A Visual Sense of Space

Divyanshu Bhartiya[1] and Amitabha Mukerjee[2]

[1] IIT Kanpur
divbhar@iitk.ac.in
[2] IIT Kanpur
amit@iitk.ac.in

Abstract
Animals effortlessly acquire a visual model for familiar spaces, enabling them to learn to use their own body, find paths and interact with objects / others. On the other hand, for robots, all poses must be calibrated against a global reference frame, and even tasks driven by vision require state estimation onto these canonical coordinates. In this paper, we propose to the idea of *Visual Generalized Coordinates*, which are a set of d parameters that describe the motion of a d degree-of-freedom system. Here we show that images captured by a camera mounted on the moving system will lie on a low-dimensional manifold homeomorphic to its motor manifold. The parametrization of such a manifold is equivalent to the traditional coordinates used in robotics, except that these can be obtained from sensory data. This provides a mechanism for explaining how cognitive systems build the allocentric map for the space around it. We demonstrate this process for a simulated robot exploring a planar space. Exploring the space with a suitable sampling strategy and image space similarity measure can be used to generate a manifold based on the similarity of images captured from nearby viewpoints, without any knowledge of the motion coordinates. We show how such a model generates structures very similar to place cells and orientation cells in mammals, and also how it can be applied for more visual approaches to robot tasks.

Keywords: Visual Manifold, Place cells, Cognitive Map

1 Introduction

Humans and animals localize themselves in familiar spaces based on visual features, the cognitive models for which are thought to be organized in two frameworks - based on the body (*egocentric*) and based on the environment (*allocentric*). Allocentric models, relevant to navigation tasks, implemented in the limbic brain via specialized classes of neurons (e.g. "place cells" and "orientation cells"), constitute a form of cognitive map [1, 2]. However, how sensory stimuli, particularly vision, is used to encode environmental features onto such structures remains unclear.

Localization in robotics differs from cognitive models in that robot positioning is based on canonical coordinates (e.g. x, y, θ, or joint angles) defined in a pre-specified frame of reference. The number of such coordinates equals the degrees of freedom of the system. For humans and animals, such a quantitative characterization is thought to be unavailable to the system (though both motor efference copy, and also proprioception, provides some noisy measures). In general, positioning is thought to be relative to local cues (e.g. presence of a wall of a different colour [3]. Pose knowledge is implicit, and not known quantitatively; locations are known only relative to other landmarks. Here we propose a new formulation, that a set of "coordinates" similar to those used in robotics can be derived by relating the view from the robot to other nearby view images. This enables the human (or robot) to reach for a pose by relating it on a neighbourhood or *chart* of nearby views from the agent. The combination of such charts is an image manifold [4], which is homeomorphic to the configuration space used in traditional robotics [5]. While we do not actually derive the coordinates on these manifolds, which can be noisy and data dependent, the existence of the manifold is used to construct local neighbourhoods which are sufficient for most spatial tasks.

Indeed, it is well known that the motion coordinates that are traditionally used in robotics are only one of many possible *generalized coordinates*. The alternate parametrization proposed here, with a bijective mapping to the canonical coordinates, is an useful characterization for mapping sensory inputs, and can suggest new approaches for longstanding problems both in spatial cognition (see a range of views in [6]), and also in robot self-learning (e.g. [7, 8]). In an earlier work we have shown how such an alternative generalized coordinate can be used to create egocentric maps for a robot's own body, based on images of its limb motions [9]. Here we extend this idea to show how such egocentric views can be "stitched" into a low-dimensional manifold to create an allocentric map for the agents' environment. This approach shows that perhaps, the detection of landmarks for allocentric map construction [10], or the use of optical flow in visual path integration [11], may not be the mechanisms, and perhaps more parsimonious explanations may exist.

A popular approach for modeling a robot's workspace is Probabilistic Roadmaps [5] (PRM), which involves sampling random poses and connecting these in terms of a neighbourhood graph. To our knowledge, a similar approach based on image view samples have not been used to propose a possible mechanism for constructing an allocentric map, though the robotics community has proposed a role of PRM in the brain [12] Here we construct a visual analogue of the PRM, which we call the *Visual Roadmap* [9]. by sampling a random set of images at different poses, and connecting them in a neighbourhood graph, which represents a sampling on the image manifold. We show that this possible representation of the allocentric map can be used to find allocentric bearings and complete routes using only the visual image sample 4.

The very possibility of alternate Generalized Coordinates opens up important directions in how the brain integrates spatial data from sensory modalities. The geometric abilities ascribed to mechanisms such as path integration are shown to be replicable in purely visual maps as well. This is also relevant for applications in robotics, where even for robots that use visual data (e.g. in visual servoing), implementations are based on state estimation which predicts the explicit global coordinates. Here we learn an alternative visual global model without requiring any knowledge priors for robot geometry or kinematics, external world geometry, camera pose, and without reference to any external reference frame. [13, 14]

1.1 Visual Manifolds as alternative Generalized Coordinates

The central idea driving this work is that under some relatively mild conditions, a set of images obtained during a motion with degrees of freedom d, would lie on a d-manifold that is homeomorphic to the manifold of the motion coordinates (*Visual Manifold theorem*). In the proposed model, we suggest that such a visual manifold may be constructed in practice by combining patches on the manifold, each based on local interpolations across images from nearby positions. Each patch is an approximation of a tangent plane or *chart* on the manifold, which can be implemented via principal components analysis, by a layer of neurons [15]. Further, this image manifold is shown to be a cartesian product of a space of translations (encoded in the allocentric map as place cells), and a space of orientations (encoded as orientation cells). Developmentally, these areas are learned, we suggest, by combining multiple modalities into a fused manifold [16, 11]). We show that such a cognitive map has some quasi-metric properties, and can be used for localization or navigation tasks, though it may sometimes violate euclidean norms[17, 18].

Several earlier approaches have attempted to combine multiple views to construct a spatial model. An early approach by Franz *et.al* [19] constructs a *view manifold* by combining omni-directional camera images (360° views). Such a setup permits a rotational image shift (similar to the shift register model for head re-orientation [20] that potentially aligns two views. This removes the rotational variance from the data. Although the manifolds discovered are of far lower dimensionality than the image space, they are not relatable to the degrees of freedom, hence it cannot serve as a generalized coordinates. The approach by Arleo [21] is closer to the present work but it also is substantially different since it uses integration of wheel rotations to correlate the images, enabling reinforcement learning to be applied to the motor torques. Image similarity is computed from gabor output at a set of radial points in every view image. The system also generates place cell like structures, but again, it does not attempt to capture a generalized coordinate model. Also, the model does not attempt to map orientation cells from the data. In [9], the system is primarily focused on looking at a robot's own limb motions. The manifold estimated from these images constitutes an egocentric model of the robot.

There are two main contributions in this work. First we present a model for how visual data may be used for learning a precise characterization for the motion of an unknown moving system. Since the parameters learned are equivalent to traditional coordinates used in robotics, the model can be used in lieu of robot coordinates to generate allocentric models of space. Such a model, of how sensory data informs a potential cognitive map, may help resolve some of the debates regarding the nature of the allocentric map - e.g. the degree of nativist priors or modularity [22].

A second contribution is that we address a longstanding problem in robotics - that of being able to work using vision, as most animals do (e.g. see [23] ch.9). This work applies to any robot that has a camera, and enables it to generate a view sample-based model that generates an alternative to the configuration space. This opens up new vistas in robot body schema learning and in motion planning. In using such a visually grounded map on a robot, we would interface it with a controller that can reach a pose corresponding to a given image; implementing such a controller given that we are in some sense *near* the goal is a well solved problem in visual servoing [24], and here we focus on the construction of the spatial map.

Unlike some previous work, we consider more than a single input modalities (though see [25]). It is now understood that place cells are abstractions that arise based on the combination of multiple modalities [2, 11]. Here, the result of interest is that when we look at the map that arises from vision, and combine it with a map based on olfactory signals, the combined model - and not the vision-only or olfaction-only map - can account for the generation of structures

similar to both place-cells and orientation cells. The fact that this can be done even without combining the map with motor signals, may support the position that "visual perception and the visual control of action depend on functionally and neurally independent systems." [17].

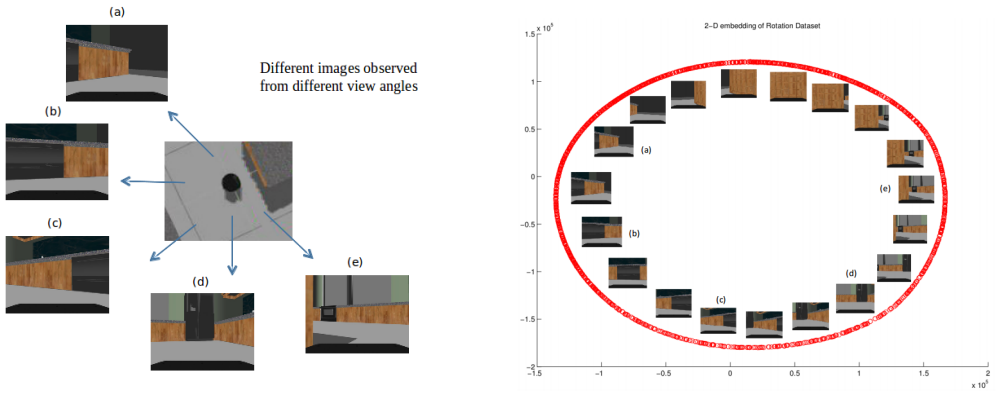


Figure 1: *Images taken from a robot turning on its own axis.* Left: Images as seen from different turn angles. Right: The image manifold discovered as a ring. The topology of the motion manifold (θ) is also a ring S^1. Thus the image manifold is "homeomorphic" - has the same topological structure - as the motion manifold. Since the robot can move along only one dimension, the images can also vary in only one way. See Table 2 below for the regression correlation between the generalized coordinates of these images and the θ parameter.

1.2 Demonstrations

We demonstrate this approach on a simulated mobile robot with a single camera, executing repeated motions in a bounded region of space. Such a robot has 3 degrees of freedom - two for position and one for orientation, and we first conduct some simulation experiments where we have it move in restricted ways, (1-DOF and 2-DOF), before having it move in full x, y, θ motions, sampling a random set of poses while it moves in a random manner that we call "brownian motion". In the 1-DOF experiment Fig 1, the robot turns on its own axis. Increasing the angle eventually returns to its original pose so it has a ring topology. The corresponding visual manifold shown (every dot is an image) - is also seen to be a 1-D ring manifold. In each case the image manifold preserves the topology of the motion space; in the last, fully general, situation, the manifold is the product of a 2D translation manifold with a rotational ring, i.e. $\mathbb{R}^2 \times S^1$. Such an integrated manifold is similar to what has been proposed by Redish and Touretzky [2].

This initial model is based only on vision, and cannot separate the information encoded in place cells (x, y) from that of orientation (θ). However, we show that when this visual model is fused with an additional sensory modality, say, olfaction, which discriminates only spatial position, ignoring orientation, then the visually derived map can be seen to be a cartesian product of a position map (bands in the 3-D map) and a orientation map (2-D patches on the 3-D map). The position maps are be analogous to place cells, encoded as a ring of images corresponding to orientation variation, and the orientation maps - sets of images obtained for motions while looking in the same direction.

In the following sections, we first present some theoretical underpinnings (section:2), followed by several demonstrations of the process, based on a simulated mobile robot (section:3), equipped with a camera that collects images at frequent intervals as it wanders about in a fixed environment. The first *rotation* experiment (section:3.1) is used to illustrate the simplest example, where the robot simply rotates on its axis. The corresponding motion manifold has a ring topology (S^1), and we show that images collected during the spin also lie on a 1-manifold, with a ring topology. In the main experiment, the robot wanders around a planar space (section:3.2), so that each pose of the robot can be described by 3 variables or degrees of freedom (DOF), corresponding to canonical coordinates like x, y for the position and θ for orientation. From the set $\mathcal{I}$ of N images captured, we can use one of many well-known Non-Linear Dimensionality Reduction (NLDR) algorithms to come up with a low dimensional mapping with the same structure as the motor manifold. Section 4 shows simulation experiments for localization (relative to "nearby" visual images), followed by generation of the Visual Roadmap for motion planning. In the last section, we add an olfactory modality and show how this helps separate the positions (places) from orientation (section:5).

2 Visual Manifold Theorem

In the absence of motion constraints, the pose of a mechanism with d degrees of freedom can be described by a d-dimensional configuration vector $\underset{\sim}{q}$. The space $\mathcal{Q}$ of all possible configurations is called the Configuration Space or C-space of the robot. However, how to determine $\underset{\sim}{q}$ is not fixed; any *generalized coordinate* that fully specifies the robot pose can be used. Let $R(\underset{\sim}{q})$ be the volume occupied by the robot in configuration $\underset{\sim}{q}$. $R(\underset{\sim}{q})$ is a subset of R_{sv} that is the volume swept by the robot in all possible configurations. For a mobile robot, the degrees of freedom $d = 3$, and traditionally, the configuration $\underset{\sim}{q} = (x, y, \theta)$ is used. We note that the topology of the C-space is $\mathbb{R}^2 \times S^1$, which is not euclidean, since as θ increases, it returns to the original pose. This causes some difficulty in the algorithm, since most NLDR manifold discovery algorithms (e.g. ISOMAP [26]) assume that the target manifold is euclidean. Now consider a camera mounted on a robot that is moving in a static world. The image obtained from the camera is a function of its pose $\underset{\sim}{q} \in \mathcal{Q}$. The imaging transform function $F()$ which maps a 3-D world point to its image point $I_R = \underset{\sim}{x}, I_R \in \mathcal{V}$, where $\mathcal{V}$ is the image space. $F()$ is parametrized by the camera imaging constants and the pose $\underset{\sim}{q}$. Under the traditional imaging situations (ignoring lens distortions etc) $F(\underset{\sim}{q})$ is a perspective transformation, which may be mapped as a linear transformation in the homogeneous coordinates space.

Assumption (Visual distinguishability assumption).*For any $\underset{\sim}{q}_1 \neq \underset{\sim}{q}_2$, it is not the case that $I(\underset{\sim}{q}_1)$ is identical to $I(\underset{\sim}{q}_2)$.*
Theorem (Visual Manifold Theorem). *Under visual distinguishability conditions, the parameter space Q and the robot image space $I_R \subset \mathcal{V}$ are homeomorphic.*

Since $F(\underset{\sim}{q})$ is linear transformation, every neighbourhood of $\underset{\sim}{q}$ maps to a neighbourhood on I_R, that is $N(\underset{\sim}{q}) \rightarrow N(I_R)$. For the other way to hold that is $N(I_R) \rightarrow N(\underset{\sim}{q})$, under the assumption of "visual distinguishability", different images must be from differing poses, so the continuous image neighbourhood must be unique as well, hence the inverse also holds. Since any neighbourhood has a bijective map in the other space, the parameter space Q is homeomorphic to the visual space I_R.

A consequence of this theorem is that for a mobile robot camera, since the motion manifold is $\mathbb{R}^2 \times S^1$, the image manifold must also have the same topology.

In creating our visualizations below, we use these expectations to guide the data-driven dimensionality reduction methods, since most NLDR methods require a target dimension to be specified. Also, most NLDR methods generate euclidean topologies, and since an S^1 subspace can only be mapped onto an euclidean $\mathbb{R}^2$ space, this necessitates a step up in the target dimension.

However, we note that in practice, we do not need to obtain the low dimensional embedding as a set of manifold coordinates. Thus, in order to localize relative to other nearby objects, or even to navigate, we do not need a set of coordinates; it is sufficient if we construct only the local tangent spaces - these are done using principal component analysis on the local neighbours in the images space - they correspond to *charts* on the manifold, and the set of all charts or the *atlas*, is a complete representation of the manifold [4].

We can gain some intuition into the image manifold by noting that although images have a dimensionality determined by the number of pixels, (320×240, say) so that each $I(\underset{\sim}{q}) \in \mathbb{R}^{80K}$. However, assigning arbitrary colours to each of the $\approx$ 80K pixels will almost never generate an image as seen from the mobile robot camera, hence the actual image space I_R is a much smaller subspace of $\mathcal{V}$. In fact, what the theorem states is that the images will ideally lie on a 3-dimensional manifold which is the I_R subspace of $\mathcal{V}$, and that it will have the topology $\mathbb{R}^2 \times S^1$.

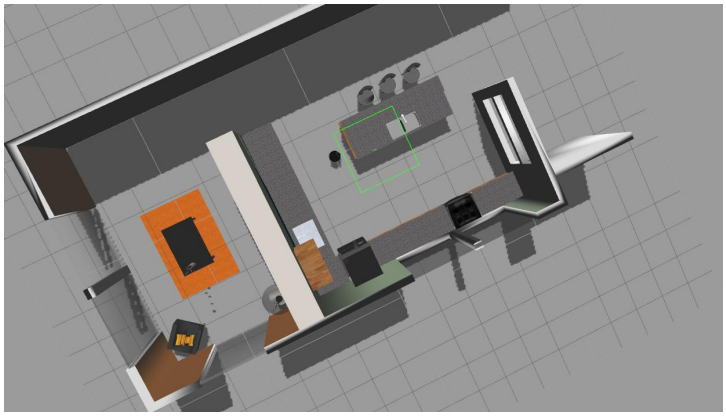


Figure 2: Simulated Environment for Experiments(top view)

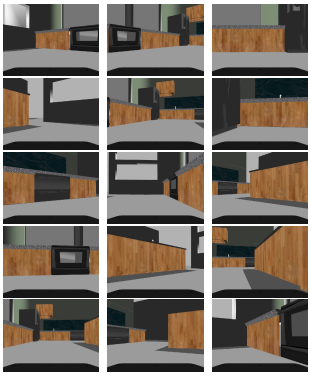


Figure 3: Sample images captured during random exploration

Datasets	Description	C-space	Expected Manifold	Images
Rotation	Rotation about a single point	$\underset{\sim}{q} = (\theta)$	S^1	5000
Cyclic	Traveling in a spiral motion	$\underset{\sim}{q} = (x, \theta)$	$\mathbb{R} \times S^1$	4000
Brownian	Random motion in a rectangle	$\underset{\sim}{q} = (x, y, \theta)$	$\mathbb{R}^2 \times S^1$	17000

Table 1: Datasets: The number of images is just an estimate of the number of points. It is not the case that rotation simulation needs more number of images than cyclic.

3 Simulation Experiments

We simulated the Turtlebot2[27] robot in an indoor environment(Fig:2) under the well known ROS (Robot Operating System)[28] with the Gazebo[29] simulation engine, along with a camera feed.

As mentioned earlier, we report three experiments in this setup. The datasets we have used in our experiments are listed down in the table 1. We sample a number of images as the robot moves under various constraints. One important aspect is that the image sample must not be very near along a path - otherwise the manifold discovery process will fail since close neighbours of any image will all be along the one-dimensional paths. Interestingly, the same patterns is observed in rats acquiring place cells; when motion is limited to a path, the place cells are more one-directional, but wandering randomly in open spaces results in general 2D distribution [2]. In this work, we attempt to emulate an independent distributed sample, by spacing out the images along the trajectories. Even this does not sample the rotation dimension adequately, so we introduce a few spin motions every now and then. The visualizations shown are based on the well-know NLDR method, Isomap [26], but k-PCA or other approaches also give similar results.

A crucial aspect of constructing image manifold is the choice of an image similarity measure (or an image distance metric). Euclidean distances do not work too well unless some objects are overlapping between the images. Earthmover's distance metrics compute the minimum transformation needed to map one image onto the other, but are very expensive to compute. As a effective compromise, we have used here a a Bag of Words (BoW) approach. Here the term "word" refers to a concept in image processing where local descriptors are obtained for a large number of images and clustered; each cluster is called a "visual word", and the set of such words is the "dictionary". An image can now be treated as a "document" containing a bunch of words. Echoing an idea that originated in natural language processing, two images are considered similar if they contain the same words.

In our case, the BoW distance measure, uses a dictionary of visual SIFT[30] features and color descriptors[31]. Then each image I_i can be represented as a binary vector $I_i = \{u_1, u_2, \ldots, u_V\}$, V being the number of clusters (size of vocabulary, 1000 in our case), $u_i = 1$ meaning that the ith feature of the vocabulary is present in the image. We modified the feature vector to account for relevance of features, as used in [32]. For each feature vector, we replace all u_is equal to 1 by its inverse document frequency *idf*, that is $u_i = \log(\frac{N}{n_i})$ where N are the total number of images in the simulation and n_i are the number of images which contain the ith feature of the vocabulary. We use cosine distance for the distance matrix in Isomap.

3.1 Constrained motion: One and Two DOFs

In the experiment reported in Fig: 1 we have considered a sample of images collected while the robot was spinning about its own axis. This is a one degree of freedom motion. Each observed image is fully determined by the rotation angle, so we expect a one-dimensional manifold. To discover the bijective map between the configuration space parameters and the generalized coordinates discovered from the images, we used a backpropagation network which is able to predict the θ parameter from the generalized coordinates on the manifold for a new image. We measure the performance of the prediction by the regression correlation between the two variables. The testing correlation(table: 2) of 0.985 signifies that there exists a robust map between the θ parameter and the low dimensional embedded space.

	Samples	R(Correlation)
Training	1350	0.993
Validation	270	0.993
Testing	180	0.985

Table 2: Neural Network Fitting Results: The fitting correlation so close to 1, shows that there exists a map between the generalized coordinates and the configuration space θ parameters.

In the next experiment, we demonstrate a two-DOF system, by giving the robot a large range of motion in one direction, with a less significant range in the other, while incorporating spinning motions at many points. Thus the motions are a cyclic pattern like a spiral (Fig:4), and the conventional C-space defined as $\underset{\sim}{q} = (x, \theta)$. The images are sampled densely along the line of motion, so that we have sufficient points to get a good manifold. The results (Fig: 5) depict a 3D embedding manifold, that is homeomorphic to a cylinder($\mathbb{R} \times S^1$). The θ values are spread across the curve of cylinder and the translation is along the height of the cylinder.

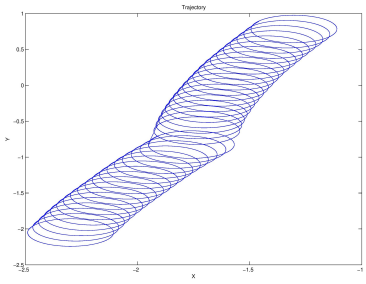


Figure 4: Cyclic Dataset Trajectory

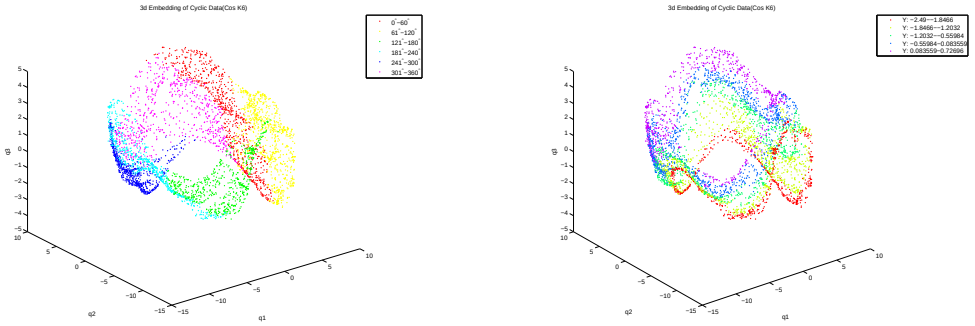


Figure 5: 3-D Embedding of Cyclic dataset. *Left*: Embedding is color coded based on variation in θ. The orientation parameter increases towards one direction along the circular fold and return after 360°. *Right*: Embedding is color coded based on the variation along the translation direction. It increases along the length of the cylindrical manifold.

3.2 Brownian Movements: 3 DOF

In the third experiment, we permit full translation and rotation motions while covering a rectangular region. This motion has all three degrees of freedom (canonical: x, y, θ). Here we expect an $\mathbb{R}^2 \times S^1$ topology. In order to discover the map, images sampled from a path must also have other neighbours sampled from other paths. This requires that the sampling have comparable density in all dimensions. This reduces to generating a sufficiently large sample with a sufficient diversity. This is what is achieved by what we call Brownian movement. The

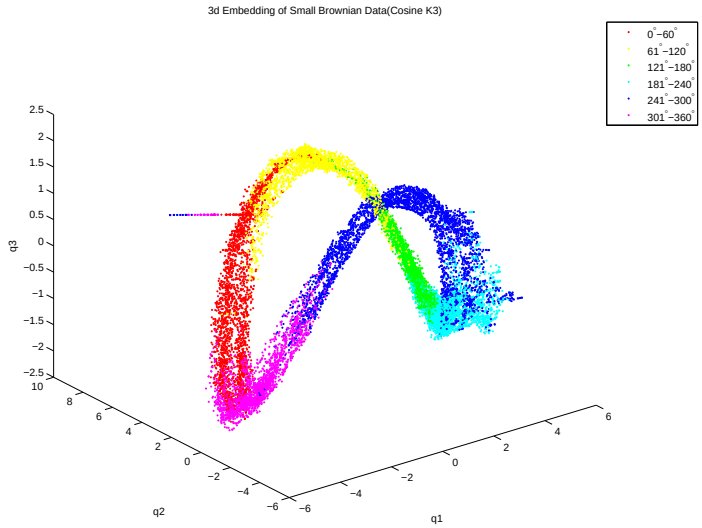


Figure 6: 3-D Embedding of Brownian Dataset

results show a manifold that seems strangely folded, but its topology is actually that of a filled-in torus, corresponding to $\mathbb{R}^2 \times S^1$ (Fig: 6). The orientations (θ) are spread across the curve of torus in rings that are coherent or somatotopic - they do not overlap, and any progression through them encounters the same sequence of orientations (colours). Thus, each section in the torus represents differing x, y with similar camera orientations, while the rings represent differing θ with the same position. Each image maps to a point on the manifold, and thus we obtain a 3-D representation for each pose in the real world. Coordinates describing a position on this structure can then act as generalized coordinates for the configurations or poses.

4 Localization and Navigation

By localization here we mean that given its (novel) view, the robot attempts to find its position relative to other images that it has encountered before. This is done based on a linear interpolation on the k nearest neighbors of the test image, i.e. on the local tangent space. If we know the manifold coordinates of these known images, then that for the unknown image can be determined using the same interpolation weights. Table 3 shows some nearest neighbors for the query images. The robot is able to build a subspace, where it can localize the visual scene it perceives, and there exists a mapping from the visual manifold to the configuration space $\mathcal{Q}$ of the system. We show that our model is robust for the environment and is a good approximation of the robot's workspace using navigation. Our path planning system using the manifold works as shown in algorithm 1.

The paths are found based on the nearest neighbours in the image space. We tested the quality of these paths by comparing them with Probabilistic Roadmap (PRM)[33] as implemented in ROS. We measure the deviation from the straight line joining source and destination points in (x, y, θ) space, for both PRM and our algorithm. We define the deviation of a path $\mathcal{L}$

Table 3: Local Neighbors for Query Image along with canonical coordinates (x, y, θ)

Query	NN#1	NN#2	NN#3	NN#4	NN#5
 **(-0.99,-1.06,243.1°)**	 (-0.99,-1.08, 270.3°)	 (-1.00,-1.09,262.5°)	 (-0.99,-1.07,268.3°)	 (-1.22,-0.98,263.6°)	 (-0.98,-1.02,267.3°)
 **(-0.82,-0.06,72.1°)**	 (-0.84,0.31,54.2°)	 (-1.31,0.37,60.9°)	 (-0.99,0.22,56.4°)	 (-1.57,-0.14,58.4°)	 (-1.33,0.17,56.4°)
 **(-2.15,-2.80,105.8°)**	 (-0.70,-2.85,87.1°)	 (-1.62,-3.13,136.4°)	 (-1.77,-3.08,133.3°)	 (-0.75,-2.97,91.4°)	 (-0.57,-2.88,90.4°)
 **(-2.17,-2.05,314.3°)**	 (-2.16,-2.04,208.5°)	 (-2.02,-3.77,226.5°)	 (-2.02,-3.77,222.8°)	 (-1.88,-4.31,225.7°)	 (-2.16,-3.41,209.4°)
 **(-1.25,-3.44,44.2°)**	 (-2.07,-2.28,297.8°)	 (-2.17,-2.05,298.1°)	 (-2.17,-2.05,301.4°)	 (-2.01,-2.30,307.7°)	 (-0.66,-1.75,326.9°)

Algorithm 1: Visual Roadmap Path Planner(VRM)

input : Source Image I_{src}, Destination Image I_{dest} and the discovered manifold $\widehat{\mathcal{M}}$ and the neighborhood graph $\widehat{G}$

output: $I_1, I_2, \ldots, I_k$, intermediate images lying on the path between I_{src} and I_{dest}

$I^*_{src} \leftarrow$FindNearestNeighbor$(\widehat{M}, I_{src})$

$I^*_{dest} \leftarrow$FindNearestNeighbor$(\widehat{M}, I_{dest})$

$P \leftarrow$DijkstraPath$(\widehat{G}, I^*_{src}, I^*_{dest})$

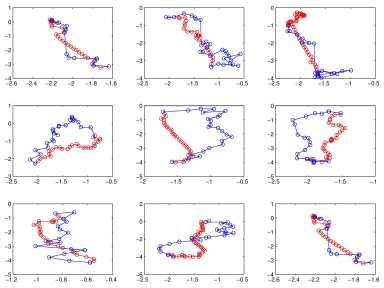
$(\mathcal{L}, \mathcal{I}) \leftarrow$FindGeneralizedCoordsImages(P)

with intermediate points L_i from straight line P as follows:

$$\mathcal{D} = \frac{\sum_{L_i \in \mathcal{L}} D(L_i, P)}{|\mathcal{L}|}$$

where D is the shortest distance between point L_i and line P. We also calculated variance$(D(L_i, P))$ and max$(D(L_i, P))$. Fig: 7 shows some paths for our algorithm, paths shown in red are obtained from PRM and blue from our Visual Roadmap motion planner. The results show that though some of the paths are as good as PRM, some paths are considerably

longer. Partly this may be because of the rather small sample size (17K points is about 11 samples per degree of freedom). But part of the reason, especially for the path traces (Fig. 7:Left) may also be that the path is being planned in the 3-D (x, y, θ) (or its analog in visual Generalized Coordinates), so it prefers motions where the θ changes gradually from start to end. On the other hand, since the trace of the path shows only x, y and has no θ information, there may be big rotations at start and end that distort our understanding of "shortest path". More than the quality of path however, the very fact that such allocentric connections can be discovered with just a sample of egocentric images, is the main result here.



Path#	$\mathcal{D}$		var$(D(L_i, P))$		max$(D(L_i, P))$	
	PRM	VRM	PRM	VRM	PRM	VRM
1	0.109	0.098	0.007	0.009	0.237	0.344
2	0.187	0.188	0.013	0.105	0.411	1.134
3	0.066	0.076	0.003	0.010	0.147	0.292
4	0.136	0.212	0.011	0.028	0.324	0.634
5	0.204	0.526	0.023	0.461	0.540	3.594
6	0.080	0.653	0.010	0.378	0.265	1.865
7	0.265	0.070	0.038	0.005	0.535	0.254
8	0.198	0.144	0.019	0.012	0.431	0.412
9	0.059	0.596	0.004	0.091	0.207	0.818
10	0.131	0.124	0.013	0.018	0.304	0.353

Figure 7: Paths in x, y space and Evaluation Results

5 Discovering Place Cells

Although we are able to do useful tasks with this visual map, there is a significant difference from mammalian visual maps, which use place cells - neurons that fire when the organism enters a particular region in a familiar environment. Our model however, cannot encode such place cells since the manifold lumps orientation and position in an integrated map. Place cells form a cognitive "map" of positions in the environment, though they are very different since metric properties are not strictly adhered to [34]. Now, it is our hypothesis that in order to form place cells, the organism must have some functional requirement that depends only on position but not on orientation - e.g. rats have a strong olfactory sense, and smell sources are often agnostic to orientation (ignoring wind). Thus, an infant rat which has learned to locate its mother based on smell, would, after it acquires vision, realize that a ring of images correspond to the same spatial position (x, y). We are not sure if this is indeed how it works, but in the computational model below, we show that we can use any task that is agnostic to orientation to reveal this structure and to decompose our integrated manifold into a cartesian product of a position space $\mathbb{R}^2$ and an orientation space S^1. The region maps in the position space can then be coded by the organism into a system of place cells.

To figure out the place cell in our manifold, we equipped the robot with a task that does not depend on orientation - we talk about a olfactory sense, but it could be some kind of homing or any positional cue which is omni-directional. We now make the robot move towards such a goal, in many repetitions, from different initial poses. The robot tries to move towards the increasing gradient in smell and approaches it from all directions. To accelerate the repetitions, on reaching the maxima of smell, we destabilize the robot, moving it away in some random direction (imagine another rat pup pushing out infant away). The robot then tries to stabilize towards the smell again. We repeat this process many times. In this process, the robot captures different images in different orientations of the same position, characterized by the goal maxima. These images,

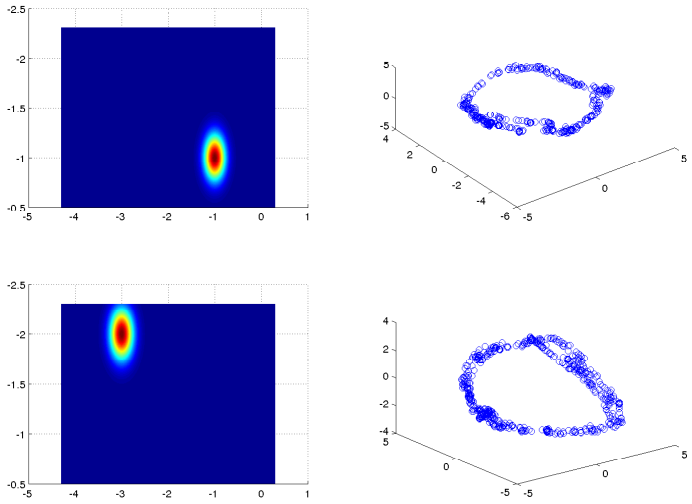


Figure 8: Discovered place cells for smell locations

when located on the manifold form a ring(Fig: 8). Hence, for each position we can discover a ring which is a particular region in the x, y space. Thus, this (or a set of these) can be coded as a "place cell" in our equivalent neural model. Finding similar images (nearest neighbors) across these rings discovers a orientation invariance - where the image remains roughly the same, while the positions change. These are like cross-sections of the torus, and can be used to code up for head orientation cells - which would respond to images corresponding to robot facing in the same direction. Each orientation cell gives us a section on the torus. Hence we have broken down our map into a cartesian product of position and orientation, the $\mathbb{R}^2 \times S^1$ topology.

6 Conclusion and Future Work

In this work, we have proposed a visual characterization for the complex motions of an unknown mobile system. By using images of the external world obtained while the system is moving, we show that we can construct a map that captures the topological and quasi-metric structure of the motor space. This image manifold also incorporates an inexact measure for distance, to the extent that the metrics for image similarity are proportional to metrics in the motor space. We have shown how the map can be used for robot localization relative to its nearby positions, and also how it can be used to navigate from one pose to another by traversing the images encoded in the image manifold.

The key idea here is that coordinates on the visual manifold are generalized coordinates that describe the motion, since a) they uniquely characterize a particular robot pose, subject to the visual distinguishability assumption, and b) given an image pose and its manifold coordinate, there is a single motor pose for it. The idea of generalized coordinates is a very powerful one and applies to a wide range of kinematics and dynamic tasks. One of the key tasks in the future would be to extend these coordinates to handle accelerations, forces and torques i.e. to tasks

in "visual" dynamics.

In practice, the visual distinguishability assumption has some interesting ramifications. This may fail in some situations - e.g. for the two diagonal corners in a rectangular room with white walls. In such a situations, infants and rats are known to confuse between the two locations [35]. In the case where one of the walls is painted blue, turning one's head just a little (taking a small step on the image manifold), results in an image which is sufficiently distinct from the others. This permits the two corners to be distinguished. In practice, violations of visual distinguishability causes problems for NLDR algorithms (short circuits across different parts of the manifold), the actual low-dimensional embeddings (coordinates) are not crucial to our enterprise. Thus, though the images of the manifolds shown here were drawn using standard NLDR algorithms, we would like to emphasize that a neural implementation would not need to compute the full embeddings. Also, the computational system for navigation and localization does not need the embeddings.

Finally, we have presented acquisition of the spatial map at two stages. First, an integrated visual map, with orientation merged with position, is acquired. For organisms where orientation differences are always crucial, this may be adequate. However, if there are some tasks that distinguishing only position but not orientation, then the system can learn maps that encode only position, by defining an equivalence class over orientations. This results in a system that has maps for specific regions in a familiar space, which is very similar to the place cell architecture.

However, though we hope this will be an important idea in this domain, it is only the first step. Much work remains to show the neural validity of these structures, and also to validate these on a wider range of robotic applications. Another important task for the future is to test the capability conferred by having a visual sense of space - the ability to dream or to imagine motions and to generate expectations - such capabilities have many roles in cognition and also increasingly, in robotic tasks.

References

[1] Tom Hartley, Colin Lever, Neil Burgess, and John O'Keefe. Space in the brain: how the hippocampal formation supports spatial cognition. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 369(1635):1–18, 2014.

[2] A David Redish. *Beyond the cognitive map: from place cells to episodic memory*. 1999.

[3] Neil Burgess. Spatial memory: how egocentric and allocentric combine. *Trends in cognitive sciences*, 10(12):551–557, 2006.

[4] W Tu Loring. An introduction to manifolds, 2008.

[5] Howie M Choset. *Principles of robot motion: theory, algorithms, and implementation*. MIT press, 2005.

[6] Jodie M Plumert and John P Spencer. *The emerging spatial mind*. Oxford University Press, 2007.

[7] D. Pierce and B.J. Kuipers. Map learning with uninterpreted sensors and effectors. *Artificial Intelligence*, 92(1-2):169–227, 1997.

[8] J. Modayil. Discovering sensor space: Constructing spatial embeddings that explain sensor correlations. In *Development and Learning (ICDL), 2010 IEEE 9th International Conference on*, pages 120–125, 2010.

[9] M. Seetha Ramaiah, Amitabha Mukerjee, Arindam Chakraborty, and Sadbodh Sharma. Visual generalized coordinates. *arXiv preprint arXiv:1509.05636*, 2015.

[10] Simon Benhamou. Place navigation in mammals: a configuration-based model. *Animal Cognition*, 1(1):55–63, 1998.

[11] Guifen Chen, John A King, Neil Burgess, and John O'Keefe. How vision and movement combine in the hippocampal place code. *Proceedings of the National Academy of Sciences*, 110(1):378–383, 2013.

[12] Jyh-Ming Lien, Marco Morales, and Nancy M Amato. Neuron prm: A framework for constructing cortical networks. *Neurocomputing*, 52:191–197, 2003.

[13] Geraldo Silveira and Ezio Malis. Direct visual servoing: Vision-based estimation and control using only nonmetric information. *Robotics, IEEE Transactions on*, 28(4):974–980, 2012.

[14] Nikolas Engelhard, Felix Endres, Jürgen Hess, Jürgen Sturm, and Wolfram Burgard. Real-time 3d visual slam with a hand-held rgb-d camera. In *Proc. of the RGB-D Workshop on 3D Perception in Robotics at the European Robotics Forum, Vasteras, Sweden*, volume 180, 2011.

[15] Erkki Oja. Simplified neuron model as a principal component analyzer. *Journal of mathematical biology*, 15(3):267–273, 1982.

[16] Thomas J Wills, Laurenz Muessig, and Francesca Cacucci. The development of spatial behaviour and the hippocampal neural representation of space. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 369(1635):20130409, 2014.

[17] A David Milner and Melvyn A Goodale. *The visual brain in action*, volume 27. England, 1995.

[18] Joan Stiles, Mark Kritchevsky, and Ursula Bellugi. *Spatial cognition: Brain bases and development*. Psychology Press, 1988.

[19] Matthias O Franz, Bernhard Schölkopf, Hanspeter A Mallot, and Heinrich H Bülthoff. Learning view graphs for robot navigation. In *Autonomous agents*, pages 111–125, 1998.

[20] BL McNaughton, CA Barnes, and J O'Keefe. The contributions of position, direction, and velocity to single unit activity in the hippocampus of freely-moving rats. *Experimental Brain Research*, 52(1):41–49, 1983.

[21] Angelo Arleo, Fabrizio Smeraldi, and Wulfram Gerstner. Cognitive navigation based on nonuniform gabor space sampling, unsupervised growing networks, and reinforcement learning. *Neural Networks, IEEE Transactions on*, 15(3):639–652, 2004.

[22] Nora S Newcombe. The nativist-empiricist controversy in the context of recent research on spatial and quantitative development. *Psychological Science*, 13(5):395–401, 2002.

[23] Joseph F Engelberger. *Robotics in practice: management and applications of industrial robots*. Kogan Page, 1980.

[24] Moslem Kazemi, Kamal K Gupta, and Mehran Mehrandezh. Randomized kinodynamic planning for robust visual servoing. 2013.

[25] Diedrich Wolter, Christian Freksa, and Longin Jan Latecki. Towards a generalization of self-localization. In *Robotics and cognitive approaches to spatial mapping*, pages 105–134. 2008.

[26] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[27] Willow Garage. Turtlebot. *Website: http://turtlebot. com/last visited*, pages 11–25, 2011.

[28] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y Ng. Ros: an open-source robot operating system. In *ICRA workshop on open source software*, volume 3, page 5, 2009.

[29] Nathan Koenig and Andrew Howard. Design and use paradigms for gazebo, an open-source multi-robot simulator. In *Intelligent Robots and Systems, 2004.(IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on*, volume 3, pages 2149–2154. IEEE, 2004.

[30] David G Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999.

[31] Joost Van De Weijer and Cordelia Schmid. Coloring local feature extraction. In *Computer Vision–ECCV 2006*, pages 334–348. Springer, 2006.

[32] Adrien Angeli, David Filliat, Stéphane Doncieux, and J-A Meyer. Fast and incremental method

for loop-closure detection using bags of visual words. *Robotics, IEEE Transactions on*, 24(5):1027–1037, 2008.

[33] Lydia E Kavraki, Petr Švestka, Jean-Claude Latombe, and Mark H Overmars. Probabilistic roadmaps for path planning in high-dimensional configuration spaces. *Robotics and Automation, IEEE Transactions on*, 12(4):566–580, 1996.

[34] Robert M Kitchin. Cognitive maps: What are they and why study them? *Journal of environmental psychology*, 14(1):1–19, 1994.

[35] Linda Hermer and Elizabeth S Spelke. A geometric process for spatial reorientation in young children. *Nature*, 370(6484):57–59, 1994.