Bootstrap Learning for Place Recognition

BENJAMIN KUIPERS AND PATRICK BEESON
Department of Computer Sciences
University of Texas at Austin
Austin, TX 78712 USA
{kuipers,pbeeson}@cs.utexas.edu

ABSTRACT

We present a method whereby a robot can learn to recognize places with high accuracy, in spite of perceptual aliasing (different places appear the same) and image variability (the same place appears differently). The first step in learning place recognition restricts attention to distinctive states identified by the map-learning algorithm, and eliminates image variability by unsupervised learning of clusters of similar sensory images. The clusters define *views* associated with distinctive states, often increasing perceptual aliasing. The second step eliminates perceptual aliasing by building a causal/topological map and using history information gathered during exploration to disambiguate distinctive states. The third step uses the labeled images for supervised learning of direct associations from sensory images to distinctive states. We evaluate the method using a physical mobile robot in two environments, showing high recognition rates in spite of large amounts of perceptual aliasing.

Introduction

Can a robot learn to recognize places (position and orientation), given a single sensory image, with high accuracy?

- Two complementary problems stand in the way of reliable place recognition.
  - *Perceptual aliasing*: different places may have similar or identical sensory images.
  - *Image variability*: the same position and orientation may have different sensory images on different occasions.
- *Bootstrap learning* is a way of composing multiple learning and inference methods to start with weak methods and build prerequisites for the application of stronger methods.
- Applying bootstrap learning to place recognition exploits the structure of the *Spatial Semantic Hierarchy (SSH)* (Kuipers 2000).
- Bootstrap learning eliminates image variability and perceptual aliasing from groups of sensor images.

distinctive state (dstate) The isolated fixed-point of a hill-climbing control law.

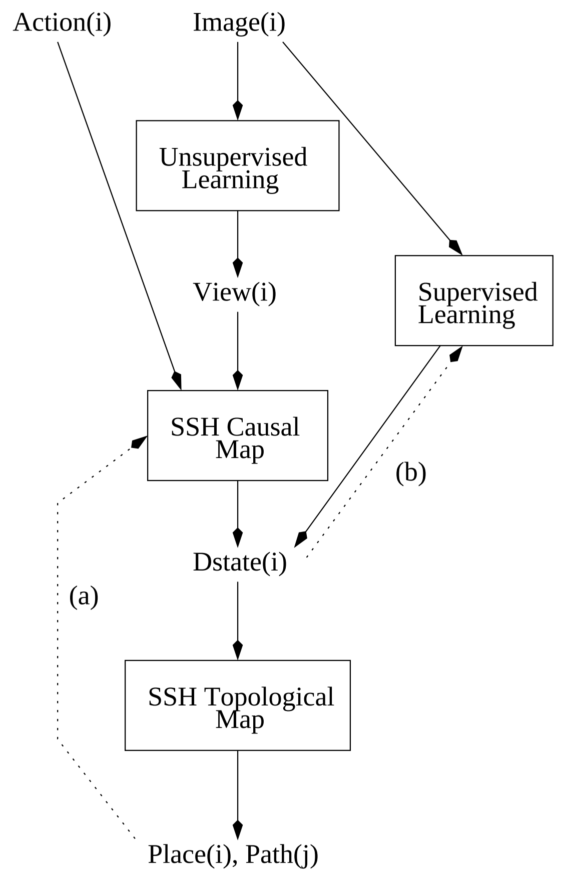
action A sequence of control laws taking the robot from one dstate to the next.

image A sensor snapshot at a distinctive state.

view Every dstate has a single view. A view can be thought of as a prototype image at a dstate.

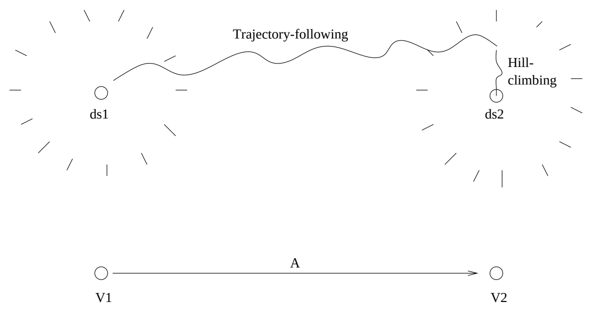
causal map A graph made of dstates (and their views) and the actions between them.

Bootstrap Learning for Place Recognition



1. Restrict attention to recognizing *distinctive states* (dstates). Distinctive states are well-separated in the robot's state space.
2. Apply an unsupervised clustering algorithm to the sensory images obtained at the dstates in the environment.
   - Reduces image variability by mapping different images of the same place into the same cluster, even at the cost of increasing perceptual aliasing.
   - We define each cluster to be a *view*.
3. Build the SSH causal and topological maps — symbolic descriptions made up of dstates, views, topological places, and paths — by exploration and abduction from the observed sequence of views and actions.
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.

Focusing on Distinctive States



- We use wall-following and hill-climbing control laws to arrive at dstates in an environment.
- We use the images from dstates to infer a causal map of views, dstates, and actions.
- By limiting global localization to dstates, we can simplify the Markov localization equation, used by many for robot localization (**?**).

$$p(x'|a,o,m)= \alpha\, p(o|x',m) \int p(x'|x,a,m)\, p(x|m)\, dx$$

  - Since actions are deterministic, if $\langle x,a,x'\rangle$, then $p(x'|x,a,m)$=1, otherwise 0.

$$p(x'|a,o,m) = \alpha\, p(o|x',m) \sum \{p(x|m) : \langle x,a,x'\rangle\}$$

    A sum over part of the topological graph is much more efficient than integration over an occupancy grid.
  - We cluster sensory images o into a small set of clusters called *views* v.

$$p(x'|a,v,m) = \alpha\, p(v|x',m) \sum \{p(x|m) : \langle x,a,x'\rangle\}$$

    Unlike $p(o|x,m)$, $p(v|x',m)$ is large enough to be meaningful.
  - For a given distinctive state x, there is a single view v such that, for every sensory image o observed at x, $o \in v$ (i.e., $p(v|x',m) \in \{0,1\}$).

$$p(x'|a,v,m) = \alpha \sum \{p(x|m) : \langle x,a,x'\rangle \wedge view(x',v)\}$$

- This Markov equation simplification clarifies the relation between our approach and previous ones.
  - When strong assumptions hold (deterministic actions, no image variability), logical inference determines the set of possible dstates x′.
  - If action determinism or image uniqueness fail, we can fall back to Markov localization, retaining the other simplifications. (See Future Work.)

Unsupervised Learning (Clustering)

- We use k-means to cluster images into views.
  - This ensures there is no *a priori* knowledge about the sensor configuration being added to the learning system.
- The robot selects the number of clusters k to maximize the *decision metric M*.

$$M=\frac{\min_{i\neq j}[min\{dist(x,y) : x \in c_i, y \in c_j\}]}{\max_i[max\{dist(x,y) : x,y \in c_i\}]}$$

- The researchers can verify that the *decision metric* works properly by using an *evaluation metric* that knows the correct state labels for each image. We use the *uncertainty coefficient* (**?**, pp. 632–635):

$$\begin{aligned} U(v|x) &= \frac{H(v)\text{-}H(v|x)}{H(v)} \\ H(v) &= -\sum_i p_{i*} \ln p_{i*} \text{ where } p_{i*}=\sum_j p_{i,j} \\ H(v|x) &= -\sum_{i,j} p_{i,j} \ln \frac{p_{i,j}}{p_{*j}} \text{ where } p_{*j}=\sum_i p_{i,j} \end{aligned}$$

- The robot chooses the maximum M. Ideally, this will correspond to the largest k where U=1. When U=1, image variability is gone (no images from the same location fall into different clusters).
- Clustering images into views eliminates image variability, but retains or increases perceptual aliasing:

$$\begin{aligned} view(x,v_1) \wedge view(x,v_2) &\rightarrow v_1=v_2 \\ view(x_1,v) \wedge view(x_2,v) &\not\rightarrow x_1=x_2 \end{aligned}$$

Map Building

- Determine the minimal set of distinctive states, topological places, and paths consistent with the observed sequence of views and actions and with the axioms for the SSH causal and topological maps.
- Remolina (2001) provides axioms and an algorithm for building the cognitive map.
- Associates each dstate with the correct image.

Supervised Learning

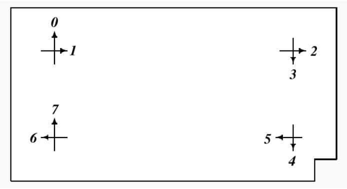
- Learn the dstate labels (given by the causal map) directly from the sensor snapshots, *images*, obtained at that dstate.
- We use nearest neighbor to do supervised learning.

Lassie

Experiments were performed with Lassie, an RWI Magellan Pro mobile robot, using a laser range finder scan for an image. The laser range finder returns 180 range points in a half-circle in front of the robot. Lassie does not have a compass.

A Simple Experiment

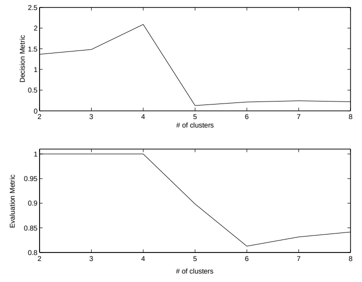
We began testing our method in the simplest environment with a distinguishing feature (the notch) small enough to be obscured by image variability.



Learning Results

After 50 clockwise cycles, 400 images.

Clustering

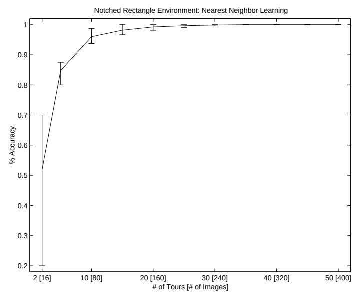


The *decision metric* $M(k)$ is maximal at k=4 views. The evaluation metric confirmed that $U(4)$=1, so all image variability is eliminated.

Map Building

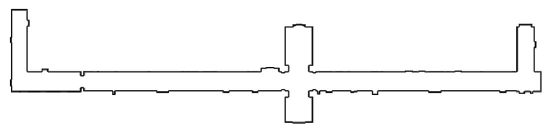
Non-monotonic reasoning found a map with 8 dstates, 4 topological places, and 4 paths.

Supervised Learning



Supervised learning from images to dstate labels reached 100% accuracy after 30 cycles around the rectangle. (10-fold cross validation was used).

A Natural Office Environment

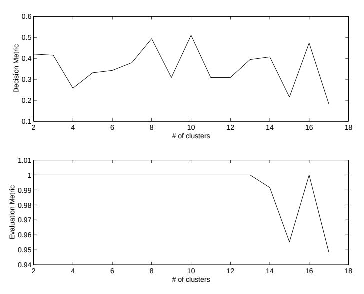


The robot explored Taylor Hall second floor.

Learning Results

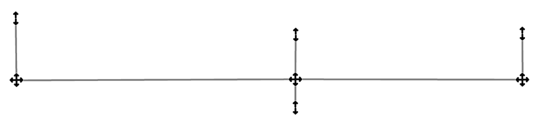
After 10 circuits, 240 images.

Clustering



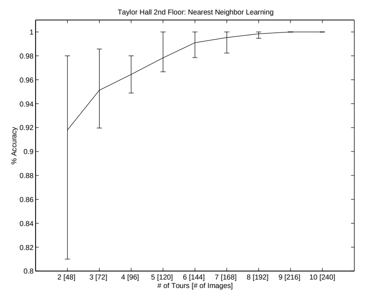
The *decision metric* $M(k)$ is maximal at k=10 views. The evaluation metric confirmed that $U(10)$=1, so all image variability is eliminated, but k=10 is not optimal, since $U(13)$=1 as well.

Causal/Topological Map Building



Using the views from clustering, non-monotonic reasoning determined there were 20 dstates, 7 topological places, and 4 paths in the environment.

Supervised Learning



When it reaches a dstate, the robot can recognize its location with 100 accuracy after only 9 training tours around the environment. (10-fold cross validation was used).

Discussion

- Confirmed the value of bootstrap learning.
  - Unsupervised clustering abstracts the world.
  - Deductive inference builds a correct model.
  - Supervised learning with accurate labels gives high performance from real inputs.

Future Work

- Eliminate need for physical hill-climbing
  - Robot will no longer return to same pose when entering a dstate neighborhood.
  - We will use *local metrical maps* as images and compare these to get views.
- Implement using vision sensors
  - Representation does not rely on range sensors.
  - cf. (**?**)
- Make learning occur online, incrementally
- Error recovery when reliable actions fail.
  - Fall back to Markov localization, temporarily.

References

Kuipers, B. J. 2000. The spatial semantic hierarchy. *Artificial Intelligence* 119:191–233.

Remolina, E. 2001. *Formalizing the Spatial Semantic Hierarchy*. Ph.D. Dissertation, University of Texas at Austin, Department of Computer Sciences.

UTCS Intelligent Robotics Lab
http://www.cs.utexas.edu/users/qr/robotics/