Sensori-motor Networks vs Neural Networks for Visual Stimulus Prediction

Ricardo Santos, Ricardo Ferreira, Ângelo Cardoso, Alexandre Bernardino
Institute for Systems and Robotics
Instituto Superior Técnico,
Lisbon, Portugal
Email: {rsantos,ricardo,acardoso,alex}@isr.ist.utl.pt

***Abstract*—This paper focuses on a recently developed special type of biologically inspired architecture, which we denote as a sensori-motor network, able to co-develop sensori-motor structures directly from the data acquired by a robot interacting with its environment. Such networks learn efficient internal models of the sensori-motor system, developing simultaneously sensor and motor representations (receptive fields) adapted to the robot and surrounding environment. In this paper we compare this sensori-motor network with a conventional neural network in the ability to create efficient predictors of visuo-motor relationships. We confirm that the sensori-motor network is significantly more efficient in terms of required computations and is more precise (less prediction error) than the linear neural network in predicting self induced visual stimuli.**

***Index Terms*—Stimulus prediction, visual and motor receptive fields, neural networks, sensori-motor maps.**

I. Introduction

Nature shows that evolution tends to improve the efficiency of organisms. Solutions found in nature are an important source of inspiration for the design of autonomous systems and bio-mimetic solutions are gaining increasing interest in the development of embedded applications where resource constraints and computational bottlenecks are the rule rather than the exception.

In terms of visual capabilities, that require a significant amount of computation, it is important to understand both the role motor actions have in visual perception and visual stimulus prediction, and its relationship with the neural circuits organization. Living organisms' visual systems are continuously trained and improved while relationships between motor actions and sensory feedback are learned by the agent during the interaction with its habitat or environment.

Without perception one is left with little criteria to decide which actions to take, while at the same time there is no purpose in having perception if you cannot act on the world. An ideal rational agent [1] always takes the actions which maximizes its performance measure based on its percepts and built-in knowledge. This definition frames perception as a component used to choose the right action, and not as a goal by itself. Under this light a broad goal is to develop sensori-motor structures which support choosing the right action. To be able to do so one crucial ability that organisms developed is the ability to discern the origin of sensory input between changes in the environment (exafference) and the result of the animal's own movements (reafference) [2]. The ability to discern between these two origins of sensory input requires a forward model [3] to predict the effect a given movement (action) has on its sensory input.

The presented adaptive model [4] learns to predict visual stimulus based on motor information resulting from self-inducing actions. This model maps motor input in a structure also processing visual stimulus, creating direct relationship between the robot's actions and its perceived visual stimuli. Following a specific learning process it was possible to minimize the prediction error evaluated by the mean square error between the predicted image and the expected image after a specific motor action. In spite of starting from an unknown topology, the proposed structure developed a topology covering the recording visual sensor and organized itself leading to a less costly prediction model.

In the sensory layer through the same developmental process an organization also emerges. Each sensory neuron's receptive field, RF, is composed of a set of retina cells which cover nearby parts of the visual field and together represent a continuous portion of it. The motor layer in the same developmental process also organizes itself. In this layer, each neuron's RF reacts to actions which produce similar results. This simultaneous development promotes a coherent representation for similar stimuli (sensory) and actions (motor), which greatly improves the effectiveness of structure by taking advantage of these organizations.

In this paper, we compare the model proposed in [4] with an artificial neural network with a standard architecture (fully-connected). For sensori-motor prediction, we show that a network with a specific structure can attain significant advantages over fully connected networks. We claim that co-developed structures yield better sensory predictions for the effects of actions, relatively to a more naive and straightforward approach which lacks a sensori-motor structure and development supporting the importance of coupling sensor and motor information.

II. Related Work

Considering a limited amount of resources an organism needs to choose which actions to represent in its motor system. A criteria which fits well with the stimulus prediction rationale is to represent actions which have predictable

effects [5]. Assuming a particular sensory structure for the simultaneous development of a motor system and a forward model (which predicts the sensory input for a given action) a topology emerges in motor system to support the predictability of the actions [6].

It has been shown that, while maximizing the sensor's self-similarity under a given set of transformations, highly regular structures emerge which resemble some biological visual systems [7]. Still, for these structures to emerge, we require apriori knowledge about the sensor spatial layout. The retinotopic structure of an unknown visual sensor has been reconstructed using an information measure, as well as the optical flow induced by motor actions [8]. A robot with the goal of estimating the distance to objects using motion parallax developed a morphology for the position of movable light sensors which was fit for the task [9].

Guiding the development of a sensori-motor system to maximize the ability of predicting the effect an action has on its sensory input (see Methods), allows for the emergence of highly regular sensory structures without any prior knowledge. To develop such ability we follow two main principles: the sensory system should capture stimuli which are relevant to motor capabilities and the actions of the motor system should have predictable effects on the sensory system [4].

These principles are related to idea of "morphological computation" in robotics and artificial intelligence, which aims at reducing the computational complexity of a problem by using a specifically designed body to solve it (e.g. [10]). The human visual system representation of the visual world is progressively differentiated from what is captured through the retina to support complex tasks, e.g. cells which are selective to objects. Also, in machine learning it is known that for recognition tasks there are huge advantages in using specific architectures [11] (e.g. convolutional) relatively to a fully-connected network.

III. Methods

We consider an agent capable of observing its environment by sensing a light field i which falls on a two dimensional sensory surface. Similarly this agent is able to interact with its environment by activating a particular motor primitive $\mathbf{q}$ on its action space. For implementation purposes, in this paper we represent the light field as a vector $\mathbf{i}$ of N_s pixels, and the action space is represented as a vector $\mathbf{q}$ with N_m elements, where a single non-zero entry represents the activated motor primitive. If the n^{th} index of $\mathbf{q}$ is 1, then the n^{th} physical action is performed (e.g. shift left by a certain amount). Note that no topological assumptions exist on the spatial locations of either the incident light field or the motor primitives.

During the learning phase, the agent interacts with the environment by randomly choosing a motor primitive $\mathbf{q}$ while collecting before and after sensor stimuli ($\mathbf{i}_0$ and $\mathbf{i}_1$). These triplets are collected for several iterations and the full batch is used as training data.

Here we consider two possible architectures for an agent, capable of predicting its interaction with the environment, and compare 1) its predicting capabilities, i.e. how well it can predict $\mathbf{i}_1$ given $\mathbf{i}_0$ and $\mathbf{q}$; 2) its simplicity, i.e. the number of parameters learned which contribute to prediction.

A. Neural Network Architecture

In this case we consider a feed-forward linear network with n_s elements in a hidden layer emulating receptive fields. The sensor input $\mathbf{i}_0$ is concatenated with the activated action $\mathbf{q}$ (working as an action identifier) and used as input to the network predicting $\mathbf{i}_1$. The optimization problem solved is thus

$$\begin{aligned}(\mathbf{W}_1^*, \mathbf{W}_2^*) = \quad & \underset{k}{\text{argmin}} \quad \sum_k \left\| \mathbf{i'}_1^k - \mathbf{i}_1^k \right\|^2, \\ & \text{s.t.} \quad \mathbf{o}^k = \mathbf{W_1} \begin{bmatrix} \mathbf{i}_0^k \\ \mathbf{q}^k \\ 1 \end{bmatrix} \\ & \qquad \mathbf{i'}_1^k = \mathbf{W_2} \begin{bmatrix} \mathbf{o}^k \\ 1 \end{bmatrix}\end{aligned} \tag{1}$$

which is represented in Fig. 1b. Here, $\mathbf{W_1}$ is an $(N_m + N_s + 1) \times n_s$ matrix, and $\mathbf{W_2}$ is $(n_s + 1) \times N_s$, where each matrix includes a constant bias term. Although the network is linear, the two matrices are required to force the dimensionality reduction emulating the existence of receptive fields.

B. Sensori-motor Network Architecture

We consider a second prediction network as described in [4], explicitly modeling the existence of light sensitive receptors represented as an $N_s \times n_s$ matrix $\mathbf{S}$ which integrate the light field $\mathbf{i}$ falling on the two dimensional sensory surface. The sensor observation is then a vector $\mathbf{o} = \mathbf{Si}$. On the motor side a dual structure exists, where a set of discrete motor movement fields modeled as a $N_m \times n_m$ matrix $\mathbf{M}$ cover the available motor primitive space $\mathbf{q}$, providing a n_m dimensional motor action representation space $\mathbf{a} = \mathbf{M}^T\mathbf{q}$. These are then fed to a predictive layer, where a predictor $\mathbf{P}^k$, for each action, is composed as a linear combination of n_m basis predictors $\mathbf{P}_j$ with linear weights given by the motor movement field activations,

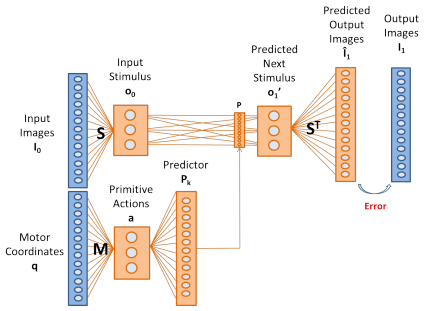
$$\mathbf{P}^k = \sum_j^{n_m} \left(\mathbf{m}_j^T \mathbf{q}^k\right) \mathbf{P}_j \tag{2}$$

where $\mathbf{m}_j^T$ represents transposed of the j^{th} column of $\mathbf{M}$ and the corresponding motor receptive field.

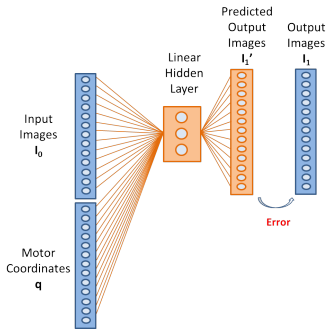
The full model description is provided in [4], resulting in the optimization problem

$$\begin{aligned}(\mathbf{S}^*, \mathbf{M}^*, \mathbf{P}^*) = \quad & \underset{k}{\text{argmin}} \quad \sum_k \left\| \mathbf{i'}_1^k - \mathbf{i}_1^k \right\|^2 \\ & \text{s.t.} \quad \mathbf{i'}_1^k = \mathbf{S}^T \left(\sum_j^{n_m} \left(\mathbf{m}_j^T \mathbf{q}^k\right) \mathbf{P}_j \right) \mathbf{Si}_0^k \\ & \qquad \mathbf{S} \geq \mathbf{0},\, \mathbf{M} \geq \mathbf{0},\, \mathbf{P_j} \geq \mathbf{0}\end{aligned} \tag{3}$$

which is represented in Fig. 1a. Unlike in the neural network architecture, you'll notice that the sensor reconstruction model is simplified to be $\mathbf{S}^T$ (instead of independent projection and reconstruction matrices). In [4] the authors argue that this simplification is justified by the particular solutions obtained from the model, particularly the fact that the matrix $\mathbf{S}$ will be nearly orthogonal.



(a) Scheme of used Sensori-motor Network.



(b) Scheme of used Neural Network.

Fig. 1: Diagram representation of (a) the sensori-motor organization method and (b) a simple neural network with a linear hidden layer. For both methods blue is used to represent the input data (set of images and action triplets) while orange represents the parameters to be trained together with the predictions (Best seen in color).

IV. Experimental Setup

To compare the proposed biologically inspired sensorimotor network architecture, here on called SNet, with the simpler linear layer neural network, NNet, we design two experiments. In the first experiment the motor space spans actions leading to translations in the image plane, whereas in the second experiment we use actions leading to centered rotations and zooms in the image. The first set of movements either mimics an agent that moves its sensor parallel to the environment surface or an agent that performs small pan-tilt rotations of the sensor when observing far objects. The second set of movements can either be seen to approximate the observations of an agent moving in a tubular structure translating and rotating along its optical axis, or the observations of an agent while actively tracking an object that rotates and changes its distance from the observer. For each case, we perform 10 runs of the training algorithms. Each run is composed of a batch of 8100 triplets, uniformly sampled from a discrete set of $N_m = 81$ canonical actions (100 triplets of each canonical action). For the first experiment the set of actions is composed of pixel translations $\mathbf{u} = \{-4:1:4\} \times \{-4:1:4\}$ and for the second experiment the set of actions combine rotations and zoom scale factors transformations $\mathbf{u} = \{-100°:25°:100°\} \times \{0.80:0.05:1.20\}$. These experiments will be referred from here on as Experiment XY and Experiment RZ, respectively.

The agent is equipped with a square retina of 15 by 15 pixels ($N_s = 225$) which is used to acquire the images. Triplets ($\mathbf{i}_0$,$\mathbf{i}_1$,$\mathbf{q}$) are obtained using a 2448 by 2448 pixels image as environment for the agent. First the agent is positioned in a random place in the environment and image i_0 is sampled. Then action u is performed and the new image i_1 is sampled. This process is illustrated in Fig.2.

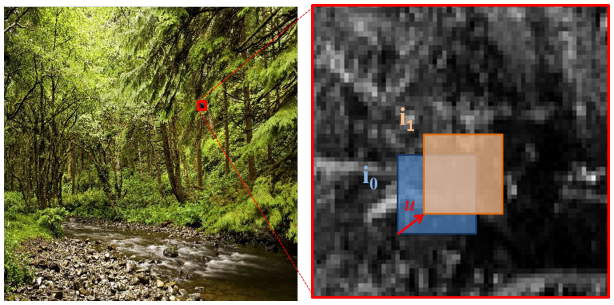


Fig. 2: Triplet acquisition process. In the left we show the full environment image. In the right we show a portion of the environment where the agent is placed to acquire the pre-action 15×15 pixel image, i_0, then transformed by action u, and acquire the post-action image, i_1 (Best seen in color).

After acquiring its exploration data in the given environment, the agent processes the data in order to obtain the network parameters for the SNet $(\mathbf{S}, \mathbf{M}, \mathbf{P})$ and for the NNet $(\mathbf{W}_1, \mathbf{W}_2)$. The optimization criteria is the average squared error in image prediction given an action as in equations (1) and (3). In both experiments, the SNet model is formed by a motor structure composed by 9 movement fields (n_m = 9) and a sensor structure composed by 9 receptive fields ($n_s = 9$), which is compared with a linear hidden layer of 9 neurons for the NNet model. The number of RFs (hidden units) can be chosen taking into account the resources available in the particular hardware used to deploy the system. In these experiments we use an identical number of sensor and motor RFs but these numbers may be different.

The optimization problem for the SNet showed in Eq. (3) is iteratively improved using a projected gradient descent method [12] within the sequential optimization of **P**, **M**, **S**, and the input triplets are considered in batches as in [4].

For SNet and NNet, while performing the optimization, the RMSE between the predicted and the expected images is computed using the training set and a validation set with half the samples,

$$\text{RMSE} = \sqrt{\frac{1}{N_s \times N_m \times l} \sum_{1}^{N_s} \sum_{1}^{N_m} \|\hat{\mathbf{i}}_1 - \mathbf{i}_1\|^2} \quad (4)$$

where l stands for the number of samples per action.

The RMSE on the validation set is used as a stopping criterion: the optimization stops when the training error becomes almost constant and the validation error starts to grow.

After training both networks, they are compared in terms of efficiency (number of parameters used) and precision (RMSE). A relative comparison regarding loss of information (information criteria) is also computed using Akaike information criterion (AIC) and Bayesian information criterion (BIC) with,

$$AIC = 2k - 2\log(L) \quad (5)$$

$$BIC = k\log(n) - 2\log(L) \quad (6)$$

where log is the natural logarithm and L is the considered likelihood function:

$$L = \exp^{-\lambda \text{RMSE}^2} \quad (7)$$

with $\lambda = 0.9$, k the number of parameters to be estimated and n the number of data samples (triplets) used for training.

V. Results

In this section we show the results obtained from the optimization of the two models under comparison (sensori-motor network *vs* neural network), using the methods and experimental setup described in the previous sections.

A. Sensori-motor Topology

Initially we revisit the emergent properties [4] with respect to SNet organization (optimization problem in Eq. (3)). These results illustrate some interesting outcomes of the optimization process in terms of the shape and distribution of the sensor and motor receptive fields. The sensor RFs (rows of **S**) organize into a regular structure (after 300 iterations) starting from a random initialization (see Fig. 3). Notice that these organize more uniformly for translation actions then for rotations and zooms. With rotations and zooms the sensor fields tend to create a group of smaller receptors in the middle of the retina and bigger fields near the boundaries (a rotation produces higher movement far from its center).

In Fig. 4 we can observe the evolution of the motor fields (columns of **M**) for both experiments. Experiment XY has its action space uniformly sampled by pixels, producing a near uniform organization of the motor fields. The performed zooms in Experiment RZ had low impact on their images in comparison with the rotations, which caused the motor fields to organize in a way that each one represents an angular range. Exception for the middle ones where no rotation is performed and zooms have weight in motor RFs organization.

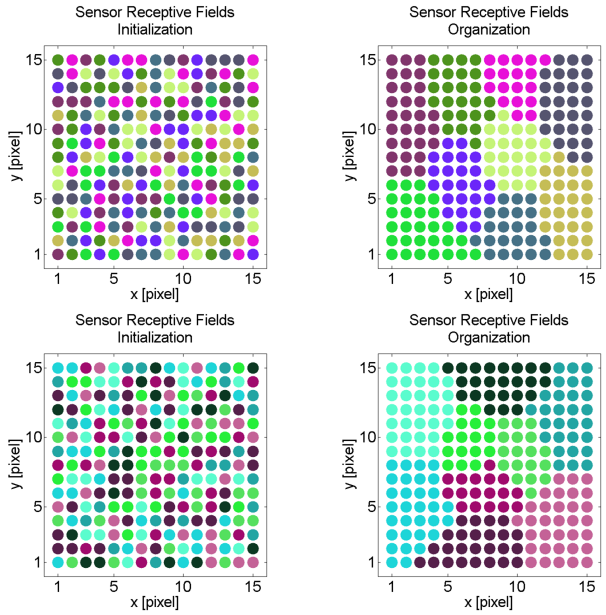


Fig. 3: Sensor RFs initialization and final organization after 300 iterations in one of the runs of Experiment XY (Top) and Experiment RZ (Bottom) (Best seen in color).

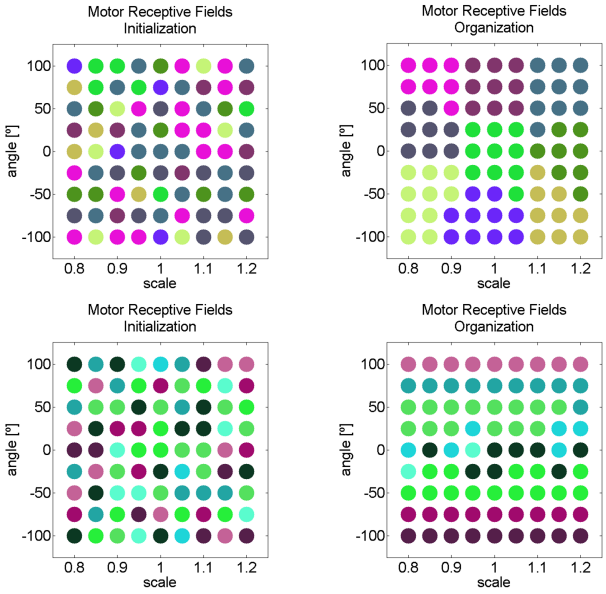


Fig. 4: Motor RFs initialization and final organization after 300 iterations in a run of Experiment XY (Top) and Experiment RZ (Bottom) (Best seen in color).

B. Quantitative Evaluation

After convergence of training on the 10 runs for each of both networks we computed several statistics in order to evaluate and compare their performance. We could observe that the SNet has significantly less RMSE (about 5 to 15% lower) and uses a much lower number of non-null parameters (about 4-6×) that the neural network, in both experiences, mainly because of its sparse solution. Different local minima

in the SNet optimization leads to some structure's variations, but yet with very similar results. The results are quantitatively expressed in Table I, where the information criteria, AIC and BIC, are also shown. As expected, being the error lower and having lower number of parameters, the SNet also has better scores in the information criteria.

In Fig. 5 we breakdown the reconstruction RMSE at each pixel of the retina, computed over all images of the test set. We can observe the localization of pixels which lead to higher error and also compare the effectiveness of reconstruction between both methods. For both experiments, the reconstruction error is higher near the retinas boundaries. Both in translations and zoom-out actions there are image regions that are not possible to reliably predict since they are out of the pre-action image. Thus it is natural to have higher reconstruction errors close to the boundaries. Anyway, this fact is exacerbated in the neural network on Experiment RZ showing its limitations on predicting this type of motions, since its prediction is a mean radial distribution of intensities showing no patterns of the expected images.

C. Stimulus Predictions

After training the sensori-motor network, we can use it for making stimulus prediction of the agent's actions. Given a certain planned motor action $\mathbf{q}$ we can compute (i) the activation of the motor fields, $\mathbf{a} = \mathbf{M}^T\mathbf{q}$; (ii) the prediction matrix $\mathbf{P}$ by Eq. (2); (iii) the predicted stimulus by $\mathbf{o_1} = \mathbf{PSi_0}$; and finally (iv) obtain the predicted image by $\hat{\mathbf{i}}_\mathbf{1} = \mathbf{S^T o_1}$. In Fig. 6 we illustrate graphically steps (i), (ii) and (iii) for the translational action $\mathbf{u} = (4, 4)$ and the rotation/zoom action $\mathbf{u} = (50, 1.0)$. In the left figures we can observe the location of the motor RFs and their activation (the shade of gray) for that specific action. In the right we can observe the sensor RFs location with arrows representing the main directions of flow of the prediction. Remember that each entry of the prediction matrix p_{mn} indicates the contribution of receptive field n before action to the value of receptive field m in the post action. The arrows thus indicate the contribution of the source RFs in the formation of the target RFs, with weights proportional to the arrows gray level. The formation of the predicted image, step (iv), is illustrated in Fig. 7. This can be interpreted as the imagination of what will appear in the agent's field of view after its action is executed. Comparing the predicted image with the actual post-action image, we can see that the former is a low pass version of the latter, i.e. the best encoding of the reality in a least squares sense, with the available computational resources (RFs).

VI. Conclusions and Future Work

In robotics, as in many other engineering fields, there are numerous problems where Nature is often the best role model to solve them. In this work, it was possible to successfully apply the proposed method [4] for post-action images reconstruction and significantly reduce the number of parameters needed to predict visual stimuli caused by self-induced actions by drawing inspiration from biological systems.

The development of visual receptive fields taking into account the changes induced by motor actions allows a good adaptability of the organism to the environment and thus a cheaper way for an agent to process and predict visual stimuli. A specialized network architecture like the SNet described in this work is advantageous for predicting the interactions between a sensorial and a motor system, as well as obtaining more reliable predictions of what agent is expecting to see after moving.

This tight relationship between perception and actions is key for guiding the development of sensory and motor systems which will support acting upon the environment. The comparison performed in this work between standard artificial neural networks and sensori-motor networks, suggests that the latter might prove useful in bringing us a step closer

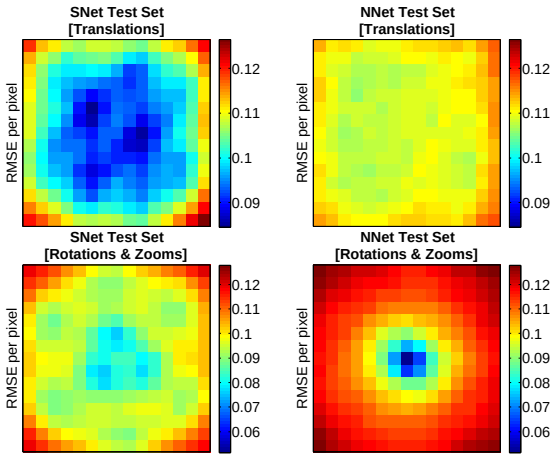


Fig. 5: Comparison between both methods regarding RMSE per pixel for reconstruction in a test set. (Top) Experiment XY run. (Bottom) Experiment RZ run (Best seen in color).

EXP. XY	Sensori-motor	Neural Network	NNet/SNet
All Parameters	3483	5013	1,44
Parameters $\neq 0$	1140	5013	4,40
Parameters $\geq 10^{-3}$	803	4910	6,11
RMSE	0.1004	0.1087	1,08
AIC	2.654	10.457	3,94
BIC	10.628	45.546	4,29
EXP. RZ	Sensori-motor	Neural Network	NNet/SNet
All Parameters	3483	5013	1,44
Parameters $\neq 0$	1053	5013	4,76
Parameters $\geq 10^{-3}$	743	4925	6,63
RMSE	0.0955	0.1100	1,15
AIC	2.442	10.467	4,29
BIC	9.817	45.556	4,64

TABLE I: Comparison between SNet and NNet in both translation and rotation experiments. The presented values result from the average from all 10 runs. As observed sensori-motor approach uses less parameters, produces a bit less reconstruction error and has less loss of information. The differences are higher between the models in Experiment RZ.

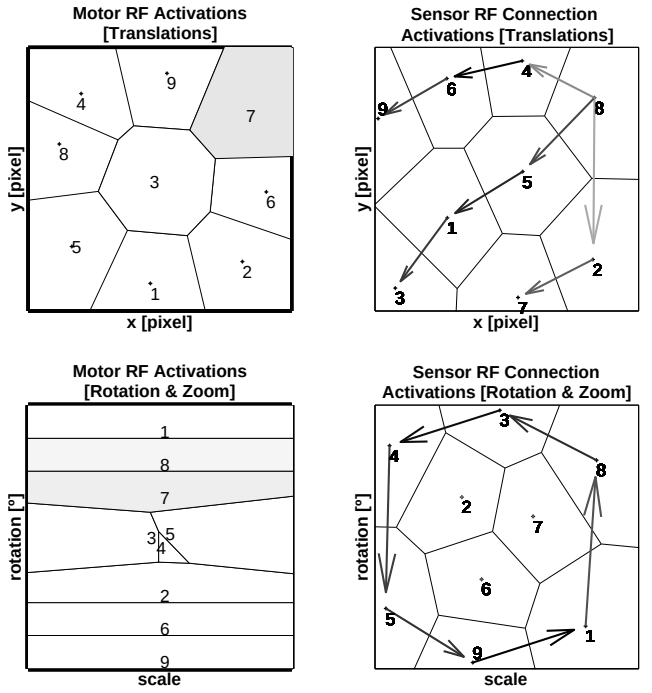


Fig. 6: (Left) Motor RF activations corresponding to particular actions. (Right) Induced prediction field in the sensory space. (Top) Action $\mathbf{u} = (4, 4)$ on the translation network (Bottom) Action $\mathbf{u} = (50°, 1.0)$ in the rotation/zoom network. The sensor RFs connections are represented by arrows intensity proportional to the corresponding prediction matrix entry (see details in text). Only prediction links with weights over 0.25 are shown. Voronoi diagrams are used to split the motor and sensor spaces into RFs.

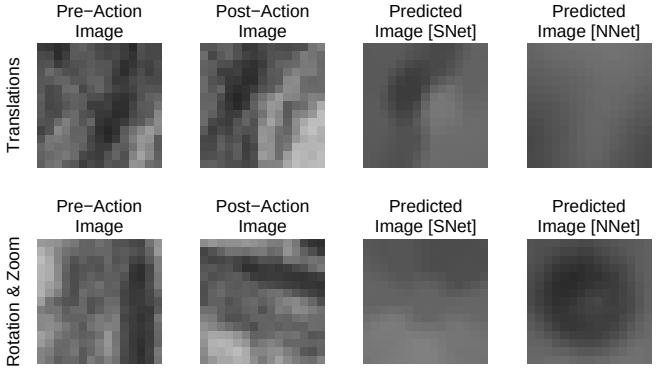


Fig. 7: Real and predicted image examples for the respective actions: (Top) Translation action example: $\mathbf{u} = (4, -4)$ and (Bottom) Rotation and zoom action example: $\mathbf{u} = (-75°, 1.20)$, using both SNet and NNet methods. As shown, reconstructions obtained by SNet optimization show a more coherent prediction of visual stimuli regarding the expected images.

to biological performance.

We plan to compare this sensori-motor network with more complex neural networks in other tasks to understand in what extent SNets can surpass the common NNet approaches. Mimicking the Human's visual and motor system, SNet can also be extended to include more layers, allowing more richer representations for complex tasks. Another component which would increase the applicability of the presented work is the notion of state to support planning tasks. An online development algorithm would simplify the application of this model to robots in different and dynamical environments.

ACKNOWLEDGMENT

This work was supported by the FCT projects BIOMORPH-EXPL/EEI_AUT/2175/2013 and Pest-OE/EEI/LA0009/2013 and also by EU Projects POETICON++ [FP7-ICT-288382].

REFERENCES

[1] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2002.

[2] T. B. Crapse and M. A. Sommer, "Corollary discharge across the animal kingdom." *Nat. Rev. Neurosci.*, vol. 9, no. 8, pp. 587 – 600, 2008. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18641666

[3] R. C. Miall and D. M. Wolpert, "Forward models for physiological motor control," *Neural networks*, vol. 9, no. 8, pp. 1265 – 1279, 1996.

[4] J. Ruesch, R. Ferreira, and A. Bernardino, "A computational approach on the co-development of artificial visual sensorimotor," *Adaptive Behavior*, vol. 21, no. 6, pp. 452 – 464, 2013.

[5] ——, "A measure of good motor actions for active visual perception," *IEEE International Conference on Development and Learning, ICDL 2011.* , vol. 2, pp. 1–6, 2011.

[6] ——, "Predicting visual stimuli from self-induced actions: an adaptive model of a corollary discharge circuit," *IEEE Transactions on Autonomous Mental Development,*, vol. 4, no. 4, pp. 290–304, 2012.

[7] S. Clippingdale and R. Wilson, "Self-similar neural networks based on a Kohonen learning rule," *Neural Networks*, vol. 9, no. 5, pp. 747 – 763, 1996.

[8] L. A. Olsson, C. L. Nehaniv, and D. Polani, "From unknown sensors and actuators to actions grounded in sensorimotor perceptions," *Connection Science*, vol. 18, no. 2, pp. 121 – 144, 2006.

[9] L. Lichtensteiger and P. Eggenberger, "Evolving the morphology of a compound eye on a robot," in *Third European Workshop on Advanced Mobile Robots, 1999. (Eurobot '99) 1999* , 1999, pp. 127 – 134.

[10] C. Paul, "Morphological computation: A basis for the analysis of morphology and control requirements," *Robotics and Autonomous Systems*, vol. 54, no. 8, pp. 619 – 630, 2006.

[11] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. [Online]. Available: citeseer.ist.psu.edu/lecun98gradientbased.html

[12] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.