Visual Learning for Reaching and Body-Schema with Gain-Field Networks

Julien Abrossimoff, Alexandre Pitti, Philippe Gaussier

Visual Learning for Reaching and Body-Schema with Gain-Field Networks

Julien Abrossimoff, Alexandre Pitti, Philippe Gaussier
ETIS Laboratory, UMR 8051, University of Paris-Seine, University of Cergy-Pontoise, ENSEA, CNRS, France.
Email: julien.abrossimoff, alexandre.pitti, gaussier@ensea.fr

***Abstract*—Perceiving our own body posture improves the way we move dynamically and reversely, motion coordination serves to learn better the position of our own body. Following this idea, we present a neural architecture toward reaching movements and body self-perception from a developmental perspective. Our framework is based on the neurobiological mechanism known as gain modulation in parietal neurons that is found to integrate the visual, motor and proprioceptive information through product-like processes. These multiplicative networks have interesting properties for learning nonlinear transformations such as the head-centered mapping in reaching tasks or the hand-centered mapping for a body-centered representation. In a simulation of a three-link arm, we perform experiments of nearby and far reach targets exploiting one or the other strategy. The later combination of the two networks generates autonomous control toward the target by processing the body-centered spatial information and the preferred visual direction for the desired motor commands.**

I. Introduction

Body perception and motion control are the two complementary sides of the same flip coin. As we acquire a better representation of where our body is, we are more precise to reach various locations in space. Reversely, knowing how to move serves to recognize more accurately the actual location of each body part, even without seeing it. Developmental studies in newborns and in 6 month-old babies have shown that self-recognition and reaching are gradually acquired following separated paths [1], [2] and progressively combined to learn a spatial representation of the body (the body schema) and a repertoire of actions (the motor synergies) by aligning the visual and the proprioceptive information. We propose to follow such paradigm for robot control in reaching tasks with a neural architecture that learns to minimize jointly any uncertainties about the robot's location (where the arm is) and about its motor commands (how to move it).

Within the brain, the parietal cortex is one important area for the development of spatial cognition and multimodal integration [3], [4]. For instance, multisensory neurons have been found to monitor nearby objects in the peripersonal space [5], [6]. These neurons combine diverse incoming information from multiple modalities to process multiple body-centered coordinate systems invariant to a motion. They are sensitive to any objects entering within the visual receptive fields anchored at specific body locations (e.g. hand-centered) even if the eyes or the hand is moving. This spatial information of objects into hand-centered coordinate frames is used then for biasing ongoing movement trajectory for grasping or for defense behaviors [5]. Similarly, multimodal neurons have been found in the motor cortex to be activated with respect to where the hand is moving [7], [8], [9]. In both regions, we observe a neural field activity sensitive to both the preferred motor activity and to the preferred visual orientation [10]. Since these neurons respond not only to one type of signal but to various information, either visual, proprioceptive, audio or tactile signals, they are called conjunctive cells or gain-field neurons and their amplitude encodes a joint distribution of various input [11], [12].

We propose to exploit the properties of these neurons in order to construct two compartmented neural networks that processes (a) a visuomotor network for inverse dynamics (reaching) and (b) visual spatial locations in body-centered coordinates based on body posture, motor information and vision, see Fig. 1. The two networks are not fully linked from each other at first but they can later iteratively update their prediction from incoming signal error to learn better how to move and where the hand is (dashed gray line). Their modeling corresponds to multiplicative Radial Basis Functions (RBFs) or sigma-pi networks [13], [14] to learn sensorimotor transformations. In image processing, these networks are known as gated networks, which have been recently re-investigated in [15], [16] for affine transformations and in developmental robotics [17], [18], [19] for multimodal integration. These multiplicative networks can serve to learn nonlinear transformations, which are common problems in robotics to compute direct mapping and inverse kinematics. In robotics, different authors have proposed several approaches in line with ours. Sturm and colleagues employed Bayesian networks to simultaneously identify a robot's kinematic structure and to learn the geometrical relationships between its body parts as a function of the joint angles [20]. Lanillos et al. used them for constructing a probabilistic body map for self-perception [21]. Besides, [22] employed self-organizing maps for learning inverse-forward kinematics for self-perception whereas [23] and [24] used them for the learning of tactile maps and body image. In comparison, gain-field networks can combine advantageously the topological self-organization property of SOM with auto-encoding and the nonlinear probabilistic mapping property of Bayesian networks based on multiplication.

After we describe the neural architecture used, we present results using gain-field networks (1) for reaching tasks (learning an inverse model) and (2) for learning a body-centered

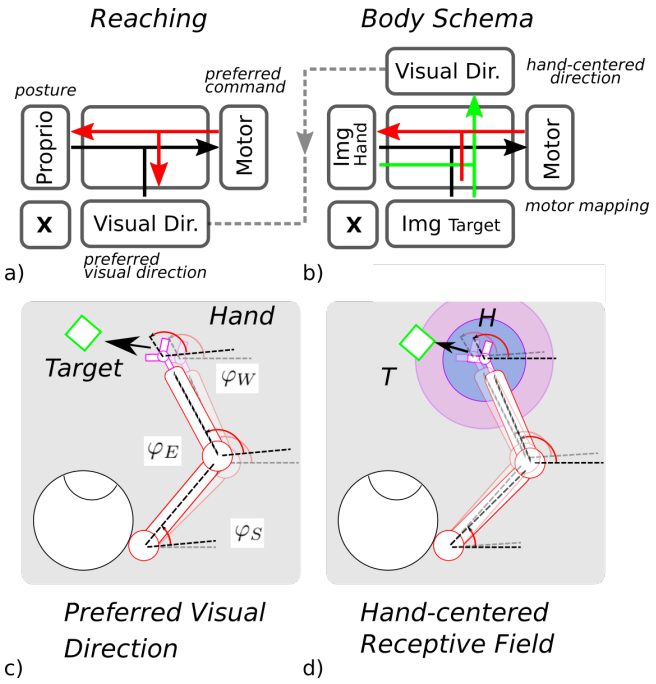

Fig. 1. Gain-field networks for reaching tasks and body schema from visuo-, motor- and proprioceptive integration. The GF network in a) learns motor commands M for reaching movements by integrating the *what* information of the arm's posture P that refers to the encoding of the triplet $\{\varphi_S, \varphi_E, \varphi_W\}$ and the *where* signal of the visual direction V. The results are reaching movements with preferred visual directions, see c). The GF network in b) learns the relative hand-centered reference frame, the direction V, from mapping of the motor commands M and the visual input of the locations of the hand H and of the target T. The results are hand-centered receptive field for grasping tasks of nearby targets, see d). Once the two networks have been learned, the two strategies can be combined for reaching targets relative to the body location (gray dashed link).

coordinate system (forward model mapping) with a three-link arm simulation in 2D. We show that few learning steps are necessary to make the robotic arm to reach various regions, binding desired visual orientation and arm posture in order to compute the most preferred motor command. Reversely, we exploit the same type of network in order to learn a body-centered coordinate system (a body schema) depending on the selected motor command and the actual visual information (forward model). The two systems can be coupled so that estimation of nearby objects in hand-centered reference frames can serve to construct a desired visual orientation vector for moving the hand in that direction. The two learning systems can work in a complementary way so that prediction error in one can serve to adapt the other for position estimation and motion control.

We discuss then the relevance of our approach for autonomous robotics (learning inverse/forward models and image-based nonlinear transformation) and its utility for more complex tasks not presented here (grasping, tool-use, body-centered representations, spatial inference, social interaction), and its implication in neurorobotics for modeling the so-called Mirror Neurons System and the Ventral Inferior Parietal (VIP) neurons.

II. Architecture and Neural Mechanisms

Gain-modulated networks are an instance of sigma-pi networks constituted of radial basis functions, pre-defined parametrically or learned, which produce a weighted sum of joint probability distributions as output [13].

The output terms Y are a linear combination of the product of the input variables X and H whose cardinalities are respectively n_Y, n_X and n_H, so that predicting $\hat{Y}$ consists on computing for all values Y_k of Y, $k \in n_Y$

$$\forall k, Y_k = \sum_i^{n_X} \sum_j^{n_H} W_{ijk} X_i \cdot H_j, \tag{1}$$

with W synaptic coefficients in $n_X \times n_H \times n_Y$. Since this matrix can be quite large, a way to reduce the dimensionality of the gain-fields networks is to categorize first each input variable with factored functions or basis functions, f_X and f_H, in order to have $f_X = \sum W_X X$ and $f_H = \sum W_H H$. The computational complexity is reduced then to $\mathbb{R}_{f_X} + \mathbb{R}_{f_H} + \mathbb{R}_{f_X} \times \mathbb{R}_{f_H}$ and the output function becomes

$$Y = \sum W_Y f_X \cdot f_H, \tag{2}$$

$$= \sum W_Y (\sum W_X X) \cdot (\sum W_H H). \tag{3}$$

The global error E is defined as the Euclidean distance calculated between Y and $\hat{Y}$ for all the input examples. The optimization function used for learning the synaptic weights of the output network is the classical stochastic descent gradient. All synaptic weights can be updated in one step with back-propagation, but in our experiments, we make to learn separately each subnetwork X and H before computing Y toward the desired value Y^* using the Widrow-Hoff learning rule. This process is more in line with our previous works [17] and differs slightly from [25]:

$$\Delta W_Y = \epsilon (Y^* - Y)(f_X \cdot f_H). \tag{4}$$

To reconstruct back one of the input variable X (or H), we can use the same network architecture as eq. 1 but implemented in mirror as an auto-encoder with now the global error E estimated from the difference between the actual variables X (resp. H) and the retrieved ones $\hat{X}$ (resp. $\hat{H}$); see also red lines in Fig. 1. The retrieved values from this second network are computed from the output $\hat{Y}$ calculated from the first network.

$$\hat{X} = \sum W_{\hat{X}} \hat{Y} \cdot H, \tag{5}$$

$$= \sum W_{\hat{X}} (\sum W_Y X \cdot H) \cdot H. \tag{6}$$

In this configuration, the two networks form a coupled system similar to an auto-encoder. Each neuron $\hat{Y}$ in the intermediate layer represents a latent representation of the input variables $X \cdot H$, a joint distribution.

This property is interesting for sensorimotor learning and multimodal integration because each hidden unit $\hat{Y}$ categorizes a nonlinear transformation, which could be caused by a motor command or a spatial mapping from one reference frame to another. Therefore, this network can be used not only for reconstructing back one missing modality from two others but it can serve also to identify which hidden variables have caused it based on the two other information. For instance, in section IV-B, we use this feature to retrieve back the motor commands Y that have generated the moving of the hand located at H to the visual goal located at X. This corresponds to the construction of a visual preferred direction in hand-centered coordinates, a body schema.

III. Experimental Setup

We set up our experiments using a 2D simulation of a three links manipulator. The arm simulator is used in all experiments of section IV from visual, motor and proprioceptive integration.

IV. Results

A. Visuo-Proprioceptive Integration for Motion Control

The GF network achieves motion coordination by learning the relationship between the proprioceptive information P and a desired visual orientation V to derive a preferred motor command C in that direction, see Fig.1 a). We give as inputs to the GF network the proprioceptive information P of the robotic arm, which corresponds to its three joint angles φ_S, φ_E and φ_W resp. coding the shoulder, elbow and wrist angles, and the preferred visual orientation V, in radian. In this section, we make the note that this desired visual orientation V in radian is computed algebraically from the coordinates of the hand position in (x_H, y_H) and the coordinates of the target (x_T, y_T) with the `arctan2` function. The `arctan2` function is replaced in sections IV-B and IV-C with the network that reconstructs the relative hand-target angle visually in hand-centered coordinates.

The feature functions f_P and f_V from eq. 3 consist on 10 units each, factorizing the input space. The desired motor command C^* sees its synaptic weights reinforced with the compositional matrix $f_P \cdot f_V$ following a gradient descent. The output vector C consists on 27 units corresponding to $3^3 = 27$ different motor synergies of the shoulder-elbow-wrist motion triplet $\{\Delta\varphi_S, \Delta\varphi_E, \Delta\varphi_W\}$ whose discrete values are comprised in the small repertoire of three speeds $\{-\Delta\varphi, 0, +\Delta\varphi\}$; i.e., going backward, release or going forward. The resulting network corresponds to an inverse model for reaching tasks guided by desired visual orientation given in absolute coordinates.

The learning stage follows the guidelines presented as previously with one thousand data of the triplet $\{P, V, C\}$ uniformly chosen in the arm workspace. Twenty epochs are necessary to make the network to converge. In order to demonstrate the capability of the GF network for a reaching task, we plot in Fig. 2 the trajectories from three different postures P to eight target locations placed in star around them. The color of each target indicates the Euclidean distance error for reaching it. In these trials, the trajectories are always smooth and curvy, which indicate that the complete repertoire of the 27 motor units contribute to the motion control. During motion, the motor synergies C change with respect to the posture of the arm P best matched by the network and the updated visual orientation V. Depending on the posture, some trajectories are preferred with respect to the learning stage, whereas for unseen postures the network poorly generalizes as for the target locations which are outside the working space beyond the dashed line Fig. 2. This behavior corresponds also to what it has been observed in real motor units that are sensitive to visual direction and proprioception [9].

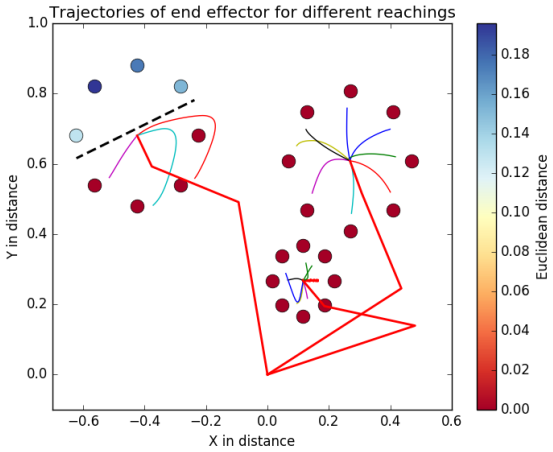

Fig. 2. Reaching task for three different positions to eight target locations placed in star around the wrist location. The red lines represent the arm in its initial configuration and the circles represent the target locations. The trajectories of the wrist are displayed (colored lines) and the color of the target indicates the Euclidean distance to the target location.

We plot in Fig. 3 the density distribution of the reach spatial error in the workspace for almost 900 examples of target locations; its corresponding histogram is plotted in Fig. 8 in blue. In this map, we can see that most of the spatial errors are made for singular configurations at far reach postures at the limit of the peripersonal space or for locations nearby the origin $\{0, 0\}$ where the arm cannot really reach it and for which the network cannot interpolate easily. Nonetheless, the system can generalize mostly in the full space, which defines also the reachable space with physical limits and visual preferred directions.

Another important point is to see how the system can switch from one synergy to another one during a reaching task. In Fig. 4, we display for two different initial postures A and goals B plotted in a) the evolution of motor neuron activities through time and the associated movement of the shoulder, elbow and wrist in c). The figure shows in c) that the system "decomposes" its reaching tasks by switching between synergies. The movement is mostly driven by one strategy in particular controlling the shoulder motion during

the first part of the phase and then by the synergy that combines the wrist and the elbow motions during a second phase. In b) we plot the evolution through time of each synergy activity for each corresponding reaching task (red corresponds to an increase, blue to a decrease), it shows the GF network is able to categorize the reaching tasks by firing one synergy, which corresponds to move the shoulder, elbow and wrist for achieving a down global shift (synergy #16 in left chart b)) or by firing and inhibiting multiple synergies dynamically through time for achieving a more complex movement on the upper left side relative to the initial position (right chart b)). This more complex strategy is represented in c) by switching between a driven shoulder movement to a wrist driven movement at $t = 50$ which corresponds to target final approach.

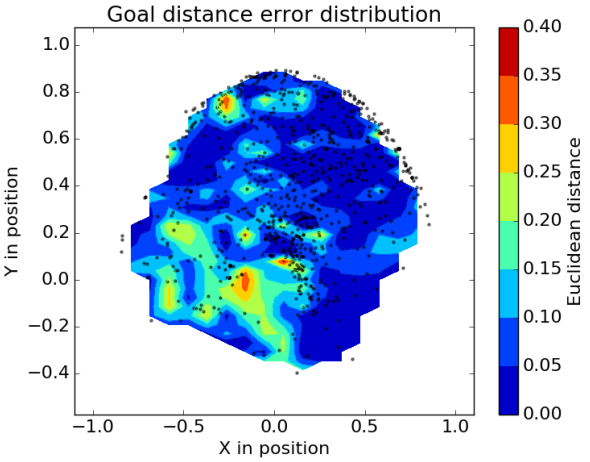

Fig. 3. Density distribution of reach distance error for 864 reach goals. The color intensity indicates the Euclidean distance error between the location of the wrist and the target.

B. Learning Target Visual Orientation in Body-centered Representation

In the previous section, we described a learning architecture for encoding the "physical goal", which is the selectivity of motor cells with the preferred visual directions V at certain postures P. V was assumed to be given.

In this section, we propose to use the same neural architecture but with different inputs and output in order to learn V, see Fig. 1 b) green line. The neural architecture is used for visual goal encoding by integrating the visual locations of the end-effector and of the nearby targets with the corresponding motor commands that permit to reach it [26]. Making a parallel with image processing, the motor command C can be envisioned as a visual transformation from hand image H to target image T. The result will be the learning by output neurons of the necessary transformation to go from the end effector to the nearby target; that is, computing the estimated distance of the reachable space based on motor command, a.k.a the peripersonal space. In the first experiment with our three-links arm simulation, the hand-centered information is

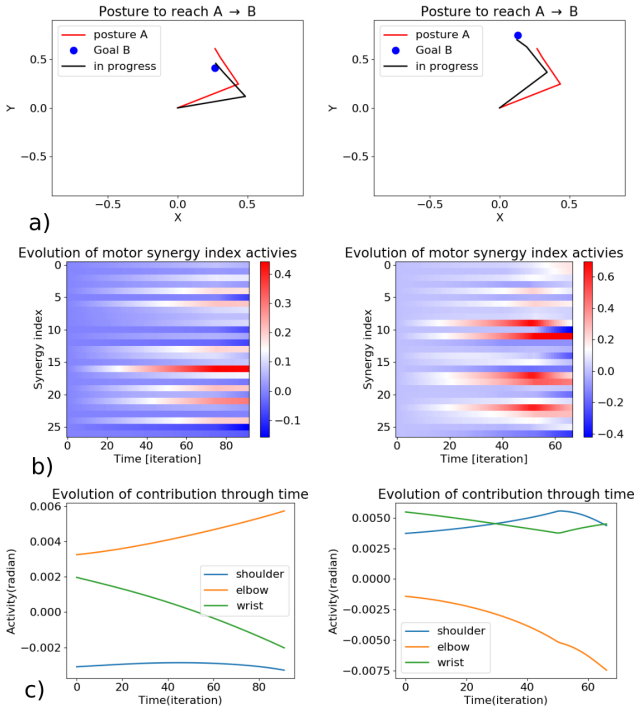

Fig. 4. Representation of activity through time for two different reaching tasks a) and the associated mouvement for each degree of freedom c).In b), the color intensity indicates the evolution of each synergy activy(red for an increase, blue for a decrease).

modeled as is when the hand H reaches the target location T with the motor command C.

The first experiment aims at testing the neural network's ability to encode a hand-centered reference frame based on the visual information of the end-effector location H and of the target location T using the motor command units C as estimators of the performed visual transformation, see Fig. 1 b) black and green lines. During the learning phase, the images H and T with (13×13) pixels are provided as input and the motor units C as output for visuomotor mapping, see Fig. 1 b) the black and red lines. This output layer is composed by 27 motor neurons –, the same number as in the previous section,– and the motor units activated during the supervised learning period are the ones that permit to reach the target position from the current end-effector location.

An example of how the network behaves after the learning stage is showed in Fig. 5. For one particular hand posture H and for different target position T, represented in the left chart column, the motor output units in the right chart column get activated more or less strongly. The activity level of these units corresponds to the estimation of the performed visual transformation necessary for going from H to T. Based on the estimated contribution of each motor unit, on the spatial location of the hand it is possible to reconstruct back the position of the target relative to the hand in the left chart column. This information can serve to model a hand-centered reference frame based on the motor units prediction, see

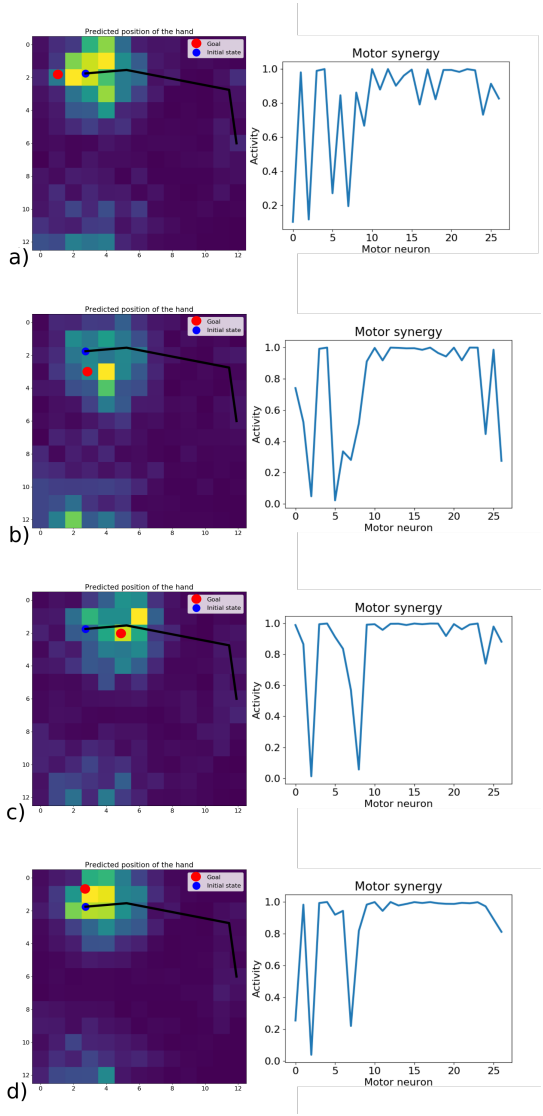

Fig. 5. Input and output information about visual locations of the end-effector H (blue dot) and of the target nearby T (red dot) are provided to the nework, see also Fig. 1 b) (red lines). Each unit in the hidden layer encodes the corresponding motor command C that performs the visual transformation from image H to image T. After learning, each motor unit performs a visual prediction about where the target is relative to the visual hand position, see Fig. 1 b) (black lines). The contribution of each motor unit is plotted in the right chart column and the location estimate in hand-centered reference frame is plotted in superposition in the left chart colum. The yellow colour corresponds to a high activity and the blue colour to a low activity.

Fig. 1 b) (black and red lines), that is, a dynamic body image. The reconstructed target position follows a gaussian distribution centered on the correct location of the target. The variance level is the result of the contribution of each motor unit for which the motor unit 5 here is the most active for this example (lower right chart column). Differently said, each motor unit gives clues about how the hand *could* move to the visual location of the target; i.e., estimating its displacement, which corresponds to the target affordance. This distribution around the hand or the target reproduces well the activity of parietal neurons seen to code the peri-personal space [27].

We plot in Fig. 6 a histogram of the estimated visual orientation reconstructed by the network from H and T images directly learned for various hand-target visual locations. This histogram shows that the estimated orientation error is small and that the network is robust for the majority of the learned examples in the workspace.

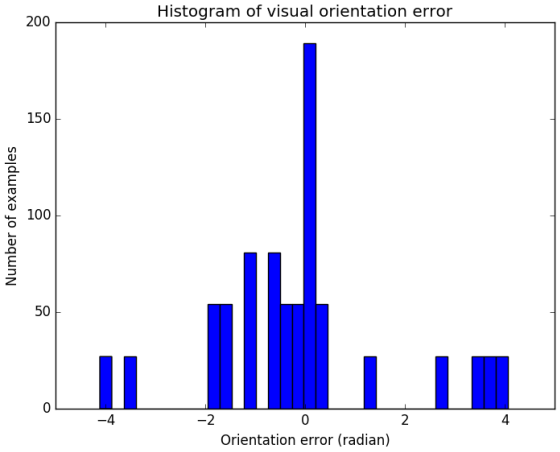

Fig. 6. Histogram of the visual orientation error for 864 target reconstructed by the GF network from hand-target images in various locations.

C. Autonomous Reaching in Body-Centered Visual Coordinates

Once the two networks have learned to reach from global visual direction and from relative hand-centered orientation, respectively in section IV-A and IV-B, it is possible then to cascade their output as examplified by the gray dashed line in Fig. 1. The body schema network in section IV-B predicts the visual orientation V and replaces the $atan2$ function of the reaching network of section IV-A. The result is the autonomous goal-directed reaching of visual targets.

In order to compare the performance of the three reaching strategies with respect to the targets location, using the visual information only, the hand-centered information and the combinaison of the two, we plot in Fig. 7 their trajectories to three nearby and distant target locations. We display also in Fig. 8 the histogram of the reaching performance for the three networks and the histogram of the reaching error depending on the goal distance in Fig. 8. These three graphs aim to explain how the three neural systems behave with respect to the reaching tasks, local or distant.

The trajectories of the Visual Reaching network (VR) plotted in blue in Fig. 7 are similar to the results found in Fig. 2 with smooth reaching directions. We can explain these trajectories as a compromise produced by the different visuo-motor synergies between the visual direction and the linearly combined motor synergies in order to generate a smooth command toward the goal. The result for the visual strategy represents somehow the ground truth as we provide the preferred visual direction information directly to the

network. This network gives also the most accurate results with nearly half percent of success to reach the targets, see Fig. 8. For the two other networks, this information of the preferred direction is learned and estimated online, which perform with less accuracy. Nonetheless, they accomplish overall good results as one-third of the reachings are accurate with a slight advantage for the hand-centered reaching in Fig. 8.

The trajectory of the hand-centered Reaching network (HCR) in green is less efficient than for the Visual network, which performs well only for the two distant targets. In contrast, the combinaison of visual and hand-centered information reaching network (VHCR) is less sensitive to the relative distance to the target as it combines efficiently the two information. In some cases however, the visual direction is not efficiently reconstructed and the trajectory is sub-optimal, see the red dashed line in Fig. 7 in order to reach the target in the bottom. However, its trajectories are smoother and straiter than the VR network.

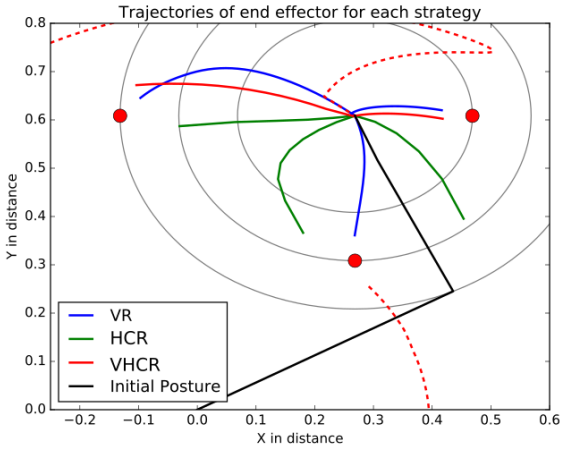

Fig. 7. Reaching trajectories to three different target distance and for the three reaching strategies: visual, hand-centered and visual and hand-centered combinaison.

V. Discussions

We proposed a framework for sensorimotor coordination and multimodal integration based on the product of independent distributions. These type of networks are known as gain-field networks in biology and as gated networks in image recognition. They present several similarities with radial basis functions [13], auto-encoders and Boltzmann networks as well [15]. The multiplicative function has interesting properties to map effective nonlinear transformations from one reference frame to another. This feature can be used to learn the effects of motor activity on different sensor maps and to construct correspondences. Each motor unit of the hidden layer thereby represents a transformation in the sensory space. At reverse, knowing the variation in the sensory maps, it is possible to estimate which transformation (hidden variable) is the most probable to have generated these

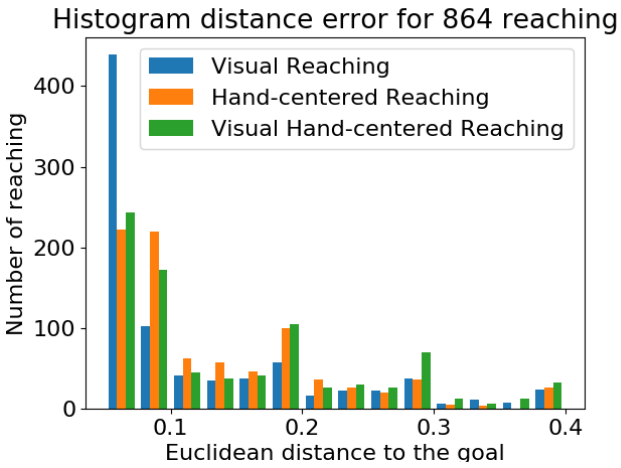

Fig. 8. Histogram of error distance for all reaching performed for the three reaching strategies: visual only, hand-centered and visual and hand-centered combinaison. The visual strategy represents the ground truth. For the later cases, the two networks have to reconstruct the missing visual modality in hand-centered reference frame.

outputs. This property of auto-encoders can serve for active inference and action observation, which are also features observed in parietal neurons and in the mirror neurons system [28], [29], [30] for affordances generatation [31], [32] and also sensorimotor adaptations as during tool-use [18]. During grasping, the prediction done by the motor units of the hidden layer of the auto-encoder can serve to "reverse-engineer" the hand preshaping based on visual information; this idea is also found in Rumelhart or Kawato's forward-inverse models [33], [34] as well as in the "virtual finger hypothesis" by Arbib who proposed to explain grasp affordance and the assignement of the orientation and of the power grip of the real fingers during grasping [31], [32], [35].

Our experiments show that GF networks can develop pointing and reaching functionalities separately, with spatial representations in hand-centered RF and visual direction preferences. In infants, these two mechanisms may grow separately and gradually during the first 6 months till their plausible integration for more robust reaching [1], [3], [2]. The possible implication of the hippocampus to the structuring of the spatial maps in the parietal cortex –, which possesses also GF types of neurons for navigation,– can even provide some hints how this framework can expand for whole body coordination in allocentric and egocentric spaces [36], [37].

In future experiments, we will extend our work with a complete humanoid robot arm with a hand and tactile sensors toward grasping objects, learning spatial representation and observing someone else actions as well. In addition, we have to perform further analysis in order to understand how the prehension synergies can be learned during babbling in a self-organised way; e.g., if the GF network develops coordinated synergies based mostly on the shoulder activity or on the wrist activity only.

ACKNOWLEDGEMENTS

This work was partially supported by grants from the EQUIPEX-ROBOTEX (CNRS), the chaire dexcellence CNRS-UCP and the project Labex MME-DII (ANR11-LBX-0023-01).

REFERENCES

[1] A. Bremner, N. Holmes, and C. Spence, "Infants lost in (peripersonal) space?" *Trends in Cognitive Sciences*, vol. 12, no. 8, pp. 298–305, 2008.

[2] D. Corbetta, S. Thurman, G. Y. Wiener, R.F., and J. Williams, "Mapping the feel of the arm with the sight of the object: on the embodied origins of infant reaching," *Frontiers in Psychology*, vol. 5, p. 576, 2014.

[3] M. Del Giudice, V. Manera, and C. Keysers, "Programmed to learn? the ontogeny of mirror neurons," *Developmental Science*, vol. 12, no. 2, pp. 350–363, 2009.

[4] A. Maravita, C. Spence, and J. Driver, "Multisensory integration and the body schema close to hand and within reach," *Current Biology*, vol. 13, no. 2, pp. R531–R539, 2003.

[5] M. Graziano and C. Gross, "Spatial maps for the control of movement," *Current Opinion in Neurobiology*, vol. 8, pp. 195–201, 1998.

[6] A. Iriki, M. Tanaka, S. Obayashi, and Y. Iwamura, "Self-images in the video monitor coded by monkey intraparietal neurons," *Neuroscience Research*, vol. 40, pp. 163–173, 2001.

[7] S. Kakei, D. Hoffman, and P. Strick, "Sensorimotor transformations in cortical motor areas," *Neuroscience Research*, vol. 46, pp. 1–10, 2003.

[8] G. Blohm, A. Khan, and J. Crawford, "Spatial transformations for eyehand coordination," *Encyclopedia of Neuroscience*, p. 203211, 2009.

[9] A. Georgopoulos, H. Merchant, T. Naselaris, and B. Amirikian, "Mapping of the preferred direction in the motor cortex," *Proc Natl Acad Sci USA.*, vol. 104, no. 26, pp. 11 068–72, 2007.

[10] P. Baraduc, E. Guigon, and Y. Burnod, "Recording arm position to learn visuomotor transformations," *Cerebral Cortex*, vol. 11, no. 10, pp. 906–917, 2001.

[11] A. Pouget and L. Snyder, "Spatial transformations in the parietal cortex using basis functions," *J. of Cog. Neuro.*, vol. 3, pp. 1192–1198, 1997.

[12] E. Salinas and T. J. Sejnowski, "Gain modulation in the central nervous system: Where behavior, neurophysiology and computation meet," *The Neuroscientist*, vol. 7, pp. 430–440, 2001.

[13] A. Pouget and L. Snyder, "Spatial transformations in the parietal cortex using basis functions," *J. of Cog. Neuro.*, vol. 3, pp. 1192–1198, 1997.

[14] D. Bullock, S. Grossberg, and F. Guenther, "A self-organizing neural model of motor equivalent reaching and tool use by multijoint arm," *Journal of Cognitive Neuroscience*, vol. 5, no. 4, pp. 408–435, 1993.

[15] R. Memisevic, "Learning to represent spatial transformations with factored higher-order boltzmann machines," *Neural Computation*, vol. 22, pp. 1473–1493, 2010.

[16] O. Sigaud, C. Masson, D. Filliat, and F. Stulp, "Gated networks: an inventory," *arXiv:1512.03201v1*, 2016.

[17] A. Pitti, A. Blanchard, M. Cardinaux, and P. Gaussier, "Gain-field modulation mechanism in multimodal networks for spatial perception," *12th IEEE-RAS International Conference on Humanoid Robots Nov.29-Dec.1, 2012. Business Innovation Center Osaka, Japan*, pp. 297–302, 2012.

[18] S. Mahe, P. Braud, R. Gaussier, M. Quoy, and A. Pitti, "Exploiting the gain-modulation mechanism in parieto-motor neurons application to visuomotor transformations and embodied simulation," *Neural Networks*, vol. 62, pp. 102–111, 2015.

[19] A. Droniou, I. Serena, and O. Sigaud, "A deep unsupervised network for multimodal perception, representation and classification," *Robotics and Autonomous Systems*, vol. 71, p. 8398, 2015.

[20] J. Sturm, C. Plagemann, and W. Burgard, "Body schema learning for robotic manipulators from visual self-perception," *Journal of Physiology-Paris*, vol. 103, no. 3-5, pp. 220–231, 2009, neurorobotics. [Online]. Available: http://www.sciencedirect.com/science/article/B6VMC-4WY6JVM-D/2/0aaabe9b7dc9628c8c818fa87c8b56e9

[21] P. Lanillos, E. Dean-Leon, and G. Cheng, "Yielding self-perception in robots through sensorimotor contingencies," *IEEE TCDS*, p. to appear, 2017.

[22] E. Escobar-Jurez, G. Schillaci, J. Hermosillo-Valadez, and B. Lara-Guzmn, "A self-organized internal models architecture for coding sensorymotor schemes," *Front. Robot. AI*, vol. 3, no. 22, p. 10.3389/frobt.2016.00022, 2017.

[23] M. Hoffmann, Z. Straka, I. Farkas, M. Vavrecka, and G. Metta, "Robotic homunculus: Learning of artificial skin representation in a humanoid robot motivated by primary somatosensory cortex," *IEEE Transactions on Cognitive and Developmental Systems*, vol. PP, no. 99, pp. 1–1, 2017.

[24] J. Born, J. Galeazzi, and S. Stringer, "Hebbian learning of hand-centred representations in a hierarchical neural network model of the primate visual system," *PLoS ONE*, vol. 12, no. 5, p. e0178304, 2017.

[25] R. Memisevic, "Gradient-based learning of higher-order image features," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1591–1598.

[26] S. Kuang, P. Morel, and A. Gail, "Planning movements in visual and physical space in monkey posterior parietal cortex," *Cereb Cortex*, vol. 2, no. 26, pp. 731–747, 2016.

[27] M. Graziano and D. Cooke, "Parieto-frontal interactions, personal space, and defensive behavior," *Neuropsychologia*, vol. 44, pp. 845–859, 2006.

[28] C. Buneo, A. Jarvis, A. Batista, and R. Andersen, "Direct visuomotor transformations for reaching," *Nature*, vol. 416, pp. 632–636, 2002.

[29] M. Brozovic, A. Gail, and R. Andersen, "Gain mechanisms for contextually guided visuomotor transformations," *The Journal of Neuroscience*, vol. 27, no. 39, pp. 10 588–10 596, 2007.

[30] R. Andersen and H. Cui, "Intention, action planning, and decision making in parietal-frontal circuits," *Neuron*, vol. 63, pp. 568–583, 2009.

[31] E. Oztop, N. Bradley, and M. Arbib, "Infant grasp learning: a computational model," *Exp. Brain Res.*, vol. 158, p. 480503, 2004.

[32] J. Bonaiuto and M. Arbib, "Learning to grasp and extract affordances: the integrated learning of grasps and affordances (ilga) model," *Biological Cybernetics*, vol. 109, no. 6, p. 639669, 2004.

[33] M. Jordan and D. Rumelhart, "Forward models: supervised learning with a distal teacher," *Cognitive Science*, vol. 16, pp. 307–354, 1987.

[34] D. Wolpert, K. Doya, and M. Kawato, "A unifying computational framework for motor control and social interaction," *Philosophical Transactions of the Royal Society*, vol. 358, pp. 593–602, 2003.

[35] A. Pitti, H. Alirezaei, and Y. Kuniyoshi, "Cross-modal and scale-free action representations through enaction," *Neural Networks*, vol. 22, pp. 144–154, 2009.

[36] A. Pitti, H. Mori, Y. Yamada, and Y. Kuniyoshi, "A model of spatial development from parieto-hippocampal learning of body-place associations," *10th International Conference on Epigenetic Robotics*, pp. 89–96, 2010.

[37] J. Hirel, P. Gaussier, M. Quoy, J.-P. Banquet, E. Save, and B. Poucet, "The hippocampo-cortical loop spatio-temporal learning and goal-oriented planning in navigation," *Neural Networks*, vol. 43, no. 0, pp. 8–21, 2013.