The Baby at One Month: Visuo-motor discovery in the infant robot

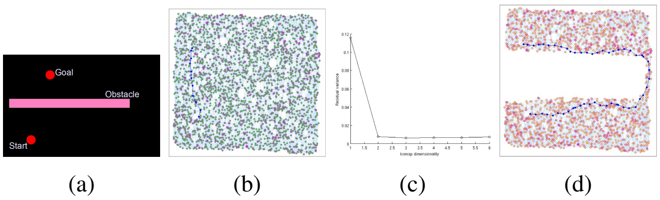
Amitabha Mukerjee M Seetha Ramaiah Sadbodh Sharma Arindam Chakraborty
E-mails: {amit, msram, sadbodh, arindamc}@cse.iitk.ac.in
Department of Commputer Science & Engineering
Indian Institute of Technology Kanpur

***Abstract*— The infant will often struggle to keep its arms in view - one reason for this visual interest may be for learning the visuo-motor map. In this work we suggest that the infant may be constructing a lower-dimensional embedding from an unordered set of sightings. We show that if each robot pose results in a distinct image, then the set of images will lie on a manifold of the same dimensionality as the robot's degrees of freedom. This implies that the number of parameters in the low-dimensional space will equal the number of joint parameters. We suggest that such a compact mapping may serve as an internal *i-representation* for the robot to reason about its actions and their effects. For example, given an object that is introduced into the workspace, it would be possible to identify reaching motions by simulating the overlap of the object with the images of the arm in that part of the image space. Poses on local patches of the manifold may be joined together to construct a *Visual roadmap*. Now, given an obstacle, one may remove nodes and edges from the roadmap that cause collisions, and obtain a free-space roadmap. We show how motion planning can be achieved on such a graph and suggest using a local planner operating by subdivision on the i-representation. The computational study suggests a possible mechanism for models in psychology that argue for high orders of dimensionality reduction in moving from task space to specific action.**

Thus uninterpreted sensory and motor data are combined to enable a naive self-representation, and to solve problems with it. The process is demonstrated for three situations with real and simulated robots. No knowledge of robot or obstacle geometry, or robot kinematics, is used at any step. In conclusion we suggest that such an i-representation may in fact, lead to generalizations that construct an agent's model for space itself.

I. Motor learning in the computational infant

The neonates conception of space arises from actions performed within it. J.J. Gibson put it succinctly when he said that “We must perceive in order to move, but we must also move in order to perceive.” (Ecology of Visual Perception, 1979). This has been underlined in the last decade by intriguing experiments on neonates (10-29 days), who were observed to be exerting to keep an arm moving so that it is visible [17]. Such actions have been taken to be indicative of the possibility that the neonate may be learning a map between vision and proprioception, and discover new possibilities for its motions [18]. When in a darkened room with a beam of light, the infant attempts to keep the arm in the light, and slows down the motion of the limb when it is about to reach the beam. This behaviour arises months before the infant can reach or grasp, and some have argued that the movements are intentional and prospective [1].



(a) (b) (c) (d)

Fig. 1: *Overview example*: (a) Disk robot in a room with a long table. (b) Initially, the robot images are acquired in the space without obstacle. Images are organized by similarity on nearest-neighbor graphs (this demo uses Isomap). (c) By comparing the residual variance at differing dimensions we note that the variance is best explained at dimension 2, indicating that the robot is likely to have 2 DOFs. The resulting low-dimensional space and its parametrization constitute the *Visual Configuration Space* (VCS). (d) Now, the obstacle is introduced and all nodes for which the robot image overlaps with the obstacle are removed; the resulting graph is the basis for the *Visual Roadmap* for motion planning. We use only uninterpreted images, and no knowledge of robot structure or kinematics is used at any stage.

One aspect of the infant repertoire that has not received sufficient attention is how rapidly she learns to avoid involuntary motions that hit its own body. The motion that brings the hand to the mouth is often well developed before parturition, but knowledge of its body continues to mature. Anecdotal evidence suggests that many parents are mortified when their newborn repeatedly pokes her own eyes. However, this behaviour disappears within a couple of days. Bimanual coordination requires that the arms do not hit each other, which is otherwise quite a likely event.

In this paper, we take a computational perspective on the development of this visuo-motor mapping. Our only input are a set of visual images of its own limbs, and attempt to develop a map that associates the various poses to proprioceptive data. While the images are correlated with proprioceptive information about how they were achieved, this information is not used in discovering the ego-motion characteristics, which is achieved purely visually. Nor do we use any other information about the kinematics or other parameters of the system. We adopt a dimensionality reduction approach, and unlike similar work in the dynamical systems tradition [7], we do not require that the motions be presented in a sequence; thus the memory of a disjoint set of observations suffices to construct the internal map.

In the process, the system a) discovers the structure underlying the motion space (e.g. the number of degrees of freedom, a low-dimensional description for each pose), b) formulates a *forward* map from this latent parameter space to the pose space, c) identifies the location of objects in this space in terms of the poses where the arm would hit the object, d) plans paths that allow it to reach around the object into other parts of the space.

This process seems ambitious, and indeed, it is possible only when certain conditions are met. The critical condition that we identify is in terms of what we call visual distinguishability - i.e. the imaging transformation must preserve the homeomorphism underlying the C-space manifold, so that the image space generated by a fixed camera also defines an n-manifold (for an n-DOF robot). This enables us to start with a sample of random pose images, and use well-known manifold-discovery processes to infer n, as well as a mapping from the discovered parameter space to both image and motor spaces. Further, one can then connect the local neighborhoods on the manifold to construct a *Visual Roadmap* which is analogous to the well-known roadmap-based algorithms used for constructing a neighbourhood graph on robot configuration spaces [3].

Detection of poses that will hit or swat at the object is easy; this is done by superposing the memory of the pose with the perceived image of the obstacle in the visual space. Thus uninterpreted sensory and motor data are combined to construct a naive self-representation that can reach for objects. In the next step, a mapping of obstacles onto this Visual Roadmap is performed by superposition with robot images. By deleting these nodes from the roadmap, one obtains a visual mapping of the free-space. We demonstrate motion planning using this algorithm in three situations with real and simulated robots. No knowledge of robot or obstacle geometry, or robot kinematics, is used at any step.

A. Relation with other work

A number of approaches developed over the last two decades attempt to discover some aspects of a robot's ego-structure from sensory data, using approaches such as the discovery of sensor–actuator patterns via linear manifolds (PCA) [12], which learns a unified structure for robot and environment. Alternately, one may discover the dimensionality of the input-output relation by analyzing locally linear tangent spaces [11]. More recent work, often termed developmental robotics, attempts to establish visuo-motor correlations by analyzing random motions (motor babbling) [2], by observing smooth patches in optical flow data [10], or via clusters of sensory data that permit the recognition of object categories and shapes [9], [8]. Other approaches have focused on discovering the topology [13], or on constructing dynamical system models [14].

While these discovered structures capture many aspects of the robot's self and also its environment, very often these models are verbose and differ significantly from what is called a knowledge representation in the traditional AI sense - an explicit model that can serve as a surrogate for testing actions on the world without having to act them out [5]. The term *explicit* is by analogy with conscious human processing, where it is used in opposition to implicit or tacit models of knowledge. For machines, explicit models were traditionally hand-constructed. In today's learning-driven systems, the model is induced from a large training set, and the internal mechanics of the model are often not accessible by virtue of the complexity of the model. The system may need to store a large matrix of weights that have been optimized, or a significant fraction of the training set (e.g. support vectors). Thus the learned structure is far from compact, and the effect of the various elements on the final behaviour are far from clear. In these senses, such models do not qualify as "explicit".

In this body of work, the sensorimotor mapping is learned for the obstacle and environment together, and any change in environment requires a complete re-learning of the map. In this sense, and also because the parameter set is large and not compact, they are difficult to make explicit, the representations learned are very different from those used in AI.

B. Illustrative example: Circular mobile robot

Let us illustrate this process via an example of a planar disk robot moving in a room. Later we shall introduce an obstacle, say a long desk (figure 1a). An unknown motor plant and unknown imaging process results in the robot $R(q)$ being mapped to the image $I(q)$. The robot has access to its own proprioceptive signals q (here taken as its x,y coordinates), but these are used only for moving the arm, and not for discovering the manifold structure.

We first consider the robot in the absence of any other objects or obstacles. Here, 2000 such images are organized into local neighborhoods based on image similarity (figure 1b). When mapped onto a lower-dimensional manifold, we find that the residual variance in the data is best explained when the manifold dimensionality is 2 (figure 1c). This dimensionality, which can be empirically discovered by trying different target dimensions, matches the degrees of freedom of the system. The resulting low-dimensional embedding (the Visual Configuration Space), is then characterized by two parameters, say $v = (v_1, v_2)$.

The neighborhood graph, obtained via image similarity, provides connectivity to different parts of the motion space of the robot. Now, when the obstacle is introduced, one may superimpose it on images of the robot and remove those nodes where the images overlap. We now need to test the remaining edges for local connectivity using a local planner, which works by subdividing the path connecting the two free configurations and testing a set of interim poses. The graph that remains is the final *Visual Roadmap*, which is used for motion planning. Given manifold nodes v, the robot is controlled between them using the corresponding control parameters q obtained using the inverse sensorimotor map. These procedures, and the variables they operate on, constitute a naive representation - they are the sensorimotor analogs of the operations and parameters in the formal C-space. We later demonstrate the same process for articulated robot arms. We observe that while no knowledge of the robot kinematics is required, the algorithm has high space requirements - the entire set of original images has to be preserved. This may be thought of as part of the robot "subconscious" that feeds into the compact model. Also, the superimposition condition is only necessary, but not sufficient. Hence it is overly conservative when the camera axis is not orthogonal to the motion directions.

II. External and internal representations

At this point, we would like to make an observation on the relation between implicit and explicit representations. Each of us has an internal model with which we move our limbs, or throw a piece of chalk at a student sleeping in class. These representations are internal to us, and are implicit and non-compact. Yet, certain aspects of the same representations - e.g. "bend your elbow at 120 degrees", are explicitizations that conform to social convention. These external descriptions define a pose (or a class of poses), so that the set of locations can be expressed in a compact manner. Different people would be able to achieve the motions required under such instructions, though their internal representations may be significantly different. We call these explicit descriptions, those that are conventionalized within a linguistic or cultural group, as *e-representations*. These differ from *i-representations* in that the former may vary from agent to agent, since each agent's experiences are different. At the same time, since these experiences are all constrained in the same physical world, there would also be significant consistencies in the actions generated by these i-representations. Whereas these i-representations in the human are implicit, we argue that if they are sufficiently compact, there is no reason why artificial agents may not use them explicitly for the purposes of reasoning (clearly, they cannot be communicated). We also posit that large models that result from machine learning may constitute the artificial agents' "subconscious". In any case, there is always a clear mapping from the i-representations to e-representations.

The e-representations for a robot are the conventionalizations which we adopt to reason about robotic motions. They include parameters such as the type of each joint, the link lengths and offsets, as well as the set of joint angles or motions. Once the link geometries and kinematics are fixed, a specific pose of the robot is determined by a set of joint parameters. In the *i-representation* model we discover here, the pose is described using the same number of parameters as the joint parameters, and hence these are just as compact. These parameters are the low-dimensional embeddings for the images, and they do not equal the joint angles. They may be thought of as transformations on the external parameters, but transformations that preserve the topological properties. Depending on the type of non-linear dimensionality reduction used, they may also approximate to some degree the metric properties. In our investigtions of different NLDR algorithms for the purpose, we find that the Isomap algorithm [15]), which attempts to preserve the geodesic distances in the low-dimensional space, often results in a map that seems to be an approximate metric. Other NLDR algorithms also preserve the topological relations, but may severely distort the metric aspects. The demonstrations used here are based on Isomap for this reason - but almost any NLDR algorithm, or even the coding layer in a deep autoencoder may be used.

This is the first part of the construction of our (internal) *i-representation* - it serves as an explicit surrogate model for testing actions on the world, and also serves as a predictor of action outcomes. We then outline how various processes can work on this *i-representation* to identify the location of objects, and also to plan paths that circumvent obstacles.

The models developed in this work apply to any robotic system (articulated or mobile) that is able to obtain visual images of its own poses (e.g. a baby observing its own limb motions). We limit ourselves to this set of uninterpreted visual images, and define certain conditions under which the system may discover a compact ego-model, resembling in many ways the conventional robotics models, but one that is derived solely based on dimensionality reduction. We observe that the images lie in a very high dimensional space e.g. 640×480 pixel images would result in a space of 3×10^5 dimensions. In this vast space, valid images of the robot comprise a vanishingly small fraction; indeed, as has been claimed earlier [11], if the robot has n degrees of freedom and the camera is static and there are no other sources of motion, these images would constitute a manifold of dimension n. In this work, we assume a static camera, but we observe that if the camera can also undergo constrained motion, then its motions will be composed with those of the robot and the total degrees of freedom will be the sum of these two motion spaces.

Visual distinguishability - the requirement that given a viewpoint and imaging parameters, no two poses of the robot should result in identical images - is not very difficult to meet in many practical systems. We first establish that under these conditions, the images of the robot poses will lie on a manifold of the same dimensionality as its degrees of freedom. This then paves the way for using NLDR algorithms for discovering this manifold. The representation of the robot is then discovered in terms of the characteristics of the low-dimensional embedding. Thus the parameters v in the discovered lower-dimensional space $\mathcal{V}$ can be mapped to the robot image space $\mathcal{I}$ and to the the control commands, $q \in \mathcal{Q}$, that generated the motions. We observe that these mappings are visual analogs for the processes of forward and inverse kinematics used in traditional robotics, which is why we call this map the *Visual Configuration Space* or VCS.

III. Visual Configuration Space (VCS)

The shape of the robot in the workspace, $R(q)$, is completely determined by its n-dimensional configuration q. Now, let us consider the imaging transformation which maps $R(q^{(i)})$ to the image region ${}^{I}R_i$ in the set of observable images $I^{(i)}$, each of which is represented by a D-dimensional vector, where $D \gg n$ (using superscripts as indices).

The basic premise of this work is that this set of observations (images) $I^{(i)} \in \mathbb{R}^D$, is drawn from an underlying m-dimensional manifold, with $m \ll D$. If the imaging

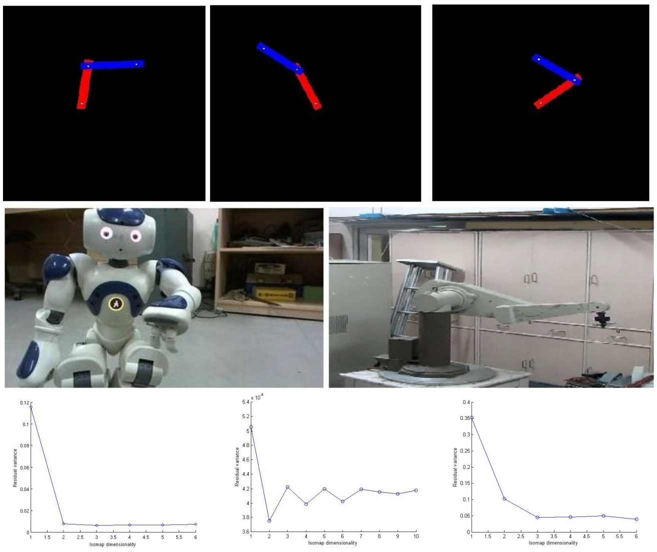


Fig. 2: *VCS dimensionality reflects degrees of freedom*. Images from two-degree of freedom planar arm (top row). Row 2: Nao humanoid robot moving arm along a linear trajectory; PUMA robot moving with 3 joints. Bottom row: Residual variance of low-dimensional embedding for different target dimensions. Planar arm variance (left) is almost completely explained at dimension 2; 99.9% of the variance in the smooth motion of the nao arm is explained at dimension 1; and the 3-DOF Puma motion requires 3 dimensions before the error drops off (right).

transformation is smooth and the robot is *distinguishable*, i.e. different parts of the robot, even if geometrically symmetric, are distinguished by colour, then each configuration will present a different view, and these views are one-on-one and an inverse mapping for the imaging transformation exists.

Let $R(\mathcal{Q}) = \{R(q) : q \in \mathcal{Q}\}$ denote the space of all possible shapes of the robot.

Lemma 1. *For an n-DOF robot, $R(\mathcal{Q})$ is an n-dimensional manifold.*

Proof. The mapping $\phi : \mathcal{Q} \rightarrow R(\mathcal{Q})$ is expressible as a composition of two matrices from the special orthogonal group over $\mathcal{Q}$, so it is bijective. Hence small changes in q result in neighboring shapes $R(q)$, and vice versa, so that $\phi^{-1} : R(\mathcal{Q}) \rightarrow \mathcal{Q}$ is a local homeomorphism. □

Note 1. (Visual distinguishability assumption) *The robot is opaque, and every point on the boundary $\delta R(q)$ is distinguishable in terms of the colour in its neighborhood.*

Let $\psi : R(\mathcal{Q}) \rightarrow \mathcal{I}$ be the imaging transformation so that $\psi(R(q)) = I_q$, the image corresponding to the shape $R(q)$ of the robot in configuration q. So the map $\psi \circ \phi : \mathcal{Q} \rightarrow \mathcal{I}$ takes a configuration into the image space.

Theorem 1. (Image Manifold Theorem) *The space $\mathcal{I}$ of robot images is an n-dimensional manifold.*

Proof. The visible part of the robot is its boundary $\delta R(q)$. For $q, q' \in \mathcal{Q}$ even if $q \neq q'$ there may be situations where the set $R(q)$ and $R(q')$ are identical, though each point is the map from a different initial point. (e.g. consider a cylinder rotating about its own axis). However, by the assumption that the the colouring at every patch on $\delta R(q)$ is distinct, the image will be different so long as every point on $\delta R(q)$ is the mapping from some different initial point so that $\forall q \in \mathcal{Q}, I_q \in \mathcal{I}$ obtained from $\delta R(q)$ is also unique. Thus, for every $I \in \mathcal{I}$, there is a unique q and for every q there is a unique I_q, i.e. the mapping $\psi \circ \phi : \mathcal{Q} \rightarrow \mathcal{I}$ is a bijection. Since the imaging transformation is a projective group, it is also smooth and hence, $(\psi \circ \phi)^{-1}$ is a local homeomorphism. □

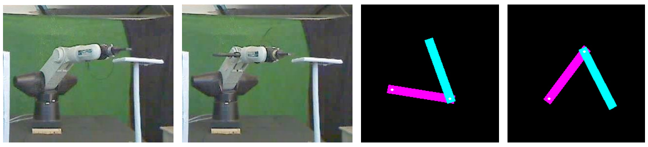


Fig. 3: *Visual symmetry*. Images of CRS A460 6-axis robot (first and second images) appear to be neighboring poses, but close observation reveals that the base joint θ_1 has rotated by nearly 180 degrees, while θ_2 and θ_3 have changed sign. Similar situation observed in a simulated planar articulated arm (third and fourth images). Under euclidean distance metrics, the two poses may be closer than many other poses with smaller angle shifts. Such situations would not arise under the visual distinguishability assumption.

We observe that the visual distinguishability assumption often causes difficulties in practice, for robots with various symmetries (see figure 3). Even in such situations however, the symmetries can be decomposed into separate submanifolds on widely separated configurations, and one may design control regimes that operate within a particular submanifold.

Given that we expect the image space to be an n-manifold, if we can discover the manifold dimensionality of the image space, this will give us a cue to the number of DOFs in the system. This can be achieved by any of a number of non-linear dimensionality reduction (NLDR) algorithms [6]. In many such algorithms, the first step is to construct a nearest neighborhood graph (based on a measure of distance in $\mathcal{I}$); the resulting graph is mapped to the embedding $v^{(i)} \in \mathcal{V}$. If the graph is connected, then distances between any two distal images can now be computed via a path on the edges connecting near neighbors. The Isomap algorithm constructs a manifold by attempting to preserve this geodesic distance [15]. We prefer to use the Isomap here for this reason, though far from exact, it gives it a closer resemblance with the global metric distances compared to other algorithms. In order to estimate the robot DOFs (n), we simply try out a range of target dimensions and choose the lowest dimension that is able to adequately explain the variance in the data (based on residual variance). Note that sometimes for images sampled on a single trajectory (as with the Nao, fig. 2), the manifold is one-dimensional, indicated by a very low residual variance even at m=1 (10^{-5}). If we are able to discover the dimensionality of the robot ($m = n$), the lower-dimensional space can be described in terms of n latent parameters $v_1 ... v_n$, which act as (state) parameters in the robot *i-representation*.

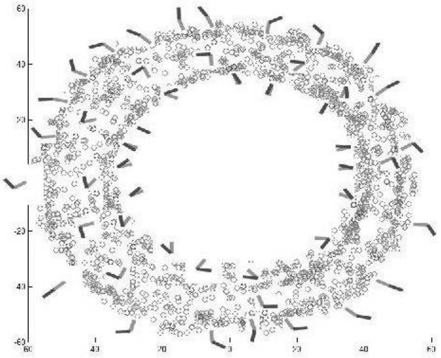


Fig. 4: *Isomap embedding of a set of images of a 2-DOF articulated robotic arm (top row of figure 2).* The embedding resembles the toroidal structure of the topology of this C-space.

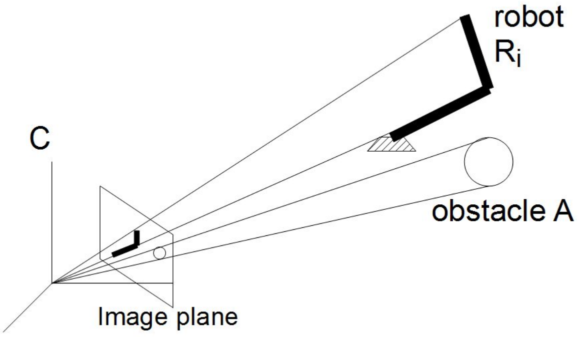


Fig. 5: *Imaging the workspace*. The robot and obstacle lie along the projection bundle from the optical center via their image regions in the virtual image plane (left). If these bundles do not intersect, $R \cap A = \emptyset$. However, the converse is not true.

A. Topology discovery in VCS

The configuration space of a freely-rotating 2-DOF articulated robot (such as the one in the top row of figure 2) is $\mathcal{S}^1 \times \mathcal{S}^1 = \mathbf{T^2}$, which is a torus embedded in $\mathbb{R}^3$ [3]. The NLDR algorithm that generates the VCS shown in figure 4, assumes that the target space for the dimensionality reduction is a euclidean space (a subspace of R^2). This means that the torus surface, which is also two-dimensional, cannot be globally fitted to this space. Hence the map, as shown, resembles the torus in capturing the variability along the θ_1 dimension of this space, but not θ_2.

We note that much of this topological complexity would be reduced for real world robots; e.g. the Scara arm demonstrated in section VI-B has a range of rotation $-135°$ to $135°$ for both θ_1 and θ_2. This implies that the mapping, though it is part of the surface of the torus, can be stretched and would fit in a R^n target topology. For this reason, we have made no attempt to map onto complex spaces.

We next describe how obstacles are mapped on the VCS.

IV. Visual Roadmap and Motion Planning

In the imaging process, robot and obstacle are mapped to a bundle of rays converging on the camera optical center (figure 5).

Let ${}^CR_i$ be the bundle subtended at camera optical center ${}^CO$ by the robot in configuration $q^{(i)}$, ${}^CA$ be the bundle subtended at ${}^CO$ by the obstacle A and ${}^IR_i$, ${}^IA$ be the image regions corresponding to the robot and the obstacle.

Lemma 2. *If* ${}^CR_i \cap {}^C A = \emptyset$ *then* $A \cap R(q^{(i)}) = \emptyset$*.*

Thus, robot configurations for which the bundles do not intersect with the obstacle bundle are guaranteed to be in the free space $\mathcal{F}$. Note that the converse is not true.

Lemma 3. ${}^CA \cap^C R = \emptyset$ *iff* ${}^IA \cap^I R = \emptyset$*.*

Theorem 2. (Visual Collision Theorem) *For a robot in a given pose* $q^{(i)}$*, if* ${}^IR_i \cap {}^IA = \emptyset$*, then* $q^{(i)} \in \mathcal{F}$*.*

We note that the above is a necessary condition, but it is often rather conservative. Indeed, the inverse condition defines *occlusion* situations: where $A \cap R = \emptyset$ but ${}^CR \cap^C A$ is non-null. This limitation is a result of the information loss in the imaging process. These can cause particular difficulties for articulated arms. In such cases, one may use multiple cameras; since the *Visual Collision Theorem* holds for all cameras, we may define any space as free if ${}^CR \cap^C A = \emptyset$ in at least one view. In this situation, both robot and obstacle are less conservatively modeled as the intersection of multiple cones.

In general, for non-orthographic projections, the higher the ratio of camera distance/focal length, the tighter the bound. (e.g the Scara robot arm in section VI-B).

Algorithm 1 describes the process of computing the VCS. The resulting **i-representation** has the following structure:

i-representation {
- Degrees-of-freedom n
- *Visual Configuration Space*: space $\mathcal{V}$ of the low-dimensional embedding. The space is discretely sampled via nodes $\{v^{(1)}, \ldots, v^{(N)}\}$ and corresponding images $\{I^{(1)}, \ldots, I^{(N)}\}$
- Sensorimotor map $f : Q \rightarrow \mathcal{V}$; inverse mapping: $f^{-1} : \mathcal{V} \rightarrow \mathcal{Q}$;
- Visual querying procedure for new image
- *Visual Roadmap*: graph G, with nodes $\{v^{(1)}, \ldots, v^{(N)}\}$ and edges that pass the local planner.

}

V. Motion planning on the VCS

Given the sensorimotor representation including the VCS, we are now in a position to introduce obstacles, and plan paths on it (algorithm 2).

The i-representation also includes the graph based on stitching together the local k-NN neighborhoods. After all obstacle-colliding nodes and edges are deleted from this graph, motion planning can be performed on the attenuated graph, which constitutes the final *Visual Roadmap* (algorithm 2).

Algorithm 1 *Visual Configuration Space* and sensorimotor representation

```
Input: Set of images $\{I^{(1)}, \dots, I^{(i)}, \dots, I^{(N)}\}$ and corresponding control parameters $\{q^{(1)}, \dots, q^{(i)}, \dots, q^{(N)}\}$; optionally, a query image $I^{query}$.
Output: The Visual Configuration Space with low-dimensional embedding $\{v^{(1)}, \dots, v^{(i)}, \dots, v^{(N)}\}$, and the sensorimotor i-representation.
Step1: Based on a suitable metric of image distance, construct the k-nearest neighbor graph $G$ on the set of images.
Step2: Construct low-dimensional manifolds for several different dimensions $n_i$ and compute the residual variance at each $n_i$. Estimate the degrees of freedom of the system $n$ as the dimensionality at which the variance explanation is maximal.
Step3: Obtain the encoding $\{v^{(1)}, \dots, v^{(i)}, \dots, v^{(N)}\}$ at this dimension. VCS: space of these latent variables (dimension = $n$).
Step4: Learn a mapping $f$ from the control parameters $q$ to the latent variables $v$. Also learn the inverse mapping $f^{-1}$ from $v$ to $q$.
Step5: Visual querying. Interpolate using k-NN in the image space to solve for set of weights $I^{query} = \sum_{j=1 \dots k} w_j * I^{(j)}$. Assign $v^{query} = \sum_{j=1 \dots k} w_j * v^{(j)}$ and obtain $q^{query}$ as $f^{-1}(v^{query})$. This procedure estimates the joint coordinates that would reach a given image configuration.
```

Algorithm 2 Motion planning on the sensorimotor representation

```
Input: Sensorimotor Representation, images $I^{goal}$ and $I^{start}$, and obstacle image $I_o$.
Output: sequence of control coordinates $q$ moving the robot from $I^{start}$ to $I^{goal}$ and optionally the estimated control parameters $q$ for these two poses.
Step1: Perform background subtraction [4] to obtain the robot as the foreground object $g_i$ from each image $I^{(i)}$. Apply the mean background also to the obstacle image to segment the obstacle foreground $g_o$. (see figure 7)
Step2: If $g_i \cap g_o \neq \emptyset$, remove node $v^{(i)}$ and all associated edges from Visual Roadmap graph $G$.
Step3: For remaining edges, use local visual planner (below) to check for collisions. If colliding, remove from $G$.
Step4: If either $I^{start}$ or $I^{goal}$ overlaps the obstacle, report failure. Else obtain $v^{start}$ and $v^{goal}$ by the Visual querying procedure given as part of algorithm 1.
Step5: Find a path from $v^{start}$ to one of its near neighbors $v^s$ on the roadmap by testing it with the local planner. Similarly find $v^g$ for $v^{goal}$.
Step6: Find a shortest path (Djikstra) from $v^s$ to $v^g$ on $G$. For each edge on the route, compute a set of intermediate $v$, and return the sequence of joint parameters $q = f^{-1}(v)$ as via points for the controller.
```

We observe that for representing the robot and performing motion planning, the discovered i-representation is quite adequate. The need to map to the e-representation arises only if the agent wishes to communicate or participate in a social interaction.

A. Interpolation for new states

We observe that the set of images $I^{(i)}$ and corresponding low-dimensional mappings $v^{(i)}$ are mapped well, but it is problematic to extrapolate the mappings for intermediate points. This is the *out-of-sample* situation: once a set of data points $X \in \mathbb{R}^{N \times D}$ has been mapped to $Y \in \mathbb{R}^{N \times m}$ ($m << D$) using any NLDR approach, it is not possible to add a new y to the embedding corresponding to a new data point x. This is why we adopt a node deletion strategy for obstacles - deleting nodes on the non-linear manifold leaves the remaining structure intact.

At the same time, we also observe that once the basic neighborhood structure has been put in place, it is possible to update the neighbourhood graph without changing the underlying manifold model. This is because local neighborhoods can be computed and expanded without updating the manifold model - some added nodes will not have their immediate counterparts.

B. Local planner

Traditional roadmap algorithms are based on a collision check for nodes, but the path from two free-space nodes needs to be checked via a suitable local planner [3]. In our situation, a local planner is one that involves actually moving the robot to intermediate configurations v (this is done by interpolating in the q space - see section V-A). A subdivision algorithm works well where we keep testing v_{mid}, recursively until a suitable precision.

C. Forward and Inverse mappings

The embedding captures the inherent regularities underlying the image space, hence it is only to be expected that the mapping from this space to the actual joint angles q would be relatively simple. This is demonstrated empirically using a multi-layer perceptron (figure 6); the convergence is faster and has an error that is two orders of magnitude less, than a learner that attempts to learn the map to q from the images directly. Both forward and inverse maps are easy to learn and are quite accurate, which is why we use these maps to interpolate for new image data in the querying process (algorithm 1, step5).

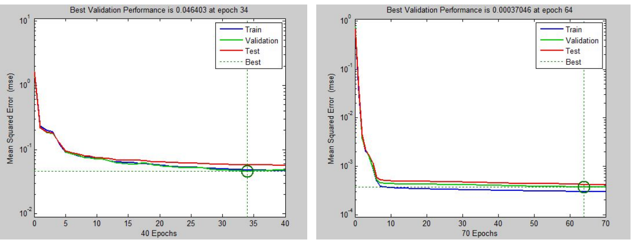


Fig. 6: Left: Comparison of the performance of a multi-layer perceptron mapping to the motor parameters q. When learned directly on high dimensional images (left), the error is two orders of magnitude higher than when learning from the manifold coordinates v (right).

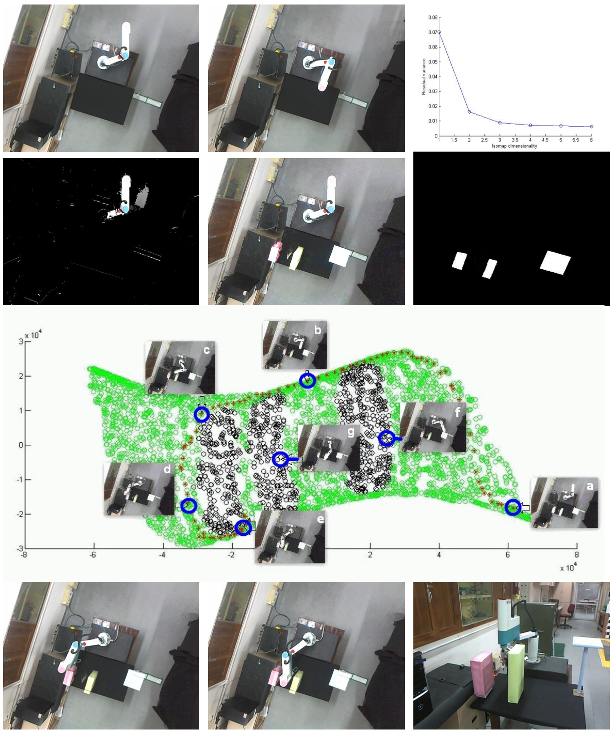


Fig. 7: *Path planning for the MTAB Scara robot.* Row 1: (a),(b) some of the 4000 images of the arm. (c) scree plot. Row 2: incorporating obstacles. (a) background subtracted image of the arm, (b) image with obstacles. (c) obstacles after image subtraction. Row 3: *Visual Configuration Space*; obstacle nodes shown in black, and showing a path plotted from start to goal images. Row 4: path being executed by Scara.

The Isomap algorithm takes $O(N^3)$ time. For N images of size K each, identifying obstacles takes $O(N * K^2)$ and the path computation on the roadmap is $O(N * log_2 N)$; in practice the former step dominates since K is much greater than N in many situations. The algorithm also requires $O(N * K)$ space. For robots with high degrees of freedom, one typically requires rapidly increasing samples, and thus N will also grow, and this often overwhelms the capacity to use an NLDR algorithm like Isomap.

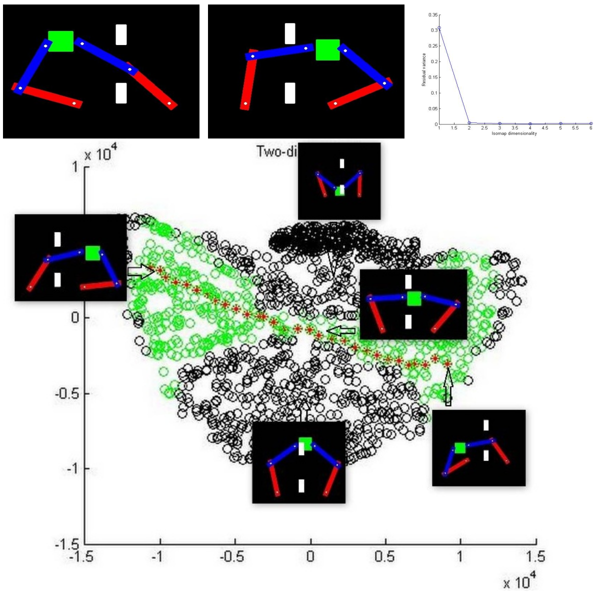


Fig. 8: *Two-armed robot carrying a box with water.* Row 1: two poses of the robot as it moves the box through a hole. Scree plot shows the dimensionality, while holding the box, as 2, though the base system has four degrees of freedom. bottom: The VCS - black nodes indicate collision; A planned path is shown.

VI. Results: Case Studies

A. *Case study 1: Mobile robot simulation*

We have presented the results for a mobile robot simulation in figure 1. The VCS here resembles the actual C-space in this situation because of the choice of Isomap for the embedding. Using other approaches to dimensionality reduction, the shape of the space gets rounded and altered considerably. Also, the fact that the v_1, v_2 axes are aligned to the global axis is a coincidence; on most trials, the axes orient in an orthogonal or mirror directions.

B. *Case study 2: Scara robot*

We now demonstrate the algorithm for a real robot, a Scara 4-DOF arm, in which two revolute joints move the first two links in a plane and it has two more joints providing translational and rotational motion at the wrist.

We observe this robot with a camera mounted on the ceiling. 4000 images are sampled from a video while the robot is moving between random poses throughout its workspace. Background subtraction is performed on each image to generate the foreground robot. Thereafter, one or more obstacles are introduced in the workspace and the obstacles are discovered via background subtraction.

Now the collision configurations are identified by superimposing the foreground robot on the foreground obstacle. Deleting these in the VCS gives us the visual roadmap on which we plan paths (figure 7).

C. *Case study 3: Two-arm planar robot*

We now consider the case of a 2-armed robot. This system has two arms, which jointly constitute a four degree-of-freedom space. This simulated robot is given the challenging task of carrying a box filled with water through a hole (figure 8).

The verticality constraint reduces the dimensionality of the overall system to two. Through trial and error in the space of dimensions, we find after NLDR that the dimension of this space is only two (scree plot in row 1 of figure 8), as expected. The resulting computation of the VCS and the *Visual Roadmap* with a path is shown in figure 8.

VII. Conclusion

In this work we consider the problem of a robot infant, and attempt to discover its sensorimotor map in a manner that draws upon some aspects of human infant cognition. We argue that for artificial agents where the conscious subconscious distinction is irrelevant, the primary hallmark of a representation should be compactness, and its effectiveness in replicating the behaviour. We show that under some normal conditions, most robot systems would be able to discover a low-dimensional mapping from the image space; the set of parameters in this space constitute a state description which are closely aligned with robot joint angles. This sensorimotor map or i-representation is an alternative, consistent, symbolic space and is as compact as the traditional e-representations.

This representation now allows the identification of objects in the workspace via visual overlap. If the object is to be reached by a given part of the robot, poses for this can be identified by overlapping the obstacle image with a series of robot images. For the purposes of obstacle avoidance, one may construct a *Visual Roadmap* from the local neighbourhoods on the manifold. Given an obstacle, putative collision poses (the images where the robot overlaps the obstacle) can be removed, and motion planning performed on the remaining free space. Additional obstacles or moving obstacles can be handled with incremental computation. Though the computation and space costs are high, we must consider that in contrast to traditional approaches, this also discovers the robot self-structure as well as obstacle structures.

An important ramification of this process would be that such a representation, after repeated application in diverse situations, may lead to a generalization which may be considered to be an internal representation of space itself. Such a role for sensori-motor development has been suggested by many [16]. As an example, given a base pose for the robot, distant parts of the space are to be reached with a greater change in the state parameters. Two locations may be close if they can be reached with similar configurations. By generalizing over a large set of such experiences, one may form many notions of space, such as its dimensionality, a hierarchical scale structure, and many other aspects based on the action-perception pairings. By identifying the configurations that reach various parts of the workspace, the system is also constructing a model for the space itself. This is a powerful argument and a possibility that such a computational model can be used to demonstrate.

What we have presented here is just an initial step. The basic idea of discovering patterns from the lower-dimensional mapping of visual images is actually more general, and can also be used for learning other regularities, as in learning the laws of physics, or for handling self-motions of a camera, based on the image space alone. These and many other matters related to this approach remain to be explored.

VIII. Acknowledgements

The authors acknowledge the discussions they have had with Binoy Raveendran and the financial support received from the EADS project and the Research-I foundation.

References

[1] K.E. Adolph and S.E. Berger. Motor development. In *Handbook of child psychology*. 2006.

[2] D. Caligiore, T. Ferrauto, D. Parisi, N. Accornero, M. Capozza, and G. Baldassarre. Using motor babbling and hebb rules for modeling the development of reaching with obstacles and grasping. In *International Conference on Cognitive Systems*, pages 1–8, 2008.

[3] Howie Choset, Kevin M. Lynch, Seth Hutchinson, George A Kantor, Wolfram Burgard, Lydia E. Kavraki, and Sebastian Thrun. *Principles of Robot Motion: Theory, Algorithms, and Implementations*. MIT Press, Cambridge, MA, June 2005.

[4] M. Cristani, M. Farenzena, D. Bloisi, and V. Murino. Background subtraction for automated multisensor surveillance: a comprehensive review. *EURASIP Journal on Advances in Signal Processing*, 2010:43, 2010.

[5] R. Davis, H. Shrobe, and P. Szolovits. What is a knowledge representation? *AI Magazine*, 14(1):17–33, 1993.

[6] John Lee and Michel Verleysen. *Nonlinear Dimensionality Reduction*. Springer, June 2007.

[7] Sridhar Mahadevan. Learning representation and control in markov decision processes. *Foundations and Trends in Machine Learning*, 1(4):403565, 2008.

[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] J. Modayil and B. Kuipers. Autonomous development of a grounded object ontology by a learning robot. In *Proceedings of the National Conference on Artificial Intelligence*, volume 22:2, page 1095, 2007.

[10] L. Olsson, C. Nehaniv, and D. Polani. From unknown sensors and actuators to actions grounded in sensorimotor perceptions. *Connection Science*, 18:2:121–144, 2006.

[11] D. Philipona, J. Kevin O'Regan, and J.-P. Nadal. Is there something out there? inferring space from sensorimotor dependencies. *Neural Computing*, 15:9:2029–2049, September 2003.

[12] David Pierce and Benjamin Kuipers. Map learning with uninterpreted sensors and effectors. *Artificial Intelligence*, 92:169–229, 1997.

[13] A. Ranganathan and F. Dellaert. Online probabilistic topological mapping. *The International Journal of Robotics Research*, 30(6):755–771, 2011.

[14] H. Shatkay and L.P. Kaelbling. Learning geometrically-constrained hidden markov models for robot navigation: Bridging the topological-geometrical gap. *Journal of Artificial Intelligence Research*, 16:167–207, 2002.

[15] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[16] E. Thelen. Motor development as foundation and future of developmental psychology. *International Journal of Behavioral Development*, 24(4):385–397, 2000.

[17] A.L. van der Meer. Keeping the arm in the limelight: advanced visual control of arm movements in neonates. *European Journal of Paediatric Neurology*, 1(4):103–108, 1997.

[18] C. Von Hofsten. An action perspective on motor development. *Trends in cognitive sciences*, 8(6):266–272, 2004.