Champ neuronal et apprentissage profond de topologies pour la fusion multimodale

Simon Forest, Jean-Charles Quinton, Mathieu Lefort

Champ neuronal et apprentissage profond de topologies pour la fusion multimodale

S. Forest[1,2], J.-C. Quinton[1], M. Lefort[2]

[1] Univ. Grenoble Alpes, CNRS, Grenoble INP, LJK, UMR 5224, F-38000, Grenoble, France
[2] 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

Résumé

Des agents artificiels tels que des robots ont souvent intérêt à fusionner des données issues de différentes modalités. Pour cela, il peut être pertinent de prendre en compte les variations dans la structure et la résolution des topologies sous-jacentes aux espaces sensoriels. Nous proposons d'utiliser un champ neuronal dynamique pour sélectionner des stimuli dans un contexte multimodal. Nous avions adapté le modèle à des topologies apprises (par des gaz neuronaux croissants notamment) pour la fusion, nous l'étendons maintenant en insérant un auto-encodeur pour réduire la dimensionnalité des données d'entrée.

Mots-clés

Fusion multimodale, champ neuronal dynamique, apprentissage de variétés, auto-encodeur, gaz neuronal croissant

Abstract

Artificial agents such as robots can often benefit from merging data from different modalities. For this purpose, it may be relevant to take into account the variations in the structure and resolution of the topologies underlying the sensory spaces. We propose to use a dynamic neural field to select stimuli in a multimodal context. We had adapted the model to learned topologies (with growing neural gas in particular) for fusion, we now extend it by inserting an auto-encoder to reduce the dimensionality of the input data.

Keywords

Multimodal fusion, dynamic neural field, manifold learning, auto-encoder, growing neural gas

1 Introduction

Quand on parle de traitement de l'information et de prise de décision comportementale, la façon dont on fusionne les données issues de différentes sources n'est pas à négliger. Prenons un exemple : un robot a la tâche de trouver et atteindre un réveil lorsqu'il se met à sonner. Au début, le robot peut faire face à plusieurs objets ressemblant à un réveil, qu'il ne devrait avoir aucune difficulté à distinguer. Lorsqu'une sonnerie se fait entendre, le robot devrait être capable de localiser son origine, mais avec généralement une faible précision. Avant d'entreprendre une action, le robot doit sélectionner un objet. Dans ce cas, il s'agit de l'objet, parmi ceux qui ressemblent à un réveil, qui coïncide le plus avec la localisation de la source sonore. Mais la manière dont les modalités visuelle et auditive doivent être pondérées dépend non seulement de la tâche (une horloge visible de face est moins importante qu'un son provenant d'un côté), mais aussi de la fiabilité des capteurs (la réverbération de la pièce peut rendre l'orientation du son moins décisive).

La tâche dans cet exemple est confrontée à de multiples défis, notamment la fusion de modalités sensorielles de disponibilité et de fiabilité différentes, et la sélection de (et l'attention vers) la cible. Pour résoudre ces problèmes, beaucoup de modèles actuels se basent exclusivement sur l'apprentissage profond. Dans cet article, nous prenons l'apprentissage profond comme un outil de pré-traitement de données et de réduction de dimension. Pour la fusion, nous nous reposons sur une autre approche à partir d'un champ neuronal dynamique (*dynamic neural field*, DNF), un modèle bio-inspiré d'activité neuronale [2]. Il s'agit d'un réseau récurrent en temps continu placé dans une topologie, où les poids sont connus et dépendent de la distance entre les neurones. Avec un mélange d'excitation à courte portée et d'inhibition à longue portée, les stimuli d'entrée sont mis en compétition jusqu'à ce qu'une bulle d'activité émerge, qui peut être interprétée comme une décision, décentralisée et dynamique, de sélection d'une cible et/ou d'action. De plus, la dynamique donne un lissage temporel à la bulle d'activité, malgré les fluctuations des entrées et les distracteurs potentiels. Le DNF a connu diverses applications, notamment en robotique [14, 34, 39]. En particulier, les propriétés d'interaction du DNF le rendent très adapté à la fusion multimodale [11, 35].

Les premières implémentations de DNF trouvent une limite dans la nature de la variété (*manifold* en anglais) sur laquelle elles évoluent. La plupart des applications de la littérature supposent l'existence d'une topologie régulière sous-jacente, le plus souvent 1D ou 2D [36]. Mais elle n'est guère représentative des disparités de l'espace sensoriel, disparités qui deviennent cruciales lors de la fusion multimodale. En effet, intéressons-nous à la forme des stimuli perçus dans l'environnement. La quantité d'informations disponibles est énorme, et les données qu'un agent reçoit

de ses capteurs n'en sont qu'une projection dans quelques dimensions données. Équipé d'une caméra standard, un robot recevra une projection en 2D de la partie de l'environnement à laquelle il fait face. Avec un seul microphone, il peut détecter des sons provenant de n'importe où autour de lui, mais il peut difficilement les localiser. Deux microphones peuvent permettre une certaine localisation sonore 1D le long de l'axe sur lequel ils sont alignés, généralement azimutal (grâce à la différence de temps ou d'intensité interaurale), et même un peu de 2D ou 3D en exploitant la forme des pavillons des oreilles, avec une fonction de transfert liée à la tête (*head-related transfer function*, HRTF) [4]. Nous devons d'abord tenir compte des spécificités de chaque modalité sensorielle avant de créer des comportements qui l'exploitent au mieux. De plus, nous devons trouver un moyen de faire correspondre des informations complémentaires provenant de différentes modalités, ce qui revient généralement à projeter des stimuli sur une variété commune.

Dans [10], nous avons proposé une manière d'adapter le DNF à des variétés de dimensionnalité et structure arbitraires. À l'aide de gaz neuronaux croissants (*growing neural gas*, GNG), une topologie multimodale est créée à l'intérieur de laquelle le DNF peut fusionner et sélectionner des stimuli. Les applications testées dans ce précédent article étaient limitées par la faible capacité d'apprentissage du GNG : s'il permet de faire ressortir un espace sous-jacent dans des données de plus grande dimension, il ne permet pas de réaliser des tâches plus complexes dans des espaces de très grande dimension. Localiser un stimulus visuel à partir d'une photographie, ou un stimulus auditif à partir d'un enregistrement audio brut, n'est pas possible avec un GNG seul. Il est nécessaire au préalable de réduire la dimensionnalité des entrées en apprenant des projections vers un espace plus facile à exploiter. Cette solution peut être apportée par des réseaux de neurones. Dans cet article, nous étendons la précédente contribution en ajoutant une méthode d'apprentissage profond de variétés, à savoir un auto-encodeur de Wasserstein en coupes (*sliced Wasserstein auto-encoder*, SWAE), en amont du modèle. Notre objectif est de vérifier si les propriétés du modèle précédent se maintiennent lorsqu'on ajoute une opération visant à réduire les dimensions de l'espace d'entrée, au risque de dégrader la réelle topologie sous-jacente pendant l'apprentissage de l'encodeur.

Cet article est structuré comme suit. Dans la section 2, nous présentons des travaux existants sur l'apprentissage de variétés et le DNF, et en particulier leurs applications à la fusion multimodale. Puis nous décrivons notre modèle complet dans la section 3, et montrons sa robustesse, ses performances et ses propriétés dans la section 4. Nous concluons et ajoutons des perspectives additionnelles dans la section 5.

2 Travaux existants

2.1 Apprentissage de variétés

Les capteurs fournissent des échantillons de haute dimension de l'environnement, mais les espaces sensoriels correspondent souvent à des variétés de dimension inférieure. L'apprentissage profond est particulièrement adapté à la génération de telles variétés (voir [5] pour une revue). Par exemple, il a été démontré que les dernières couches d'un réseau neuronal profond contiennent une dimensionnalité intrinsèque inférieure au nombre de descripteurs dans les données [3]. Des méthodes spécifiques telles que les auto-encodeurs variationnels [16] peuvent apprendre une structure sous-jacente de manière non supervisée. D'autres types d'auto-encodeurs existent, notamment le SWAE [18]. Ce dernier utilise dans l'apprentissage une distance de Wasserstein, qui compare la distribution des données encodées dans l'espace latent à une distribution choisie. On peut ainsi imposer, par exemple, que l'espace latent suive une distribution uniforme en 2D. Il est ensuite possible d'exploiter les propriétés géométriques de la topologie ainsi encodée [19].

Une approche moins contraignante, privilégiée dans notre précédente contribution [10], repose sur les méthodes d'auto-organisation. Dans les cartes auto-organisatrices (*self-organizing maps*, SOM), par exemple le modèle de Kohonen [17], chaque neurone représente une entrée prototypique dans l'espace sensoriel à haute dimension, de sorte que l'espace d'entrée est projeté sur un treillis neuronal de forme et de taille fixes. Dans le cas du gaz neuronal (*neural gas*, NG), les neurones ne sont pas disposés sur un treillis, mais sont connectés selon une règle Hebbienne, de sorte que les neurones de prototypes proches sont reliés entre eux [25]. Le gaz finit par remplir l'espace d'entrée, d'une manière qui imite la distribution des stimuli. Le gaz neuronal croissant (*growing neural gas*, GNG) [12] est un dérivé du NG, dans lequel des neurones sont ajoutés (ou supprimés) au fur et à mesure jusqu'à ce qu'une condition spécifique soit remplie, s'adaptant ainsi à la topologie indéterminée de l'espace d'entrée.

Variétés en fusion multimodale

De multiples articles ont montré des résultats prometteurs en fusion multimodale à l'aide d'apprentissage profond. L'apprentissage profond non supervisé peut être utilisé pour projeter des données multimodales sur une variété de faible dimension pour une utilisation en robotique [8, 21]. Les entrées peuvent être mélangées pendant l'entraînement du réseau de neurones pour exploiter les corrélations entre les modalités [40, 43]. Récemment, des extensions du réseau Transformer ont été proposées, permettant de recevoir des entrées multimodales pondérées par un module d'attention [15, 28]. Cependant, la plupart de ces travaux partent du principe que toutes les données multimodales sont corrélées entre elles. De plus, les architectures profondes sont dédiées à une tâche spécifique et aucun paradigme générique n'émerge [29].

Nous cherchons à créer une nouvelle topologie multimodale sur laquelle de nouvelles propriétés dynamiques pourraient être appliquées. Dans un premier temps, l'auto-organisation offre des solutions pour un coût bien moindre [13,20,22,27,31,41]. Les SOM et leurs dérivés sont utilisés depuis longtemps comme modèles de fusion multimodale,

mais les façons de combiner les modalités peuvent être très diverses. Les architectures composées de SOM peuvent être divisées en deux catégories. Dans la première, une SOM est formée pour chaque modalité, puis toutes les cartes unimodales sont connectées en fonction d'une règle d'apprentissage spéciale [13, 22]. Dans la seconde, les cartes unimodales sont reliées à une nouvelle SOM [20, 27] ou un NG [41] multimodal qui combine toutes les informations. Des couches supplémentaires de SOM peuvent également être envisagées pour créer un flux hiérarchisé d'informations [31]. De plus, les modèles peuvent être rendus plus adaptatifs aux tâches dépendantes du temps à l'aide de cartes « croissantes au besoin » [31], une alternative au GNG conçue pour les distributions d'entrées dynamiques [24]. Certains de ces modèles ont déjà été testés pour des modalités visuelles, auditives et/ou proprioceptives sur des robots [13, 20].

2.2 Champ neuronal dynamique

Après l'apprentissage de cartes multimodales et/ou de cartes unimodales interconnectées, nous avons besoin d'un paradigme pour dicter la manière dont la perception va se produire. La perception multimodale peut être considérée comme une forme de décision prenant en compte des entrées sensorielles de fiabilité et de pertinence variables. Nous suivons le choix d'architecture fait dans [27] et [22], où le champ neuronal dynamique (DNF) est utilisé comme paradigme qui régit la fusion ou la ségrégation des stimuli dans l'espace topologique multimodal. Le DNF a de nombreuses propriétés utiles pour la perception multimodale.

Originaire du domaine des neurosciences, le DNF a diverses applications en robotique [36]. Par exemple, l'attention visuelle peut être cumulée avec un contrôle moteur pour qu'un robot fixe de manière autonome les objets de son environnement et apprenne une carte sensorimotrice [34]. Le DNF repose sur une population d'unités connectées topologiquement, à une échelle mésoscopique, où l'activité apparente (ou potentiel membranaire moyenné sur des groupes de neurones) peut être lue pour interpréter des décisions à un niveau comportemental. L'activité évolue dans le temps en fonction d'une somme de stimulations externes et d'interactions latérales entre les neurones. Les neurones stimulés vont envoyer une forte excitation à leurs voisins les plus proches, et une inhibition modérée à leurs voisins plus éloignés, conduisant à l'émergence d'une bulle d'activité stable. En fonction du paramétrage, cela peut conduire à plusieurs types de comportements [36]. Avec une forte excitation locale, la bulle peut être auto-entretenue, agissant comme une mémoire à long terme [34]. L'inhibition à longue portée créera une compétition entre les stimuli conflictuels, jusqu'à ce que l'un d'entre eux domine les autres, ou qu'ils soient fusionnés en une seule bulle à une position interpolée [11, 38]. Ensuite, la bulle auto-entretenue peut être utilisée pour une attention sélective robuste, capable d'ignorer le bruit et les faibles distracteurs [9]. Enfin, la sortie du DNF peut être directement exploitée pour générer une commande motrice [33, 34].

Les propriétés du DNF peuvent être très utiles à la fusion multimodale. Il fournit des moyens non seulement pour améliorer la robustesse des décisions lorsque les modalités sont congruentes [35], mais aussi pour résoudre les conflits entre modalités [11]. C'est là que le choix de la variété sous-jacente peut être très important, car tout le processus de décision se base dessus, et sa structure peut déterminer entre autres la fiabilité des stimuli, qui peut avoir une influence sur la sélection.

La grande majorité des travaux utilisant le DNF supposent que la dynamique se déroule sur une topologie complètement régulière, par exemple un treillis 2D dans le cas de la vision. Cependant, il n'existe pas de moyen clair de projeter deux modalités ou plus sur le même treillis. Dans [35] et [11], des hypothèses fortes sont faites sur la forme des stimuli dans une modalité afin qu'ils s'intègrent dans la topologie de l'autre. Pour résoudre ce problème, [22] propose d'utiliser des variétés distinctes pour chaque modalité, chacune appris par une SOM, et d'appliquer un DNF sur chacun d'eux. La communication entre les modalités est assurée par un ensemble de connexions topographiques.

Une contribution intéressante de cette dernière référence est l'utilisation d'une variété apprise comme siège de la dynamique neuronale. Par ailleurs, certaines tentatives visant à modifier la projection des entrées dans la variété ont donné des résultats satisfaisants : [38] et [11] ont ainsi reproduit des comportements biologiques après avoir appliqué aux stimuli visuels une transformation logpolaire, qui modélise les variations de résolution de la rétine humaine [30]. Dans le cas de [22], les projections reçues par les neurones sont modifiées, bien qu'elles soient toujours organisées en un réseau rectangulaire. Étant donné que le DNF est fortement dépendant de la topologie et qu'il repose sur un noyau d'interaction symétrique [1], rompre la régularité de la topologie sous-jacente doit être fait avec prudence. Une démonstration de ceci a été faite dans [10], où le DNF est adapté à des variétés de forme et de dimension non contraintes, avec des résultats probants. Cependant, les topologies qui y sont présentées sont relativement simples, et le DNF n'a pas encore été testé sur des espaces réduits par apprentissage profond, qui risquent d'être beaucoup plus irréguliers. C'est l'objet de cet article.

3 Méthodes

Cet article reprend en partie les méthodes développées dans [10], où nous utilisons des GNG pour apprendre des variétés de l'espace sensoriel dans chaque modalité (sous-section 3.3), avant de les assembler en un graphe multimodal (3.4), puis d'utiliser un DNF pour créer des comportements dans cette nouvelle topologie (3.5). Nous ajoutons une étape préalable d'apprentissage d'un auto-encodeur, dont l'encodeur servira à réduire la dimension des données d'entrées avant l'apprentissage des graphes (3.2), et le décodeur servira à l'analyse des sorties du DNF (3.6). Les

1. Il a été proposé de rompre la symétrie du côté du DNF, soit par des noyaux asymétriques [6], soit par des distorsions de la topologie à l'aide de renforcements prédictifs [32], mais les deux requièrent une étape d'apprentissage supplémentaire.

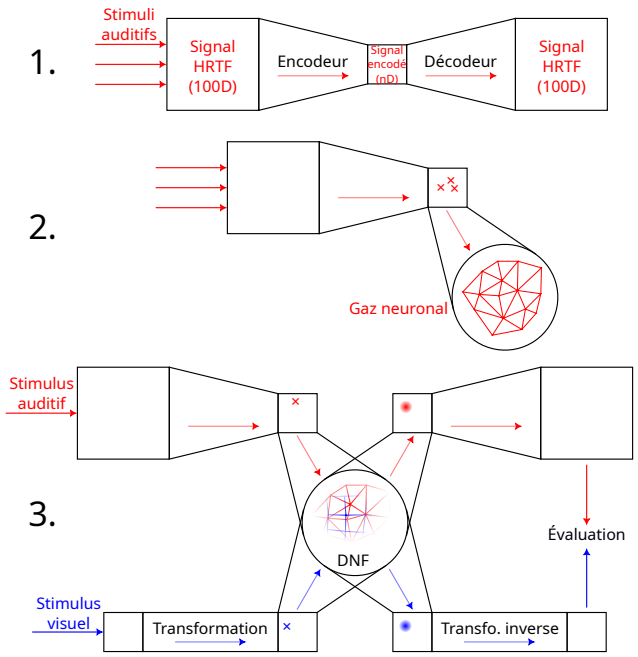

FIGURE 1 – Résumé des principales étapes de cette contribution. **1.** Apprentissage de l'auto-encodeur (sous-section 3.2). **2.** Apprentissage des graphes topologiques (sous-sections 3.3 et 3.4). **3.** Stimulation, émergence d'une activité dans le graphe bimodal et évaluation (sous-sections 3.5 et 3.6).

grandes étapes de cette contribution sont synthétisées dans la figure 1.

3.1 Données

Pour cette contribution, nous nous concentrons sur une architecture bimodale, même si ces travaux pourraient également être applicables à trois modalités ou plus. Nous prenons l'exemple d'un robot devant localiser un signal (e.g., la position d'une personne qui l'appelle) en recevant simultanément des données visuelles (détection de visage) et auditives (spectre de fréquences sonores).

Audition. Une façon de localiser les sources sonores pour les robots consiste à calculer une HRTF, une fonction qui associe des caractéristiques fréquentielles (causées par les interférences de la tête et des pavillons des oreilles sur le signal) à l'orientation de l'origine du signal [4]. Les données procurées par [1] comprennent les réponses enregistrées par un robot équipé de pavillons artificiels, à un son émis depuis différents angles. Étant donnée la position d'un stimulus externe en 2D, nous pouvons interpoler les réponses reçues par les deux oreilles robotiques. Nous calculons ensuite leur transformée de Fourier et faisons la différence entre les oreilles pour obtenir une HRTF. À la fin, chaque entrée audio est à 100 dimensions.

Cette HRTF 100D porte implicitement les informations 2D de la localisation du signal dans le référentiel de la tête du robot : azimut et élévation. C'est le signal 100D que nous donnerons en entrée de l'auto-encodeur.

Vision. La vision artificielle a très généralement une résolution homogène. Cependant, nous pouvons concevoir des cas où la perception visuelle n'est pas parfaitement régulière, par exemple à cause d'une tâche sur l'objectif de la caméra. Nous n'ajoutons pas d'étape d'apprentissage pour la vision ici. Mais, pour tester la robustesse de la fusion dans des modalités de résolution changeante, nous traitons les stimuli visuels comme des points dans un espace 2D, auxquels nous appliquons une transformation bio-inspirée, pour rester dans le même niveau de réalisme que la HRTF.
Nous modifions donc l'espace visuel par une transformation logpolaire. Cette transformation a originellement été utilisée pour décrire chez l'humain la façon dont un stimulus capté par la rétine est projeté sur le colliculus supérieur, une région du cerveau impliquée dans la génération de mouvements oculaires [30]. Elle permet notamment de reconstituer la différence de résolution entre le centre de la rétine et sa périphérie. Elle a déjà été appliquée à des systèmes artificiels, par exemple pour améliorer le contrôle du regard chez les robots [23], ou pour renforcer l'apprentissage d'un réseau de neurones sur des données visuelles [7]. Et elle a déjà été couplée avec un DNF [11, 38].

3.2 Encodage

La réduction de données est effectuée par un auto-encodeur. Nous utilisons en particulier un SWAE (*sliced Wasserstein auto-encoder*), car il permet de former un espace latent euclidien muni d'une distance directement exploitable par le GNG. Il est entraîné sur les données HRTF 100D, avec un espace latent 2D, 5D ou 20D, dans lequel les entrées encodées doivent suivre une distribution uniforme. L'encodeur et le décodeur sont faits d'un réseau de neurones entièrement connecté. Les couches cachées sont séparées par un ReLU pour l'encodeur, et un LeakyReLU pour le décodeur. Le nombre de neurones dans chaque couche est listé dans la table 1.

TABLE 1 – Taille des couches de neurones des encodeurs de SWAE. Les mêmes hyper-paramètres sont utilisés pour le décodeur, en sens inverse.

Entrée	Couches cachées			Sortie
100	90	70	50	20
100		50	20	5
100	64	32	16	2

3.3 Topologies unimodales

L'objectif de cette étape est de créer des graphes représentatifs de l'espace latent de chaque modalité. Ces graphes seront ensuite fusionnés pour servir de support à la prise de décision dans un environnement multisensoriel.
Le choix de l'auto-encodeur peut permettre d'imposer la structure de l'espace latent et la distribution des stimuli encodés dans cet espace. Connaissant ceci, il serait aisé d'échantillonner la distribution par un ensemble de nœuds, et de les connecter en fonction de la structure de la topologie sous-jacente. Il n'est pas indispensable d'ajouter une

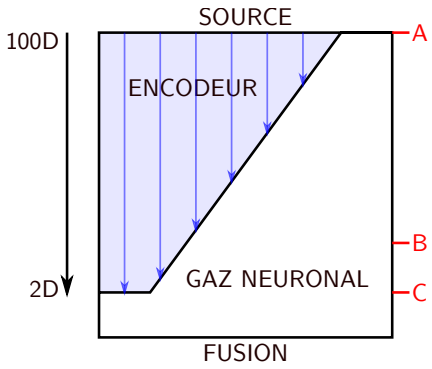

FIGURE 2 – Pré-traitement des données de HRTF avant la fusion. Plusieurs degrés de compression par l'encodeur sont possibles avant la création d'un graphe (ici, par gaz neuronal) utilisé pour la fusion et la prise de décision. Nous testons différents niveaux dans cet article. A : Pas d'encodage. C : Réduction maximale du nombre de dimensions de l'espace d'entrée. B : Réduction intermédiaire, par exemple 20D ou 5D.

étape d'apprentissage. Cependant, nous reprenons le choix de [10] de créer ces graphes à l'aide de GNG. L'algorithme en question n'introduit pas d'hypothèse supplémentaire sur la structure de l'espace latent. Il a l'avantage de fonctionner à la fois sans et avec auto-encodeur, pouvant aussi apprendre une structure sous-jacente de l'espace latent qui diffère de celle imposée dans l'apprentissage du SWAE.
L'algorithme complet du GNG est décrit dans [12]. Pour résumer, le GNG se forme à partir de stimuli tirés aléatoirement. À chaque fois, les deux nœuds dont l'entrée prototypique correspond le mieux au stimulus sont connectés entre eux. Ensuite, l'unité qui correspond le mieux (*best-matching unit*, BMU) et ses voisins topologiques directs voient leur prototype déplacé vers le stimulus. Les connexions qui n'ont pas été mises à jour depuis longtemps sont supprimées, de même que les nœuds isolés. Puis, à intervalles fixes, un nouveau nœud est inséré [2]. Son entrée prototypique va au milieu de la connexion la plus activée.
Cette étape diffère de l'article précédent dans la mesure où le GNG peut recevoir des données plus ou moins compressées par l'encodeur (figure 2). Dans le cas des signaux auditifs, le GNG peut opérer aussi bien sur des HRTF brutes en 100 dimensions que sur un encodage en 2D dans le cas le plus extrême. Pour les signaux visuels, l'encodage est remplacé par un pré-traitement manuel.

3.4 Topologie multimodale

Cette étape est directement reprise de [10] : nous créons un nouveau graphe bimodal qui contient tous les nœuds et arêtes d'une modalité et de l'autre. Pour créer de nouvelles arêtes bimodales, nous connectons des neurones de chaque modalité qui s'activent ensemble, d'une façon inspirée d'un apprentissage Hebbien. Plus précisément, nous tirons une entrée multimodale aléatoire, et si elle se trouve à portée des deux modalités, nous récupérons la BMU de chaque GNG et nous les connectons. Un seul changement est fait par rapport à l'article précédent : afin d'éviter que des nœuds trop épars dans un des GNG se connectent à énormément de nœuds de l'autre modalité, le nombre de connexions intermodales est limité à deux par nœud.

3.5 Sélection d'une activité

L'adaptation du DNF aux graphes créés précédemment est une des contributions centrales de [10]. Nous la récapitulons dans cette sous-section.
Une fois le graphe bimodal créé, les neurones qui lui sont associés peuvent être stimulés par des entrées sensorielles (via leur modalité respective), et nous pouvons utiliser un DNF pour sélectionner un stimulus. Le DNF s'exprime généralement sous la forme d'une équation intégro-différentielle dans un champ continu de neurones, qui est ensuite discrétisée et calculée par la méthode d'Euler (voir eq. 2). L'intégration d'un DNF est comparable à la simulation de réseaux de neurones récurrents en temps continu. Dans le DNF, la distance entre les neurones joue un rôle important, car elle détermine s'ils vont s'exciter ou s'inhiber mutuellement. Notre modèle diffère des autres modèles de la littérature dans la mesure où tous les neurones ne partagent pas un système de coordonnées commun. Nous devons donc adapter l'équation du DNF, afin que les distances soient définies sur le graphe, et seulement cela. Nous nous basons sur la distance standard de la théorie des graphes, c'est-à-dire le nombre d'arêtes sur le chemin le plus court entre deux sommets quelconques [42].
Dans notre modèle, chaque neurone est lié à une modalité spécifique. Ainsi, l'entrée externe reçue individuellement sera spécifique à la modalité (bien que le reste des opérations du DNF ne le soit pas). Pour s'assurer que la quantité totale de stimulation externe soit indépendante de la résolution locale d'une modalité, nous rangeons tous les neurones d'une modalité par ordre de proximité au stimulus (en utilisant la distance euclidienne dans le système de coordonnées de cette modalité), et les excitons en fonction de leur rang par ordre décroissant. Pour chaque neurone indexé k, étant donné un stimulus indexé i, on note $r_{k,i}$ le rang de proximité entre l'entrée prototypique de k et les coordonnées de i. La stimulation externe I_k reçue par k est donnée par :

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

où $\lambda_{m,i}$ est l'intensité du stimulus i par rapport à la modalité m de k. Un neurone ne peut recevoir que des entrées externes provenant de sa propre modalité.
Ensuite, nous calculons l'évolution de l'activité dans le graphe au cours du temps. Ce qui suit est complètement indépendant de la modalité. Le potentiel U_k du neurone k est initialisé à 0 et mis à jour de façon incrémentale par [3] :

2. Dans cet article, le nombre de nœuds est plafonné à 500 dans chaque modalité.

3. Dans cette équation, seul U_k est incrémenté dans le temps, et les entrées I_k sont statiques. Cependant, aucune de nos hypothèses n'empêche les entrées d'évoluer au cours du temps. Nous faisons ce choix car les entrées dynamiques ne sont pas nécessaires pour les résultats présentés dans cet article. Sinon, l'équation (2) pourrait être réécrite en exprimant $U_k(t)$ comme une fonction de $U_*(t - \Delta t)$ et $I_k(t)$.

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

où Δt est le temps de simulation entre chaque étape, τ une constante de temps qui détermine la vitesse de mise à jour du DNF, f une fonction d'activation (ReLU), et h un potentiel de repos négatif. $< \cdot, \cdot >$ désigne la distance minimale en nombre d'arêtes entre deux nœuds dans le gaz neuronal bimodal, et W est une fonction de poids exprimée comme suit :

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

avec des amplitudes $\lambda_+ > \lambda_- > 0$ et des largeurs $\sigma_+ < \sigma_-$. W peut être vu comme un noyau en forme de chapeau mexicain [2].

3.6 Évaluation

Une façon possible d'interpréter la décision est de lire la sortie $f(U)$. Il est courant de prendre un barycentre de la sortie comme estimateur de la position sélectionnée par le modèle. En l'occurence, les coordonnées des nœuds sur lesquels le DNF évolue ne sont pas directement exploitables. Nous devons d'abord décoder ces coordonnées, soit en inversant la transformation logpolaire pour la modalité visuelle, soit en utilisant le décodeur appris précédemment pour la modalité auditive (figure 1, étape 3). Les HRTF décodées sont reliées à des coordonnées 2D par interpolation dans la base de données de signaux audio. La somme de toutes les coordonnées 2D pondérées par l'activation $f(U)$ donne la position perçue du signal.

À des fins d'évaluation, nous comparons la position perçue à la position réelle du stimulus. Ceci nous donne un indicateur de la précision du modèle de fusion, même si cette étape n'est pas indispensable à ce stade. Le modèle pourrait fonctionner sans que nous ne possédions de manière supervisée de replacer, pour chaque modalité, chaque nœud de l'espace latent dans un système de coordonnées 2D intelligible (azimut–élévation). En effet, si l'espace latent contient des représentations internes de l'espace de décision, une décision prise par le DNF pourrait être transformée directement en action sur l'environnement, sans décodage explicite. Le décodage n'est pas non plus nécessaire à la fusion. Nous le faisons donc principalement pour l'évaluation et la visualisation.

4 Résultats

4.1 Topologies apprises

Un graphe est créé à partir de données audio encodées par un SWAE (figure 3). Plusieurs niveaux de compression des dimensions ont été testés (voir figure 2) : pas d'encodage (A), une réduction intermédiaire (B) vers 20 ou 5 dimensions, ou une réduction maximale (C) vers 2 dimensions.

Le premier GNG est le même qu'obtenu dans [10] (figure 3, première ligne). Une fois replacé dans des coordonnées 2D, il paraît assez régulier. L'allongement apparent du graphe le long de la direction azimutale est dû aux conditions matérielles de la perception des sons, qui rendent une discrimination gauche/droite plus facile à mener qu'une discrimination haut/bas. Une analyse plus poussée est proposée dans l'article cité.

En ajoutant un SWAE avec un espace latent imposé en 20 dimensions (figure 3, deuxième ligne), on constate que le GNG produit est très similaire au premier. Les données pertinentes dans notre évaluation (azimut et élévation) n'ont presque pas été dégradées. Cela signifie qu'il n'y a aucune perturbation à anticiper dans les propriétés du modèle de fusion, et ce, même après l'insertion d'un réseau de neurones qui transforme les données sans avoir connaissance des descripteurs les plus pertinents dans cette tâche.

Cependant, la conservation de ces propriétés n'est pas systématique en fonction de l'encodage employé. La forte dégradation des GNG 5D (figure 3, troisième ligne) et 2D (quatrième ligne) montre qu'une compression trop forte peut aboutir à une perte partielle des informations d'azimut et d'élévation. En effet, le SWAE n'a pas de raison de privilégier ces deux caractéristiques. Il sélectionne seulement les descripteurs qui permettent de mieux encoder les données en quelques dimensions et de les reconstruire. Dans les cas 5D et 2D, on voit que le GNG reste assez régulier en périphérie, et est particulièrement dégradé entre -40 et 40 degrés environ. C'est cohérent avec le fait que la localisation des sons est plus facile sur les côtés de la tête. Le SWAE a ainsi appris que la HRTF était mieux encodée par les informations de localisation sur les côtés, mais pas au centre du champ perceptif, où ces informations sont perdues au profit d'autres caractéristiques, moins pertinentes dans notre tâche, mais plus utiles pour la reconstruction des données par le décodeur.

Cette dégradation involontaire ne rend pas le nouveau GNG inexploitable, d'autant plus dans un contexte de fusion de données, où une autre modalité peut apporter des informations complémentaires. Nous le testons avec l'audition encodée en 2D (le cas le plus critique) et la vision. Comme le GNG 5D est qualitativement proche du GNG 2D, nous nous attendons à ce que les performances soient similaires avec ce nombre de dimensions. De même, nous n'attendons pas de différence dans le comportement du modèle entre des données encodées en 20D et des données non encodées, puisque les GNG obtenus sont qualitativement très proches.

4.2 Évaluation en deux dimensions

Pour cette expérience, nous apprenons un second GNG à partir de données visuelles 2D altérées par transformation logpolaire (figure 4). Comme prévu, la résolution est beaucoup plus élevée au centre du champ de vision [4] qu'en périphérie. Ensuite, comme décrit en section 3.4, nous connectons ce GNG visuel au GNG auditif 2D pour former un nouveau graphe bimodal (figure 5).

Sur ce graphe multimodal, nous pouvons envoyer des stimuli audiovisuels et faire opérer un DNF pour sélection-

4. L'absence de nœuds autour d'un azimut de zéro est due à la présence d'un log dans la transformation. Le modèle sur lequel elle s'appuie est originellement prévu pour deux hémi-champs disjoints.

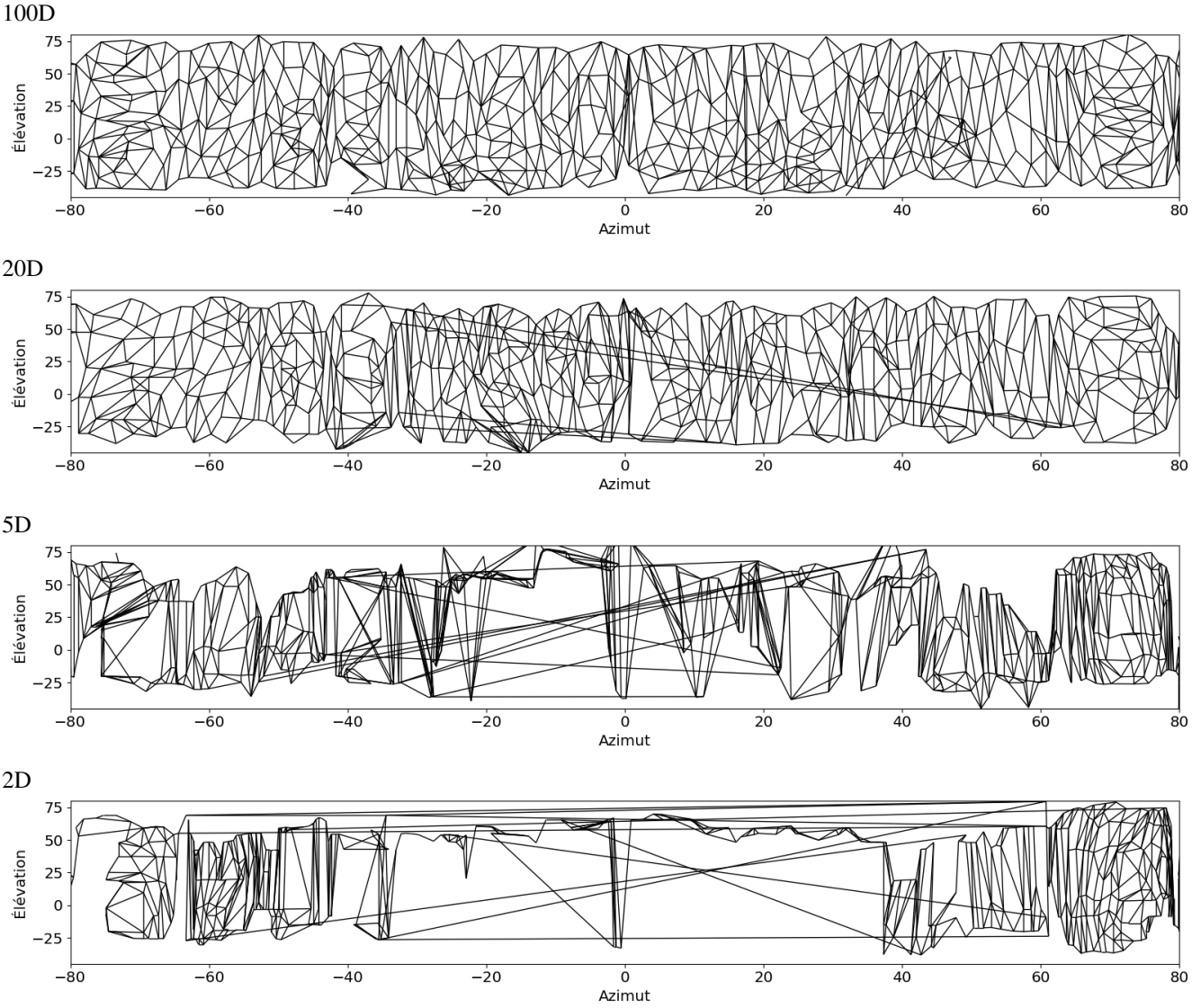

FIGURE 3 – GNG créés à partir de données audio. Le premier est appris à partir des données HRTF 100D sans encodage. Les autres sont appris à partir d'un encodage vers un espace latent 20D, 5D ou 2D. L'affichage des nœuds en 2D est réalisé en décodant leurs entrées prototypiques avec le SWAE le cas échéant, puis en interpolant les azimut et élévation d'après la base de données de HRTF. Notez que dans cet affichage, les deux axes n'ont pas la même échelle.

ner une localisation. Nous évaluons la distance entre la position trouvée et la position réelle de la source pour plusieurs azimuts (figure 6). Afin de mitiger l'effet du choix de l'élévation sur la sélection (la précision peut varier selon que la position du stimulus coïncide par hasard avec l'entrée prototypique d'un des nœuds, ou qu'au contraire elle soit très éloignée de la BMU), nous testons les élévations $[-30, -25, -20, \ldots, 30]$ et gardons la valeur de distance moyenne.

Les performances dans les cas unimodaux confirment nos observations précédentes. Hormis un artefact au centre, la perception visuelle est plus précise vers le centre qu'en périphérie, et inversement pour l'audition. Dans le cas bimodal, la perte est généralement entre les pertes subies par chaque modalité seule. Même si l'amélioration n'est pas franche, elle reste intéressante étant donnée la forte dégradation de la topologie par un SWAE qui a comprimé la HRTF en 2D sans avoir de raison explicite de conserver les informations utiles à la localisation.

5 Conclusion et perspectives

Notre modèle étend une précédente contribution, en ajoutant un SWAE pour encoder des données en amont de la création d'un graphe multimodal, sur lequel un DNF peut faire de la fusion. Dans une certaine mesure, il est possible de réduire la dimensionnalité des données d'entrée par un auto-encodeur sans aucun impact à anticiper sur l'efficacité du modèle de fusion. Toutefois, il apparaît une borne à la force de la compression, au-delà de laquelle des informations pertinentes risquent d'être perdues. Mais la perte sera en partie compensée par l'autre modalité après la fusion.

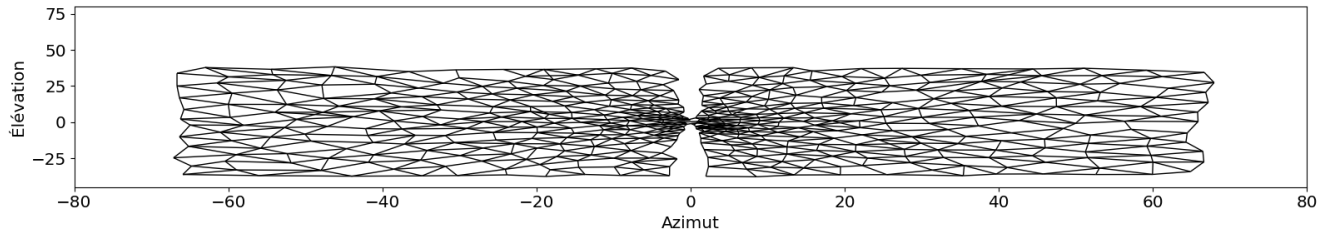

FIGURE 4 – GNG appris sur des données visuelles 2D déplacées par une transformation logpolaire

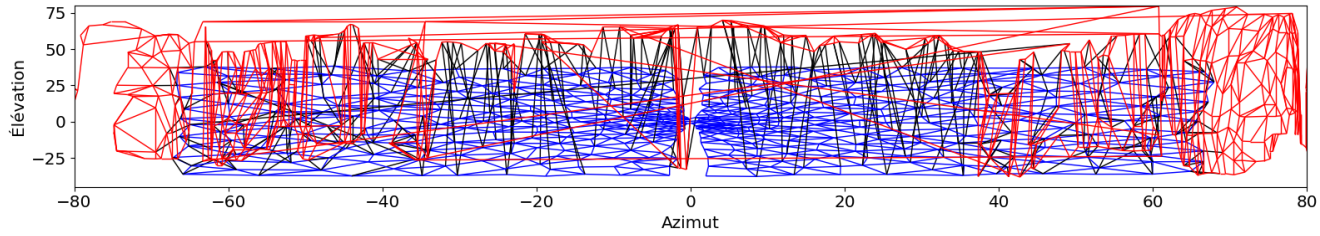

FIGURE 5 – Graphe audiovisuel obtenu en associant un GNG visuel et un GNG auditif 2D. Les connexions intra-visuelles sont affichées en bleu, intra-auditives en rouge, et inter-modales en noir.

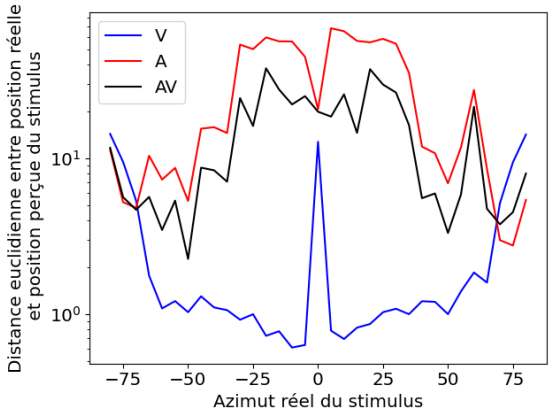

FIGURE 6 – Perte de précision dans la localisation d'un stimulus dans une topologie uniquement visuelle (V), uniquement auditive (A), ou bimodale (AV). L'ordonnée représente la distance en échelle logarithmique entre les positions réelle et perçue du stimulus. Pour chaque azimut, la distance est moyennée sur plusieurs élévations.

Les résultats présentés ici pourraient être améliorés de plusieurs façons. D'abord, dans la figure 6, il aurait été souhaitable que la fonction de distance dans le cas audiovisuel suive la meilleure des deux modalités, voire fasse mieux que les deux. Les résultats présentés précédemment dans [10] tendaient à montrer qu'il était possible de conserver les meilleurs propriétés des deux modalités. Mais il semblerait que la dégradation des informations de localisation auditives par le SWAE, quand l'espace latent est imposé en 2D sans supervision sur la pertinence des dimensions à conserver, soit trop détrimentale pour que la fusion des GNG rende les modalités vraiment complémentaires. C'est pour cette raison que nous avons eu besoin de limiter le nombre de connexions intermodales pour chaque nœud, car sinon les nœuds situés en haut du GNG auditif se connecteraient à toute une colonne de noœuds visuels, effaçant indirectement la perception de l'élévation dans tout le centre du champ de vision. Bien entendu, il serait envisageable de contraindre la structure de l'espace latent pour que, même en 2D, le SWAE apprenne à garder les propriétés de localisation sonore qui nous intéressent. L'optimisation du réseau de neurone pour accomplir cette tâche précise serait une perspective de prolongement de nos travaux. Cependant, des tests préliminaires que nous avons menés sur d'autres types d'auto-encodeurs (notamment VAE) semblent confirmer que les propriétés géométriques de l'espace latent formé par un SWAE sont bien indispensables à la création d'un graphe compatible avec notre méthode de fusion, même s'il se posera inévitablement la question de la superposition d'espaces latents de modalités différentes.

La principale nouveauté de notre modèle est que nous sommes désormais capables de coupler, d'une part, un apprentissage profond, à, d'autre part un modèle de fusion peu coûteux en apprentissage et ayant accès à des propriétés intéressantes (i.e., les capacités de sélection, attention, mémoire, etc. qui ont longuement été développées dans la littérature du DNF). L'ajout de réseaux de neurones ouvre la porte à la manipulation d'espaces d'entrée bien plus complexes. Un exemple serait la reconnaissance d'émotion. Les informations sur l'émotion d'un individu peuvent être perçues par plusieurs canaux : reconnaissance visuelle des expressions du visage, reconnaissance auditive du timbre de la voix, traitement du langage naturel… De nombreux travaux proposent de fusionner ces modalités dans le domaine de l'apprentissage profond, mais pas avec un DNF, car il

n'est pas capable d'intégrer des données aussi complexes. Notre méthode crée une opportunité de le faire.

Parmi les extensions possibles, et notamment dans le cas de tâches aussi complexes que la reconnaissance d'émotions, il serait intéressant d'étudier l'apprentissage simultané de plusieurs auto-encodeurs pour des modalités différentes. Une première piste, qui serait plus utile dans notre exemple de localisation, serait d'utiliser une modalité pour superviser l'autre. Nous entraînerions un auto-encodeur sur les données auditives, pour se conformer non pas à une distribution fixée arbitrairement, mais à la distribution latente des données visuelles, dont on sait qu'elles sont apprises avec une meilleure précision. Ce type de solution a déjà été exploré dans la littérature [26,37]. Ce serait une manière de privilégier lors de l'encodage les informations qui correspondent d'une modalité à une autre.

Cependant, il ne faut pas compter sur une correspondance systématique entre les modalités. Par exemple, il n'est pas garanti que tous les descripteurs d'une émotion soient accessibles aussi bien par reconnaissance visuelle que par traitement du langage : dans le langage seul, il est notoirement difficile de distinguer si un compliment est sarcastique ou non, là où l'expression du visage permet de faire plus facilement la distinction entre un sentiment heureux ou en colère. En forçant une corrélation visuo-langagière, nous risquerions d'entraîner un auto-encodeur (côté traitement du langage) à reconnaître dans du bruit un descripteur (existant en vision) auquel il n'a en fait pas accès. Une piste alternative serait de modifier la distance de Wasserstein pour qu'à l'entraînement, un auto-encodeur accorde moins d'importance à une dimension sur laquelle une autre modalité est plus efficace. En entraînant les modalités en parallèle, nous pourrions nous assurer de leur bonne complémentarité, et laisser ensuite le DNF réaliser la fusion et la prise de décision sans apprentissage supplémentaire.

Remerciements

Ces travaux ont été soutenus par la région Auvergne-Rhône-Alpes via l'initiative Pack Ambition Recherche (projet AMPLIFIER), ainsi que l'Agence Nationale de la Recherche – institut 3IA – MIAI@Grenoble Alpes (ANR-19-P3IA-0003).

Une partie des calculs présentés dans cet article ont été réalisés grâce aux infrastructures de GRICAD (`https://gricad.univ-grenoble-alpes.fr`), soutenue par les communautés de recherche de Grenoble.

Références

[1] 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)*, pages 99–102. IEEE, 2001.

[2] S.-I. Amari. Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, 27(2) :77–87, 1977.

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

[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*, 34(1) :87–112, 2015.

[5] Y. Bengio, A. Courville, and P. Vincent. Representation learning : A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8) :1798–1828, 2013.

[6] M. Cerda and B. Girau. Asymmetry in neural fields : a spatiotemporal encoding mechanism. *Biological cybernetics*, 107(2) :161–178, 2013.

[7] G. Dabane, L. U. Perrinet, and E. Daucé. What you see is what you transform : Foveated spatial transformers as a bio-inspired attention mechanism. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2022.

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

[9] J. Fix, N. Rougier, and F. Alexandre. A dynamic neural field approach to the covert and overt deployment of spatial attention. *Cognitive Computation*, 3(1) :279–293, 2011.

[10] S. Forest, J.-C. Quinton, and M. Lefort. Combining manifold learning and neural field dynamics for multimodal fusion. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2022.

[11] S. Forest, J.-C. Quinton, and M. Lefort. A dynamic neural field model of multimodal merging : application to the ventriloquist effect. *Neural Computation*, 34(8) :1701–1726, 2022.

[12] B. Fritzke. A growing neural gas network learns topologies. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*, volume 7. MIT Press, 1995.

[13] 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)*, pages 38–42. IEEE, 2021.

[14] Q. Houbre, A. Angleraud, and R. Pieters. Balancing exploration and exploitation : a neurally inspired mechanism to learn sensorimotor contingencies. In *Human-Friendly Robotics 2020 : 13th International Workshop*, pages 59–73. Springer, 2021.

[15] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, and J. Carreira. Perceiver : General perception with iterative attention. In M. Meila and T. Zhang,

editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4651–4664. PMLR, 2021.

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

[17] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1) :59–69, 1982.

[18] S. Kolouri, P. E. Pope, C. E. Martin, and G. K. Rohde. Sliced wasserstein auto-encoders. In *International Conference on Learning Representations*, 2019.

[19] S. Krishnagopal and J. Bedrossian. Preserving data manifold structure in latent space for exploration through network geodesics. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2022.

[20] S. Lallee and P. F. Dominey. Multi-modal convergence maps : from body schema and self-representation to mental imagery. *Adaptive Behavior*, 21(4) :274–285, 2013.

[21] M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg. Making sense of vision and touch : Self-supervised learning of multimodal representations for contact-rich tasks. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8943–8950. IEEE, 2019.

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

[23] L. Manfredi, E. S. Maini, and C. Laschi. Neurophysiological models of gaze control in humanoid robotics. In B. Choi, editor, *Humanoid Robots*, chapter 10. IntechOpen, Rijeka, 2009.

[24] S. Marsland, J. Shapiro, and U. Nehmzow. A self-organising network that grows when required. *Neural networks*, 15(8-9) :1041–1058, 2002.

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

[26] S. Moon, S. Kim, and H. Wang. Multimodal transfer deep learning with applications in audio-visual recognition. *arXiv preprint arXiv :1412.3121*, 2014.

[27] O. Ménard and H. Frezza-Buet. Model of multi-modal cortical processing : Coherent learning in self-organizing modules. *Neural Networks*, 18(5) :646–655, 2005.

[28] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34 :14200–14213, 2021.

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

[30] F. P. Ottes, J. A. V. Gisbergen, and J. J. Eggermont. Visuomotor fields of the superior colliculus : A quantitative model. *Vision Research*, 26(6) :857–873, 1986.

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

[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*, pages 1629–1636. IEEE, 2011.

[33] J.-C. Quinton and L. Goffart. A unified dynamic neural field model of goal directed eye movements. *Connection Science*, 30(1) :20–52, 2018.

[34] Y. Sandamirskaya. Dynamic neural fields as a step toward cognitive neuromorphic architectures. *Frontiers in Neuroscience*, 7 :276, 2014.

[35] 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)*, volume 4, pages 2523–2528, 2004.

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

[37] S. Stojanov, A. Thai, and J. M. Rehg. Using shape to categorize : Low-shot learning with an explicit shape bias. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1798–1808, 2021.

[38] 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*, 109(4) :549–559, 2015.

[39] J. Tekülve, A. Fois, Y. Sandamirskaya, and G. Schöner. Autonomous sequence generation for a neural dynamic robot : scene perception, serial order, and object-oriented movement. *Frontiers in neurorobotics*, 13 :95, 2019.

[40] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 247–263, 2018.

[41] M. Vavrečka and I. Farkaš. A multimodal connectionist architecture for unsupervised grounding of spatial language. *Cognitive Computation*, 6(1) :101–112, 2014.

[42] D. B. West et al. *Introduction to graph theory*, volume 2. Prentice hall Upper Saddle River, 2001.

[43] 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)*, pages 5066–5074, 2017.