Combining Manifold Learning and Neural Field Dynamics for Multimodal Fusion

Simon Forest, Jean-Charles Quinton, Mathieu Lefort

Combining Manifold Learning and Neural Field Dynamics for Multimodal Fusion

Simon Forest*†, Jean-Charles Quinton*, Mathieu Lefort†
*Univ. Grenoble Alpes, CNRS, Grenoble INP, LJK, UMR 5224, F-38000, Grenoble, France
†Univ. Lyon, UCBL, CNRS, INSA Lyon, LIRIS, UMR 5205, F-69622, Villeurbanne, France
{simon.forest, quintonj}@univ-grenoble-alpes.fr, mathieu.lefort@univ-lyon1.fr

***Abstract*—For interactivity and cost-efficiency purposes, both biological and artificial agents (e.g., robots) usually rely on sets of complementary sensors. Each sensor samples information from only a subset of the environment, with both the subset and the precision of signals varying through time depending on the agent-environment configuration. Agents must therefore perform multimodal fusion to select and filter relevant information by contrasting the shortcomings and redundancies of different modalities. For that purpose, we propose to combine a classical off-the-shelf manifold learning algorithm with dynamic neural fields (DNF), a training-free bio-inspired model of competition amid topologically-encoded information. Through the adaptation of DNF to irregular multimodal topologies, this coupling exhibits interesting properties, promising reliable localizations enhanced by the selection and attentional capabilities of DNF. In particular, the application of our method to audiovisual datasets (with direct ties to either psychophysics or robotics) shows merged perceptions relying on the spatially-dependent precision of each modality, and robustness to irrelevant features.**

***Index Terms*—multimodal fusion, growing neural gas, manifold learning, dynamic neural field, selective attention**

I. Introduction

When it comes to information processing and behavioral decision-making, the way we merge data coming from inputs of mixed nature is becoming increasingly important. Let us start with a toy example. A robot is given a task, for example: "touch the alarm clock when it goes off". At first, the robot might be facing several objects resembling an alarm clock, which it should have no difficulty distinguishing. When a sound goes off, the robot should be able to locate its origin, but it is usually achieved with a low precision. Before taking an action, the robot has to select an object. Here, it should be the one clock-looking object that coincides most with the sound source localization. But how the modalities should be weighted depends not only on the task (a clock visible on the front has lower priority than sound coming from the side), but also on the reliability of the sensors (room reverbation can make sound orientation irrelevant).

The task in this example faces multiple challenges, starting with two: the fusion of sensory modalities of different availability and reliability, and the selection of (and attention towards) a target. To tackle these problems, most of model nowadays are based on deep learning. In this article, we propose another approach based on dynamic neural fields (DNF), a bio-inspired model of neural activity [1]. It is a topologically-grounded continuous-time recurrent network, where weights are known and depend on the distance between neurons. With a mixture of short-range excitation and long-range inhibition, input stimuli are put in competition until a bubble of activity emerges, which can be interpreted as a decision of target selection and/or action. Additionally, temporal dynamics allows the bubble to remain stable despite input fluctuations and robust to potential distractors. DNF have seen various applications, including in robotics. In particular, the interaction properties of DNF make them very suitable for multimodal fusion [2], [3].

One limit that previous DNF implementations have faced lies in the nature of the manifold they evolve on. Most applications in the literature assume the existence of an underlying regular topology, most often 1D or 2D. But it is hardly representative of the disparities in the sensory space, disparities which become crucial when performing multimodal fusion. Indeed, let us take a look at the shape of stimuli perceived from the environment. The quantity of information available is huge, and the data an agent receives from its sensors is only a projection of it in a few given dimensions. Equipped with a standard camera, a robot will receive a 2D projection of the part of the environment it is facing. With one microphone, it can detect sounds from anywhere around it, but it can hardly locate them. Two microphones may enable some 1D sound localization along the axis on which they are aligned, usually azimuthal (with interaural time/level difference), and even a bit of 2D or 3D by exploiting the shape of pinnae with a head-related transfer function (HRTF) [4]. We must first account for the specificities of each sensory modality before we create behaviors that exploit it at best. Additionally, we must find a way to match complementary information from different modalities, which usually boils down to projecting stimuli onto a common manifold.

So, our first step will consist in learning unimodal manifolds. For this purpose, we will use growing neural gas (GNG) [5], a standard manifold learning algorithm which is quite parcimonious in light of the possible complexity of the sensory space. Then, we will suggest an easy-to-implement solution to create a multimodal manifold suitable for fusion. The main novelty of our work is that we will run DNF directly on this new topology, even though it lacks the regularity and low dimensionality of classical implementations. We will show that

This work was funded by French region Auvergne-Rhône-Alpes as part of the project AMPLIFIER.

properties of DNF in selection and attention are compatible with such fabricated manifolds, and that this coupling allows new possibilities for multimodal fusion taking into account the relative resolution of the modalities.

Our article is structured as follows. In section II, we will review the existing literature on manifold learning and DNF, and in particular their applications to multimodal fusion. Then we will describe our model in section III, and demonstrate its capabilities through three applications in section IV. We will conclude and discuss additional perspectives in section V.

II. Previous Work

A. Manifold Learning

Sensors provide high-dimensional samples of the environment, but sensory spaces can often be projected onto manifolds of lower dimension. Deep learning methods are particularly suited for learning such manifold (see [6] for a review). For example, the last layers of a deep neural network have been shown to contain an intrinsic dimensionality that is smaller than the number of features in the data [7]. Dedicated methods such as variational autoencoders [8] learn structured embedding in an unsupervised manner. As our focus in this article is the study of coupling between DNF and irregular multimodal manifold, we will use simpler methods (i.e. self-organizing neural networks) that will provide more control and insight for the study.

In self-organizing maps (SOM), e.g. the Kohonen model [9], each neuron represents a prototypical input in the high-dimensional sensory space, so that the input space is projected onto a neural lattice of fixed shape and size. In neural gas (NG) [10], neurons are not arranged on a lattice, but are connected following a Hebbian rule, thus neurons with close prototypes are linked together. Eventually, the gas fills the input space in a way that matches the stimulus distribution. Growing neural gas (GNG) [5] is a derivative of NG, in which neurons are added (or deleted) over time until a chosen condition is met, thus adapting to the unknown input space spread.

Manifolds in multimodal fusion: Numerous articles have shown promising results in multimodal fusion using deep learning. Deep unsupervised learning can be used to project multimodal data on a low-dimensional manifold for use in robotics [11]. Inputs can be mixed during neural network training to exploit the correlations between modalities [12]. Reference [13] proposes a new type of deep neural network receiving multimodal inputs allocated through an attention module. Unfortunately, most of these works make the assumption that all multimodal data are related. Also, deep architecture are dedicated to one specific task and no generic architecture emerges [14].

We aim to create a new multimodal topology over which new dynamic properties could be applied, and self-organization offers solutions for a much lower cost [15]–[23]. SOM and their derivatives have long been used as models of multimodal fusion, but the ways modalities are combined can be very diverse. Map architectures can be divided in two categories. In the first, one SOM is trained for each modality, then all unimodal maps are connected depending on a special learning rule [15]–[17]. In the second, unimodal maps link to a new multimodal SOM [18], [19] or NG [20] that combines all information. Additional layers of SOM can also be considered to create a hierarchical flow of information [21]–[23]. Additionally, models can be made more adaptive to time-dependant tasks with the help of "growing when required" maps [22], [23], an alternative to GNG designed for dynamic input distributions [24]. Some of these models have already been proof-tested for visual, auditory and/or proprioceptive modalities on hardware setups [21], [23] and robots [17], [19].

After multimodal maps and/or interconnected unimodal maps have been learned, we need a paradigm to dictate the way perception will occur. Multimodal perception can be seen as a form of decision pondering sensory inputs of different reliability and relevance. We follow the architectural choice made in [18] and [15], where dynamic neural fields (DNF) are used as the paradigm that governs fusion or segregation of stimuli in the multimodal topological space. DNF come with many useful properties for multimodal perception.

B. Dynamic Neural Fields

Originally stemming from neuroscience, DNF have various applications in robotics [25]. For example, visual attention may be cumulated with motor control to make a robot autonomously gaze at objects in its environment and learn a sensorimotor map [26]. DNF rely on a population of topologically connected units at a mesoscopic scale, where the apparent activity (or average membrane potential over assemblies of neurons) can be read to infer decisions at a behavioral level. The activity evolves through time depending on a sum of external stimulations and lateral interaction between neurons. Stimulated neurons will send strong excitation to their nearest neighbors, and moderate inhibition to neighbors located further apart, leading to the emergence of a stable bubble of activity. Depending on the parametrization, this can lead to several types of behavior [25]. With strong local excitation, the bubble can be self-sustaining, acting as long-term memory [26]. Long-range inhibition will create a competition between conflicting stimuli, until either one dominates the others, or they are merged in a single bubble at an interpolated position [3], [27]. Then, the self-maintaining bubble can be used for robust selective attention, able to ignore noise and minor distractors [28]. Ultimately, the output of DNF can be directly exploited to generate motor command [26], [29].

The properties of DNF can benefit greatly to multimodal fusion. It provides the tools not only to enhance robust decisions when modalities are congruent [2], but also to solve conflicts between modalities [3]. This is where the choice of the underlying manifold can be very important.

The vast majority of works using DNF assume the dynamics take place on a completely regular topology, e.g. a 2D lattice in the case of vision. However, there is no clear way of projecting two or more modalities onto the same lattice. In [2] and [3], strong assumptions are made on the shape of stimuli in a modality so that they fit in the topology of the other. To

tackle this issue, [15] proposes using separate manifolds for each modality, each learned by SOM, and apply DNF on each of them. Communication between modalities is ensured by a specific set of topographic connections.

The latter reference is actually one of the first to suggest using a learned manifold as the theater of neural dynamics. Otherwise, some attempts to alter the projection of inputs into the manifold have lead to satisfying results: [27] and [3] successfully reproduce biological behaviors after applying a logpolar transformation to visual stimuli, which models the discrepancies in the resolution of the human retina [30]. In [15], the projections received by neurons are altered, although they are still organized in a rectangular lattice. Since DNF are strongly dependant to the topography, and rely on a symmetrical interaction kernel[1], one may fear that breaking the regularity of the underlying topology may make DNF completely unpredictable.

An ensuing question would be how far from regular and/or rectangular can the underlying topology be for DNF to remain viable. If DNF could be made to operate on manifolds of unconstrained shape or dimension (easily accessible through GNG), then this would open the door to adding the properties of DNF to a new range of applications, starting with new capabilities in multimodal fusion like the ability to take into account the different resolution and reliability of all modalities. To our knowledge, this has not been tested. At best, suggestions have been made to approximate DNF activity using gaussian mixtures, sparsifying the space on which they operate to make them applicable in more complex topologies [33]. Yet, this latter approach still relies on a continuous regular space on which the lateral connectivity kernel function and Gaussian mixtures can be defined, which remains a strong limitation when processing high dimensional inputs.

III. Model

In this article, we use GNG to learn manifolds of the sensory space in each modality. We then assemble them into one multimodal graph, on which we use a DNF to produce behaviors that have, to our knowledge, never been implemented on this kind of manifold. These three steps are summarized in figure 1 and explained below.

A. Unimodal Topology Learning

In this part, we process modalities separately. As our focus in this article is not on tuning the unimodal topology learning on a specific task, we use the standard GNG algorithm with its original parameter values, as described in [5]. To summarize, GNG are trained by receiving a succession of randomly selected stimuli. Every time, the two neurons whose prototypical input match the stimulus best get a fresh connection. Then the best-matching unit (BMU) and its direct topological neighbors have their prototype moved towards the stimulus. Connections that have not been updated in a long time are removed, and isolated neurons as well. Then at fixed intervals, a new neuron is inserted. Its prototypical input is placed at the middle of the most activated connection.

[1]There have been suggestions to break the symmetry from the DNF side, either through asymmetrical kernels [31] or through distortions of the topology by predictive reinforcements [32], but both require an additional learning step.

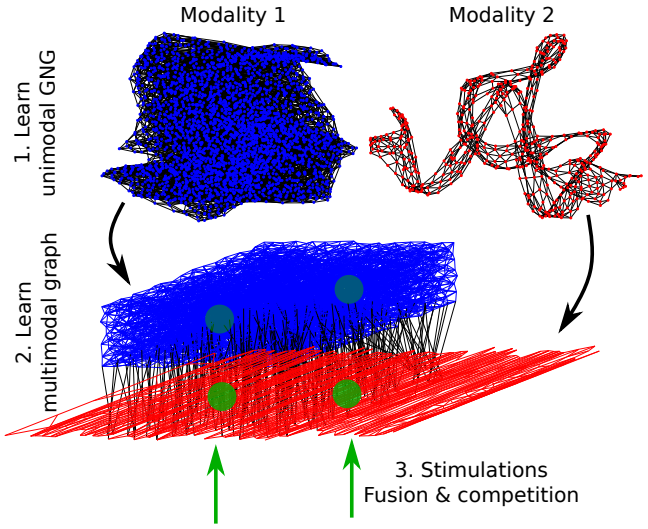

Fig. 1. Recap of the steps taken in this article. 1. Learn a growing neural gas in each modality. 2. Assemble them into one single graph by creating multimodal connections. 3. Present stimuli and compute multimodal activity.

B. Multimodal Topology Learning

For a first milestone, we will focus on bimodal architectures in the rest of this article. As a reminder, bimodal architectures in self-organization literature often merge data in one of two ways: a multimodal map is created that receives information from the unimodal ones, or new connections are added between the unimodal maps, each having its own processing unit. We propose an intermediate solution that is the most economical of all: we create a new bimodal graph that contains all nodes and edges from one modality, and all nodes and edges from the other. To create the crossmodal edges, we connect neurons of the two modalities that fire together, which is similar to an Hebbian learning. More precisely, the algorithm is: We draw a random multimodal input. If it lies in the sensory range of both modalities, we find the BMU in each GNG and connect them (if they are not already connected). We repeat until a certain proportion of nodes have at least one crossmodal edge.

C. Selection of Activity

Once the bimodal graph is created, its associated neurons can be stimulated by sensory inputs (through their respective modality), and we can use DNF to select and attend to a stimulus. DNF are usually expressed as an integro-differential equation in a continuous field of neurons, that is later discretized and computed using the Euler method. The integration of DNF is comparable to the simulation of continuous-time recurrent neural networks. In DNF, the distance between neurons plays an important role, as it determines whether they will excite or inhibit one another. Our model differs from others in the literature in that all neurons do not share a

common coordinate system. So, we need to adapt the DNF equation, so that the distances are defined on the graph, and only that. We rely on the standard distance from graph theory, i.e. the number of edges on the shortest path between any two vertices.

In our model, each neuron is tied to a specific modality. So, the external input received individually will be modality-specific (although the rest of DNF operations will not be). To ensure that the total amount of external stimulation is independant from the local resolution of a modality, we will order all neurons of a modality by their proximity to the stimulus (using the euclidian distance in the coordinate system of that modality), and stimulate them descendingly according to their rank. For each neuron indexed k, given a stimulus indexed i, we note $r_{k,i}$ the rank of proximity between the prototypical input of k and the coordinates of i. The external stimulation I_k received by k is given by:

$$I_k = \lambda_{m,i}\, \mathrm{e}^{\frac{-r_{k,i}^2}{2\sigma_I^2}} \tag{1}$$

where $\lambda_{m,i}$ is the intensity of stimulus i with regards to k's modality m. A neuron can only receive external inputs from its own modality.

Next, we compute the evolution of activity in the graph over time. The following is completely modality-agnostic. The potential U_k of neuron k is initialized as 0 and updated incrementally by[2]:

$$\Delta U_k = \frac{\Delta t}{\tau}\left(-U_k + I_k + \sum_{k'} W(<k,k'>)\, f(U_{k'}) + h\right) \tag{2}$$

where Δt is the simulated time between steps, τ a time constant that determines the speed of DNF updates, f an activation function (ReLU), and h a negative resting level. $<\cdot,\cdot>$ designates the minimal distance in number of edges between two nodes in the bimodal neural gas, and W is a weight function expressed as:

$$W(\delta) = \lambda_+\, \mathrm{e}^{\frac{-\delta^2}{2\sigma_+^2}} - \lambda_-\, \mathrm{e}^{\frac{-\delta^2}{2\sigma_-^2}} \tag{3}$$

with amplitudes $\lambda_+ > \lambda_- > 0$ and widths $\sigma_+ < \sigma_-$. W can be seen as a kernel shaped like a mexican hat [1].

One possible way to interpret the outcome is to read the output $f(U)$. It is common to take a barycenter of the output as an estimator of the position targeted by the model. While we are not supposed to know an euclidian topology in which the positions of GNG nodes can be averaged, we can still use the input data to interpolate a corresponding location in a 2D euclidian space for each neuron. We will do that for our experimentations, but please note that this interpolation will not always be possible. Similarly, for the GNG, we will plot them by putting all nodes to their asserted location, only for visualization purposes.

[2]In this equation, only U_k is incremented over time, and the inputs I_k are static. However, none of our hypotheses prevent the inputs from being updated over time. We make this choice because dynamic inputs are not necessary for the results presented in this paper. Otherwise, equation (2) could be written by expressing $U_k(t)$ as a function of $U_*(t-\Delta t)$ and $I_k(t)$.

TABLE I
RANGES OF INPUTS IN THE EXTERNAL ENVIRONMENT

Section	Modality	X-range	Y-range	Z-range
IV-A	vis.	$[0, 90]$	$[-45, 45]$	–
	aud.	$[0, 90]$	$[-45, 45]$	–
IV-B	vis.	$[-45, 45]$	$[-45, 45]$	–
	aud.	$[-90, 90]$	$[-45, 85]$	–
IV-C	vis.	$[-45, 45]$	$[-45, 45]$	$[0, 45]$
	aud.	$[-90, 90]$	$[-45, 85]$	–

IV. RESULTS

Our results will be divided in three parts, with a common protocol for all. For this article, we will consider two modalities, vision and audition. That can correspond for example to a robot asked to locate a visual and/or audible stimulus. We test three setups that take into account challenges that might happen in the robot perception: differences of resolution within the same sensory space (section IV-A), high-dimensional feature space (IV-B), and non-relevant features (IV-C).

So, the main difference between the setups will be in the first step of our model, the generation of the unimodal manifolds (described in section III-A). For the GNG training, a stimulus location will be drawn within the subspace of the environment that is accessible to the appropriate sensors. For example, a robot's visual perception might be restricted to the space in front of them, while their auditory range might be all around them. Input ranges are listed in table I. Then, we simulate the information that would be received from the sensors if a real stimulus was sent from this position. The way they are preprocessed will be defined in each subsection.

We have set an upper limit to the number of neurons in the GNG. Otherwise, the resolution could become excessively high, increasing the computational cost for no valid reason. Once the limit is reached, the GNG is trained like a regular NG, except that nodes that have become irrelevant can still be removed and replaced. This is still more efficient than starting with all neurons and training a NG from the beginning.

The creation of a bimodal manifold is roughly the same in all setups. For the DNF, input stimuli will be specified in each scenario, depending on the properties to showcase. For the same reasons, parameters might need to be adjusted slightly from one setup to the next. All values are given in table II.

A. Bio-inspired Model of Audiovisual Processing

Our first experimentation is inspired from observations in neurophysiology. Human visual perception is affected by the heterogeneous distribution of sensors in the retina, giving a higher resolution in the center of the field of view (the fovea) than in its periphery. This disparity can be observed in brain regions processing visual information, such as the superior colliculus [30]. A mathematical model of the disparity between fovea and periphery, using a logpolar transformation, has been suggested by [30], and previous works have coupled it with DNF for visual [27] and audiovisual processing [3].

TABLE II
PARAMETERS USED IN OUR DNF IMPLEMENTATION. SPREAD PARAMETERS ARE EXPRESSED IN ARBITRARY UNIT THAT DENOTES THE MINIMAL NUMBER OF EDGES THAT SEPARATE TWO NEURONS.

Parameter	Value		Meaning
	IV-A	IV-B & IV-C	
			Simulation settings
Δt	0.01	0.01	Time step
σ_I	2.5	2.5	Spread of stimulus
$\lambda_{\text{vis, A}}$	2	2	Strength of visual bottom stimulus
$\lambda_{\text{vis, B}}$	2.4	2.02	Strength of visual top stimulus
$\lambda_{\text{aud, A}}$	2.4	1.5	Strength of audio bottom stimulus
$\lambda_{\text{aud, B}}$	2	0	Strength of audio top stimulus
			DNF parameters
τ	0.1	0.1	Time constant
λ_+	0.4	0.55	Amplitude of lateral excitation
σ_+	2.5/3/3.5	1.5	Spread of lateral excitation
λ_-	0.3	0.3	Amplitude of lateral inhibition
σ_-	$+\infty$	10	Spread of lateral inhibition
h	-1	-1	Resting level

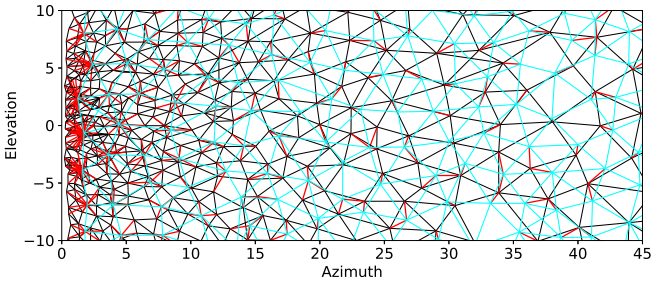

Fig. 2. Sample representation of a bimodal graph. Edges are colored depending on the modalities of the neurones they connect. Visual-visual: black. Auditory-auditory: cyan. Visual-auditory: red.

Models of the superior colliculus are not only useful for computational neuroscience. While cameras used by robots are supposed to have a homogeneous resolution, they might happen to have blurry spots because of dirt or wear. Other modalities may also have a high variance in resolution. The logpolar transformation is a straightforward way of testing these variations in a controlled setting. Additionally, even when the camera sensory space is perfectly regular, it has been suggested that adding a logpolar transformation on top of it could improve gaze control in robots [34].

1) Sensory space: In light of the aforementioned hypothesis, we take coordinates of a visual stimulus in a regular 2D visual hemifield, and displace them following the logpolar transformation in [30]. The new 2D coordinates are used as inputs for the visual GNG. Since we study the effect of variable resolutions in one modality, the other modality, audio, will be modeled as a regular 2D space as in [3], with the same range as vision (table I), so that it does not interfere with the analysis. Both GNG are given 1000 nodes maximum.

2) Produced manifolds: A sample of the bimodal graph is shown in figure 2. For visualization, visual nodes are placed according to a reverse logpolar transformation of their features, and auditory nodes according to their raw coordinates. The unimodal GNG are superposed with different colors.

As expected, the visual GNG has a much higher resolution around the fovea (0°), as can be presumed by the high density of nodes. It gradually decreases as the azimuth augments. On the contrary, the auditory GNG has roughly the same resolution everywhere. Connections between neurons of different modalities are shown in red[3]. For azimuths between 0° and approximately 30°, vision has a better resolution than audition: most nodes from the audio GNG are connected to multiple visual nodes. The trend is reversed for higher azimuths.

3) Resulting properties: After the bimodal manifold is created, we are interested in seeing what a DNF would select when confronted to conflicting bimodal stimulus. It is expected that near the fovea, vision is more reliable, so it should have a bigger weight in the fusion than audition. To test this, we put two conflicting stimuli A and B at a common azimuth x, and elevations $-5°$ and $5°$ respectively. Both stimuli can be both seen and heard, but A is 20% more auditively salient than B, and B is 20% more visually salient than A.

When tested on a unimodal manifold, the DNF has no trouble selecting either A or B. Every time, the most salient stimulus in its respective modality has a higher chance of being selected. Occasionally, the DNF forms a bubble in-between the stimuli. This is mostly visible for higher azimuths in the visual GNG. The reason is that the resolution is so low that A and B are separated by only a few edges. The DNF does not have access to the corresponding inputs of its neurons viewed from the exterior. Thus, when viewed from inside the model, they are topologically very close to each other. So, the DNF treats the stimuli as if they were right next to each other, and merges them into a bubble of activity located at their center of mass.

In the bimodal manifold, the stochasticity in the creation of the GNG starts having an impact, as it may seemingly give a locally higher resolution to a modality when it is not expected. A might be selected instead of B, when B is more salient, just because B stimulates a region with fewer neurons or connections than average. To separate the random effect caused by the creation of the GNG, we create 50 bimodal manifolds, and test a run of DNF on 90 different azimuths for each of them. The results are aggregated in figure 3. As we suspect that the distance at which stimuli are merged depends on the width of the DNF kernel, we couple in our analysis the effect of resolution with the value of σ_+. We test three different values of σ_+, represented by three different colors: green, red, blue from thinnest to widest.

The curves represent the outcome of two mixed logistic regressions. The fit of the black curve is obtained after

[3]For this model, we initially observed that a lot of visual neurons close to the fovea were never connected to auditory ones. Because there are so many of them in a very close space, a huge number of random draws is required before they are all visited. To ensure that the merging task would not be hindered by a lack of connectivity, we biased the draw of external stimuli so that the prototypical input of every neuron was drafted. We found that this manual bias has no effect on the graph connectivity outside the fovea. This draw method is not applicable to most scenarios, since we are not supposed to know the actual coordinates of the neurons in the external environment.

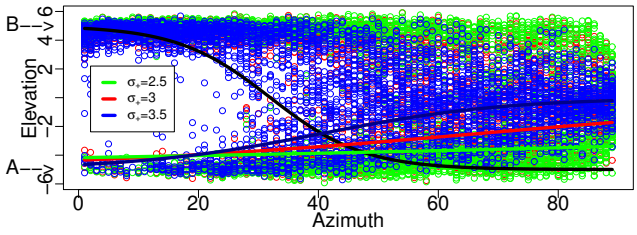

Fig. 3. Statistical model of the modality priority change (in black) and the stimulus merging. One point represents the barycenter of the output of one of the 3 differently parametrized DNF (green: $\sigma_+ = 2.5$, red: $\sigma_+ = 3$, blue: $\sigma_+ = 3.5$), on one of the 50 randomized GNG, with two bimodal stimuli A and B at azimuth x and elevations $\pm 5°$. The black curve shows a logistic regression of the switch between preferred stimuli. Colored curve show logisitic regressions of the stimulus merging effect depending on values of σ_+.

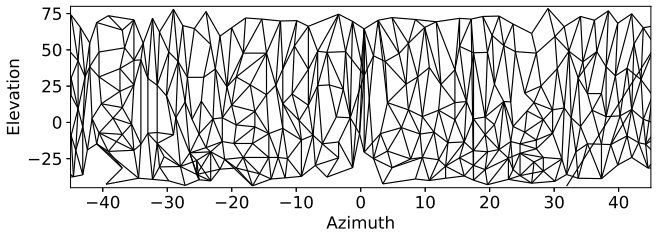

Fig. 4. Sample of the auditory graph obtained from HRTF data. The 2D location of neurons is not known by the GNG, it has been interpolated from their prototypical input in HRTF space, for visualization purposes only. Note that the x-axis and y-axis have different scales.

cancelling the merging effect, and shows a clear switch of preference from B to A centered on 32°. B is more likely to be selected than A when the visual modality is the most reliable, and vice versa. Logically, this effect is independant of σ_+ variations. This amounts to the DNF automatically selecting a stimulus according to the most reliable sensor.

The fit of the colored curves are obtained by canceling the switch effect. We can see a convergence from $\pm 5°$ to $0°$ elevations, although for lower values of σ_+, the limit at $0°$ is not reached before the end of the field of view. Only the lower curves are displayed but the effect is symmetrical.

The results show two trends. First, from the higher concentration of points at the $5°$ elevation in the leftmost part of the figure, we can see that B (visually stronger) is more often selected in lower azimuths than A. Then A is preferred for higher azimuths. Second, we see that the probability of A and B being merged (manifesting as an increasing concentration of points around $0°$) increases with the azimuths. As we expected, the distance at which they are merged depends a lot on the value of σ_+. The larger the interaction kernel, the sooner the merging seems to happen.

B. Real-world Robotic Sensory Data

In the previous section, we used manufactured data to showcase DNF selection properties in manifolds of variable resolution, favoring the most reliable modality. In this section, we will partly use real experimental data and show that these properties are still available in more complex sensory spaces. Our main change will be on auditory preprocessing. One way of performing sound source localization for robots is to compute a HRTF, a function that associates spectral features (caused by interferences on the signal by the head and pinnae) to source orientations [4]. Meanwhile, vision is less of a challenge nowadays, as extracting the position of an object from an image is easily achievable, and one can reasonably expect to have a homogeneous resolution in most cases.

1) Sensory space: Data provided by [35] includes head-related impulse responses of a robot equipped with artificial pinnae, to a sound located at different angles. Given an external stimulus position in 2D, we can interpolate the responses received by the two robotic ears within a specific range (table I). We then compute their Fourier transform and make the difference between the ears to obtain a HRTF. In the end, each audio input is 100-dimensional.

For vision, we will consider a robot with an intact camera and assume it can roughly estimate the 2D coordinates of an object in front of it. We do not need visual and auditory perception to have the same range. Realistically, stimuli can be heard from more orientations than they can be seen. To keep resolutions approximately balanced, we will use respectively maximum 500 and 200 nodes for auditory and visual GNG.

2) Produced manifolds: The visual GNG is very similar to the auditory GNG in the previous section, which also directly received stimuli drawn from a regular 2D space. The new auditory one, however, has a distinct shape. Figure 4 shows what the GNG looks like after placing each node at the source location that would match its audio (100D) coordinates best. The graph appears to be stretched vertically.

3) Resulting properties: Like in the previous scenario, we test the DNF with two stimuli A and B. This time, they are separated both horizontally and vertically. Stimulus A has congruent audio and visual components, while B is not audible but visually more salient by 1%. It is expected that A should be selected over B, as A is consistent over modalities. Results are synthesized in figure 5.

In the visual-only manifold, B largely takes precedence. A is mostly inhibited, with some (negative) residual activity left. This is expected, as B is more visibly salient, but it is worth noting that the 1% difference between $\lambda_{\text{vis, A}}$ and $\lambda_{\text{vis, B}}$ matters. While not shown here, we have tested swapping the intensity values, and A does take precedence in that inverted case. We are in a situation where both stimuli are considered equally by the DNF, and a very small difference in intensity is enough to bias the competition towards one or the other. This is a very standard observation in DNF literature, but it is still worth noting considering the topology is not entirely regular.

In the audio-only manifold, A is trivially selected, but we can see some loss of precision in elevation: the barycenter is found 7° higher than the actual stimulus. This is very consistent with the general lack of elevation-wise precision in auditory perception.

The precision is improved in the bimodal manifold. As would be expected, audiovisual congruent stimulus A is selected over visual-only B. But the barycenter is also closer

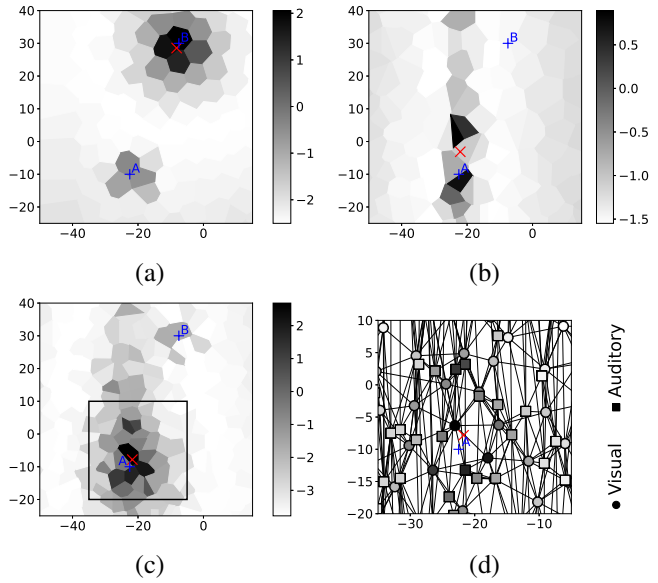

Fig. 5. Results of stimulus selection by DNF unimodal and bimodal GNG. These 2D depictions use neuron positions interpolated from the source data (for visualization). Shades of gray reflect neuron potential U. Red crosses indicate the barycenter of output activation $f(U)$ in the reconstructed 2D projection. (a) Visual-only neural gas with two stimuli located at A and B, with B slightly more salient. Nodes are represented by Voronoi cells, edges connecting nodes are not shown. (b) Auditory-only neural gas, with only one input at A. (c) Bimodal neural gas. Its input is the sum of the ones used for (a) and (b). (d) Zoom on (c) around A, where all nodes and edges are shown.

to the actual stimulus position than in the audio-only case, meaning the visual elevation-wide better precision had a positive impact. Again, the enhanced multimodal precision is a classical observation in either neuroscience or machine learning, but it is worth noting that it persists when working with a complex underlying topology.

When we look more closely at the nodes around A, we can see than despite there being a lot of edges in all directions, a few neurons form a discernable bubble. It is interesting that these neurons come indiscriminately from both modalities. One could have feared an outcome where only visual neurons interact with each other, and auditory neurons, less regularly distributed, only serve to transmit a little bit of auditory stimulation. On the contrary, the crossmodal connections play an important part, so that the DNF does not leave out one modality for the other. When both are useful, both are used.

C. Dealing with a Superfluous Dimension

1) Sensory space: This setup is similar to the previous one, except the visual sensory space is now 3D. We add a dimension that is not relevant to the task, e.g. color when a robot is asked to select an object designated by shape only. Since the visual space expands, and GNG are not advanced enough to reduce the dimensionality when the amount of possible inputs increases brutally, we also increase the number of neurons in the visual GNG to 3000. The rest of the setup remains the same.

2) Resulting properties: We did the same experiments as in section IV-B. Stimuli A and B are given the same color, so that

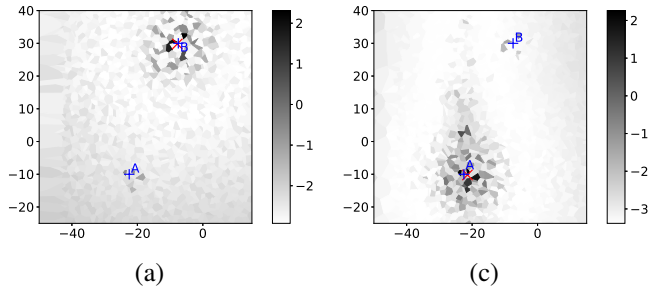

Fig. 6. Same as figure 5 with a supplementary dimension in the visual modality. The third dimension is orthogonal to the plane used in this representation.

their distance in the external environment remains the same as before. According to our preliminary tests, the conclusion would be the same with stimuli of different colors. Results are displayed in figure 6. Only the visual-only and audiovisual conditions are shown, since the auditory-only condition is the same as before, and the zoom-in picture with edges is hardly readable. As a reminder, the visualizations are still made using x- and y-axes, meaning the new color axis is completely flattened. These presentations are akin to looking at a cube from a side, hence the dense Voronoi tessellation and the scattered activity.

We find that the outputs are strikingly similar, i.e. a preference for multimodal consistent inputs and improving audio precision, despite a big increase in the number of neurons and edges, many of which are irrelevant to the task. This shows robustness of the model to distracting dimensions.

V. Conclusion and Perspectives

Our model consists in two unimodal GNG, trained using the standard algorithm by [5], then connected to form one new multimodal manifold with a simple Hebbian rule. This manifold is used as a support for neural dynamics that are implemented by adapting the DNF paradigm [1]. Our model was tested on multiple setups, including real data. The main novelties of our work are twofold. First is the use of neural dynamics in a multimodal manifold of unspecified dimensionality or regularity, a capability of DNF that has not been showcased before. The field applies on a learned manifold that is faithful to each unimodal sensory space, and is not hindered by irrelevant dimensions. Second is the combination of the multimodal topology with DNF to obtain interesting properties such as the contribution of different modalities that depends on their respective learned resolution, the selection of the most relevant multimodal stimulus by using the best information each modality had to offer, and the filtering of irrelevant informations. These results are scalable to applications with more than two modalities.

As we have seen when adding a dimension, the number of neurons in the GNG necessary to keep the same resolution, and consequently the computational cost of the model, may increase drastically when the sensory space is broadened. This would not be an issue with deep neural networks, that are very

effective at finding intrinsic dimensions in data [7]. It would be interesting to see whether manifolds created by deep learning are also suitable vectors of neural dynamics. This would be complementary to existing approaches to encode topological maps with neural networks [36], [37].

In our model, learning of the multimodal topologies and their use for multimodal fusion are decoupled. An interesting perspective would be to perform them simultaneously, which raises some challenges like making the model robust to the temporal dynamics and to the detection of relevant features for learning and fusion. Another perspective is to study multimodal active perception, where the internal perception will be related to motor actions to explore the environment. DNF are well suited to model saccades [29]. This raises open questions related to multimodal attention and active perception.

ACKNOWLEDGMENT

Most of the computations presented in this paper were performed using the GRICAD infrastructure (https://gricad.univ-grenoble-alpes.fr), which is supported by Grenoble research communities.

REFERENCES

[1] S.-I. Amari, "Dynamics of pattern formation in lateral-inhibition type neural fields," *Biological Cybernetics*, vol. 27, no. 2, pp. 77–87, 1977.

[2] C. Schauer and H. M. Gross, "Design and optimization of Amari neural fields for early auditory-visual integration," in *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541)*, vol. 4, 2004, pp. 2523–2528.

[3] S. Forest, J.-C. Quinton, and M. Lefort, "A dynamic neural field model of multimodal merging: application to the ventriloquist effect," *Neural Computation*, in press. [Online]. Available: https://hal.archives-ouvertes.fr/hal-03600794

[4] S. Argentieri, P. Danès, and P. Souères, "A survey on sound source localization in robotics: From binaural to array processing methods," *Computer Speech & Language*, vol. 34, no. 1, pp. 87–112, 2015.

[5] B. Fritzke, "A growing neural gas network learns topologies," in *Advances in Neural Information Processing Systems*, G. Tesauro, D. Touretzky, and T. Leen, Eds., vol. 7. MIT Press, 1995.

[6] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[7] A. Ansuini, A. Laio, J. H. Macke, and D. Zoccolan, "Intrinsic dimension of data representations in deep neural networks," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019.

[8] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2014.

[9] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological Cybernetics*, vol. 43, no. 1, pp. 59–69, 1982.

[10] T. Martinetz and K. Schulten, "A "neural-gas" network learns topologies," *Artificial neural networks*, vol. 1, pp. 397–402, 1991.

[11] A. Droniou, S. Ivaldi, and O. Sigaud, "Deep unsupervised network for multimodal perception, representation and classification," *Robotics and Autonomous Systems*, vol. 71, pp. 83–98, 2015.

[12] X. Yang, P. Ramesh, R. Chitta, S. Madhvanath, E. A. Bernal, and J. Luo, "Deep multimodal representation learning from temporal data," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5066–5074, 2017.

[13] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, and J. Carreira, "Perceiver: General perception with iterative attention," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 2021, pp. 4651–4664.

[14] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *ICML*, 2011.

[15] M. Lefort, Y. Boniface, and B. Girau, "SOMMA: Cortically inspired paradigms for multimodal processing," in *The 2013 International Joint Conference on Neural Networks (IJCNN)*, 2013, pp. 1–8.

[16] L. Khacef, L. Rodriguez, and B. Miramond, "Brain-inspired self-organization with cellular neuromorphic computing for multimodal unsupervised learning," *Electronics*, vol. 9, no. 10, 2020.

[17] N. Gonnier, Y. Boniface, and H. Frezza-Buet, "Input prediction using consensus driven SOMs," in *2021 8th International Conference on Soft Computing & Machine Intelligence (ISCMI)*. IEEE, 2021, pp. 38–42.

[18] O. Ménard and H. Frezza-Buet, "Model of multi-modal cortical processing: Coherent learning in self-organizing modules," *Neural Networks*, vol. 18, no. 5, pp. 646–655, 2005.

[19] S. Lallee and P. F. Dominey, "Multi-modal convergence maps: from body schema and self-representation to mental imagery," *Adaptive Behavior*, vol. 21, no. 4, pp. 274–285, 2013.

[20] M. Vavrečka and I. Farkaš, "A multimodal connectionist architecture for unsupervised grounding of spatial language," *Cognitive Computation*, vol. 6, no. 1, pp. 101–112, 2014.

[21] M. Johnsson, M. Martinsson, D. Gil, and G. Hesslow, "Associative self-organizing map," in *Self Organizing Maps*, J. I. Mwasiagi, Ed. Rijeka: IntechOpen, 2011, ch. 30.

[22] G. I. Parisi, J. Tani, C. Weber, and S. Wermter, "Emergence of multimodal action representations from neural network self-organization," *Cognitive Systems Research*, vol. 43, pp. 208–221, 2017.

[23] K. Huang, X. Ma, R. Song, X. Rong, X. Tian, and Y. Li, "An autonomous developmental cognitive architecture based on incremental associative neural network with dynamic audiovisual fusion," *IEEE Access*, vol. 7, pp. 8789–8807, 2019.

[24] S. Marsland, J. Shapiro, and U. Nehmzow, "A self-organising network that grows when required," *Neural networks*, vol. 15, no. 8-9, pp. 1041–1058, 2002.

[25] G. Schöner, J. Spencer, and DFT Research Group, *Dynamic Thinking: A Primer on Dynamic Field Theory*, ser. Oxford Series in Developmental Cognitive Neuroscience. Oxford University Press, 2015.

[26] Y. Sandamirskaya, "Dynamic neural fields as a step toward cognitive neuromorphic architectures," *Frontiers in Neuroscience*, vol. 7, p. 276, 2014.

[27] W. Taouali, L. Goffart, F. Alexandre, and N. P. Rougier, "A parsimonious computational model of visual target position encoding in the superior colliculus," *Biological Cybernetics*, vol. 109, no. 4, pp. 549–559, 2015.

[28] J. Fix, N. Rougier, and F. Alexandre, "A dynamic neural field approach to the covert and overt deployment of spatial attention," *Cognitive Computation*, vol. 3, no. 1, pp. 279–293, 2011.

[29] J.-C. Quinton and L. Goffart, "A unified dynamic neural field model of goal directed eye movements," *Connection Science*, vol. 30, no. 1, pp. 20–52, 2018.

[30] F. P. Ottes, J. A. V. Gisbergen, and J. J. Eggermont, "Visuomotor fields of the superior colliculus: A quantitative model," *Vision Research*, vol. 26, no. 6, pp. 857–873, 1986.

[31] M. Cerda and B. Girau, "Asymmetry in neural fields: a spatiotemporal encoding mechanism," *Biological cybernetics*, vol. 107, no. 2, pp. 161–178, 2013.

[32] J.-C. Quinton and B. Girau, "Predictive neural fields for improved tracking and attentional properties," in *The 2011 International Joint Conference on Neural Networks*. IEEE, 2011, pp. 1629–1636.

[33] ——, "A sparse implementation of dynamic competition in continuous neural fields," in *Brain Inspired Cognitive Systems 2010 - BICS 2010*, Madrid, Spain, 2010.

[34] L. Manfredi, E. S. Maini, and C. Laschi, "Neurophysiological models of gaze control in humanoid robotics," in *Humanoid Robots*, B. Choi, Ed. Rijeka: IntechOpen, 2009, ch. 10.

[35] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575)*. IEEE, 2001, pp. 99–102.

[36] P. Hartono, P. Hollensen, and T. Trappenberg, "Learning-regulated context relevant topographical map," *IEEE transactions on neural networks and learning systems*, vol. 26, no. 10, pp. 2323–2335, 2014.

[37] P. Hartono, "Mixing autoencoder with classifier: conceptual data visualization," *IEEE Access*, vol. 8, pp. 105 301–105 310, 2020.