the LPM. When the agent sees the occupied parking space again, will it recognize it as a new space and add it to the tree of topological maps with the "new" place along with the current robot position? If so, it seems like the tree of topological maps can grow quite large in a dynamic environment. Is there a more efficient approach to model places in an environment that changes, or is this already handled? More research can be done to improve the efficacy and efficiency of the LPM and the tree of topological maps.

References

Beeson, P., Modayil, J., & Kuipers, B. (2010). Factoring the mapping problem: Mobile robot map-building in the hybrid spatial semantic hierarchy. *The International Journal of Robotics Research*, *29*(4), 428-459. doi: 10.1177/0278364909100586

Kuipers, B. (2000). The spatial semantic hierarchy. *Artificial Intelligence*, *119*(1), 191 - 233. doi: https://doi.org/10.1016/S0004-3702(00)00017-5

Kuipers, B. (2004, January). Control tutorial.

Kuipers, B. (2008). *An intellectual history of the spatial semantic hierarchy"*. Springer Berlin Heidelberg.