A Framework for Robust Cognitive Spatial Mapping

Andrzej Pronobis, Kristoffer Sjöö, Alper Aydemir, Adrian N. Bishop and Patric Jensfelt

I. Introduction

Spatial knowledge constitutes a fundamental component of the knowledge base of a cognitive, mobile agent. The typical spatial knowledge representations are purely metrical and rely on information extracted from simple, but accurate metric sensors. However, in large-scale, dynamic environments, metrical global maps become harder to control and observe. The agent should be able to exploit sensory information that might be complex and non-metric, yet reflect crucial aspects of the environment [1]–[3], [5]. Moreover, it is not clear that the level of detail offered by metric maps is necessary, or even desirable, when the agent is a cognitive system intended to interact with the world in a human-like way [6].

This work introduces a rigorously defined framework for building an abstracted cognitive spatial map that permits high level reasoning about space along with robust navigation and localization while maintaining a description that permits formal proofs and derivations. Although the literature contains many algorithms for spatial mapping [1], [2], there is little work on the formal analysis of their fundamental requirements and properties. The idea of this work, is to take a step back and see how a rigorous formal treatment can lead the way towards a powerful spatial representation.

II. Overview of the Framework

Our framework is built around the assumption that the role of a cognitive map is not to represent the world as accurately as possible, but rather to allow the agent to act in an environment despite uncertainty and dynamic variations. Such a map does not need to provide perfect global consistency as long as the local spatial relations are preserved with sufficient accuracy. In our framework, the map is represented as a collection of basic spatial entities called *places*.

A *place* is defined by a subset of values of arbitrary, possibly complex, distinctive *features* and *spatial relations* reflecting the structure of the environment. Consider a set $\{f_i\}_{i=1}^{n_f}$ of features $f_i\left(\boldsymbol{x},t\right) : \mathcal{C} \times \mathbb{R} \rightarrow \mathcal{F}_i \in \mathbb{R}^n$, where $\mathcal{C}$ represents the *configuration space* of the agent, $t \in \mathbb{R}$ represents time, and $\mathcal{F}_i$ is the range of values of the feature f_i. Together, the $\mathcal{F}_i$'s give rise to the definition of the *feature space* $\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2 \times \ldots \times \mathcal{F}_{n_f}$, in which each tuple $(\zeta_1, \ldots, \zeta_{n_f}) \in \mathcal{F}$ of the feature values $\zeta_i = f_i(\boldsymbol{x}, t)$ corresponds to a single point. The features provide information about the world as it can be perceived by an agent when at a place and can range from simple metric features (e.g. Euclidean distances to landmarks) to complex appearance-based descriptors typically used in visual topological localization (e.g. the "gist" of a scene [5], SIFT [3] or SURF [1]).

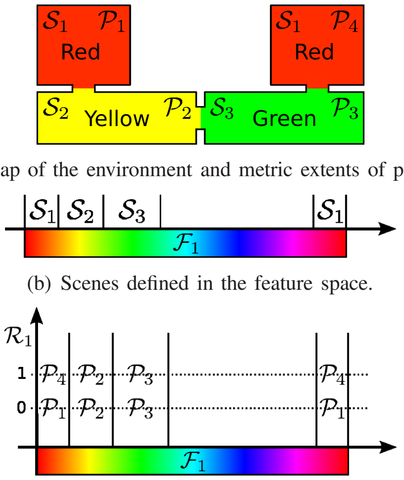
Additionally, we introduce the concept of a *scene* which facilitates generation of places and serves as a basis for defining spatial relations by providing groupings of similar feature values. Consider a set $\{\mathcal{S}_i\}_{i=1}^{n_s}$ of *scenes* $\mathcal{S}_i = \left\{(\zeta_1, \ldots, \zeta_{n_f})\right\} \subseteq \mathcal{F}$, such that each scene is a non-empty and non-overlapping collection of tuples of features that could be perceived by the robot. Intuitively, the definition of scenes leads to a division of metric space into regions based on properties such as appearance. However, this segmentation may not reflect the spatial relationships in the world that constitute a large portion of the spatial knowledge. For instance, two distant disconnected regions could share similar properties (see e.g. Figure 1(a)). Additional power to distinguish between such regions can be attained using knowledge about spatially neighboring regions.

Consider a set $\{r_i\}_{i=1}^{n_r}$ of *spatial relations* $r_i\left(\boldsymbol{x},t\right) : \mathcal{C} \times \mathbb{R} \rightarrow \mathcal{R}_i \in \mathbb{R}^m$, where $\mathcal{R}_i$ is the range of values of the relation r_i. Each *spatial relation* r_i is defined with respect to the set of scenes $\{\mathcal{S}_i\}_i$ and describes the spatial relation of the point $\boldsymbol{x}$ in the configuration space $\mathcal{C}$ at time t to some or all of those scenes (e.g. adjacency to a scene). In many cases, the values of relations can be estimated in practice by performing a dynamic action in the environment (e.g. the agent moving between points in configuration space that correspond to different scenes).

Analogously to the feature space, we introduce the *place descriptor space* $\mathcal{D} = \mathcal{F} \times \mathcal{R}_1 \times \mathcal{R}_2 \times \ldots \times \mathcal{R}_{n_r}$, in which each tuple $D = (\zeta_1, \ldots, \zeta_{n_f}, \rho_1, \ldots, \rho_{n_r}) \in \mathcal{D}$ of the feature values and relation values $\rho_i = r_i(\boldsymbol{x}, t)$ corresponds to a single point. The *place map* can now be defined as a set $\mathcal{M} = \left\{\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_{n_p}\right\}$ of *places* $\mathcal{P}_i = \{D\} \subseteq \mathcal{D}$, such that $\forall_i \mathcal{P}_i \neq \emptyset$ and $\forall_{i \neq j} \mathcal{P}_i \cap \mathcal{P}_j = \emptyset$. In other words, similarly to scenes, places are groups of values of features; however, they encompass additional knowledge about the structure of the world encoded in the values of relations. In this sense, the places build on the perception of the agent and are based on its perceptual capabilities. Details on the properties of places that can be derived from the above definitions and segmentations of the descriptor space in terms of real-valued functions which encode the degree of belonging to the different places can be found in [4].

Consider a simple example of a small environment presented in Figure 1(a) consisting of 4 rooms characterized by the color of the floor. We define a single feature $f_1(\boldsymbol{x}, t) : \mathcal{C} \times \mathbb{R} \rightarrow \mathcal{F}_1$ that corresponds to the hue of the floor color at

The authors are with the Centre for Autonomous Systems at the Royal Institute of Technology (KTH), Stockholm, Sweden. This work was supported by the Centre for Autonomous Systems (CAS) and the EU FP7 project CogX and the Swedish Research Council, contract 621-2006-4520 (K. Sjöö) and 2005-3600-Complex (A. Pronobis).


$\mathcal{S}_1$ $\mathcal{P}_1$
Red
$\mathcal{S}_1$ $\mathcal{P}_4$
Red
$\mathcal{S}_2$ Yellow $\mathcal{P}_2$
$\mathcal{S}_3$ Green $\mathcal{P}_3$
$\mathcal{S}_1$ $\mathcal{S}_2$ $\mathcal{S}_3$ $\mathcal{S}_1$
$\mathcal{F}_1$
$\mathcal{R}_1$
1
0
$\mathcal{P}_4$ $\mathcal{P}_2$ $\mathcal{P}_3$ $\mathcal{P}_4$
$\mathcal{P}_1$ $\mathcal{P}_2$ $\mathcal{P}_3$ $\mathcal{P}_1$
$\mathcal{F}_1$

(a) Map of the environment and metric extents of places.

(b) Scenes defined in the feature space.

(c) Places defined in the descriptor space.

Fig. 1. Illustrative example of an environment consisting of 4 rooms characterized by the color of the floor and definitions of scenes and places in the feature and descriptor space.

the location $\boldsymbol{x} \in \mathcal{C} = \mathbb{R}^2$. Then, the feature space is simply defined by the range of the hue values e.g. $\mathcal{F}_1 = [0, 255]$. If we divide the feature space into regions as presented in Figure 1(b), we can differentiate between three scenes $\mathcal{S}_1$-$\mathcal{S}_3$. The scene $\mathcal{S}_1$ corresponds to two separate rooms which could be distinguished if we consider their relations to other scenes. Let us define an adjacency relation with respect to the scene $\mathcal{S}_3$, $r_1(\boldsymbol{x}, t) : \mathcal{C} \times \mathbb{R} \to \mathcal{R}_1 = \{1, 0\}$ that takes the value of 1 if the point $\boldsymbol{x}$ is adjacent to the scene $\mathcal{S}_3$ and create the place descriptor space $\mathcal{D} = \mathcal{F}_1 \times \mathcal{R}_1$. In that space, we can create four non-overlapping places $\mathcal{P}_1$-$\mathcal{P}_4$ by dividing the scene $\mathcal{S}_1$ into two places, one of which is adjacent to the scene $\mathcal{S}_3$ and the other is not. This division is reflected in the clustering of the descriptor space presented in Figure 1(c).

III. Localization and Navigation

According to the definition of places, given the true values of place descriptors D_t (features and spatial relations) obtained at time t for location $\boldsymbol{x}(t)$, the place to which that $\boldsymbol{x}$ corresponds is uniquely identified and given by $L_t \triangleq i : D_t \in \mathcal{P}_i$. However, in the real world an agent is moving through space and needs to maintain its localization in the face of uncertainty. Let us denote the observation of all descriptors at time t as $\hat{D}_t = D_t + e_t$, where e is an error. We view the agent's progress from place to place as a Markov process with L_t the state at (discrete) time t and $\hat{D}_t$ the measurement. Localization is then carried out iteratively according to the following formula:

$$\begin{aligned} p(L_t \mid \{\hat{D}\}_t, \{\alpha\}_{t-1}) &= \sum_{L_{t-1}} p(L_t \mid L_{t-1}, \hat{D}_t, \alpha_{t-1}) \\ &\times \; p(L_{t-1} \mid \{\hat{D}\}_{t-1}, \{\alpha\}_{t-2}) \end{aligned}$$

where $\{\hat{D}\}_t$ represents all measurements up until time t, and α_t is the action taken at time t. As described below, this corresponds to the path between two places that the robot follows. The probability update is given by:

$$p(L_t \mid L_{t-1}, \hat{D}_t, \alpha_{t-1}) = \gamma \cdot p(\hat{D}_t \mid L_t) p(L_t \mid L_{t-1}, \alpha_{t-1}),$$

where the factors represent respectively a normalization constant, the measurement integration step, and the prediction step of the localization update. Details on each of the steps can be found in [4].

A map must, besides allowing for localization, provide a means for navigating through it. We do this in terms of *paths*, which represent the movement from one place to another. Just as places are defined by descriptors, so each path is associated with a single *path precept* being a mapping from low-level sensory inputs available to the agent to a low-level control output: $\pi_i : \mathcal{S} \mapsto \mathcal{O}$. $\mathcal{S}$ is given by the system instantiation, and may include virtual sensor modalities, such as local metric maps built over a period of time. It is in general a richer representation than the feature space $\mathcal{F}$, and allows for low-level considerations such as obstacle avoidance and other reactive behaviours. At the same time, as described in detail in [4], the form of the precept can naturally arise from the descriptors that define places.

IV. Summary

The structure of the framework represent a certain view on a cognitive map. First, the map is defined in terms of the agent's perception of space and adapts to its perceptual capabilities. Second, the perceived features can be abstract and non-metric and describe for instance visual properties of the world. In this sense, the map is subjective and robocentric as the robot's observations do not have to be expressed in terms of any objectively defined quantities or any global coordinate system. The map is discretized (consists of a set of independent places), topological and does not require maintaining global spatial consistency. Despite that, the framework can accommodate a very wide range of different methods for abstracting space into places as specific cases. For instance, the topological map constituting a part of the Multi-Layered Conceptual Spatial Representation presented in [6] can easily be expressed in terms of our framework given properly chosen features.

The framework is designed so that a robot can build from the bottom-up a cognitive map of the environment which follows certain cognitive principles. As shown in practical demonstrations (see e.g. [1], [2]), such principles can lead to better performance in localization, navigation and loop-closing for robots moving in large-scale environments.

References

[1] M. Cummins and P. Newman. FAB-MAP: Probabilistic localization and mapping in the space of appearance. *IJRR*, 27(6), 2008.

[2] M. Milford and G. Wyeth. Mapping a suburb with a single camera using a biologically inspired SLAM system. *IEEE Tr. Rob.*, 24(5), 2008.

[3] A. Pronobis, O. Martinez Mozos, and B. Caputo. SVM-based discriminative accumulation scheme for place recognition. In *Proc. of ICRA'08*.

[4] A. Pronobis, K. Sjöö, A. Aydemir, A. Bishop, and P. Jensfelt. A framework for robust cognitive spatial mapping. In *Proc. of ICAR'09*.

[5] A. Torralba, K. Murphy, W. Freeman, and M. Rubin. Context-based vision system for place and object recognition. In *Proc. of ICCV'03*.

[6] H. Zender, O. Mozos, P. Jensfelt, G. Kruijff, and W. Burgard. Conceptual spatial representations for indoor mobile robots. *RAS*, 56(6), 2008.