# From State Transitions to Sensory Regularity: A Topological Grounding of Naive Sensorimotor Experiences

Jean-Merwan Godon, Loïc Goasguen, Sylvain Argentieri

# From State Transitions to Sensory Regularity: A Topological Grounding of Naive Sensorimotor Experiences

Jean-Merwan Godon*, Loïc Goasguen* and Sylvain Argentieri

***Abstract*—How could a naive agent build some internal notions of continuity in its sensorimotor experiences? This is a key question for all sensorimotor approaches to perception when trying to make them face realistic interactions with an environment, including noise in the perceived sensations, errors in the generation of motor trajectories, or uncertainties in the agent internal representation of this interaction. This paper proposes a detailed formalization, but also some experimental assessments, of the structure a naive agent can leverage from its own sensorimotor flow to capture a subjective sensory continuity, making it able to discover some notions of closeness or regularities in its experience. The precise role of the agent action is also questioned w.r.t. the spatial and temporal dynamics of its exploration of the environment. On this basis, the previous authors contribution on sensory prediction is extended to successfully handle noisy data in the agent sensorimotor flow.**

***Index Terms*—Sensorimotor contingency theory, topological grounding, sensory regularities.**

## I. Introduction

It is certainly the case that we deem our sensory experience to be "continuous". Indeed, one crucial property of many psychological perceptual processes is that they generally *seem* continuous [1]; in point of fact, this intuition is strong enough that it is the converse situations where it visibly is not, that earn explicit mentions, such as that of Categorical Perception [2], [3]. However such continuity does not trivially follow from our knowledge of how perceptual processes are materially –e.g. neurally– mediated [4], [5], [6]. In the instance of visual perception, for example, it is known that the eye only acquires very partial snapshots of visual information due to the sparse layout of discrete photoreceptors on its retina as well as the typical trajectories of ocular saccades.

Nevertheless, the continuity of perception subjectively experienced by sensorimotor agents is undeniably useful, allowing for formulation and exploitation of several powerful ideas. One such idea, for instance, is that of inter and extrapolation. If an agent hopes to infer properties of an unknown situation from a structure it has learned on previous experiences, this agent should have a way to quantify in what way this new experience relates to the data it already knows. One very common way to deal with this is thus to *a priori* assign close properties to experiences that are themselves close: the agent should then have the capabilities to distinguish "similar" things, be it external objects, sensory attributes, or even perceptual items. These capabilities may in turn provide grounds for the emergence of its felt continuity of perception: in the end, the agent should then be able to assert that "Red is closer to Pink than it is to Blue, and it is certainly closer to Blue than it is to the sound of a bell" [7]. Such closeness properties are usually leveraged in robotic settings through the well-known mathematical notion of continuity of maps $\mathbb{R}^n \rightarrow \mathbb{R}^m$ since the data available to the robotic agent is usually represented numerically. More generally, the modern examination of continuity and related problems is the subject of *topology* [8], a field of mathematics which is precisely devoted to the study of what it means for something to be continuous. This field has indeed proved a powerful tool for bootstrapping [9], or for modeling geometric ideas in several sensorimotor works [10], [11], [12], in particular those that attempted internally establishing properties of external space [13]. Such approaches allow e.g. motion planning in the internal sensorimotor body representation of an agent through the generation, by interpolation, of continuous motor trajectories [14], or the emergence of a topological representation of the sensor poses from the sensorimotor flow [15]. In most of these works, the (almost) only assessments initially available to the agent are entirely categorical: the agent is indeed only able to perform comparison at a "symbolic" level denoted by a strict equality operator between e.g. sensory values. While these works certainly proved that these operations allow for the extraction of interesting features or meaningful internal representations from a naive form of sensorimotor flow, they also share limitations related to the absence of a "closeness" concept: what about their robustness w.r.t. noise, imperfect repetition of motor paths, etc.? The very same limitation is also shared by the previous work by the authors [16]; in this contribution, the interlink between motor actions and sensory prediction is explored, through the demonstration of the existence of a group isomorphism between them. But predicting the sensory outcome of an action is only accessible to the agent by detecting the exact shift of values inside its own sensor array. Endowing the agent with some internal notion of sensory closeness would then make it able to assess its own prediction, and more generally might allow these sensorimotor approaches to perception –so far mainly restricted to simulated territories– to deal with more realistic conditions.

In this paper, we then propose to examine how some internal notions of continuity in sensorimotor experiences can emerge for a naive agent. To that end, some formal considerations are first introduced in §II. After evaluating a purely topological

* Jean-Merwan Godon and Loïc Goasguen have both equally contributed to this paper. All authors are with Sorbonne Université, CNRS, Institut des Systèmes Intelligents et de Robotique, ISIR, F-75005 Paris, France.

approach, a metric approach is proposed instead and the probability of transition between sensory symbols is used to define some appropriate notion of sensory distance. On this basis, some simple simulations are introduced in Section III to illustrate how an agent could leverage some structure from its own sensory observation. This is illustrated for visual perception through the building by a naive agent of the grayscale or some RGB color model. Then, the role of the agent action in this framework is questioned in §IV. More precisely, the spatial and temporal dynamics of the agent exploration is shown critical to obtain a meaningful and useful structure of its own sensory symbols. Next, some experiments initially proposed in [16] are reproduced in Section V to illustrate how the proposed framework could allow an agent to actually build some sensory prediction functions even in the presence of sensory noise. Finally, a conclusion ends the paper.

## II. Towards a topology of sensory values

This first section aims at defining a topology of sensory values, built on the basis of the agent sensorimotor experience. After a short subsection devoted to the required definitions and notations, a time variable is added to the formalism in the second subsection, so as to account for explicit time dependency of the agent experience, allowing us to introduce a first time-inherited topology. While being possibly sufficient, arguments for the introduction of an explicit metric are then discussed. The third subsection thus proposes the definition of an internal probabilistic metric and highlights the benefits and limits of the proposed approach. Section 3 then exploits these elements in a simple experimental framework to illustrate these elements and demonstrate their actual exploitation.

### A. A short reminder on notations

Let us consider in the following an agent endowed with motor and sensory capabilities. Its internal sensorimotor configuration is classically noted as (**m**,**s**), where $\mathbf{m} \in \mathcal{M}$ (resp. $\mathbf{s} \in \mathcal{S}$) represents the internal motor (resp. sensory) agent configuration, both being elements of their corresponding motor $\mathcal{M}$ (resp. sensory $\mathcal{S}$) set. As shown in [16], the agent motor description can be enriched from $\mathbf{m} \in \mathcal{M}$ to $\mathbf{b} \in \mathscr{B}$, where $\mathbf{b} = (\mathbf{m}, \boldsymbol{\tau})$ depicts the absolute agent motor configuration. $\mathbf{b}$ is made of the agent internal (and thus known to it) motor configuration $\mathbf{m}$ and of its absolute external (and thus unknown from it) pose $\boldsymbol{\tau}$ in its ambient space. Importantly, as discussed in [16], such a change in the notations allows to keep a functional relation between motor and sensory data, even in the case where the agent can freely move in its environment. Next, the environment state is characterized as a function $\epsilon : \mathcal{X} \to \mathcal{P}$, i.e. as a state $\epsilon \in \mathscr{E}$ linking the geometrical space $\mathcal{X}$ (classically endowed with some rigid transformations group $\mathcal{G}(\mathcal{X})$) to the set of the physical properties $\mathcal{P}$ observable by the agent, where $\mathscr{E}$ denotes the set of environmental states. Then, $\epsilon(\mathbf{x})$ represents the observable physical properties at point $\mathbf{x} \in \mathcal{X}$. On the basis on the previous definitions, one can now define the sensorimotor map $\psi$ as the function $\psi : \mathscr{B} \times \mathscr{E} \to \mathcal{S}$, such that $\mathbf{s} = \psi(\mathbf{b}, \epsilon)$. One can notice here that the sensorimotor law does not explicitly depends on time, as is the case of most other contributions in the fields [10], [17], [14]. We will now enrich this formalization with an explicit time dependency. It will then constitute our gateway towards continuity in the sensory experience of the agent, much as in J. Elman's famous 1990 paper [18].

### B. All is well in continuous land

*1) Introducing time in the sensorimotor experience:* The definitions we recalled in the previous subsection actually described *snapshots* of the agent sensorimotor interaction. Nevertheless, these can be easily enriched with an explicit dependency of the various states with a time variable $t \in \mathscr{T}$. Thus, the environmental state $\epsilon \in \mathscr{E}$ can now be written

$$\begin{aligned} \epsilon \colon \mathscr{T} \times \mathcal{X} &\to \mathcal{P} \\ (t, \mathbf{x}) &\mapsto \epsilon(t, \mathbf{x}). \end{aligned} \tag{1}$$

With this notation, one can express an instantaneous snapshot of the environmental state as the partial function

$$\epsilon_t : \mathbf{x} \in \mathcal{X} \mapsto \epsilon_t(\mathbf{x}) = \epsilon(t, \mathbf{x}) \in \mathcal{P}. \tag{2}$$

Therefore any temporal succession of environment states can be described as a trajectory

$$\gamma_\epsilon : t \in \mathscr{T} \mapsto \epsilon_t \in \mathscr{E}. \tag{3}$$

Correspondingly, the agent's *absolute* configuration trajectories and sensory one are respectively denoted by

$$\gamma_{\mathbf{b}} : t \in \mathscr{T} \mapsto \mathbf{b}_t \in \mathscr{B}, \tag{4}$$

and

$$\gamma_{\mathbf{s}} = \gamma_{\mathbf{b},\epsilon} : t \in \mathscr{T} \mapsto \mathbf{s}_t = \psi(\gamma_{\mathbf{b}}(t), \gamma_\epsilon(t)) \in \mathcal{S}. \tag{5}$$

In the following, we will consider a particular but arbitrary subset $\mathcal{S}_\mathscr{T}$ (resp. $\mathscr{B}_\mathscr{T}$ and $\mathscr{E}_\mathscr{T}$) of these sensory trajectories as ones the agent can *effectively* experience. This is intended to denote that some "structural" constraints (which we would externally call "physical" constraints, e.g. limitations on joints' velocity and their smoothness as actuated by the agent) shape the content of its sensorimotor experience.

*2) Towards a sensory topology:* Let us now get back to the intuition of the sensory experience being continuous, as discussed in the introduction of this paper. More precisely, this continuity is that of the agent's sensory experience unfolding with the time $\mathscr{T}$ during which it occurs. In (purely) topological settings, an argument examined e.g. in [19] shows that searching for (*formal*) continuity of the $\gamma_\mathbf{s}$ sensory experiences is entirely dual to searching for topological constraints on the sensory values $\mathbf{s} \in \mathcal{S}$. These two viewpoints intersect at the *final topology* of the $\gamma_\mathbf{s}$ [8], a topology on $\mathcal{S}$ which precisely encodes which structural constraints on the **s** sensory values is needed to make (all) the $\gamma_\mathbf{s}$ experiences continuous. While this final topology seems to solve –at least from a purely topological point of view– the initial problem, one has to keep in mind that most robotic setups rely on discrete time computations. The resulting final topology thus makes $\mathcal{S}$ discrete. Intuitively, this occurs because if the agent only experiences jumps in times such that no instant follows

continuously from the previous one, then it does not need to introduce new continuities in its sensations to make their succession continuous. So how can we solve this issue? One proposes to turn to the setting of metric geometry, which although less general is more suited, in the next subsection.

### C. Introduction of a statistical sensory metric

Introducing corresponding metric considerations however raises new issues: given an abstract sequence of points in a (metrized) point cloud, how can one determine whether it represents a regular/continuous trajectory? For example, how can one decide that a jump in values across a distance of e.g. $5$ units corresponds to a *regular* transition, or instead represents a break in continuity? Without *a priori* assumptions about the expected reasonable dynamics of the experience, it seems these numbers are entirely arbitrary, and related to some *external* knowledge the agent aims to do without. Instead we propose to define a statistical sensory metric, for which the agent ought to set to zero any distance between sensory values that *immediately* (and not *continuously*) follow one another. Thus, the temporal length between successive sensory samples is now central to how the agent perceive them. Consequently, one should first assume that the agent is able to compute distances (or durations) between two timesteps in $\mathscr{T}$. On this basis, we will assume in all the following that the laws of the sensorimotor experiences the agent can observe are *time homogeneous*. This hypothesis then indicates that no statistical measurement the agent can empirically obtain from its sensorimotor experience may depend on the absolute value of the timestep indexing its interaction.

Let us now define the likelihood $P_{\mathbf{s},\mathbf{s}'}$ over all experiences that the sensory value $\mathbf{s}'$ immediately follows $\mathbf{s}$ in the sensorimotor flow of the agent along

$$P_{\mathbf{s},\mathbf{s}'} = \mathbb{P}(\gamma_{\mathbf{s}}(t+1) = \mathbf{s}' \mid \gamma_{\mathbf{s}}(t) = \mathbf{s}). \tag{6}$$

Importantly, from the previous time homogeneity assumption, $P_{\mathbf{s},\mathbf{s}'}$ does not depend on the current time $t$ it is computed. From there and following the intuition that "closeness" of sensory values $\mathbf{s}$ and $\mathbf{s}'$ should increase whenever the probability of the transition $\mathbf{s} \to \mathbf{s}'$ does, we propose to define a simple metric prototype via

$$\delta_f(\mathbf{s},\mathbf{s}') = f(P_{\mathbf{s},\mathbf{s}'}) \quad \forall \mathbf{s},\mathbf{s}' \in \mathcal{S}, \tag{7}$$

where $f$ should verify the two conditions:

1) $f : [0;1] \to \mathbb{R}_+$: $f$ only needs to map probabilities in $[0;1]$ to nonnegative values, i.e. dissimilarity values;
2) $f$ is non-increasing: probable transitions (i.e. $P_{\mathbf{s},\mathbf{s}'}$ close to 1) should result in low dissimilarities.

These conditions do not make $\delta_f$ a metric since it only verifies the non-negativity property. We therefore extend it via minimal paths considerations, i.e. by defining a distance $\mathrm{d}_f$ along

$$\mathrm{d}_f(\mathbf{s},\mathbf{s}') = \inf\left(\sum_{k=0}^{n-1} \delta_f(\mathbf{s}^{(k)},\mathbf{s}^{(k+1)})\right), \mathbf{s}^{(0)} = \mathbf{s} \text{ and } \mathbf{s}^{(n)} = \mathbf{s}'. \tag{8}$$

This in turn enforces the properties of *triangular inequality* and *reflexivity*. In the case where $\mathcal{S}$ is finite, this reduces to the familiar computational form of finding minimal paths on a finite graph with nonnegative weights (corresponding to the $\delta_f(\mathbf{s},\mathbf{s}')$ edge from $\mathbf{s}$ to $\mathbf{s}'$). One should also note that this does *not* guarantee *symmetry* at its core because $P_{\mathbf{s},\mathbf{s}'}$ may differ from $P_{\mathbf{s}',\mathbf{s}}$. Then the $\delta_f$ weights naturally define a *directed* graph (*digraph*), which do not impair the search for minimal paths but do however lead to a non symmetric $\mathrm{d}_f$ function. While there exist several ways to obtain a closely related undirected graph from any given digraph, we hypothesize instead that symmetry should occur as a contingency of the sensorimotor exploration in most real world examples. Therefore, we do not enforce such corrections for now and will instead assess this hypothesis in the resulting graph.

## III. Building the sensory topology from statistics

The previous section was devoted to the mathematical roots of the approach. We will now illustrate how these points can be exploited inside a simple experimental framework which could allow a naive agent to leverage some structure from its own sensory observation. To begin with, a detailed description of the simulation setup is proposed. On this basis, two main experiments are conducted: the first one deals with the construction of a probabilistic sensory metric and the corresponding low-embedding representation for a grayscale camera sensor; the second one extends the reasoning on more complex representation when using RGB image sensors.

### A. Experimental setup and sensory distance estimation

*1) Experimental setup:* In all the following, we consider an agent endowed with a camera sensor observing a 3D scene. Since we are for now dealing with sensory values and their transitions only, the visual perception is basically simulated by playing a video file $\mathbf{v}[n]$ of size $W \times H$, where $n$ represents the video frame number. This is a (temporary) very restrictive setup, which will be enriched later when discussing influence of movement of the agent (see §IV). Also the experience occurs in discrete time, for which each timestep verifies$t = t_n = nT_s$ with $T_s$ the sampling period. In practice, one have $\mathbf{v}[n] = (v_{ij}[n])_{i,j}$, with $i \in [0; W-1]$, $j \in [0; H-1]$, and where $v_{ij}[n]$ depicts the pixel value of the video at frame $n$, row $i$ and column $j$. Each pixel $v_{ij} = (R_{ij}, G_{ij}, B_{ij})$ is represented as a traditional color tuple $\in [\![0; 255]\!]^3$. The agent sensory state $\mathbf{s}_n$ is then simulated by applying some instantaneous function $g : [\![0; 255]\!]^3 \to \mathcal{S}$ to the video, i.e.

$$\mathbf{s}_n = (s_{ij}[n])_{i,j}, \text{ such that } s_{ij}[n] = g(v_{ij}[n]), \tag{9}$$

where $s_{ij}[n]$ represents the $(i,j)$ sensel value at time $n$, row $i$ and column $j$ of the agent camera sensor. Introducing $g(.)$ in (9) allows to explain formally how a physical state of the environment (which can be envisaged here as the pixel values of the video) is turned into the internal sensory state of the agent. But one has to keep in mind that the agent does not know the relation (9), it does not even have any knowledge about the meaning of these numerical values: they are only *uninterpreted symbols* to it, with no a priori structure, order, not any way to actually *compare* them. In addition, the set $\mathcal{S}$ may well be isomorphic to the set of actual pixel values,

but there may also have a lower number $S$ of symbols than pixel values, resulting in a compressed representation. Without loss of generality, $\mathcal{S}$ will then be defined as the finite set of positive integers $\{0,\dots,S-1\}$ and we will adopt a traditional $s_{ij} \in [\![0; S-1]\!]$ coding convention for the numerical values of each $(i,j)$ sensel, where $S = 256$ for traditional camera sensors. As outlined in §II-C, it is then proposed to look at the relationship between those $S$ uninterpreted (numerical) symbols through the statistics of their transitions. Let us now detail how these transitions are actually captured.

*2) Description of the experiment:* In all the following, we will assume that all $W \times H$ agent's sensels contribute equally to the building of the same representation. Then, we define a $S \times S$ matrix $M = (m_{kl})_{k,l}$ counting all the transitions of sensels values along observations, with

$$m_{kl}[n+1] = m_{kl}[n] + \sum_{i,j} \zeta_{kl}(i,j)[n], \tag{10}$$

with $(i,j) \in [\![1; W \times H]\!]^2$, $m_{kl}[0] = 0$, and $k,l$ both representing two symbols in $\mathcal{S}$ (that is, sensor output values $\mathbf{s}_k$ and $\mathbf{s}_l \in \mathcal{S}$). $\zeta_{kl}(i,j)[n]$ aims to capture the existence of a change of value of the $(i,j)$ sensel from value $k$ at time $n$ to value $l$ at time $n+1$, i.e.

$$\zeta_{kl}(i,j)[n] = \begin{cases} 1 & \text{iff } s_{ij}[n] = k \text{ and } s_{ij}[n+1] = l, \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

From (10), one can then compute the probability of transition of sensels values gathered in a $S \times S$ matrix $P = (p_{kl})_{k,l}$ with

$$p_{kl}[n] = \frac{m_{kl}[n]}{\sum_{q=0}^{S-1} m_{ql}[n]} \tag{12}$$

the probability at time $n$ for any sensel to see its value changing from symbol $k$ to $l$. Obviously, $p_{kl}[n]$ is expected to converge towards $P_{\mathbf{s}_k,\mathbf{s}_l}$ as time $n$ tends to infinity. Then, once the estimation of the matrix $P$ has converged after a fixed number frames $N$, it is turned into a $S \times S$ metric prototype matrix $\Delta = (\delta_{kl})_{k,l}$ according to Eq. (7) where $f = -\log$[1] is selected, with

$$\delta_{kl} = -\log(p_{kl}[N]). \tag{13}$$

Again, any function verifying the two conditions in §II-C could have been selected. Then, Dijkstra's algorithm [20] is applied on the $\Delta$ matrix along Eq. (8) to produce the $S \times S$ distance matrix $D = (\mathrm{d}_{kl})_{k,l}$, providing the agent with the result metric $\mathrm{d}$ we set out to discover

$$\mathrm{d}_f(\mathbf{s},\mathbf{s}') = \mathrm{d}_{-\log}(\mathbf{s}_k,\mathbf{s}_l) = \mathrm{d}_{kl}, \tag{14}$$

which is finally visualized in 2D or 3D through a multi-dimensional scaling projection method (MDS [21], [22], [23] or ISOMAP [24]).

### *B. Results for a grayscale perception*

The $W \times H = 856 \times 480$ video used to conduct the experiments comes from a slightly stabilized camera filming an evening walk in Midtown New York City in the rain[2]. It consists in a natural city scene filmed in real time from a first person point of view. A grayscale (cropped) preview of the video is shown in Figure 1a. To begin with, one will consider a function $g$, mapping the $(R_{ij}, G_{ij}, B_{ij})$ color coding of the video pixels $v_{ij}$ to the sensel values $s_{ij} \in [\![0; 255]\!]$ of the agent, such that

$$s_{ij} = g(v_{ij}) = h(\texttt{round}(\texttt{mean}(R_{ij}, G_{ij}, B_{ij}))), \tag{15}$$

where $h$ is a function that can be tuned to artificially modify the agent's perception. Two cases for $h$ are discussed in the following: either $h() = id()$, corresponding to the case where the agent grayscale perception exactly matches the grayscale version of the video, or $h() = \text{sawtooth}()$ for which the perception is altered on purpose to exhibit some properties of the agent internal representation of sensory values.

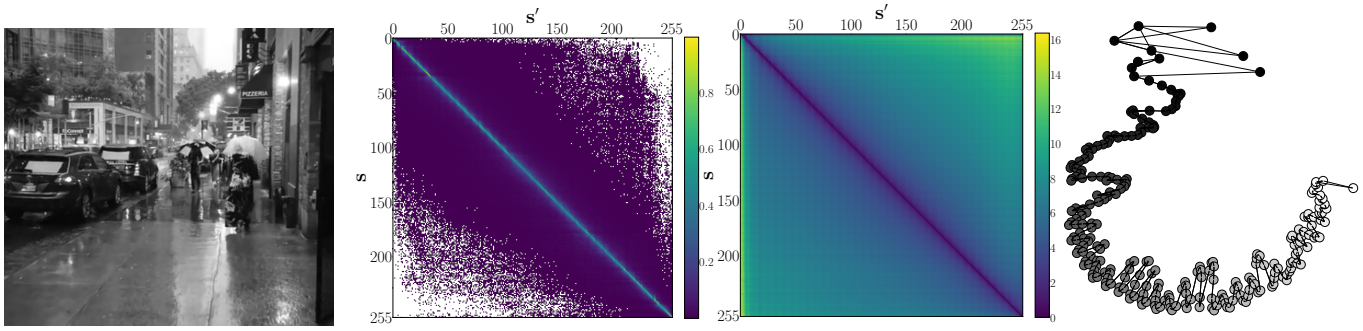
*1) First case: $h()$ is the identity function:*

*a) Estimation of the probability of transition between symbols:* Since $h() = id()$ in Eq. (15), the agent sensory values are made of $S = 256$ uninterpreted symbols, whose values along frames can be used to compute their probability of transition along Equation (12). The resulting $S \times S$ matrix $P$ is shown in Figure 1b for frame number $n = 10$. Note that the $S$ symbols are ordered in the figure according to their numerical values: this is something the agent can not actually do for now, but this ordering has no effect on the reasoning and helps in understanding the process. From Figure 1b, one can see that the most probable transitions are all placed along the diagonal of the matrix P, meaning that the most probable sensory output at the next time step is the very same symbol. Further, the *a priori* ordering of symbols allows to observe that the diagonal is thick and fades away as the symbols values are distant: this clearly indicates that the most probable transitions are the one to symbols that are *close, from an external point of view* (again, the a priori ordering is unknown to the agent). Conversely the least probable transitions are the one to *distant* symbols. Those results are in accordance with the intuition that close time intervals lead to close sensory outputs, and that some regularity of the sensory experience has been captured. Note that since the probability estimation is evaluated on occurrences, the case where no transitions at all between two symbols is observed leads to a probability of 0 (represented in white on the Figure 1b); this appears at the beginning of the experiment only and mainly concerns *distant* symbols with a very low transition probability, i.e. in the two corners of the Figure 1b.

*b) Computation of the distance matrix:* On the basis on the previous probability of transitions between symbols, one can compute the metric prototype in the form of the $S \times S$ matrix $\Delta$ whose elements are given by Eq. (13). Then, Dijkstra's algorithm [20] is performed on $\Delta$ to obtain the $S \times S$ distance matrix $D$. The resulting matrix $D$ is represented in Figure 1c for $n = 10^4$. Obviously one should note that when direct transitions between symbols are missing in $P$ (and thus in $\Delta$), Dijkstra's algorithm will nonetheless generally find an alternate path towards those symbols by finding adequate successive transitions; consequently the $D$ matrix is expected to be fully defined (i.e. with all coefficients finite) as long as the agent has experienced enough sensory symbols transitions.

[1] If a probability of transition is equal to 0, the corresponding distance is set to `NaN` by convention.

[2] https://youtu.be/eZe4Q_58UTU




(a) $v[n]$ (b) $P$ at $n = 10$ (c) $D$ at $n = 10^4$ (d) $\texttt{MDS}_2(D)$ at $n = 10^4$

Figure 1: Building of the internal organization of sensory values. (a) Grayscale version of one frame of the video used in the experiment. (b) Estimated probability matrix $P$ at $n = 10$, i.e. at the very beginning of the experiment. (c) Estimated distance matrix $D$ at $n = 10^4$, i.e. at the end of the experiment. (d) Corresponding low-dimensional embedding of $D$: one can see the intuitive grayscale organization of pixel values, discovered by the agent from its sensory values transitions.

One can see from Figure 1c that previous low transition probabilities are now associated to high distances (and vice versa). One also recognizes the same diagonal pattern, which now corresponds to low distances. One can also see that $D$ is *almost* symmetric, except in the corners where lie most of the high distances, corresponding to the least probable transitions of sensory symbols. This is not an encoded property of the agent experience but instead seems to appear as a contingency of the sensorimotor exploration, as outlined in §II-C.

*c) Visualization of the representation:* Finally, one can qualitatively assess the shape of the captured sensory symbols topology by projecting the resulting distance matrix $D$ into a space of lower dimension. The 2D visualization of the matrix $D$ through a MultiDimensional Scaling (MDS) projection is represented in Figure 1d. Note that such a method requires the input matrix to be symmetric; hopefully, we qualitatively showed it was almost the case so that MDS can be actually applied on the symmetrized matrix $1/2 \times (D + D^T)$. In Figure 1d, each circle represents a single symbol where the inner color corresponds to the color perceived from an external point of view (color that also matches the classical gray-level scale in this case, since $f = id()$). One can see from this representation that the obtained manifold is almost one dimensional, and captures the classical gray scale from white to black in a continuous manner. This can be evaluated by looking for the 2 nearest neighbor of each symbol in the internal metric; these neighbors are then linked together in the projection by an arrow drawn in the figure. Browsing the manifold by following these arrows allows to go from white (coded as the number 255) to black (coded as a 0) almost without any discontinuity in the symbol order. From this graph, one can conclude that the agent has been able, starting only from the probability of transition between uninterpreted sensory symbols, to discover the gray level scale. Such a capability will be further exploited for different applications, like sensory prediction, see §V.

*2) 2nd case: $h()$ is a sawtooth function:* We will now consider a case where the agent's sensory output does not exactly match the original grayscale world as per Eq. (15), where $h() = sawtooth()$. With such a change, a world grayscale value between 0 and 127 will be encoded by (uninterpreted) sensory symbols between 0 and 254 (with a step of 2) for the agent, along with the world grayscale values between 128 and 255. Consequently, a single internal sensory symbol (e.g. 54) will now correspond to two possible world grayscale values (27 and 155). Intuitively, such a change is expected to *create* continuity that does not exist initially between symbols through a stronger proximity between values representing dark and light shades. The previous process is then repeated and the resulting 2D MDS embedding is depicted in Figure 2: as expected, one identifies a looping monodimensional manifold. In the figure, each sensory symbol is depicted as a circle whose color represents its *internal* coding (i.e. a numerical value from 0 to 254 with a step of 2), represented as grayscale values for convenience. This color no longer matches the grayscale values of the world it represents because of the introduction of the sawtooth function. But the continuity initially captured in the previous experiment leads to a looping

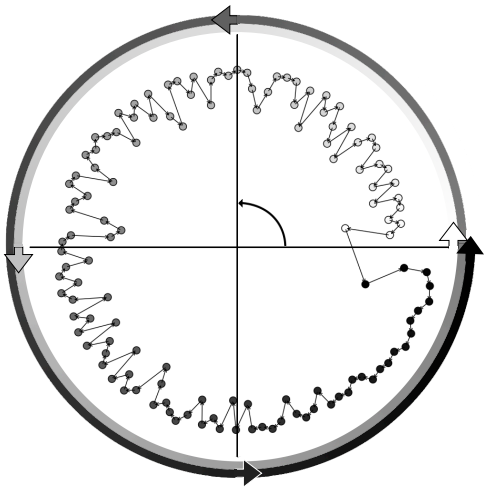


Figure 2: 2D MDS projection of the sensory symbols when a sawtooth function links together world gray values to sensory symbols. Each symbol is represented as a circle whose color represents the *internal* coding. The corresponding symbols *in the outside world* are represented as a looping arrow around the projection. *Internal* Black (symbol 0) and white (symbol 255) symbols are now close to each other, differently from Fig. 1d.

representation where the two opposite symbols 0 and 254 are now close to each other in the internal representation as they both correspond to close grayscale values in the environment. Such a conclusion might be obvious in this specific case, but it highlights that the *internal* representation of the sensory symbols' topology might actually highly differ from our initial intuition as it depends on the way the agent's sensors encode sensory information. The same remark could apply to faulty sensors, which output symbols could be modified or rearranged because of some failure in the information acquisition process; the proposed approach could then allow the agent to (re)build an adequate internal representation, though still intrinsically limited by its own (limited) sensory capabilities.

### C. Results for color perception

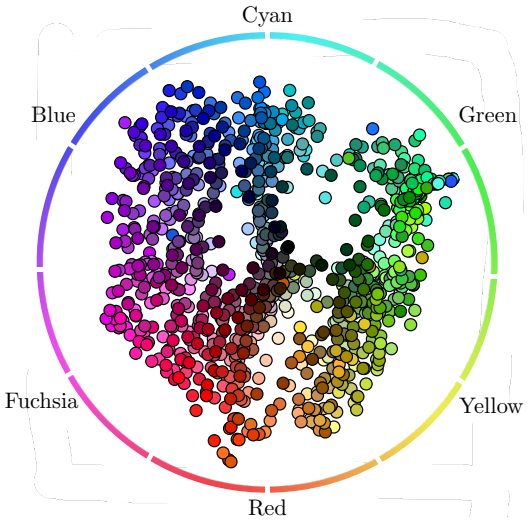
To further illustrate the approach, one will now endow the agent with some color perception capabilities. Then, in this subsection, the initial color tuples $(R_{ij}, G_{ij}, B_{ij}) \in [\![0; 255]\!]^3$ coding the video pixels values $v_{ij}$ are now mapped to the $S = \alpha^3$ agent sensels values $s_{ij} \in [\![0; \alpha^3 - 1]\!]$ along

$$s_{ij} = g(v_{ij}) = Q_\alpha(B_{ij}) + \alpha Q_\alpha(G_{ij}) + \alpha^2 Q_\alpha(R_{ij}), \quad (16)$$

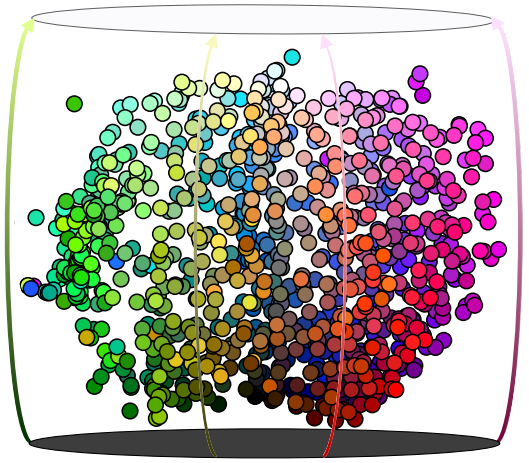
with $Q_\alpha(.)$ a quantification function defined by

$$Q_\alpha : X \mapsto Q_\alpha(X) = \texttt{round}\left(\frac{X}{255} \times (\alpha - 1)\right), \quad (17)$$

with $X \in [\![0; 255]\!]$ and $Q_\alpha(X) \in [\![0; \alpha - 1]\!]$. Note that while the symbol ordering was quite obvious for grayscale values from an external point of view (e.g. the natural order from 0 to 255) for the various matrices $M$, $P$, $\Delta$, and $D$, this no longer holds for these color sensory output symbols. Nevertheless, the order in which they appear as line or column indices in these matrices is not relevant since the only relevant information of closeness between them is entirely independent on how these symbols are actually ordered. In all the following, $\alpha = 10$ is selected, so that the agent sensory space is made of $S = \alpha^3 = 1000$ uninterpreted (numerical) symbols. On this basis, all the previous steps are successively applied. The resulting $D$ matrix can then be visualized through a low dimensional embedding technique like ISOMAP [24]. The result of this projection performed in 3D is shown in Figure 3. The obtained representation is pretty much in line with some classical representations of RGB color models, like the HSL or HSV coding of color. Indeed, the 3D points cloud first appears to capture some color order very similar to the classical hue color wheel where pure colors are represented through an angular position on a circle, as depicted in Figure 3a. But the 3D projection also exhibits a third axis linking very dark to very light shades for each color of the hue wheel, similar to the lightness axis in the HSL color coding, see Figure 3b. Then, with such a representation, the agent is now able to assess if the sensory symbol associated to the rose color is *closer* to the one associated to the red color than it is to the green one thanks to its internal metric matrix $D$.




(a) 3D ISOMAP projection seen as a 2D color wheel.



(b) The same 3D projection seen as a cylinder, with the lightness axis drawn as arrows from black to white.

Figure 3: Interpretation of the 3D ISOMAP projection of the matrix $D$ when the agent is endowed with color perception capabilities. (a) Representation obtained when viewing the projection "from below": one can notice that all the sensory symbols are arranged by colors, matching the intuitive color wheel which has been added to the graph. (b) Another point of view on the 3D sensory symbols representation: in addition to the color ordering highlighted in subfigure (a), a third axis is supporting the variation of lightness. The obtained projection can thus been understood as analog to the HSL cylindre or biconic representation of the RGB color model.

## IV. Locomotive motives: a case for fitting exploration and exploitation dynamics

The previous developments were largely devoted to the relationship between two internal observations: the transition probabilities and the resulting metric. We showed how this information allows the agent to build some notion of closeness between sensory symbols –which could be understood as some subjective notion of sensory continuity– from certain successions of sensory experiences being more likely, or *typical*, than others. But such considerations clearly rely on the idea that typical environment states also display certain typical patterns themselves. From an external point of view, one would certainly declare that "environment states are (mostly) contin-

uous", both in time and space. This underlying assumption has not been dealt with so far, especially since the agent was passively observing sensory symbols changing along time in the previous experiments, and not actively exploring its environment. This section thus aims to study which external structure in the states of the environment could explain the relationships between the agent's motor actions associated to a sensory experience and the observed regularity, effectively giving action a defining role in this internal assessment. To that end, additional formal considerations are introduced in a first subsection. On this basis, some new experiments are proposed to highlight the importance of movement in building this subjective continuity.

### A. Fitting spatial and sensory dynamics in the exploration

*1) Spatial and temporal coherence:* The results obtained in Section III were based on a purely passive observation of a changing "natural" visual scene –where the word *natural* here refers to our own usual and intuitive sensorimotor experience– allowing the agent to build a representation of its own sensory symbols organization. But this representation should highly depend on the environment states and the successive configurations with which the agent samples it along time. More precisely, this implies that the environment state should exhibit some typical patterns, both in space and time, in line with the manner in which the agent conducts its interaction, to make apparent the notion of certain sensory symbols transition being "more typical" than others. Thus, one first condition to fulfill is *spatial*, mandating that e.g. immediately next to a red region $\mathcal{X}'$ of ambient space $\mathcal{X}$ it is more likely to be another region $\mathcal{X}''$ that is orange than cyan itself. In other words, we would generally expect the two events

$$\{\gamma_\epsilon(t)_{|\mathcal{X}'} = \epsilon_0\} \text{ and } \{\gamma_\epsilon(t)_{|\mathcal{X}''} = \epsilon_1\} \tag{18}$$

to largely depend on one another when $\mathcal{X}'$ and $\mathcal{X}''$ denote close (and small) regions of space. Furthermore, one second condition is *temporal*, so that the environment state at any localization $\mathcal{X}'$ does not immediately change too randomly so that the two events

$$\{\gamma_\epsilon(t)_{|\mathcal{X}'} = \epsilon_0\} \text{ and } \{\gamma_\epsilon(t+\Delta t)_{|\mathcal{X}'} = \epsilon_1\} \tag{19}$$

are conditioned to another when $\Delta t$ remains sufficiently small. One should insist on the fact that this coherence property however should only be local and relative to the agent exploration dynamics. It is clear that the color of a point $x \in \mathcal{X}$ and time $t \in \mathcal{T}$ does not depend on which colors appears two kilometers, one and a half days from there. On the other hand, should the agent instead perform a two kilometers long movement between two successive time samples, it should not be able to infer any relationship between successive sensory readings from the sole spatial coherence constraints.

*2) A formal account of spatiotemporal coherence:* Let us now generalize the previous sensory transition probabilities (6) by introducing, for any (sub)collection of motor trajectories $\mathscr{B}'_\mathcal{T} \subset \mathscr{B}_\mathcal{T}$,

$$P^{\mathscr{B}'_\mathcal{T}}_{\mathbf{s},\mathbf{s}'} = \{\gamma_\mathbf{s}(t+1) = \mathbf{s}' \mid \gamma_\mathbf{s}(t) = \mathbf{s} \text{ and } \gamma_\mathbf{b} \in \mathscr{B}'_\mathcal{T}\}, \tag{20}$$

for which $P^{\mathscr{B}_\mathcal{T}}_{\mathbf{s},\mathbf{s}'} = P_{\mathbf{s},\mathbf{s}'}$. Such a (slight) generalization allows to highlight how a specific set of motor trajectories actually condition the sensory transitions available in the agent sensorimotor flow. More precisely, we used in [16] the *sensor receptive field* notion to define the specific region of space which environment state suffices to fully determine the agent sensory state $\mathbf{s}$. Formally, a sensor receptive field can be seen as a function $F : \mathbf{b} \in \mathscr{B} \mapsto F(\mathbf{b}) \subset \mathcal{X}$ verifying

$$\begin{aligned}&\forall \epsilon_1, \epsilon_2 \in \mathcal{E}, \forall \mathbf{b} \in \mathscr{B},\\ &\qquad \epsilon_{1|F(\mathbf{b})} = \epsilon_{2|F(\mathbf{b})} \Rightarrow \psi(\mathbf{b}, \epsilon_1) = \psi(\mathbf{b}, \epsilon_2) = \mathbf{s}.\end{aligned} \tag{21}$$

Then, let us now consider $\mathscr{B}'_\mathcal{T}$ as a set of motor explorations $\gamma_\mathbf{b}$ such that the receptive fields $F(\gamma_\mathbf{b}(t))$ and $F(\gamma_\mathbf{b}(t+1))$, which condition successive sensory outputs $\gamma_\mathbf{s}(t)$ and $\gamma_\mathbf{s}(t+1)$, fall *far apart* from one another. Then, the corresponding local environment states $\gamma_{\epsilon|F(\gamma_\mathbf{b}(t+1))}(t+1)$ and $\gamma_{\epsilon|F(\gamma_\mathbf{b}(t))}(t)$ are independent: the physical properties available to the agents in the environment, restricted to the regions of space it would sample at time $t$ and $t+1$ by following a motor trajectory $\gamma_\mathbf{b} \in \mathscr{B}'_\mathcal{T}$, do not depend on each other. It then follows that $\gamma_\mathbf{s}(t+1) = \gamma_{\gamma_\mathbf{b}(t+1),\epsilon|F(\gamma_\mathbf{b}(t+1))}(t+1)$ and $\gamma_\mathbf{s}(t) = \gamma_{\gamma_\mathbf{b}(t),\epsilon|F(\gamma_\mathbf{b}(t))}(t)$ are independent themselves. As a result, we have

$$\begin{aligned}P^{\mathscr{B}'_\mathcal{T}}_{\mathbf{s},\mathbf{s}'} &= \{\gamma_\mathbf{s}(t+1) = \mathbf{s}' \mid \gamma_\mathbf{s}(t) = \mathbf{s}, \gamma_\mathbf{b} \in \mathscr{B}'_\mathcal{T}\},\\ &= \{\gamma_\mathbf{s}(t+1) = \mathbf{s}' \mid \gamma_\mathbf{b} \in \mathscr{B}'_\mathcal{T}\}.\end{aligned} \tag{22}$$

Thus, the probability $P^{\mathscr{B}'_\mathcal{T}}_{\mathbf{s},\mathbf{s}'}$ –where $\mathscr{B}'_\mathcal{T}$ is a motor trajectory for which the coherence properties mentioned before are not verified– does not depend on previous sensory output $\mathbf{s}$ anymore; instead, it simply replicates the *un*conditional probability that the agent experience the particular sensory value $\mathbf{s}'$. To the agent, this means that the knowledge of which sensation $\mathbf{s}$ it experiences at timestep $t$ does not give it *any* information on which sensation $\mathbf{s}'$ it is poised to experience at $t+1$. Importantly, this shows that suitable choices of motor explorations are required for building a valid sensory metric, as well as giving an internal observation to assess whether this condition fails through Equation (22). The influence of this motor exploration is experimentally studied in the next subsection to illustrate these developments.

### B. An experimental assessment of the influence of the movement amplitude

We propose in this subsection to assess the effect movement of the agent has on the internal representation of sensory symbols through simple simulations, where an agent is now allowed to move in a fixed environment. To begin with, details about the experiment setup are given. Next, the resulting representations are analyzed and discussed.

*1) Experimental setup:* let us consider in all the following a very simple agent, whose body is made of a planar, rectangular, camera sat atop one actuator allowing the agent to only move in one direction, see Figure 4. The pixels of the camera are sensitive to the luminance of the ambient stimulus, which is a fixed grayscale image placed in front of the moving camera. In such a case, the ambient space $\mathcal{X}$ is then the plane $\mathbb{R}^2$, and the state of the environment is a function $\epsilon$ mapping a position

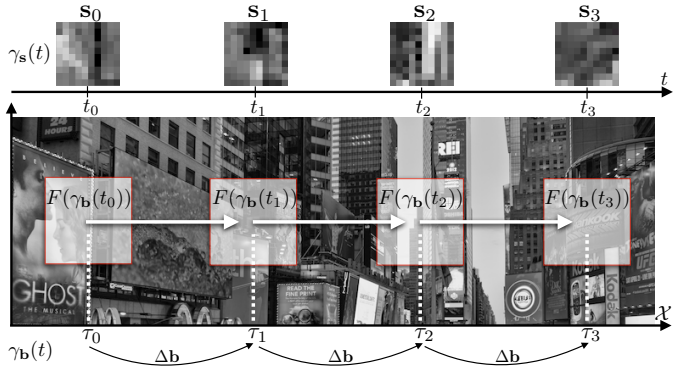



Figure 4: Experimental setup to assess the effect of the agent movement in the internal sensory symbol topology. A camera, whose field of view –or receptive field– is drawn as a square with red borders, faces a grayscale image and moves from one position $\tau$ to another thanks to an action of amplitude $\Delta\mathbf{b}$. The corresponding sensory states $\mathbf{s}_t$ are then captured along time to build the statistics of the sensory symbol transitions.

$(x, y)$ in the plane to luminance values $\epsilon(x, y) \in [\![0; 255]\!]$ as encoded in the grayscale image. Those values are then converted into a sensory vector $\mathbf{s} \in [\![0; 255]\!]^{W \times H}$ directly capturing the corresponding grayscale value in the environment (the function $h()$ in Equation (15) is thus the identity function). In the forthcoming simulations, $W = H = 100$. The agent is able to move in its environment by applying a single *action* $a$ [16], i.e. by applying a function $a$ to its current absolute configuration $\mathbf{b} = (\mathbf{m}, \tau)$ to go to another configuration $\mathbf{b}' = (\mathbf{m}', \tau')$. In this section, we will mainly study the influence of the amplitude $\Delta\mathbf{b}$ of this action, which is supposed to produce a movement of the camera in only one direction and with the same amplitude, as illustrated in Figure 4. This is obviously a very particular and restrictive action, at least in comparison with the more generic motor actions framework presented by the authors in [16], but it will still allow a comprehensive study of the effect of movement on the internal representation built by the agent. The different action amplitudes $\Delta\mathbf{b}$ used in the simulations will all be equal to a multiple of $\sigma_{\mathbf{b}}$, a particular amplitude which causes a shift of the perceived information in $\mathbf{s}$ of exactly 1 pixel. This actually corresponds to a displacement of the camera receptive field $F(\mathbf{b})$ in $\mathcal{X}$ (represented as squares with red borders in Figure 4) of the width of 1 pixel in the plane supporting the grayscale image.

In practice, the experiment is conducted the following way. To begin with, the environment observed by the agent is a grayscale image of a crowded street, partially shown in Figure 4. Then, starting from a fixed (random) position $\tau_0$ in the environment, the agent follows a motor trajectory $\gamma_{\mathbf{b}}(t)$ made of jumps of fixed amplitude $\Delta\mathbf{b}$. This produces a displacement of the agent sensor receptive field in the environment at which the agent gathers samples $\mathbf{s}_t$ of its corresponding sensory trajectory $\gamma_{\mathbf{s}}(t)$. After having generated $N_a$ times the same action $a$, the camera is put in one other random position in the image; then, the action $a$ is used again to move the camera $N_a$ times in the image. This process is repeated $N_r$ times, so that $N_r \times N_a$ sensory samples $\mathbf{s}_t$ are collected. These samples then allow one to build the matrix $P$ as in Equation (12). Then, the corresponding MDS projection of the distance matrix $D$ can be computed to visualize the captured sensory symbols topology. The experience is finally repeated for various amplitudes $\Delta\mathbf{b}$.

*2) Results and discussions:* The experiment has been conducted for $N_a = 500$ and $N_r = 100$, so that $50.10^3$ sensory transitions are used to build the matrix $P$ for each action amplitude $\Delta\mathbf{b}$ chosen among $\{\sigma_{\mathbf{b}}, 25\sigma_{\mathbf{b}}, 250\sigma_{\mathbf{b}}, 1000\sigma_{\mathbf{b}}\}$. Note that being greater than the size $N_s = 100$ of the sensor, $\Delta\mathbf{b} = 250\sigma_{\mathbf{b}}$ leads to the sampling of an area in the environment that does not overlap with its previous receptive field position. Figure 5 represents successively (in each row) the matrices $P, D$ and its corresponding embedding $\texttt{MDS}_2(D)$ for each of the 4 selected amplitudes (in each column). Let us first consider the evolution of the probability matrix as a function of the movement amplitude (first row). For a very small action amplitude, the probability matrix $P$ exhibits a clear diagonal pattern indicating that close sensory symbols (in terms of gray levels) correspond to high transition probabilities; qualitatively, one faces here more or less the same conditions than in §III-B1 where the observation of the environment changes matches the changes in perception induced by action of the agent. These two scenarios no longer correspond when the action amplitude rises: the higher the amplitude, the wider the probability distribution. For the largest amplitude, the diagonal pattern cannot even be seen anymore in $P$ and high probabilities do not correspond to close gray levels anymore. This tendency is clearly confirmed when computing the distance matrix $D$ from $P$ (2nd row in Figure 5): for high amplitudes, mostly all symbols are now close to each other. Obviously, this results in very different $2D$ projections of the matrix $D$ (3rd row). For lowest amplitudes, one still clearly sees a one-dimensional manifold, folding on itself when the action amplitude grows. But the dimensionality of the manifold is not sufficient to tell if the agent correctly captured or not the sensory symbols topology. Like we did in Figure 1d, a *k*-NN algorithm is computed on $D$ and displayed in Figure 5 to link each symbol to its closest neighbor, this link being represented as an arrow between two symbols in the projection. Looking at the smallest amplitude, the sensory symbols manifold can be browsed in the usual grayscale order by following the aforementioned arrows. On the contrary this proves impossible for larger amplitudes, where arrows link together e.g. symbols associated to clear and dark gray levels. It is then clear that the conditions written as Equations (19) and (18) are not verified anymore, with two successively sampled environment states associated to two distant positions in space, leading to the loss of perception by the agent of the spatial and temporal coherence in the environment.

Equation (22) states that, for a specific set of motor trajectories $\mathscr{B}'_{\mathscr{T}}$ making successive receptive fields falling apart from one other, the probability of transition between successive sensory symbols tends to an *un*conditional probability $P^{\mathscr{B}'_{\mathscr{T}}}_{\mathbf{s},\mathbf{s}'} = P_{\mathbf{s}'}$. Importantly, this phenomenon can be *internally* assessed by the agent since both probability distributions are only based on sensory symbols observations; this then constitute some *internal* way for the agent to rate the spatial and temporal coherence of its interaction with the environment.

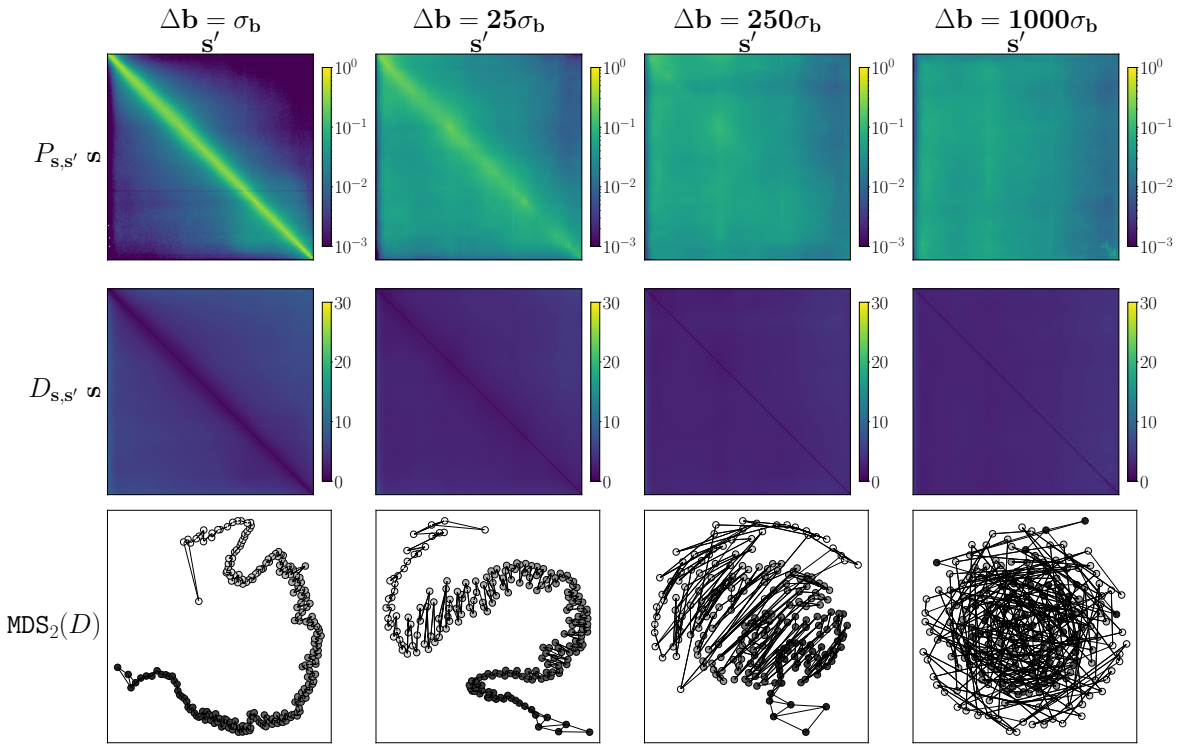



Figure 5: Evolution of the transition probability measure $P$ (displayed with logarithmic norm), the distance measure $D$ and the representation of this distance projected using 2-dimensional MDS for increasing movement amplitudes. Each column represents a motor trajectory for a fixed amplitude $\Delta\mathbf{b}$ described relatively to a 1 pixel shift of its sensor's field of view. One can see that the diagonal pattern for $P$ and $D$ as well as the uni-dimensional grayscale manifold are deteriorating as the movement amplitude get bigger indicating the inability to capture spatial coherence properties. The links that connect each symbol on the MDS representation are a $k$-NN like algorithm that assess how the agent perceive its symbol continuity.

To that end, we propose to compare the two probability distributions $P_{\mathbf{s}=s_i,\mathbf{s}'}$, the probability of every sensory value to succeed to a specific sensory value $s_i$, and $P_{\mathbf{s}'}$ by using the Jensen-Shannon distance $D_{\text{JS}}$ [25], a metric based on the Kullback–Leibler divergence [26] defined along

$$D_{\text{JS}}(P_{\mathbf{s}=s_i,\mathbf{s}'} \| P_{\mathbf{s}'}) = \sqrt{\frac{\text{KL}(P_{\mathbf{s}=s_i,\mathbf{s}'} \| M) + \text{KL}(P_{\mathbf{s}'} \| M)}{2}},$$
$$\text{with } M = \frac{1}{2}\left(P_{\mathbf{s}=s_i,\mathbf{s}'} + P_{\mathbf{s}'}\right), \tag{23}$$

with the KL divergence for two discrete probabilistic distributions $A$ and $B$ defined on the probability space $\mathcal{W}$ as

$$\text{KL}(A \| B) = \sum_{x \in \mathcal{W}} A(x) \log_2 \left(\frac{A(x)}{B(x)}\right).$$

This results in a distance $D_{\text{JS}}(s_i)$ between 0 and 1, computed for each sensory symbol $\mathbf{s}'$, that is expected to converge towards 0 when both distributions are identical, i.e. when the motor trajectory of the agent leads to having $\mathbf{s}$ and $\mathbf{s}'$ independent. Again, $D_{\text{JS}}$ is computed for 4 different amplitudes $\{\sigma_\mathbf{b}, 5\sigma_\mathbf{b}, 25\sigma_\mathbf{b}, 125\sigma_\mathbf{b}\}$ with corresponding graphs in Figure 6. The results displayed in Figure 6 show that the JS distance for every probability distribution systematically decreases when the amplitude of agent's movement increases, i.e. the conditional distribution tends towards the unconditional one as described by Eq. (22). This internal comparison between the two probability distributions could then provide the agent

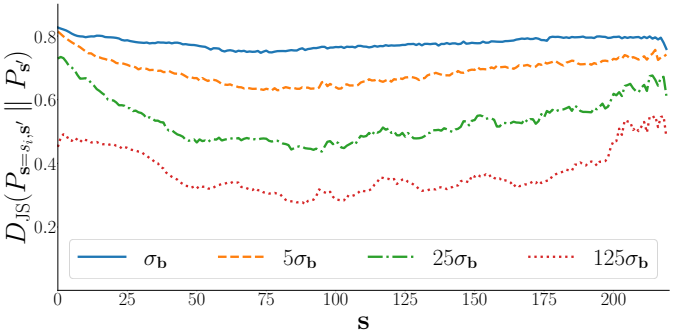



Figure 6: JS distance of every conditional probabilities relatively to the unconditional probability $P(\mathbf{s}')$ for different movement amplitudes. Each point of the plot represents a JS distance for a single conditional probability to $P(\mathbf{s}')$. As the amplitude increases the divergence of every symbol decreases, getting closer to the unconditional probability.

with a way to rate the adequation of its motor exploration performed by applying actions with (at least for now) unknown consequences, or even an internal signature of the amplitude of actions.

## V. Using the metric to get an internal assessment of sensory regularity

Now that we have been able to quantify how and why the agent action modulates its sensory symbol topology, let

us focus on a more experimental use of the obtained representation. Intuitively, and thanks to the introduction of the metric $\mathrm{d}_f$, the agent should now be able to assess if a sensory transition is typical or not. This could be used as a way to deal with the presence of noise in the raw sensory data, i.e. by being able to discriminate close (but not strictly equal) sensory values from irregular sensory transitions due to the presence of specific events in the environment (movement of an object in the scene, changes in the illumination conditions, etc.). This section thus aims to present how the agent could internally assess its sensory regularity by first depicting some simple formal elements in a first subsection. Then, a second subsection shows how a naive agent could actually be capable of performing a sensory prediction task, even in the presence of noise, in the vein of the sensorimotor action framework presented by the authors in [16].

### A. Internally rating the sensory regularity

*1) Some formal considerations:* to begin with, let us consider again Eq. (7) by which $\delta_f(\mathbf{s}, \mathbf{s}')$ is defined in terms of the sensory transition probabilities $P_{\mathbf{s},\mathbf{s}'}$. It can be trivially rewritten as

$$\forall \mathbf{s}, \mathbf{s}' \in \mathcal{S}, \mathbb{P}(\gamma_{\mathbf{s}}(t+1) = \mathbf{s}' \mid \gamma_{\mathbf{s}}(t) = \mathbf{s}) = f^{-1}(\delta_f(\mathbf{s}, \mathbf{s}')), \tag{24}$$

when $f$ is injective. But because $f$ is also necessarily non-increasing, so must $f^{-1}$ be; this obviously entails that the probability of any sensory transition from $\mathbf{s}$ to $\mathbf{s}'$ is as expected a decreasing function of the sensory distance between them. However, we also know from the definition of the metric $\mathrm{d}_f$ from shortest paths in Eq. (8) that

$$\forall \mathbf{s}, \mathbf{s}' \in \mathcal{S}, \mathrm{d}_f(\mathbf{s}, \mathbf{s}') \leq \delta_f(\mathbf{s}, \mathbf{s}'). \tag{25}$$

One then has immediately

$$\forall \mathbf{s}, \mathbf{s}' \in \mathcal{S}, \mathbb{P}(\gamma_{\mathbf{s}}(t+1) = \mathbf{s}' \mid \gamma_{\mathbf{s}}(t) = \mathbf{s}) \leq f^{-1}(\mathrm{d}_f(\mathbf{s}, \mathbf{s}')). \tag{26}$$

Then, Eq. (26) guarantees that, from any sensory value $\mathbf{s}$, the probability to land on $\mathbf{s}'$ at a distance $\mathrm{d}_f(\mathbf{s}, \mathbf{s}') = \lambda$ is therefore *less than* $f^{-1}(\lambda)$. This property thus gives an intrinsic way of quantifying the *regularity* of a transition in the sensory experience. Indeed, providing some "metric rejection threshold" $\tau_r$, the agent might be able to deem all sensory transitions $\mathbf{s}$ to $\mathbf{s}'$ of corresponding distance $\mathrm{d}_f(\mathbf{s}, \mathbf{s}')$ as irregular (resp. regular) whenever $\mathrm{d}_f(\mathbf{s}, \mathbf{s}') \geq \tau_r$ (resp. $\mathrm{d}_f(\mathbf{s}, \mathbf{s}') < \tau_r$).

Still, one should notice that Eq. (26) is merely an inequality, as opposed to the corresponding equality in Eq. (24). To the agent, this means that there may be some particular transitions from $\mathbf{s}$ to $\mathbf{s}'$ which are still unlikely even if the agent found $\mathrm{d}_f(\mathbf{s}, \mathbf{s}')$ to be small: basically, this criterion can allow *false positives*, while it guarantees that all transitions rejected on the basis of this metric verify the occurrence probability inequality, that is it does not cause *false negatives*.

*2) Example:* we selected in previous sections (see Eq. (13)) the function $f = -\log$ to map the estimated transition probabilities $p_{kl}$ to the metric prototype $\delta_{kl}$. In such a case, still with a threshold $\tau_r$, an irregular transition should then typically occur with a probability $\mathbb{P}(\gamma_{\mathbf{s}}(t+1) = \mathbf{s}' \mid \gamma_{\mathbf{s}}(t) = \mathbf{s}) \leq e^{-\tau_r}$. Then, selecting for instance the classical threshold values $\tau_r \in \{1, 3, 5\}$ will allow the agent to reject transitions that occur in less than about $\{37, 5, 1\}\%$ of occurrences.

### B. Exploiting the sensory regularity for sensory prediction

We now propose to exploit the agent's capability to decide whether a sensory transition is regular in a sensory prediction task in the presence of noise inside the sensory data. To that end, the experimental setup used in [16] is replicated and shortly recalled in a first paragraph. Then, the resulting sensory prediction function, taking the form of a permutation matrix between pixels of the agent's camera, is computed for different scenarios: (i) with no noise in the sensory data (exactly like in [16]), or with noise, but (ii) without or (iii) with rating the sensory regularity. A discussion comparing these scenarios is then proposed in a second paragraph.

*1) Experimental setup:* the proposed simulation setup is very close to the one already presented in §IV-B1. The agent is still made of a moving camera facing a fixed grayscale image, as shown in Figure 4. This time, the agent is endowed with a $W \times H = 10 \times 10$ sensor, and is now able to move in four orthogonal directions by applying 5 different actions $a_{id}$, $a_f$, $a_b$, $a_r$ and $a_l$ making the camera receptive field respectively stay still, move in the left, right, up or down directions in $\mathcal{X}$. Importantly, the agent has no clue about the incidence of a given action nor about their possible relationship. All it can do is perform an action and observe its consequence in its sensory data [16]. The proposed experimentation then relies on the two following steps.

*a) Step 1: building of the sensory symbol topology*. The agent explores its environment by randomly selecting an action in $\mathcal{A} = \{a_{\mathrm{id}}, a_f, a_b, a_r, a_l\}$ with identical amplitudes $\Delta\mathbf{b} = \sigma_{\mathbf{b}}$ (apart from $a_{\mathrm{id}}$), and then infers the distance matrix $D$, in line with §IV-B where the number $N_a$ of draws of actions is set to $N_a = 25$, and the number of repetition is selected to $N_r = 2.10^3$. As opposed to the previous case, some artificial noise $n_{ij}$ is now added to the pixel value $v_{ij}$ of the image to form the agent's sensel values[3] $s_{ij} = v_{ij} + n_{ij}$ before computing the matrix $D$, with $n_{ij}$ a random integer drawn from a centered discrete uniform distribution of width $2\sigma_n$.

*b) Step 2: building of the sensory prediction function*. Once the matrix $D$ is obtained, the agent performs a second exploration of its environment so as to build a sensory prediction function for each of its actions in $\mathcal{A}$. In practice, these functions take the form of binary permutation matrices [16] $\Pi_{a_p} = (\pi_{kl}^{(p)})_{k,l}$ of size $N_s \times N_s$, with $a_p \in \mathcal{A}$ and $N_s = W \times H$, as each pixel value in the sensory array is expected to shift in different positions depending on the spatial effect of the performed action. For this experiment, $N_a = 50.10^3$ and $N_r = 1$. Initially, every element $\pi_{kl}^{(p)}$ of the permutation matrices $\Pi_{a_p}$ is initialized to 1, meaning that all permutations between the agent sensels are possible for action $a_p$. Then, each time this action is drawn from $\mathcal{A}$, the agent can discard in $\Pi_{a_p}$ some permutations by observing that some

[3] which is further clamped if need be, i.e. if $s_{ij}$ exceeds 0 or 255, the sensel value is set to the closer bound.

sensels values do not switch with one others, then updating the corresponding matrix elements to 0 as per the update rule

$$\pi_{kl}^{(p)}[n+1] = \begin{cases} 1 \text{ iff } s_l[n] = s_k[n+1] \text{ and } \pi_{kl}^{(p)}[n] = 1 \\ 0 \text{ else,} \end{cases} \quad (27)$$

where $s_k$ and $s_l$ represent the sensel values associated to the element at the position $(k, l)$ in the permutation matrix $\Pi_{a_p}$. One can notice in Eq. (27) that the elements in these matrices are set to 0 as soon as a permutation is not detected by the strict equality between sensory values. This limitation, already outlined in [16], makes this approach fall apart when dealing with noise in the sensory data. Benefiting from the previous developments, one instead proposes a revised update rule of the permutation matrix along

$$\pi_{kl}^{(p)}[n+1] = \begin{cases} 1 \text{ iff } \mathrm{d}_f(s_l[n], s_k[n+1]) < \tau_r \text{ and } \pi_{kl}^{(p)}[n] = 1 \\ 0 \text{ else,} \end{cases} \quad (28)$$

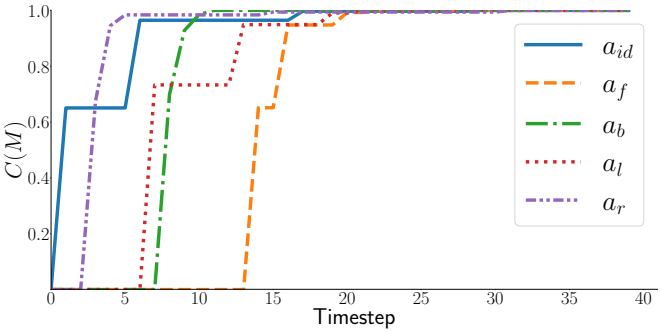
where $\tau_r$ is a manually chosen threshold applying on the built matrix distance $D$. In the following, $\tau_r$ is tuned so as to correspond to the smallest threshold that allows for permutations matrices to converge. It is clear that this a strong *a priori*, and the way the agent can autonomously set this threshold is still an ongoing work, discussed in the conclusion.

*c) Evaluation: convergence of the permutation matrices.* To evaluate the influence of the added noise on the convergence of permutation matrices $\Pi_{a_p}$, one proposes an (external) criterion $C(\Pi_{a_p}) = C_H(\Pi_{a_p}) \times C_D(\Pi_{a_p})$ adapted from [16] to account for the added noise to the data and defined along

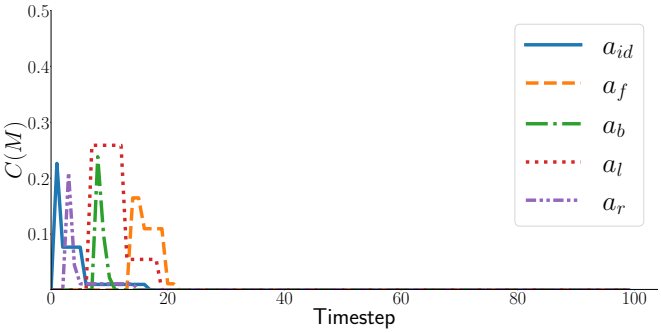
$$\begin{aligned} C_D(\Pi_{a_p}) &= \frac{\sum_{kl} \pi_{kl}^{(p)} \bar{\pi}_{kl}^{(p)}}{\sum_{kl} \bar{\pi}_{kl}^{(p)}}, \text{ and} \\ C_H(\Pi_{a_p}) &= 1 - \frac{1}{N_s \log_2(N_s)} \sum_{i=1}^{N_s} H_i, \\ \text{with } H_i &= -\sum_{l=1}^{N_s} \frac{\pi_{kl}^{(p)}}{\mu_k} \log_2 \left( \frac{\pi_{kl}^{(p)}}{\mu_k} \right), \\ \text{and } \mu_k &= \max \left( 1, \sum_{l=1}^{N_s} \pi_{kl}^{(p)} \right), \end{aligned} \quad (29)$$

where $\bar{\pi}_{kl}^{(p)}$ represents the (binary) coefficients of the ideal matrix $\bar{\Pi}_{a_p}$ associated to the action $a_p$. Basically, $C_H$ can be understood as an average measure of certainty in the discovery of the permutations, weighted by the percentage $C_D$ of the correctly identified permutations w.r.t. the ground truth to account for the noise possibly discarding some of them. In the end, criterion $C$ lies between 0 –i.e. the matrix is full of 1's (initialization) or 0's (all permutations have been discarded)– and 1 –i.e. the permutation has been correctly discovered–.

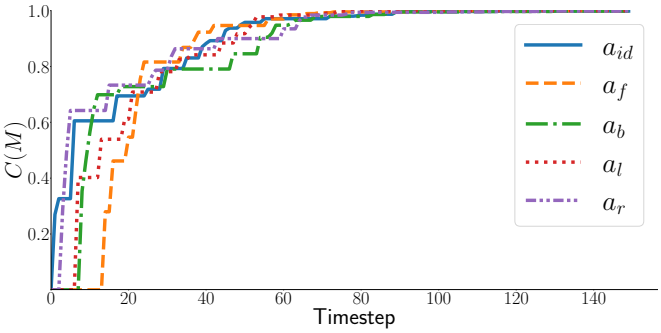
*2) Results:* As outlined in the introduction of Section V, three different scenarios are evaluated. To be begin with, one first considers the case where there is no noise in the agent's perception by setting $\sigma_n = 0$. Then, using the update rule (27) should allow the agent to correctly build all of its permutation matrices, exactly as in in [16]. As expected, Figure 7a shows that criterion $C$ converges towards its maximal value 1 for all actions in $\mathcal{A}$. $C$ plots also exhibit sparse jumps at random




(a) Evaluation criterion $C$ with $\sigma_n = 0$ and strict equality update rule.




(b) Evaluation criterion $C$ with $\sigma_n = 2$ and strict equality update rule.




(c) Evaluation criterion $C$ with $\sigma_n = 2$ and a threshold in $D$.

Figure 7: Evolution of the evaluation criterion $C$ for the 5 considered actions in $\mathcal{A}$. (a) With no noise and the update rule (27), $C$ converges towards 1 in a very short number of realization of each action. (b) In the same scenario, but with $\sigma_n = 2$, the update rule (27) do not allow to detect permutations anymore, resulting in the criterion falling down to 0. (c) When selecting a correct threshold $\tau_r$ in Eq. (28), the agent is now able to build the 5 sensory prediction functions correctly, but with more realization of each action in comparison with (a).

times, corresponding to the steps where the action was actually drawn in $\mathcal{A}$ during the experiment. More importantly, one can see in Figure 7a that only a few realizations of each action $a_p$ (about 4 to 6 here) is required for $C(\Pi_{a_p})$ to almost reach 1, showing how easy it is for the agent to discover the existence of such permutations in its own perception. In the second scenario, a noise of amplitude $\sigma_n = 2$ is now added to the sensation. Obviously the strict comparison of sensel values in (27) in the presence of such noise (however small) entirely breaks the approach , as shown in Figure 7b. As expected, the criterion $C$ now converges to 0: each $\Pi_{a_p}$ matrices converges to null matrices as all possible permutations of values have been (including erroneously) discarded in the process. Finally,

the new update rule (28) is now used to judge of the closeness of sensels values on the basis on the built distance $D$, resulting in the evolution of the criterion $C$ represented in Figure 7c. For this scenario, $\sigma_b = 1$ and $\tau_r = 1.63$. In the presence of noise, the ability for the agent to assess if a sensation is now close to others allows it to correctly discover the existence of permutations in its perception. But clearly, this task is not as easy as in the first scenario: the number of required actions for correctly evaluating their corresponding permutation matrices is significantly higher. This is apparent in Figure 7c, not only in the slower convergence time of the criterion $C$, but also in the smaller jumps of values in $C$. Indeed, each generation of action brings less information in the prediction process because of the noise included in the agent sensation. But still, the important structures anchoring the sensorimotor interaction the agent has with its environment are still available, allowing it e.g. to build an image of its body [14] or of its peripersonal space [15], at least at the cost of a longer interaction in time.

## VI. Conclusion

In this paper, and after purely topological considerations, a metric-based approach is proposed to formalize the ability for a naive agent to build some subjective sense of sensory continuity. An experimental framework is then proposed, illustrated and assessed in the context of visual perception for the discovering of gray or color scales. Then the importance of the dynamic of the agent exploration relatively to that of the environment is studied, highlighting an important spatiotemporal coherence principle of this exploration. Finally, a sensory closeness notion being now available to the agent, a sensory prediction task is proved accessible even in the presence of noise, thus extending the robustness of this sensorimotor framework to realistic conditions.

Nevertheless, it is clear that this work still suffers from some limitations. For instance, the scalability of the proposed experimental framework is certainly limited. Indeed, although it was not the objective of this paper, the way the regularities are extracted from the raw sensations is certainly not computationally effective, considering the possibly very high number of sensory symbols involved in e.g. color perception for traditional camera sensors. Hierarchical approaches might be preferred [27], but still remain to be explored in the context of sensorimotor approaches to perception. Another limit concerns the notion of sensory neighbors: while being now formally accessible to the agent thanks to the proposed contribution, it still practically requires a threshold to be set w.r.t. the task to be performed. In this paper this threshold has been manually tuned, but one could instead rely on a closed-loop approach mixing the discovery of the sensory regularities with the corresponding sensory prediction task: as long as the prediction is not correctly built, the threshold must be adapted accordingly. Still, should the agent be able to perform some sensory prediction task, so should it be able to *quantitatively* compare its prediction with its actual perception. This should make it capable of detecting outliers in its environment, and in particular changes in its perception which are not directly correlated to its own action. This might be the way towards some internal notion of sensorimotor objects, and thus would undoubtedly extend the scope of these approaches to more potential applications.

## References


[1] B. Dainton, "Temporal Consciousness," in *The Stanford Encyclopedia of Philosophy*, Winter 2018 ed., E. N. Zalta, Ed. Metaphysics Research Lab, Stanford University, 2018.

[2] J. M. B. Fugate, "Categorical perception for emotional faces," *Emotion Review*, vol. 5, no. 1, pp. 84–89, 2013.

[3] J. C. Toscano, B. McMurray, J. Dennhardt, and S. J. Luck, "Continuous perception and graded categorization: Electrophysiological evidence for a linear relationship between the acoustic signal and perceptual encoding of speech," *Psychological Science*, vol. 21, no. 10, pp. 1532–1540, 2010.

[4] A. Herwig, "Transsaccadic integration and perceptual continuity," *Journal of Vision*, vol. 15, no. 16, pp. 7–7, 12 2015.

[5] J. M. Stroud, "The fine structure of psychological time," *Annals of the New York Academy of Sciences*, vol. 138, no. 2, pp. 623–631, 1967.

[6] R. VanRullen and C. Koch, "Is perception discrete or continuous?" *Trends in Cognitive Sciences*, vol. 7, no. 5, pp. 207–213, 2003.

[7] J. O'Regan, *Why Red Doesn't Sound Like a Bell: Understanding the feel of consciousness*. Oxford University Press, 2011.

[8] S. A. Morris, *Topology without tears*, 2020.

[9] A. Censi, "Bootstrapping vehicles: A formal approach to unsupervised sensorimotor learning based on invariance," Ph.D. dissertation, California Institute of Technology, June 2012.

[10] D. Philipona, J. K. O'Regan, and J.-P. Nadal, "Is there something out there?: Inferring space from sensorimotor dependencies," *Neural Comput.*, vol. 15, no. 9, pp. 2029–2049, 2003.

[11] A. Laflaquière, J. K. O'Regan, S. Argentieri, B. Gas, and A. Terekhov, "Learning agents spatial configuration from sensorimotor invariants," *Robotics and Autonomous Systems*, vol. 71, pp. 49–59, September 2015.

[12] A. Laflaquiere, "Grounding the experience of a visual field through sensorimotor contingencies," *Neurocomputing*, vol. 268, no. C, 2017.

[13] A. V. Terekhov and J. K. O'Regan, "Space as an invention of active agents," *Frontiers in Robotics and AI*, vol. 3, p. 4, 2016. [Online]. Available: https://www.frontiersin.org/article/10.3389/frobt.2016.00004

[14] V. Marcel, S. Argentieri, and B. Gas, "Building a sensorimotor representation of a naive agent's tactile space," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 9, no. 2, pp. 141–152, June 2017.

[15] V. Marcel, S. Argentieri, and B. Gas, "Where do i move my sensors? emergence of a topological representation of sensors poses from the sensorimotor flow," *IEEE Transactions on Cognitive and Developmental Systems*, pp. 1–1, 2019.

[16] J.-M. Godon, S. Argentieri, and B. Gas, "A formal account of structuring motor actions with sensory prediction for a naive agent," *Frontiers in Robotics and AI*, vol. 7, p. 179, 2020. [Online]. Available: https://www.frontiersin.org/article/10.3389/frobt.2020.561660

[17] A. Laflaquière, S. Argentieri, O. Breysse, S. Genet, and B. Gas, "A non-linear approach to space dimension perception by a naive agent," in *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, Oct 2012, pp. 3253–3259.

[18] J. L. Elman, "Finding structure in time," *Cognitive Science*, vol. 14, no. 2, pp. 179–211, 1990. [Online]. Available: https://www.sciencedirect.com/science/article/pii/036402139090002E

[19] J.-M. Godon, "A structuralist formal account for sensorimotor contingencies in perception," Ph.D. dissertation, Sorbonne Université, 2022.

[20] E. W. Dijkstra *et al.*, "A note on two problems in connexion with graphs," *Numerische mathematik*, vol. 1, no. 1, pp. 269–271, 1959.

[21] J. B. Kruskal, "Nonmetric multidimensional scaling: A numerical method," *Psychometrika*, vol. 29, no. 2, pp. 115–129, 1964.

[22] ——, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, no. 1, pp. 1–27, 1964.

[23] I. Borg and P. J. Groenen, *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

[24] J. B. Tenenbaum, V. d. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[25] D. Endres and J. Schindelin, "A new metric for probability distributions," *IEEE Transactions on Information Theory*, vol. 49, no. 7, pp. 1858–1860, 2003.

[26] S. Kullback and R. A. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, March 1951.

[27] D. H. Ballard and R. Zhang, "The hierarchical evolution in human vision modeling," *Topics in Cognitive Science*, vol. 13, no. 2, pp. 309–328, 2021.