Visuo-Motor Control Using Body Representation of a Robotic Arm with Gated Auto-Encoders

Julien Abrossimoff, Alexandre Pitti, Philippe Gaussier

Visuo-Motor Control Using Body Representation of a Robotic Arm with Gated Auto-Encoders

Julien Abrossimoff[1], Alexandre Pitti[1] and Philippe Gaussier[1]

***Abstract*— We present an auto-encoder version of gated networks for learning visuomotor transformations for reaching targets and representating the location of the robot arm. Gated networks use multiplicative neurons to bind correlated images from each others and to learn their relative changes. Using the encoder network, motor neurons categorize the induced visual displacements of the robot arm when applying their corresponding motor commands. Using the decoder network, it is possible to infer back the visual motion and location of the robot arm from the activity of the motor units, aka body image. Using both networks as the same time, near targets can simulate a fictious visual displacement of the robot arm and induce the activation of the most probable motor command for tracking it. Results show the effectiveness of our approach for 2 d-of and 3 d-o-f robot arms. We discuss then about the network and its use for reaching task and body representation, future works and its relevance for modeling the so-called gain-field neurons in the parieto-motor cortices for learning visuomotor transformation.**

I. INTRODUCTION

To move in space, we need to correlate the variations occuring between the sensory space and the motor space. Within the brain, neural circuits in the parieto-motor areas are involved in learning these co-variations across multimodal signals. For instance, observations of the motor neurons have showed their tuning to particular visual motion direction in space [1], [2], [3], [4]. These visuomotor primitives can serve then toward aligning one body limb to one orientation with respect to a target [5] or to minimize the relative distance to it [6]. Interestingly, the way it is done is by multiplicative interaction across the different sensorimotor signals to convolve conditionally variables from each other [7], [8], [9]. As for probability variables, the resulted joint distributions can serve then to construct body-centered reference frames (e.g., eye-, head-, torso-, or hand-centered) to ease the locating and the controlling of the body limbs toward targets [10], [11].

There is some mathematical advantages to manipulate product of variables for learning transformations through multiplicative networks. For instance, it has been emphasized that they can discriminate better than deep networks affine transformations [12]; see Fig. 1 a). In computer vision, this technique has been applied extensively by Memisevic to the learning of optical flow, of rotational shifts as well as of spectral filters and spatio-temporal patterns for action recognition [12], [13], [14].

Nonetheless, in these researches, the transformation from unseen images is assumed to be hidden to the experimenter and is estimated afterwards [15]. However, in the case of robotics, we know exactly *which* actions have been performed and *which* effects have been caused on sensors. In comparison, knowing the extra information of having motors can serve multiplicative networks to estimate better which transformation has been performed between co-variating sensorimotor signals (inverse model) and to predict better how the sensory signals will evolve based on the learned transformations and current motor activity (forward model); resp. the blue and red networks in Figs. 1 b-c) and Fig. 2. On more advantage is that these learned transformations can be applied to totally new data and contexts, making them similar to generative networks.

We propose to exploit these properties in robotics especially in form of multiplicative auto-encorders, in order to learn inverse and forward models, see Fig. 2. On the one hand, the encoder part of the network will learn the different visuomotor transformations by estimating which motor primitive caused the visual displacement; e.g. sensor-to-motor mapping. On the other hand, the decoder part of the network will serve to reconstruct the visual signal by anticipating the visual displacement and location of the arm caused by the selected motor primitive; e.g. motor-to-sensor mapping, something similar with motor simulation, body image or action observation. Combining the encoder and decoder networks will permit to control the arm using its generated body image, in a similar fashion with inverse-forward networks in [16], [17], [18].

In this line, [19], [20] proposed to employ multi-layered perceptrons as auto-encoders for the learning of body image and motor simulation whereas [21], [22], [23], [24], [25], [26] used other techniques focusing ranging on Bayesian networks and gaussian mapping with the detection of images difference or of the extra tactile information for learning the effects of actions on the body or on objects. Considering the robot learning of body image and motor simulation, [21] and [23] proposed that affordances as learning the effects of actions on objects; for instance, by using a "speed of movement" feature. To our knowledge, only few teams proposed the use of gated networks in robotics [17]; recently Sigaud and colleagues for categorization and retrieving of motor sequences with the ICub robot [27], [28] and Pitti and colleagues for audio-visual and visuomotor integration [29], [30] and adaptation as during tool-use [31]. A precedent work has been done in [32] applied on a 3-link robot arm simulation, showing the validy of our method for reaching tasks. The current research is extending it to a real robot.

In this paper, we propose to exploit these characteristics

[1] University Paris-Seine, ETIS lab, UMR 8051 / ENSEA, University of Cergy-Pontoise, CNRS, F-95000, Cergy-Pontoise, France surname.name@u-cergy.fr

of multiplicative networks in an auto-encoder architecture for visuomotor control and body image; see Fig. 2. We will present three robotic experiments performed on the Kinova Jaco arm controlling the shoulder and elbow joints and a camera. The first experiment consists in learning the motor transformations from seen visual displacements of the robot arm; e.g. learning the inverse model (encoder). The second experiment consists in mapping back the learned motor transformations for estimating the spatial location of the robot; e.g. learning the forward model (decoder) for prediction of the body image. The third experiment consists in using the full auto-encoder for reaching nearby targets in the visual space by using the body image for motor control. Our results show that (1) the first network can effectively learn motor transformations from local displacements of the arm in the complete visual space with few reconstruction errors, (2) the second network can estimate the visual displacement of the body (body image) without any tags on it using the motor units to predict its location, and (3) the full auto-encoder can reach objects nearby by merging target position as body displacement (induced body image) selecting the most appropriate motor unit.

We will discuss then the relevance of our approach for reaching, body representation and action observation tasks as well as its current caveats and our future works.

II. METHODS AND EXPERIMENTAL SETUP

A. Methods

Gated networks are an instance of radial basis functions networks pre-defined parametrically or learned that produce a weighted sum of joint probability distributions as output [8].

The output terms Z_k are a sum of the product of the input variables X_i and Y_j weigthed by coefficients W^{in}_{ijk} whose cardinalities are respectively $z \in n_Z$, $i \in n_X$ and $j \in n_Y$. Since this matrix can be quite large, a way to reduce drastically the dimensionality of the gated networks is to factorize the inputs, for instance, by multiplying term by term each X_i and Y_j with $i = j$, if we assume that X and Y are of same dimension; see the blue network in Fig. 1 c) and Fig. 2. This reduces the dimension of coefficients W^{in}_{ik} for the encoding network, with $\{i, k\} \in \{n_X, n_Z\}$. This is equivalent in image processing to mapping the energy function between two related images for image transformation [14]. A second operation is to remove the mean field X in order to amplify the energy amplitude; this reduction might be not true for other types of inputs than images. The output Z_k becomes

$$\forall k \in n_Z, Z_k \quad = \quad \sum_{i}^{n_X} W^{in}_{ik}(X_i \times Y_i - X_i), \qquad (1)$$

The global error E is defined as the euclidean distance calculated between Z and Z^* for all the input examples. The optimization function used for learning the synaptic weights of the output layer Z is the classical stochastic descent gradient without a sparsity cost, which differs slightly from [13], because our motor variables are not latent and we know exactly which ones have caused the visual transformation from input images.

To reconstruct back the input variables X and Y, we can use a second network architecture as eq. 1 but in mirror as for an auto-encoder to estimate the input distribution $X \times Y - X$; see the red network in Fig. 1 c) and Fig. 2. The retrieved values $(X_k \times Y_k - X_k)'$ and Y' from this second network

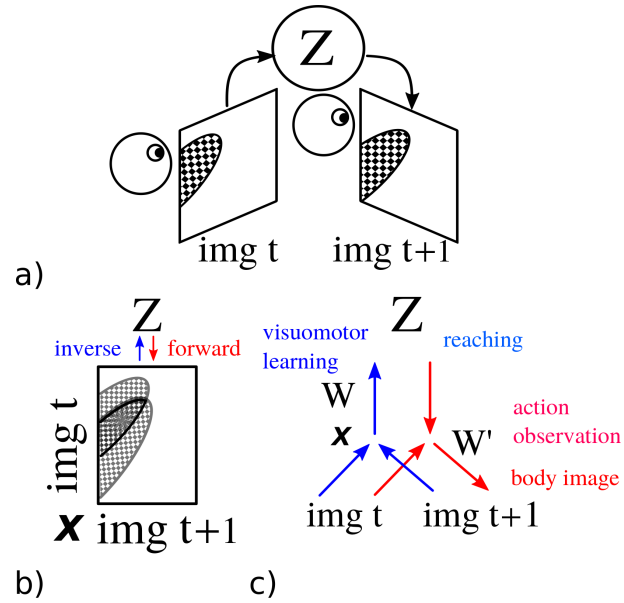

Fig. 1. Learning the visual transformation of two occuring images and its use for motor mapping, body image and visuomotor control. (a) The transformation **Z** between two related images at time t and $t+1$ can be learned and estimated by processing the pixel-wise multiplication between the two images, which is seen in (b). An auto-encoder can be used to learn the motor mapping of the visual displacement (the encoder part or inverse model) and reconstruct back a predicted image (the decoder part or forward model); resp., the blue and red lines in (c). In robotics, the estimated **Z** function corresponds to visual changes between two images at time t and $t+1$ induced by the applied motor control (for example hand motion). The encoder network discriminates the applied motor transformation **Z** whereas the decoder network estimates the visual displacement (aka body image). The full auto-encoder network can be used then to reach one desired target in the space nearby the robot by simulating a fictious visual displacement of the robot and by triggering the learned motor command in that direction.

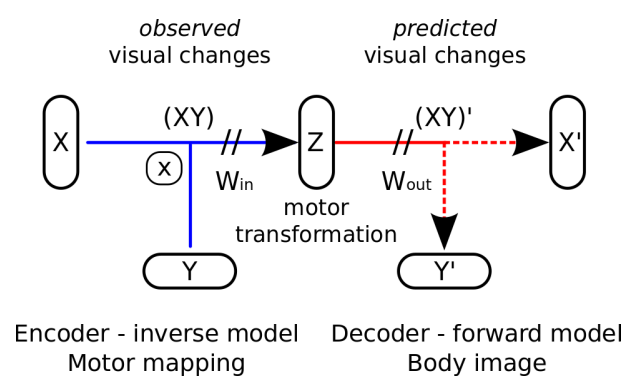

Fig. 2. Architecture of gated network applied to motor transformation and body image. The first part of the auto-encoder is a mapping of the observed visual changes induced by the applied motor commands, the co-occuring images are multiplied pixel-wise. The second network predicts the visual displacement of the arm that the motor commands will generate (aka body image).

Motor index	$\Delta\phi_S$	$\Delta\phi_E$
0	-0.1	-0.1
1	-0.1	0
2	-0.1	+0.1
3	0	-0.1
4	0	+0.1
5	+0.1	-0.1
6	+0.1	0
7	+0.1	+0.1

TABLE I

MOTOR UNIT INDICES AND CORRESPONDING MOTOR SYNERGIES.

are computed from the output values of Z calculated from the first network or those given by the experimenter Z^* and synaptic weights W^{out}:

$$\forall k, (X_k \times Y_k - X_k)' = \sum_i^{n_Z} W_{ik}^{out} Z_i, \quad (2)$$

$$Y'_k = ((\sum_i^{n_Z} W_{ik}^{out} Z_i) + X_k)/X_k \quad (3)$$

In this configuration, the two networks form a coupled system similar to an auto-encoder [16]. Each neuron Z in the intermediate layer represents a latent representation of the input variables $X \times Y - X$: the motor transformation responsible for the observed visual displacement.

B. Experimental Setup

The experimental setup is as follows. The input of the network is given by the multiplication of one image at instant t (image X) and one image at instant $t+1$ (image Y). The hidden layer of the gated network can be seen as a factored layer of visual transformations (e.g rotation, translation), which can be categorized in order to retrieve the motor command Z.

As described in Fig. 1 and Fig. 2, the first network receives the pixel-wise multiplication of two images X and Y, resp. at instant t and $t+1$ of dimension $[24 \times 32]$, which corresponds to $(X \times Y - X)$, and the hidden layer consists in eight motor units associated to the motor command Z defined by different angular displacements of the shoulder-elbow pair $\{\Delta\phi_S, \Delta\phi_E\}$ of the robot arm Jaco from Kinova. Each motor unit is defined by discrete values in the angular interval $\Delta\phi = \{-0.1, 0, +0.1\}$ corresponding to eight possible motor synergies, see Tab. I. Finally, the second network reconstructs back the $[24 \times 32]$ matrix with neurons approximating the input image $(X \times Y - X)$'. We make the note that these eight motor synergies are purposely exhaustive in order to test our algorithm. This is not a limitation of our network for which the hidden layer can be factorized and learned in a sparser code for a higher number of degrees-of-freedom robot.

The learning set consists on nine different postures $\{\phi_S, \phi_E\} \in \{[0,1],[0,1]\}$ taken in the robot arm space on which the eight motor synergies $\{\Delta\phi_S, \Delta\phi_E\}$ are applied. The trajectories describe a eight-branch star pattern around the nine locations of the end effector in order to ease the generalization from local transformations.

The dataset collected represents 7000 samples of $\{X, Y\}$ consecutive image pairs with their corresponding motor activity Z for the learning of the two networks in a supervised manner. We used for the learning of the weights the stochastic gradient method with sets of 30 samples each, during 900 epochs. The two networks are tested with different samples and configurations in Section. III.

III. EXPERIMENTS

A. Experiment 1 – Visuo-Motor Control for Reaching

This experiment consists on learning the visuomotor transformations for a reaching task. For this, we provide to the first gated network the input images $(X \times Y - X)$ for various visual displacements of the robot arm in the eight motor primitives given in Table. I and in different locations (see experimental setup for details). After categorization of the visual transformations into motor commands, the first gated network is then capable to be used for reaching tasks.

In order to describe the behaviour of the first gated network after the learning stage, we display in Fig. 3 a) and b) respectively the prediction of the motor units Z for visual displacements at different locations in the visual space for the same motor unit $\#4$ (global generalization) and for visual displacements around the same location in the visual space for different motor units (local discrimination). The left chart corresponds to the input received by the gated network (i.e $X \times Y - X$) as indicated in eq. 1; in order to better visualize the information, here the input is displayed with a $[240 \times 320]$ image but the network uses a scale down version of this image $[24 \times 32]$, as explained in the experimental setup. The upper right chart shows the motor neurons' response associated to the eight motor transformations and the lower right chart indicates the pairs value $\{\Delta\phi_S, \Delta\phi_E\}$ associated to the current motor transformation as described in table. I.

As shown in the figures, the visual displacement between two related images is exhausted by the multiplication localised around the visual displacement of the arm and for which the corresponding motor units are the most sensitive with. The motor unit $\#4$ in Fig. 3 a) is associated mostly to visual displacements in the upper right direction, for most arm postures. It is noteworthy that the relationship learned by the first network between visual displacement and motor command is nonlinear: depending on the spatial location where the robot arm is, the visual displacements will be different for the same commands performed in the shoulder and the elbow. Conversely, when the robot arm is located at the same posture and different visual displacements are seen as in Fig. 3 b), the first network is capable to predict the motor unit associated with this local visual displacement. The generalization done by the motor units supports the construction of a repertoire of general movements. In the figure, the temporal delay used to extract images X and Y is of 17 ms but we used normally the network with higher sampling rates, which permits to get more robust results.

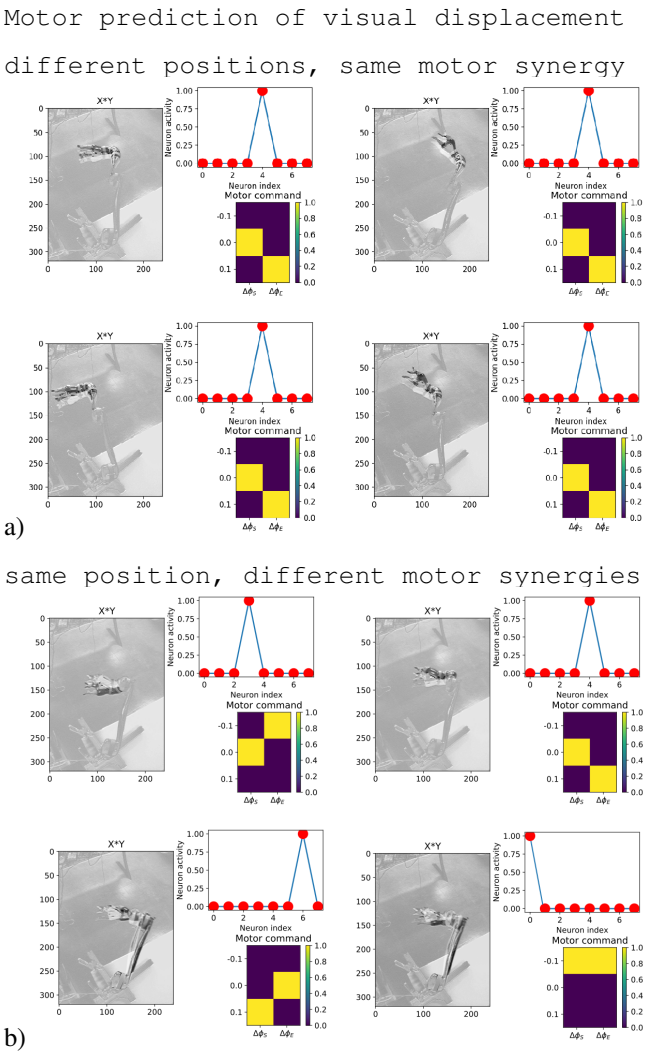

Fig. 3. Categorisation of visual displacements by motor units. The first gated network discriminates the visual motion between two related images X and Y (left chart). However, the correspondence between the type of visual displacements and the motor units are nonlinear (see text). In a) for different postures of the robot arm but for same motor command, the motor unit #4 has been correctly predicted (top chart); ie global transformation. In b) for same initial posture of the robot arm but for different motor command, various motor units have been predicted; ie local transformation. The red dots corresponds to real motor unit and the blue line corresponds to predicted motor unit.

We display in Fig. 4 the error matrix for the eight motor synergies and for 300 samples of $\{X, Y\}$ pairs taken from various location and visual displacements. They correpond to the configuration of the robotic arm in nine locations on which we apply the eight motor synergies (e.g., eight visual directions). We can see that most of the motor categories are well correlated with some slight errors for two of them, motor units #0 and #2 –, resp. synergies $[-0.1, +0.1]$ and $[+0.1, +0.1]$,– which can be considered as a linear combination of other motor units. For example, motor unit #7 is the sum of motor units #4 and #6, which is in terms of synergy equal to $[0, +0.1] + [+0.1, 0]$. Errors on this map are due to the use of the max strategy for motor category selection, however despite theses errors, the max strategy still

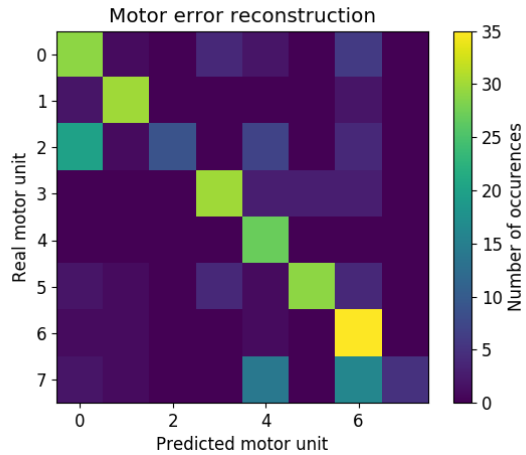

Fig. 4. Error matrix on motor prediction for 300 samples. High activity on the diagonal corresponds to correct motor prediction, values outside the diagonal correspond to motor errors.

permits a good mapping for other categories with low error.

We plot in Fig. 5 the spatial density distribution of the prediction error estimated by the motor units. As expressed above, they correspond to the displacement of the robotic arm in eight-branch star for the eight motor synergies (e.g., eight visual directions) and in nine locations. Each black dot corresponds to the end-effector location in cartesian space and the intersection of the lines corresponds to one of the nine locations picked up randomly from the dataset. The color scale indicates the euclidean distance error made by the motor units to estimate the visual displacement between $\{X, Y\}$ images. All data are projected in the cartesian 2D plan.

In this map, we can see that we have mainly three singularities (red bumps) due to the visual overlapping between several postures for which the motor selection cannot be interpolated easily. We can also notice that we have an area in the lower right of the working space not learned because we did not provide any samples due to avoid collision between the robot with its support. Nevertheless, the density map shows the robustness of the network for categorizing the visual transformation with its associated motor unit.

This shows also how each motor unit discriminates visual transformation and how the first gated network may serve then to combine linearly these transformations into a forward model; e.g., motor simulation, see Sec. III-B.

B. Experiment 2 – Body Spatial Representation

Once the learning of the visuomotor transformations in the first gated network has been done in section III-A, a second gated network can be used then as a forward model for action-based prediction of visual information; see the red network in Fig. 1 c) and Fig. 2. We train directly this second network using eqs. 3 from the motor neurons Z in order to predict the images product Y', the visual displacement. After the learning stage, the second network can estimate the density probability of the visual displacement Y' from

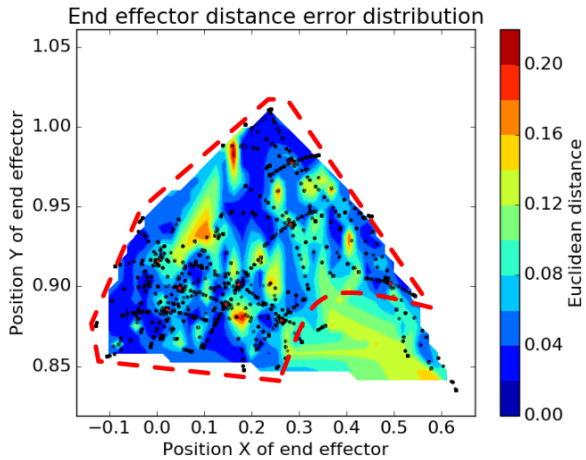

Fig. 5. Spatial density distribution of reach distance error for nine locations in space (red dots) and eight motor directions. The color intensity indicates the Euclidean distance error between the location of the wrist and of the desired target. Local errors in the workspace (red dashed lines) are due to subsampling of the dataset on these locations. High level errors outside the workspace are due to unmapping.

the motor neurons Z and image X of the current location of the robot arm, see Fig. 6. In this figure, we super-imposed to the image X the filtered neurons output Y' in four arm postures. The neurons' activity corresponds directly to the location where the body is estimated to be. Depending on where the arm is located, different density maps of Y' are generated. These density maps represent therefore the current body image of the robot arm in the visual space. Each motor neuron is not necessary relevant for the image reconstruction as they did not correspond to a correct motor primitive, however with respect to the ones found, a good estimation of the spatial location of where the robot is can be done without visual tags.

C. Experiment 3 – Reaching in Body-Centered Reference Frames

The previous experiments in section III-A and III-B showed that an encoder network can categorize motor units Z from $\{X, Y\}$ and that a decoder network can estimate back images Y' from motor units Z. In this section, an integrated inverse and forward model combining both networks is designed for reaching targets using the body image.

In order to make it works, we need to construct the peripersonal space, which is the space within reach, so that when we will have a target close enough to the robot arm, the network will assimilate it as a visual displacement of the arm, which can serve in return to estimate the most suited motor units for control.

In the previous experiment in section III-B, the space surrounding the robot arm was strongly filtered out from the neural activity so that only the arm location was observed. Here, we lowered the filtering threshold so that the second network can estimate the arm location as well as its surrounding space. By doing so, the two networks can estimate

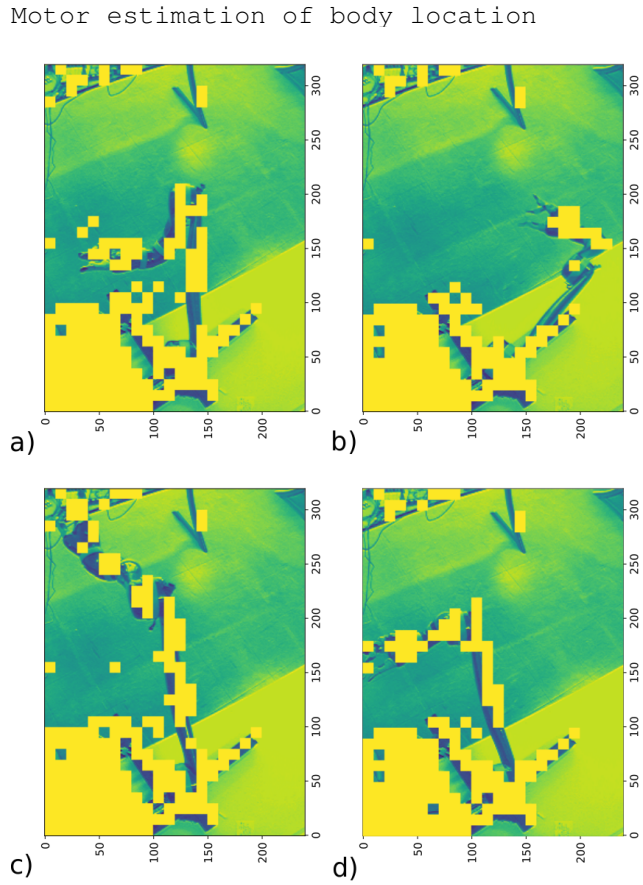

Fig. 6. Motor-based prediction of body location. Output neurons of the second gated network estimate the visual displacements and locations Y' in four different locations. A threshold has been applied to the output units. The estimation is relatively precise on the location of the robot arm.

dynamically the most suited motor commands that would fit in return the desired visual displacement to the target.

In this section, we use a different experimental setup with 27 motor neurons to control 3 d-o-f of the robot arm, which can estimate better this peri-personal space and control better the arm movement, see Fig. 7. The 27 motor synergies correspond to the exhaustive combination of the 3 d-o-f of the robot arm, including the wrist. In this figure, we superimposed the activity level of the output network Y' on the image X. As it can be seen, the visual prediction of the peri-personal is fuzzier than the one in the previous section due to use of more motor units to weight the integration.

We display also in Fig. 7 the trajectory of one target (red square) moving from the bottom of the image to the center with the robot arm tracking it for ten iterations. The activity of the motor neurons Z is presented in the bottom chart for the 27 units. Using the two networks, the robot arm use its reconstructed body image to follow the target. During the motion tracking, we can observe that the motor units are changing dynamically as so for the estimation of the body image. The precision of the body representation is dependent on the coherence of the learned motor space, which is represented sparsely.

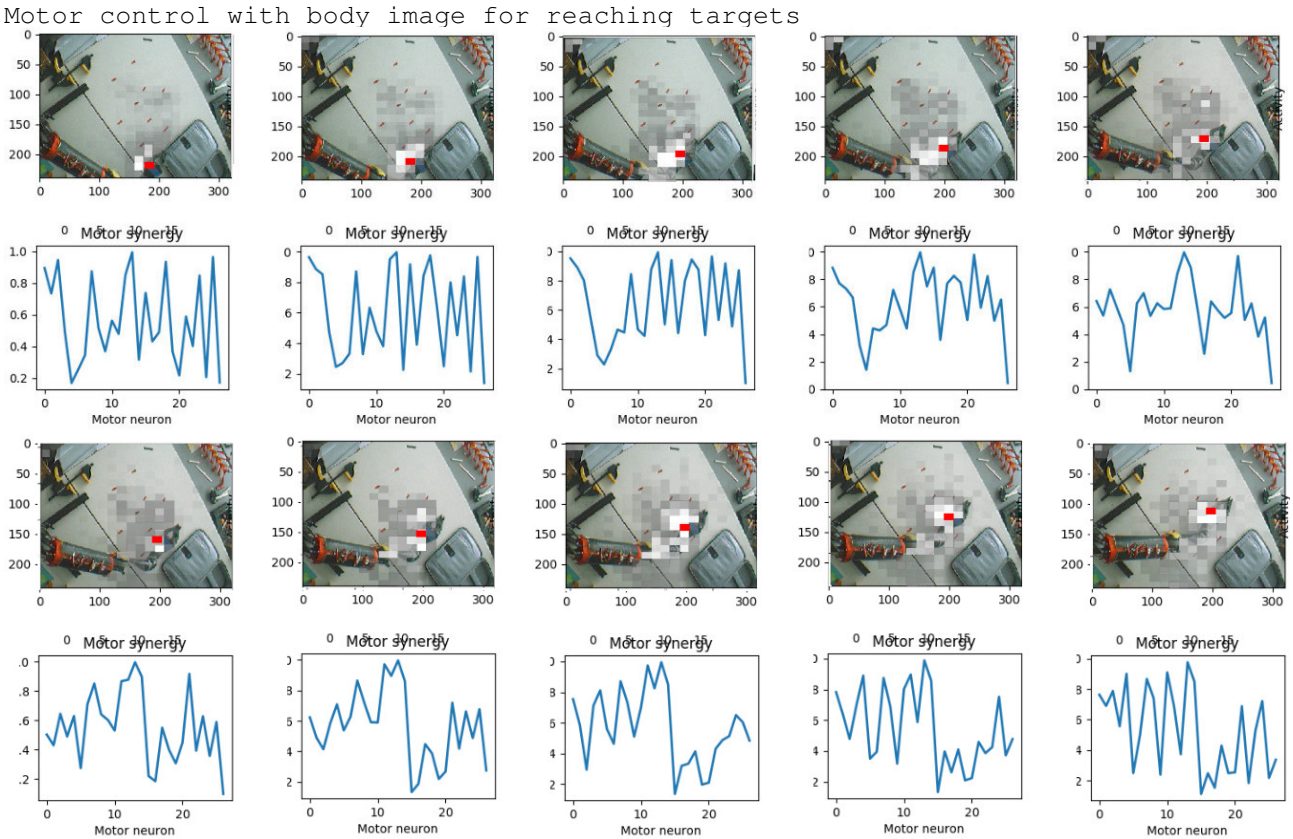

Fig. 7. Reaching targets with body image. Sequence of images with a moving target (red dot) from the bottom to the center of the image. The target is given as input Y to the first network and estimation by the second network of the body image (super-imposed pixels). Experiment done on a 3d-o-f robot with 27 motor units. Targets near the arm represent ambiguous information of visual displacement of the robot arm between $\{X, Y\}$ image pairs (top chart). The inferred motor categories (bottom chart) are generating a dynamic body image to follow the visual displacement (top chart).

IV. CONCLUSION

We have presented a gated auto-encoder network for learning visuo-motor transformations for reaching targets with the construction of a body image. Gated networks are based on multiplicative interaction between related images for infering the corresponding transformation. In robotics, these transformations can be discriminated easily as they correspond to the robot motor activity, which can be learned through supervised learning. Using an inverse model, one network predicts the most probable motor command within a repertoire of eight motor primitives from its visual motion (27 units in section III-C). This network performs few motor errors and can be employed for a predictor for categorizing visual displacements. Using a forward model, another network uses its motor repertoire as input in order to estimate the spatial location and displacements of the robot arm, which corresponds to its peripersonal space. The combination of the first and second networks permits to track objects within reach using a dynamic body image, without having any visual tags on the robot arm.

Interestingly, the cortical parieto-motor neurons are involved in visuomotor transformations and the way they are coding motion is by binding the multimodal signals with multiplication [1], [2]. These units, known as gain-modulated neurons, serve to translate sensory signals from one reference frame (e.g. retina) into another one (hand-centered) [8]. For instance, Georgopoulos found that the motor neurons were aligned to the visual orientation of the arm direction [3]. Graziano studied body-centered receptive fields in retina coordinates sensitive to the distance of objects nearby and active even in the dark [33], [34]. Sakata found motor neurons sensitive to visual motion depth and rotation [35]. These neurons are therefore particularly important for 3D perception and hand manipulation tasks.

In future works, we will extend our current research to more complex tasks such as infering 3D directions, depth and 3D rotations for hand manipulation. We believe that such network is effective enough for permitting reaching and grasping tasks with a higher degrees of freedom robotic system. As explained earlier, we purposely used an exhaustive number of motor units for our task but a factorized version of it can be designed for representating with a sparse code motor synergies of a higher degrees of freedom robot.

ACKNOWLEDGMENT

We would like to acknowledge the ROBOTEX EQUIPEX Ile-de-France grant.

REFERENCES

[1] A. Georgopoulos, J. Kalaska, R. Caminiti, and J. Massey, "On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex," *J. Neurosci.*, vol. 2, p. 15271537, 1982.

[2] D. Zipser and R. Andersen, "A back propagation programmed network that simulates response properties of a subset of posterior parietal neurons," *Nature*, vol. 331, pp. 679–684, 1988.

[3] A. Georgopoulos, "Current issues in directional motor control," *Trends in Neurosciences*, vol. 18, no. 11, p. 506510, 1995.

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

[5] S. Kakei, D. Hoffman, and P. Strick, "Sensorimotor transformations in cortical motor areas," *Neuroscience Research*, vol. 46, p. 110, 2003.

[6] C. stn, "A sensorimotor model for computing intended reach trajectories," *PLoS Comput Biol*, vol. 12, no. 3, p. e1004734, 2016.

[7] E. Salinas and P. Thier, "Gain modulation a major computational principle of the central nervous system," *Neuron*, vol. 27, pp. 15–21, 2000.

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

[9] S. Deneve and A. Pouget, "Bayesian multisensory integration and cross-modal spatial links," *J. of Phys.–Paris*, vol. 98, pp. 249–258, 2004.

[10] P. Baraduc, E. Guigon, and Y. Burnod, "Recording arm position to learn visuomotor transformations," *Cerebral Cortex*, vol. 11, p. 906917, 2001.

[11] S. Deneve, A. Pouget, and J. Duhamel, "A computational perspective on the neural basis of multisensory spatial representations," *Nature Rev. Neurosciences*, vol. 98, pp. 741–747, 2002.

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

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

[14] ——, "Learning to relate images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1829–1846, 2013.

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

[16] M. Jordan and D. Rumelhart, "Forward models: Supervised learning with a distal teacher," *Cognitive Science*, no. 16, pp. 307–354, 1992.

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

[18] 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.

[19] G. Schillaci, B. Lara, and V. Hafner, "Internal simulations for behaviour selection and recognition," *Lecture Notes in Computer Science, , eds A. Salah, J. Ruiz-del Solar, . Merili and P. Y. Oudeyer (Berlin, Heidelberg: Springer) Proceedings of the Third international conference on Human Behavior Understanding*, vol. 7559, p. 148160, 2012.

[20] G. Schillaci, V. Hafner, and B. Lara, "Online learning of visuo-motor coordination in a humanoid robot. a biologically inspired model," *Joint IEEE International Conferences on Development and Learning and Epigenetic Robotics (ICDL-Epirob)*, pp. 130–136, 2014.

[21] B. Moldovan, P. Moreno, M. van Otterlo, J. Santos-Victor, and L. De Raedt, "Learning relational affordance models for robots in multi-object manipulation tasks," *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4373–4378, 2012.

[22] J. Galeazzi, B. M. W. Mender, M. Paredes, J. Tromans, B. Evans, L. Minini, and S. Stringer, "A self-organizing model of the visual development of hand-centred representations," *PLoS ONE*, vol. 8, no. 6, p. e66272, 2013.

[23] E. Ugur and J. Piater, "Emergent structuring of interdependent affordance learning tasks," *International Conference on Development and Learning and on Epigenetic Robotics*, pp. 1–6, 2014.

[24] A. Roncone, M. Hoffmann, U. Pattacini, and G. Metta, "Automatic kinematic chain calibration using artificial skin self-touch in the icub humanoid robot," *IEEE International Conference on Robotics and Automation, ICRA*, pp. 2305–2312, 2014.

[25] J. Born, J. M. Galeazzi, and S. M. Stringer, "Hebbian learning of hand-centred representations in a hierarchical neural network model of the primate visual system," *PLOS ONE*, vol. 12, no. 5, pp. 1–35, 05 2017. [Online]. Available: https://doi.org/10.1371/journal.pone.0178304

[26] P. Lanillos, E. Dean-Leon, and G. Cheng, "Yielding self-perception in robots through sensorimotor contingencies," *IEEE TCDS*, vol. 9, no. 2, pp. 100–112, 2017.

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

[28] O. Sigaud and A. Droniou, "Towards Deep Developmental Learning," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 8, no. 2, pp. 99–114, 2016.

[29] 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.

[30] 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.

[31] R. Braud, A. Pitti, and P. Gaussier, "A modular dynamic sensorimotor model for affordances learning, sequences planning and tool-use," *IEEE TCDS*, p. to appear, 2017.

[32] J. Abrossimoff, A. Pitti, and P. Gaussier, "Visual Learning for Reaching and Body-Schema with Gain-Field Networks," *Joint IEEE International Conferences on Development and Learning and Epigenetic Robotics (ICDL-Epirob)*, pp. 1–7, 2018.

[33] M. Graziano, X. Hu, and D. Cooke, "Visuospatial properties of ventral premotor cortex," *Journal of Neurophysiology*, vol. 77, pp. 2268–2292, 1997.

[34] ——, "Coding the locations of objects in the dark," *Science*, vol. 277, pp. 239–241, 1997.

[35] H. Sakata, M. Taira, M. Kusunoki, and A. Murata, "The parietal association cortex in depth perception and visual control on hand action," *Trends in Neurosciences*, vol. 20, pp. 350–357, 1997.