# Toward Bootstrap Learning for Place Recognition

**Benjamin Kuipers and Patrick Beeson**
Computer Science Department
The University of Texas at Austin
Austin, Texas 78712

## Abstract

We present a method whereby a robot with no prior knowledge of its sensors, effectors or environment 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). Previous work showed how such a robot could learn from its experience a useful set of sensory features, motion primitives, and local control laws to move from one distinctive state to another. Such progressive learning of a hierarchical representation is called *bootstrap learning*. The first step in learning place recognition eliminates image variability in two steps: (a) focusing on recognition of distinctive states defined by the robot's control laws, and (b) 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 cognitive 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 large amounts of perceptual aliasing and high resulting recognition rates.

## 1 Bootstrap Learning

Suppose an agent awakes in an unknown environment with an uninterpreted set of sensors and effectors. How can it learn the nature of its own sensorimotor system and then learn the structure of its environment?

This problem is important in practical terms because we want robots with very rich sensorimotor systems to be able to adapt to new senses or to changes in its existing sensors. Future robot sensors based on MEMS technology may also have irregular structures similar to biological sensors, rather than being (for example) a rectangular array of pixels. This problem is an aspect of the fundamental question of how symbols in a knowledge representation gain their meaning by being grounded in sensorimotor interaction with the world.

Pierce and Kuipers [1997] explored how such an agent could progressively learn: (1) properties of its sensors, (2) a basis set of motor commands, (3) sets of sensory features useful as local state variables, and (4) control laws for trajectory-following and hill-climbing. These control laws are sufficient to support travel among *distinctive states* (dstates), and hence to support creation of the causal, topological and metrical levels of the Spatial Semantic Hierarchy (SSH) [Kuipers & Byun, 1991; Kuipers, 2000]. We use the term *bootstrap learning* for this kind of progressive creation of a hierarchy of representations.

In this paper, given that the agent has learned to move reliably among distinctive states, we ask how it can learn to recognize places.

## 2 Place Recognition

We want a robot to learn from experience to recognize the place it is at and its orientation at that place. Together, the robot's position and orientation constitute its *state* in the environment. Without contextual information, this recognition problem is unsolvable even with perfect sensors, since different places may have identical sensory images. Realistically, sensors are imperfect, so even if subtle distinguishing features are present, they may be buried in sensor noise. There are two difficulties that must be overcome for effective 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, for example at different times of day.

These two difficulties trade off against each other. With relatively impoverished sensors (e.g., a sonar ring) many places have similar images, so the dominant problem is perceptual aliasing. With much richer sensors such as vision or laser range-finders, discriminating features are more likely to be present in the image, but so are noise and dynamic changes, so the dominant problem for recognition becomes image variability.

## 3 A Hybrid Solution

In order to bootstrap to an effective solution to the place recognition problem, we combine several different learning

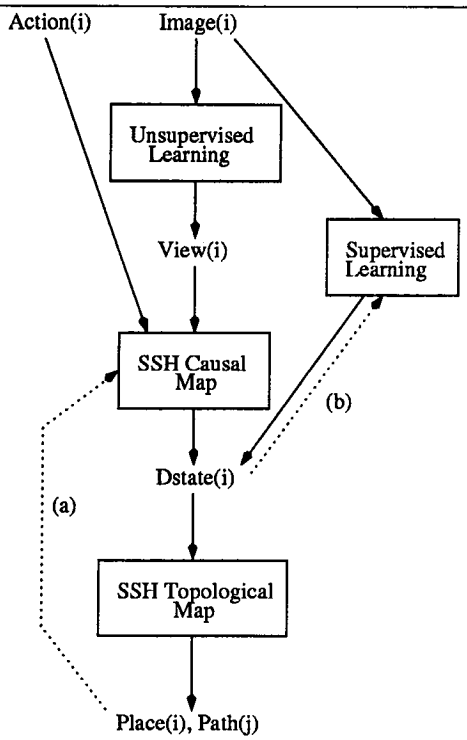


Figure 1: Bootstrap learning of place recognition. Solid arrows represent the major inference paths, while dotted arrows represent feedback.

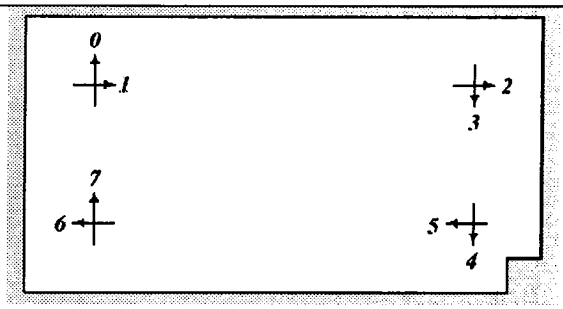


Figure 2: Lassie explores a rectangular room whose only distinguishing feature is a small notch out of one corner. Image variability arises from position and orientation variation when Lassie reaches a distinctive state, and from the intrinsic noise in the laser range-finder. Perceptual aliasing arises from the symmetry of the environment, and the lack of a compass. The notch is designed to be a distinguishing feature that is small enough to be obscured by image variability.

methods (Figure 1). We start by attacking the problem of image variability, first by focusing on distinctive states and second by clustering to eliminate noise. Then we apply a relatively expensive deductive and exploration method to disambiguate perceptual aliasing, some of which may be caused by the first step. And finally, we use our expensively-bought knowledge of distinctive states to learn direct associations from sensory images to dstates. The resulting association may not be perfect, in the case of dstates with identical sensory images, but the more expensive contextual methods remain available.

The steps of our solution are the following.

1. Restrict attention to recognizing *distinctive states* (dstates). Distinctive states are well-separated in the robot's state space, and their sensory images tend to be either well-separated or very similar.
2. Apply an unsupervised clustering algorithm to the sensory images obtained at the dstates in the environment. This reduces image variability by mapping different images of the same place into the same cluster, even at the cost of increasing perceptual aliasing by mapping images of different states into the same cluster. We define each cluster to be a *view*, in the sense of the SSH [Kuipers & Byun, 1991; Kuipers, 2000].
3. Build the SSH causal and topological map — a symbolic description made up of dstates, places and paths — by exploration and abduction from the observed sequence of views and actions [Kuipers, 2000; Remolina & Kuipers, 2001]. If needed to disambiguate the current distinctive state, use history-based methods such as the rehearsal procedure [Kuipers & Byun, 1991] or homing sequences [Rivest & Schapire, 1989]. While theoretically tractable, these are expensive both in computation and in additional travel. This is feedback path (a) in Figure 1.
4. Now the sequence of images is classified into views and labeled with the correct dstate. Apply a supervised learning algorithm to learn a direct association from sensory image to dstate. The added information in supervised learning makes it possible to identify subtle discriminating features that would be indistinguishable from noise by an unsupervised clustering algorithm. This is feedback path (b) in Figure 1.

We describe the individual steps in more detail along with a simple experiment. Lassie is an RWI Magellan robot. It perceives its environment using a laser range-finder: each sensory image is a point in $R^{180}$. Although each image represents the ranges to obstacles in the 180° arc in front of Lassie, because we are doing bootstrap learning the robot does not have this knowledge, so it cannot apply powerful spatial modeling methods like occupancy grids [Moravec, 1988].

As Lassie performs clockwise circuits of its environment, it encounters eight distinctive states, one immediately before, and one immediately after the turn at each corner (Figure 2).

## 4 Focus on Distinctive States

The SSH [Kuipers, 2000] is a hierarchy of distinct but closely related representations for knowledge of large-scale space. It shows how the cognitive map can be robustly acquired during exploration and used for problem-solving even in the face of resource limitations and incomplete knowledge.

A distinctive state is the isolated fixed-point of a hill-climbing control law. Travel among distinctive states elim-

inates cumulative estimated position error. A sequence of control laws taking the robot from one dstate to the next is abstracted to an *action*.

For a typical mobile robot, the state variable $\mathbf{x} = [x, y, \theta]$ is three-dimensional with components for position and orientation, and a two-dimensional motor vector $\mathbf{u} = [v, \omega]$ specifies linear and angular velocities. In contrast, a realistic robot of the present and future will have a very high-dimensional sense vector $\mathbf{s}$, including such sensors as binocular vision, laser, sonar and IR range-finders, bump sensors, odometry, compass and GPS.

In a qualitatively uniform segment of the environment, the robot governs its behavior by selecting a reactive control law $\chi_i$. The robot and its environment, coupled through the sensorimotor system, are described as a dynamical system which evolves to a fixed-point $\mathbf{x}$ (the distinctive state) where $\dot{\mathbf{x}} = 0$.

$$\dot{\mathbf{x}} = \Phi(\mathbf{x}, \mathbf{u}) \quad (1)$$

$$\mathbf{s} = \Psi(\mathbf{x}) \quad (2)$$

$$\mathbf{u} = \chi_i(\mathbf{s}) \quad (3)$$

$\Phi$ represents the physics of the robot and the constraints of the environment. $\Psi$ represents the sensory system of the robot. $\chi_i$ is the reactive control law for the current segment of the robot's behavior. We define an *image* as being the high-dimensional sensory input $I = \mathbf{s} = \Psi(\mathbf{x})$ when $\mathbf{x}$ is a distinctive state.

Any implementation of the dynamical system (1-3) will have finite tolerances, so the values of $\mathbf{x}$ when $\dot{\mathbf{x}} = 0$ on different occasions will not be precisely equal. However, they will be clustered very closely, and they will be separated from other distinctive states by the basin of attraction of the system (1-3). The same will be true of the dependent variable $\mathbf{s} = \Psi(\mathbf{x})$.[1]

## 5 Cluster Images Into Views

An environment contains a relatively small discrete set of dstates. In a high-dimensional sensory space, their sensory images are likely to be well-separated, so a clustering algorithm can eliminate amounts of variation that are small compared with the separation between groups of images. When images from different dstates happen to be very close (due to highly structured environments or weak sensors), they can simply be mapped to the same cluster, resulting in perceptual aliasing.

Distinctive states are significantly easier to recognize than places selected at regularly spaced intervals in the environment [Yamauchi & Langley, 1997; Duckett & Nehmzow, 2000]. Regularly spaced states are unlikely to be as well separated in sensory space so it will be difficult to eliminate all image variability by clustering without incurring large amounts of perceptual aliasing.

The SSH assumes that each dstate is associated with a single view, though different dstates may have the same view, so clustering must eliminate image variability.

[1]Note that $\Psi$ cannot behave pathologically in the neighborhood of a dstate $\mathbf{x}$, or the dynamical system would not converge stably to $\mathbf{x}$, and it couldn't be a distinctive state.

We cluster images into $k$ clusters using $k$-means [Duda, Hart, & Stork, 2001], searching for the value of $k$ that maximizes our "internal measure" of clustering quality,

$$M = \left[ k \sum_{c=1}^{k} \sum_{i=1}^{n_c} |x_{c,i} - \bar{x_c}|^2 \right]^{-1} \quad (4)$$

where $n_c$ is the number of elements and $\bar{x_c}$ is the mean of cluster $c$, and $x_{c,i}$ is the $i$th element of cluster $c$. $M$ rewards tight clusters but penalizes larger numbers of clusters.

An "external measure" of cluster quality uses knowledge of the true dstate associated with each image. The *uncertainty coefficient* $U(V|S)$ measures the extent to which knowledge of $S$ predicts the view $V$ [Press *et al.*, 1992, pp. 632–635]. ($p_{i,j}$ is the probability that the current view is $V_i$ and the current dstate is $S_j$.) The largest value of $k$ with $U = 1$ corresponds to the greatest discriminating power while completely eliminating image variability.

$$U(V|S) = \frac{H(V) - H(V|S)}{H(V)}$$

$$H(V) = -\sum_i p_{i\cdot} \ln p_{i\cdot} \text{ where } p_{i\cdot} = \sum_j p_{i,j}$$

$$H(V|S) = -\sum_{i,j} p_{i,j} \ln \frac{p_{i,j}}{p_{\cdot j}} \text{ where } p_{\cdot j} = \sum_i p_{i,j}$$

In 50 circuits of the notched rectangle environment (Figure 2), Lassie experiences 400 images. Applying the internal measure (4) of cluster quality, Lassie determines that $k = 4$ is the clear winner (Figure 3(a)). Figure 3(b) shows that $k = 4$ is also optimal to the external measure.

The notch in the rectangle is clearly being treated as noise by the clustering algorithm, so diagonally opposite dstates have the same view. In this environment, the four views correspond to the following table of distances perceived to the robot's left, front, and right.

| | left | front | right |
|---|---|---|---|
| $V_0$ | 0.5 | 0.5 | 4.5 |
| $V_1$ | 0.5 | 4.5 | 2.5 |
| $V_2$ | 0.5 | 0.5 | 2.5 |
| $V_3$ | 0.5 | 2.5 | 4.5 |

## 6 Build the Causal and Topological Maps

As the robot travels among distinctive states, its continuous experience is abstracted, first to an alternating sequence of images $I_k$ and actions $A_k$, then images are clustered into views $V_k$, and finally views are associated with dstates $S_k$. The SSH Causal map can be represented as a simple tableau.

| | | | |
|---|---|---|---|
| $t_0$ | $I_0$ | $V_0$ | $S_0$ |
| | $A_0$ | | |
| $t_1$ | $I_1$ | $V_1$ | $S_1$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| | $A_{n-1}$ | | |
| $t_n$ | $I_n$ | $V_n$ | $S_n$ |

(5)

Since clustering images into views has eliminated image variability, but leaves perceptual aliasing, the problem is to

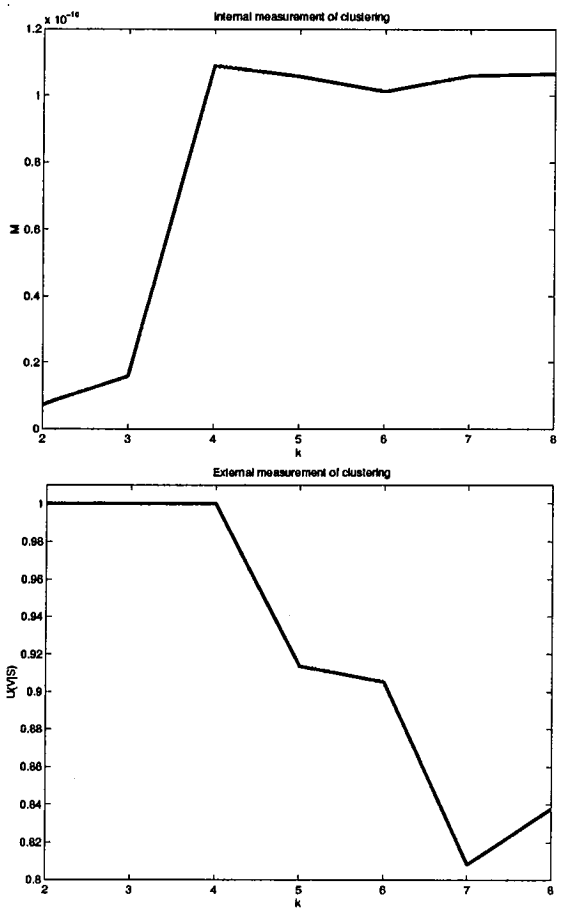


Figure 3: After Lassie's exploration of the notched rectangle, $k = 4$ is selected as the best number of clusters by the internal measure $M$ (top), and is confirmed as optimal by the external measure $U$ (bottom).

determine the correct distinctive states $S_k$. We resolve this ambiguity using the "rehearsal procedure" [Kuipers & Byun, 1991], or methods for learning deterministic finite automata (DFA) [Rivest & Schapire, 1989], even in the face of stochastic uncertainty [Basye, Dean, & Kaelbling, 1995]. However, by focusing on distinctive states and clustering to eliminate image variability, we believe that the remaining uncertainty is not stochastic. Perceptual aliasing is simply different states of the DFA having the same output (i.e., view).

The basic idea for identifying distinctive states is to assume that two dstates with the same view are equal, unless they can be proved to be different. They can be proved different with experiences stored in the SSH causal level when the two dstates are directly linked by a non-null action. This can also be done when the SSH topological level includes a relation incompatible with dstate identity, for example when the two dstates are associated with different places, or when they must lie on different sides of a path serving as a dividing boundary. In case discriminating information is not already in the cognitive map, the rehearsal procedure proposes an exploratory sequence of actions likely to produce the relevant experience.

At the SSH topological level, actions are classified as *turns* and *travels*. Sets of distinctive states that are connected by turns without travel define *places*, and sets of dstates that are connected by travels without turns (except TurnAround) define *paths*. The SSH causal and topological levels describe the environment at the discrete set of distinctive states, when the robot is at a dstate $S$, a place $P$, and on a directed path $(Pa, dir)$, where $dir \in \{pos, neg\}$ is the one-dimensional orientation along the path.

As Lassie explores the notched-rectangle environment, it creates the following tableau of experience at the SSH causal and topological levels. Note that the sequences of time-points $t_k$, images $I_k$ and actions $A_k$ are not periodic. The sequence of views $V_k$ has period four, but the sequence of distinctive states $S_k$ has period eight.

$$
\begin{array}{c|cccccc}
t_0 & I_0 & V_0 & S_0 & P_0 & Pa_0 & pos \\
 & A_0 & & (turn\ -90°) & & & \\
t_1 & I_1 & V_1 & S_1 & P_0 & Pa_1 & pos \\
 & A_1 & & (travel\ 4m) & & & \\
t_2 & I_2 & V_2 & S_2 & P_1 & Pa_1 & pos \\
 & A_2 & & (turn\ -90°) & & & \\
t_3 & I_3 & V_3 & S_3 & P_1 & Pa_2 & pos \\
 & A_4 & & (travel\ 2m) & & & \\
t_4 & I_4 & V_0 & S_4 & P_2 & Pa_2 & pos \\
 & A_5 & & (turn\ -90°) & & & \\
t_5 & I_5 & V_1 & S_5 & P_2 & Pa_3 & pos \\
 & A_6 & & (travel\ 4m) & & & \\
t_6 & I_6 & V_2 & S_6 & P_3 & Pa_3 & pos \\
 & A_7 & & (turn\ -90°) & & & \\
t_7 & I_7 & V_3 & S_7 & P_3 & Pa_0 & pos \\
 & A_8 & & (travel\ 2m) & & & \\
t_8 & I_8 & V_0 & S_0 & P_0 & Pa_0 & pos \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots
\end{array}
\qquad (6)
$$

$$
\begin{array}{lcl}
t_0 & \rightarrow & on(P_0, Pa_0) \\
t_1 & \rightarrow & on(P_0, Pa_1) \\
t_2 & \rightarrow & on(P_1, Pa_1) \wedge PO(Pa_1, pos, P_0, P_1) \\
\vdots & \vdots & \vdots
\end{array}
$$

The connectivity of the graph of places and paths is derived from the tableau above. $(on(P_0, Pa_0)$ means that place $P_0$ is on path $Pa_0$, and $PO(Pa_1, pos, P_0, P_1)$ means that in the place order associated with path $Pa_1$ in the *pos* direction, place $P_0$ precedes $P_1$.) The detailed rules are beyond the scope of this paper. The concepts are described in [Kuipers, 2000] and the formal axioms are given in [Remolina & Kuipers, 2001]. Roughly, we conclude that $S_4 \neq S_0$ because $S_4$ is at $P_2$, which is to the right of boundary $Pa_0$, while $S_0$ is at $P_0$, which is on $Pa_0$. Similarly for $S_5$, $S_6$ and $S_7$. We conclude that $S_8 = S_0$ by a minimality argument, since there is no necessity for them to be different.

Thus, by constructing the SSH causal and topological maps, Lassie determines that the four views correspond to eight distinctive states, four places and four paths.

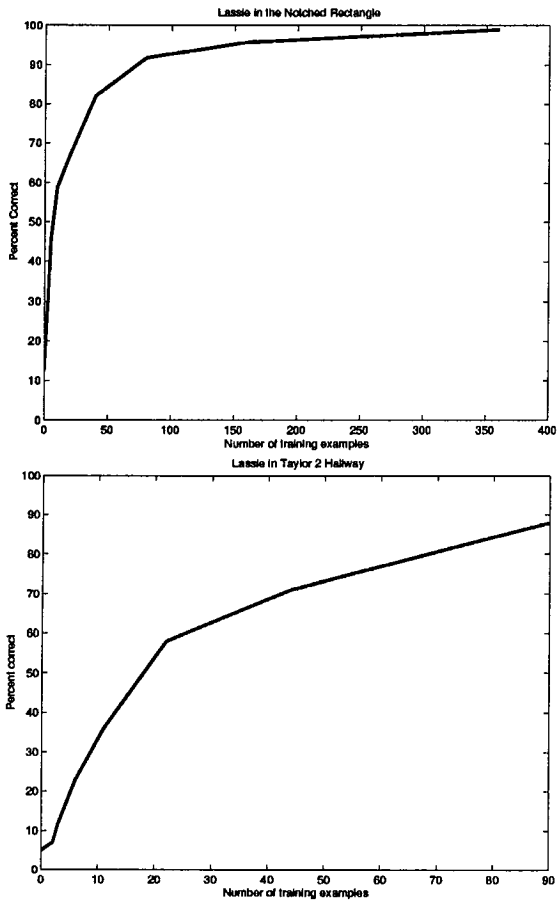


Figure 4: Learning curve (using 10-fold cross validation) for nearest neighbor classification of dstates given sensory images for (a) the notched-rectangle environment, or (b) second floor of Taylor Hall. In both cases, recognition is approximately 90% after 90 training examples.

## 7 Supervised Learning to Recognize Dstates

With unique identifiers for dstates, the supervised learning step quickly learns to identify the correct distinctive state directly from the sensory image with high accuracy. The supervised learning method is the nearest neighbor algorithm [Duda, Hart, & Stork, 2001]. Experienced images are represented in the sensory space, labeled with their true dstate. When a new image is experienced, the dstate label on the nearest stored image in the sensory space is proposed, and the accuracy of this guess is recorded. Then the image is stored with its correct dstate label. Figure 4 shows the rate of correct answers as a function of number of images experienced. In both cases, accuracy rises near 90% at about 90 images.

In general, of course, recognition of dstates from sensory images cannot be done perfectly, since there can be different dstates whose distinguishing features, if present at all, cannot be discerned by the robot's sensors. In such cases, the robot can fall back on the historical context of its travel, or on further exploration.

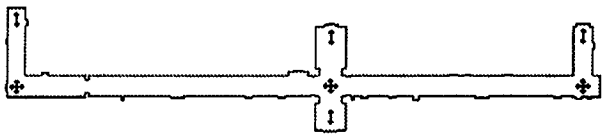

Figure 5: Taylor Hall, second floor hallway. The actual environment is 80 meters long and includes trash cans, lockers, benches, desks and a portable blackboard.

## 8 A Natural Office Environment

A natural environment, even an office environment, contains much more detail than the simplified notched-rectangle environment. To a robot with rich sensors, images at distinctive states are much more distinguishable. Image variability is the problem, not perceptual aliasing.

Lassie explored the main hallway on the second floor of Taylor Hall (Figure 5). It collected 100 images from 20 distinctive states. The topological map linking them contained seven places and four paths. When clustering the images, the internal measure $M$ had its maximum at $k = 8$. The external measure $U$ confirms that this is optimal (Figure 6). By building the causal and topological map the robot is able to disambiguate all twenty distinctive states, even though there are only eight views. Given the correct labeling with dstates, the supervised learner reaches 88% accuracy within 90 trials (Figure 4(b)).

## 9 Conclusion and Future Work

We have established that bootstrap learning for place recognition can achieve high accuracy with real sensory images from a physical robot exploring among distinctive states in real environments. The method starts by eliminating image variability by focusing on distinctive states and doing unsupervised clustering of images. Then, by building the causal and topological map, distinctive states are disambiguated and perceptual aliasing is eliminated. Finally, supervised learning of labeled images achieves high accuracy direct recognition of distinctive states from sensory images.

The current unsupervised and supervised learning algorithms we use are $k$-means and nearest neighbor. We plan to experiment with other algorithms to fill these roles in the learning method. Other clustering techniques may be more sensitive to the kinds of similarities and distinctions present in sensor images. Supervised learning methods like backprop may make it possible to analyze hidden units to determine which features are significant to the discrimination and which are noise. Using methods like these, it may be possible to identify and explain certain aspects of image variability, for example the effect of time of day on visual image illumination.

## References

[Basye, Dean, & Kaelbling, 1995] Basye, K.; Dean, T.; and Kaelbling, L. P. 1995. Learning dynamics: system iden-

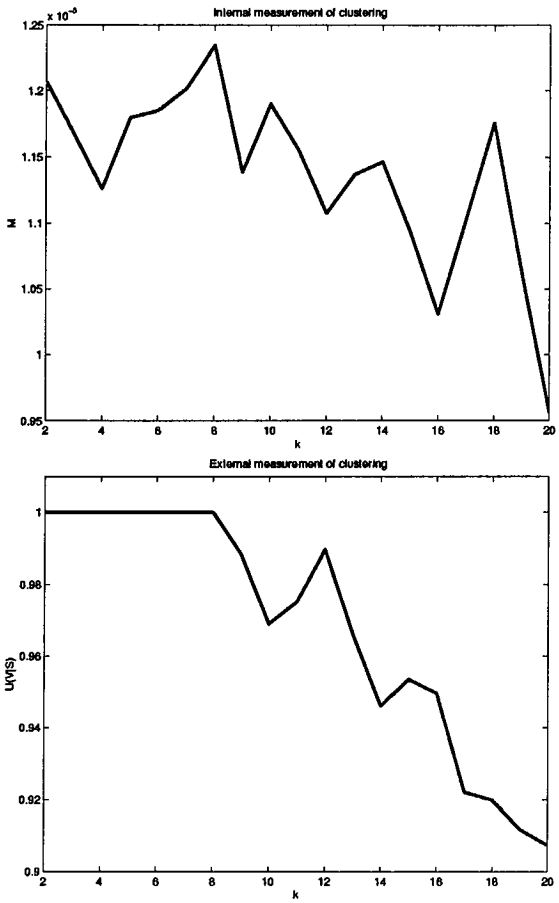


Figure 6: After Lassie's exploration of the Taylor hallway, $k = 8$ is selected as the best number of clusters by the internal measure $M$ (top), and is confirmed as optimal by the external measure $U$ (bottom).

tification for perceptually challenged agents. *Artificial Intelligence* 72:139–171.

[Duckett & Nehmzow, 2000] Duckett, T., and Nehmzow, U. 2000. Performance comparison of landmark recognition systems for navigating mobile robots. In *Proc. 17th National Conf. on Artificial Intelligence (AAAI-2000)*, 826–831. AAAI Press/The MIT Press.

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

[Kuipers & Byun, 1991] Kuipers, B. J., and Byun, Y.-T. 1991. A robot exploration and mapping strategy based on a semantic hierarchy of spatial representations. *Journal of Robotics and Autonomous Systems* 8:47–63.

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

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

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

[Press *et al.*, 1992] Press, W. H.; Teukolsky, S. A.; Vitterling, W. T.; and Flannery, B. P. 1992. *Numerical Recipes in C: The Art of Scientific Computing.* Cambridge University Press, second edition.

[Remolina & Kuipers, 2001] Remolina, E., and Kuipers, B. 2001. A logical account of causal and topological maps. In *Proc. 17th Int. Joint Conf. on Artificial Intelligence (IJCAI-01)*, 5–11. San Mateo, CA: Morgan Kaufmann.

[Rivest & Schapire, 1989] Rivest, R. L., and Schapire, R. E. 1989. Inference of finite automata using homing sequences. In *Proceedings of the 21st Annual ACM Symposium on Theoretical Computing*, 411–420. ACM.

[Yamauchi & Langley, 1997] Yamauchi, B., and Langley, P. 1997. Place recognition in dynamic environments. *Journal of Robotic Systems* 14:107–120.