This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2638885, IEEE Transactions on Cognitive and Developmental Systems

SNet: Co-Developing Artificial Retinas and Predictive Internal Models for Real Robots

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 work focuses on a recently developed biologically inspired architecture, here denoted as Sensorimotor Network (SNet), able to co-develop sensorimotor structures directly from data acquired by a robot interacting with its environment. Such networks learn efficient internal models of the sensorimotor system, developing simultaneously sensor and motor representations as well as predictive models of the sensorimotor relationships adapted to their operating environment. Here we describe our recent model of sensorimotor development and compare its performance with neural network models in predicting self-induced stimuli. In addition, we illustrate the influence of available resources and environment characteristics in the development of the Sensorimotor Network structures. Finally, a Sensorimotor Network is trained using real data recorded during a quadricopter drone flight.**

***Index Terms*—Stimuli prediction, Neural Networks, Sensorimotor Network, forward model, receptive fields, movement fields.**

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. At the same time 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 that motor actions have in visual perception and visual stimulus prediction, as well as their 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 maximize 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 sensorimotor structures which support choosing the right action. To be able to do so one crucial faculty that organisms developed is the ability to discern the origin of sensory input between changes in the environment (exafference) and the result of their 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.

A recently proposed adaptive model [4] learns to predict visual stimuli based on motor information resulting from self-induced actions. This model maps motor input in a visual predictive network, creating direct relationship between the robot's actions and its perceived visual stimuli. Following a specific learning process it was possible to minimize the mean square prediction error between the predicted image and the expected image after a specific motor action.

In the proposed Sensorimotor Network, we consider a visual sensor where each neuron's receptive field (RF) collects information from arbitrary retina cells, without any predefined shape or topology. Those will emerge from the developmental process as the agent explores the environment. Simultaneously, the motor layer organizes into movement fields (MF) that will cluster actions which produce similar perceptual results. This simultaneous development promotes a coherent representation for similar stimuli (sensory) and actions (motor), which greatly improves the effectiveness of the model.

A key issue of our model is its specialized structure, that exploits the most of the limited computational resources to enable the best possible prediction of future perceptions, in the least squares sense. The presented work compares the performance of the proposed model with other common sensorimotor mapping models, with more general purpose structures, such as the multilayer perceptron. Because of its specialized topology, the SNet can attain significant advantages over fully connected networks. Taking one step further, we trained the SNet structures, for the first time with real visual and motor data acquired by a flying drone navigating in an outdoors natural environment.

This paper follows our original formulation for sensorimotor learning presented in [4] and the constrained gradient descent based optimization algorithm detailed in [5]. In [6] we have presented the comparison of the proposed model with a standard neural network with simulated data and introduced the interpretation and visualization of the learned predictive structures as a set of motion fields overlayed in the sensor topology. In the current paper, we extend [6] by providing novel experiments to illustrate the influence of the environment in the derived sensorimotor structures, both in a real environment with images acquired during a quadricopter flight, and in an artificial environment with images of strongly organized visual patterns. Furthermore, we perform experiments to assess the influence of available computational resources in the learned topologies. The experiments with a real robotic platform, validate the method's ability to self-organize relevant sensorimotor structures for real-world applications.

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 [7]. 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 the motor system to support the predictability of the actions [8].

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 [9]. Still, for these structures to emerge, apriori knowledge is required 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 [10]. In [11], information measures are also exploited to define similarity metrics between sensorimotor experiences with extended temporal ranges. 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 [12].

Guiding the development of a sensorimotor system to maximize the ability of predicting the effect an action has on its sensory input (see III-B2), 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. [13]). 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 [14] (e.g. convolutional) relatively to a fully-connected network.

III. Predictive Architectures

A. Prediction

We consider an agent capable of observing its environment by sensing a light field i_0 which falls on a sensory surface. Additionally this agent is able to interact with its environment when performing certain actions (movements), each resulting from the activation of a particular motor primitive $\mathbf{q}$ on its motor coordinates, which will produce a new visual stimulus i_1. The agent should learn to predict the effects of its actions $\mathbf{q}$ in its visual space ($\mathbf{i}_1$). In order to give the agent the ability to predict the future visual stimulus ($\mathbf{i}_1$) we consider two possible predictive architectures: a general purpose Neural Network (Multi-Layer Perceptron) and our Sensorimotor Network (Figures 1 and 2).

The Neural Network is organized in a classical architecture. Its input image data $\mathbf{i}_0$ and action data $\mathbf{q}$ project directly to an hidden layer by a set of weights $\mathbf{W_1}$ (refer to Fig. 1). Then, the activation of the hidden units project to the output layer via weights $\mathbf{W_2}$, to create a prediction i'_1 of the future image i_1. The hidden layer plays the role of a joint sensorimotor encoder, receiving directly the raw sensor and motor data.

Instead, the proposed Sensorimotor Network , follows a more complex organization, inspired by the role and connections between the *superior colliculus* and frontal-eye field structures of the human brain [8]. Separate encoders for sensor and motor data are considered (refer to Fig. 2). A set of weights $\mathbf{S}$ encodes the input image into a compact visual representation $\mathbf{o}_0$. A different set of weights $\mathbf{M}$ encode the motor coordinates in a compact action representation $\mathbf{a}_0$. This can be interpreted as a clustering of the motor commands that correspond to similar visual effects. Each cluster k represents a *canonical action*, and is associated to a *canonical predictor* represented by a set of weights $\mathbf{P_k}$. These weights are used to convert the encoded input stimulus $\mathbf{o}_0$ into a predicted output code $\mathbf{o}'_1$. Finally a decoding layer $\mathbf{S^T}$ reconstructs the predicted image i'_1. .

For implementation purposes, we represent the light field as a vector $\mathbf{i}$ of N_s pixels, and the action space is represented as a vector q with N_m elements (number of motor actions), 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} motor action is performed (e.g. shift left by a certain amount). Note that no topological assumptions exist on the spatial locations of either the sensors or the motor primitives, i.e. the order of pixels and actions in their vector representations is arbitrary.

Among the many existing types of supervised learning machines, we chose a Multi-Layer Perceptron as a comparison baseline for our method for its simplicity and biological relevance. Other biologically inspired learning machines such as Deep Neural Networks (DNN) [15] or Recurrent Neural Networks (RNN) [16] also have the ability to address our

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2638885, IEEE Transactions on Cognitive and Developmental Systems

problem but are significantly more complex in terms of the number of parameters and amount of data required for its training. Multi-Layer Perceptrons are mature and standard learning machines that can be implemented by off-the shelf tools and libraries and easily customized to have a similar number of parameters to the proposed model, so as to present roughly the same computational complexity. Also they have the ability to fit arbitrary non-linear functions. According to the universal approximation theorem [17], a standard multilayer feed-forward network with as few as a single hidden layer and arbitrary bounded and non-constant activation function are universal approximators provided only that sufficiently many hidden units are available. Beyond non-linear activation units, we consider also linear units so as to match the activation functions of the units in the proposed model.

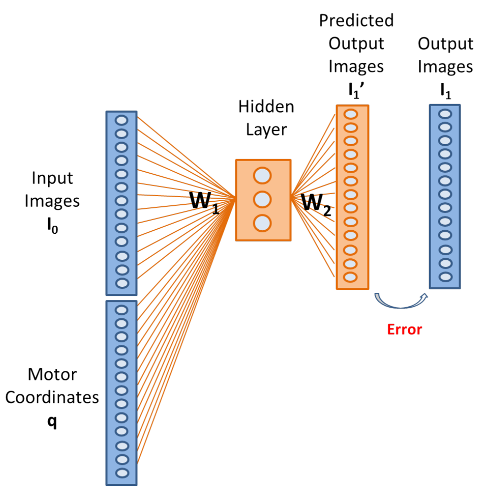

Fig. 1: Neural Network: schematic diagram representing the total data triplets ($\mathbf{I}_0$,$\mathbf{I}_1$,$\mathbf{q}$) used to train the model (blue) its parameters $(\mathbf{W}_1, \mathbf{W}_2)$ and predicted stimulus $\mathbf{I'}_1$ (orange). The hidden layer can have different transfer functions (linear or non-linear).

The Neural Network and SNet are then compared in terms of: 1) their predictive capabilities, i.e. how well they can predict $\mathbf{i}_1$ given $\mathbf{i}_0$ and $\mathbf{q}$; 2) their simplicity, i.e. the number of parameters learned.

B. Learning Algorithms

During the learning phase, the agent interacts with the environment by choosing a motor primitive $\mathbf{q}$ while collecting pre-action and post-action sensory stimuli ($\mathbf{i}_0$ and $\mathbf{i}_1$, respectively). A set of ($\mathbf{i}_0$,$\mathbf{i}_1$,$\mathbf{q}$) triplets is collected and the full batch is used as training data. In the current work, the process that chooses the actions is considered independent from the current state of the sensorimotor system and, thus, not adapted during the learning process. In the current work we only address the learning of the sensory predictive system and undeling structures, leaving the learning of the action selection policy for future work.

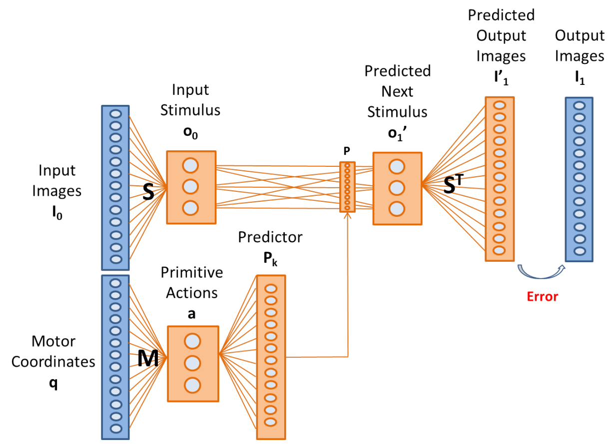

Fig. 2: Sensorimotor Network: schematic diagram representing data triplets ($\mathbf{I}_0$,$\mathbf{I}_1$,$\mathbf{q}$) used to train the model (blue), its parameters $(\mathbf{P}, \mathbf{M}, \mathbf{S})$ and predicted stimulus $\mathbf{I'}_1$ (orange).

1) Neural Network: In this case we consider a feedforward network with n_s elements in its hidden layer emulating sensor 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

$$\underset{\mathbf{W_1},\mathbf{W_2}}{\text{argmin}} \sum_k \left\| \mathbf{W_2} \begin{bmatrix} f(\mathbf{W_1} \begin{bmatrix} \mathbf{i}_0^k \\ \mathbf{q}^k \\ 1 \end{bmatrix}) \\ 1 \end{bmatrix} - \mathbf{i}_1^k \right\|^2 \tag{1}$$

which is illustrated by the neural network represented in Fig. 1. 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 and f represents the activation function (linear or non-linear).

2) Sensorimotor Network: The sensory prediction system described in [4], explicitly models the existence of light sensitive receptors represented as a $N_s \times n_s$ matrix $\mathbf{S}$ which integrates the light field $\mathbf{i}$ falling on the 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 modelled as a $N_m \times n_m$ matrix $\mathbf{M}$ cover the available motor primitive space $\mathbf{q}$, providing an n_m dimensional motor field activation vector $\mathbf{a} = \mathbf{M}^T\mathbf{q}$, where $\mathbf{a}_j$ is a scalar representing the activation of motor field j. These activations are then fed to a predictive layer, where a predictor $\mathbf{P}^k$ is composed as a linear combination of n_m canonical predictors $\mathbf{P}_j$ with linear weights given by the motor movement fields activations,

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

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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2638885, IEEE Transactions on Cognitive and Developmental Systems

The full model description is provided in [4]. The network parameters are obtained by minimizing the mean squared reconstruction error under positivity and topological constraits:[1] The optimization problem solved is thus

$$\underset{\mathbf{S}\geq\mathbf{0},\mathbf{M}\geq\mathbf{0},\mathbf{P}\geq\mathbf{0}}{\text{argmin}} \sum_k \left\| \mathbf{S}^T \left(\sum_j^{n_m} \left(\mathbf{m}_j^T \mathbf{q}^k\right) \mathbf{P}_j \right) \mathbf{S}\mathbf{i}_0^k - \mathbf{i}_1^k \right\|^2 \tag{3}$$

A diagram representing this optimization problem in a network-like view is shown in Figure 2. Unlike in the artificial neural network architecture, the sensor reconstruction model is simplified to be $\mathbf{S}^T$. 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. The algorithm to solve the optimization problem is shown in Algorithm 1. To impose positivity constraints the projected gradient method is used. The detailed calculation of the gradient for each variable is presented in [5].

The computational complexity of the proposed architecture is dependent on the complexity of the matrix product which for matrices A(n x p)*B(p x m) is $\mathcal{O}(npm)$. The time complexity of the model is $\mathcal{O}(n_m{*}n_s^2{+}n_s{*}N_s)$, i.e. quadratic on the number of sensor fields and linear on the number of motor fields

Data: Triplets $(\mathbf{i}_0, \mathbf{i}_1, \mathbf{q})$.
Result: Trained model for visual stimuli prediction.
initialization;
for *each sequential iteration* **do**
    for *each **P** iteration* **do**
        apply gradient step to **P** with **Data**;
        $\mathbf{P} \leftarrow \texttt{max}(\mathbf{P}, 0)$;
    end
    for *each **M** iteration* **do**
        apply gradient step to **M** with **Data**;
        $\mathbf{M} \leftarrow \texttt{max}(\mathbf{M}, 0)$;
    end
    $\mathbf{M} \leftarrow \mathbf{M}/\texttt{norm}(\mathbf{M})$;
    for *each **S** iteration* **do**
        apply gradient step to **S** with **Data**;
        $\mathbf{S} \leftarrow \texttt{max}(\mathbf{S}, 0)$;
    end
    $\mathbf{S} \leftarrow \mathbf{S}/\texttt{norm}(\mathbf{S})$;
end

Algorithm 1: Pseudo-Code: Sensorimotor Optimization

IV. Experiments

In this section we describe experiments performed to illustrate the properties of the proposed model, both on simulated data and on data acquired from a real robotic platform.

Simulated experiments use real imagery but virtual actions to produce the training, test and validation data for experiments. Actions are predefined and chosen from a finite set. Simulations are performed to evaluate the proposed model on several dimensions: (i) compare its performance with linear and non-linear neural network models; (ii) analyse the topology of the sensor and motor spaces after development; (iii) evaluate the influence of sensor size and environment type in the topology of the sensor and motor spaces, and (iv) will assess the predictive ability of the derived models.

Real experiments were performed with a drone flying freely in a forest environment. Actions are selected by the drone navigation system while travelling along a predefined GPS trajectory. A large dataset of real images and corresponding motor commands was acquired. Training for the first time our SNet model with real data, we were able to assess the ability of the model to derive apropriate sensorimotor models with noisy data and operate in challenging outdoor scenarios.

A. Simulation Environment

In our simulation environment, a virtual agent is equipped with a square retina of 15 by 15 pixels ($N_s = 225$) which is used to acquire grayscale images with intensity ranging from 0 to 1. Triplets ($\mathbf{i}_0$,$\mathbf{i}_1$,$\mathbf{q}$) are sampled from a large (2448 by 2448 pixels) image representing the full environment. First, the agent is positioned in a random place in the environment and an image i_0 is sampled by its 15 by 15 pixel retina. Then action u is performed and the new image i_1 is sampled. This process is illustrated in Figure 3.

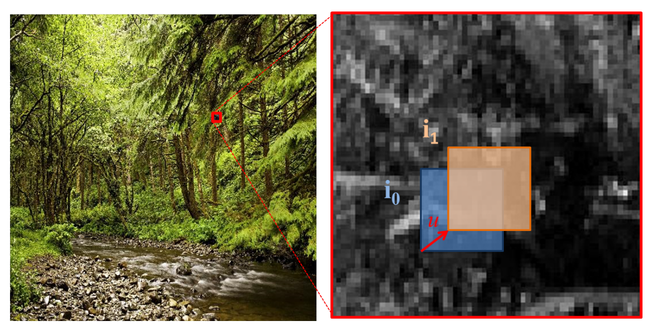

Fig. 3: Triplet acquisition process. In the left we show the full environment image. In the right we show a portion of the environment where the agent is placed to acquire the pre-action 15×15 pixel image, i_0, then transformed by action u, and acquire the post-action image, i_1 (best seen in color).

We consider two different motor spaces. One composed of translation actions (ActXY) and another composed of rotations and zooms (ActRZ). Both motor spaces are two dimensional so their topology can be easily visualized. ActXY is composed of translation actions on the set $\mathbf{u} = \{-4:1:4\} \times \{-4:1:4\}$, and ActRZ combines rotations and zoom

[1]To remove gauge freedom and avoid numerical problems the matrix norm of S and the matrix norm of M are constrained to a constant through a scalar division (see Algorithm 1)

actions $\mathbf{u} = \{-100^\circ : 25^\circ : 100^\circ\} \times \{0.80 : 0.05 : 1.20\}$. Thus, each set consists of 81 different motor actions.

The first set of actions mimic an agent that either moves its sensor parallel to the environment surface or performs small pan-tilt rotations of the sensor when observing far objects. The second set of movements can approximately represent 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 to the observer.

To train our sensorimotor models, we use training sets with 8100 triplets ($\mathbf{i}_0$,$\mathbf{i}_1$,$\mathbf{q}$) generated by random initialization of the sensor position in the simulated environment (100 triplets per each one of the $N_m = 81$ primitive actions). Test sets to evaluate generalization ability, and validation sets to define stopping criteria are generated similarly. The test set has the same number of triplets and the validation set has half the number of samples.

Unless otherwise stated, all simulated experiments are performed by executing ten independent runs of model learning, and results are averaged over the ten runs. Experiments defined in such way using the ActXY and ActRZ sets of actions are denoted ExpXY and ExpRZ, respectivelly.

B. Model Comparison

We compare the proposed Sensorimotor Network architecture with both a linear and a non-linear artificial Neural Network (Multi-Layer Perceptron with one hidden layer).

After acquiring its exploration data in the given environment, using experimental protocols ExpXY and ExpRZ, the agent processes the data in order to obtain the network parameters for the Sensorimotor Network $(\mathbf{S}, \mathbf{M}, \mathbf{P})$ and for the Neural Network $(\mathbf{W}_1, \mathbf{W}_2)$. The optimization criteria is Mean Squared Error (MSE) between the predicted and observer images, in Eqs. (1) and (3). In both experiments, the SNet model is formed by a motor structure composed by 9 motor movement fields (n_m = 9) and a sensor structure composed by 9 sensor receptive fields ($n_s = 9$). The SNet is compared with both a linear Neural Network (NNet) and a non-linear Neural Network (nNNet) with a hyperbolic tangent sigmoid transfer function, each with a hidden layer of 9 neurons. The neural networks were implemented using the Neural Network Toolbox from Matlab$^{\mathrm{TM}}$. In these experiments an identical number of sensor receptive fields (RF) and motor movement fields (MF) are used for the SNet model but they can differ. The number of hidden units can be chosen taking into account the resources available in the particular hardware used to deploy the system. Also, a higher sensor resolution should be followed by a higher number of sensor RFs or a higher number of actions should be followed by a higher number of motor MFs.

The optimization problem for the Sensorimotor Network showed in Equation (3) is iteratively improved using a projected gradient descent method [18] within the sequential optimization of $\mathbf{P}$, $\mathbf{M}$, $\mathbf{S}$, and the input triplets are considered in batches as in [4] (see Algorithm 1). For both Sensorimotor Network and Neural Networks, the RMSE between predicted and expected images is computed,

$$\mathrm{RMSE} = \sqrt{\frac{1}{N_m \times L \times N_s} \sum_{k=1}^{N_m} \sum_{l=1}^{L} \sum_{p=1}^{N_s} \left(\mathbf{i'}^k_{1_{(l,p)}} - \mathbf{i}^k_{1_{(l,p)}} \right)^2} \tag{4}$$

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 convergence of training on the 10 runs for all models, we obtained the performance statistics for evaluation. Results are shown in Tables I and II. In both experiments, (ExpXY, ExpRZ), we can observe that the SNet has significantly less RMSE (about 5 to 15% lower) and uses a much lower number of effective (non-zero) parameters (about 4-6$\times$) than the other models. Because SNet promotes sparsity in the solution, we obtain a much lower number of non-zero parameters, that lead to a much higher computational efficiency. Most likely, the regularization properties of sparse coding also help in reducing overfit and, thus, achieving a better generalization error. The distribution of the RMSE on the 10 runs is illustrated in Figure 4. There we can observe that the RMSE difference between the sensorimotor and neural networks is significant in both the translation and rotation-zoom experiments.

In Figure 5 we graphically illustrate the average RMSE at each pixel of the retina over all images of the test set. We can observe the localization of the pixels that contribute to a higher error and compare the effectiveness of the reconstruction between both methods. For both experiments the prediction error is higher near the retina's boundary. These image regions cannot be reliably predicted for some actions because they rely on information outside $\mathbf{i}_0$ image boundaries; there are image regions which are not possible to predict because they are out of the pre-action image. Anyway, this fact is exacerbated in the NNet, showing its limitations in this problem.

C. Emergent Sensorimotor Topologies

Here we revisit the emergent properties [4] of the sensor and motor spaces after the optimization problem (3) with SNet. These results illustrate some interesting outcomes of the optimization process in terms of the shape and distribution of the sensor and motor fields. After the development process described in the previous experiment, the sensor receptive fields (rows of $\mathbf{S}$) organize into a regular structure (after 500 iterations) starting from a random initialization (see Fig. 6). Notice that sensor organization is more uniform for translation actions than for rotations and zooms. With rotations and zooms the sensor RFs tend to create a group of smaller receptors in the middle of the retina and bigger fields

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2638885, IEEE Transactions on Cognitive and Developmental Systems

ExpXY	SNet	NNet	NNet/SNet
All Parameters	3483	5013	1,44
Parameters $\neq 0$	1140	5013	4,40
Parameters $\geq 10^{-3}$	803	4910	6,11
Average RMSE	0.1004	0.1087	1,08
ExpRZ	SNet	NNet	NNet/SNet
All Parameters	3483	5013	1,44
Parameters $\neq 0$	1053	5013	4,76
Parameters $\geq 10^{-3}$	743	4925	6,63
Average RMSE	0.0955	0.1100	1,15

TABLE I: Comparison between Sensorimotor Network (SNet) and linear Neural Network (NNet) in both translation and rotation experiments. The presented values are the average over 10 runs. As observed, the sensorimotor approach uses less parameters and produces less reconstruction error.

ExpXY	SNet	nNNet	nNNet/SNet
All Parameters	3483	5013	1,44
Parameters $\neq 0$	1140	5013	4,40
Parameters $\geq 10^{-3}$	803	4992	6,22
Average RMSE	0.1004	0.1241	1,24
ExpRZ	SNet	nNNet	nNNet/SNet
All Parameters	3483	5013	1,44
Parameters $\neq 0$	1053	5013	4,76
Parameters $\geq 10^{-3}$	743	4993	6,72
Average RMSE	0.0955	0.1233	1,29

TABLE II: Comparison between Sensorimotor Network (SNet) and non-linear Neural Network (nNNet) in both translation and rotation experiments. The presented values are the average over 10 runs. nNNet uses the same number of parameters with higher RMSE than NNet (see Table I).

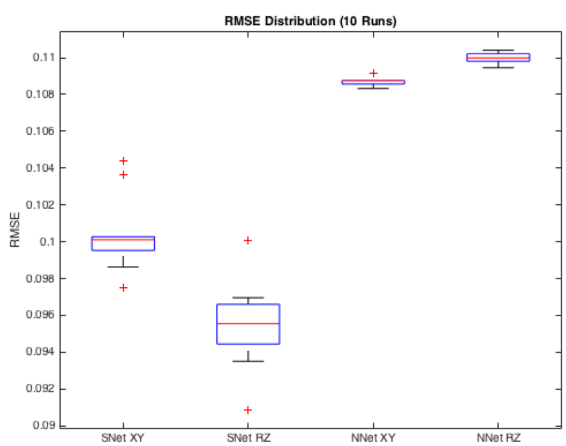

Fig. 4: Box plots of the RMSE distribution for the tested conditions. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually. Notice the significant differences between the SNet and NNet structures, bot for translations and rotation-zoom experiments.

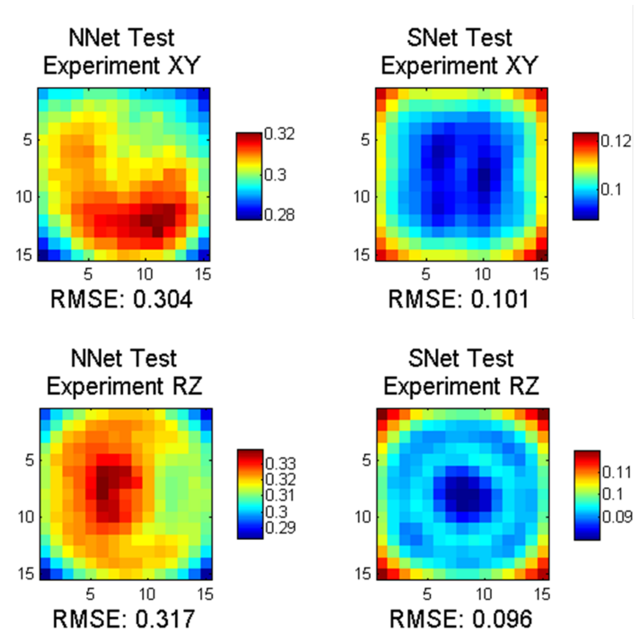

Fig. 5: Comparison between Sensorimotor Network (SNet) and linear Neural Network (NNet) methods regarding RMSE per pixel for reconstruction in a test set. (Top) Taken from 10 runs average in ExpXY. (Bottom) Taken from 10 runs average in ExpRZ. Vertical and horizontal axis represent sensor pixels (Best seen in color).

near the boundaries (a rotation produces a bigger movement far from the center).

In Figure 7 we can observe the evolution of the motor MFs (columns of $\mathbf{M}$) for both experiments. Again, ExpXY has its action space uniformly sampled by pixels, producing a near uniform organization of the motor MFs. The performed zooms in ExpRZ had low impact on their images in comparison with the rotations, which caused the MFs 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 influence in their organization.

D. Emergent Predictive Structures

After training the Sensorimotor Network we can use it for making stimulus prediction of the agent's actions. For 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 Equation (2); (iii) the predicted stimulus by $\mathbf{o_1} = \mathbf{PSi_0}$; and finally (iv) obtain the predicted image by $\mathbf{i'_1} = \mathbf{S^T o_1}$.

Here we use the SNet model trained in the previous experiments with 9 sensor receptive fields and 9 motor movement fields. In Figure 8, steps (i), (ii) and (iii) are graphically illustrated for the translational action $\mathbf{u} = (4, 4)$ and the rotation/zoom action $\mathbf{u} = (50°, 1.0)$. On the left, the resulting predictor $\mathbf{P}^k$ for the activated action is represented.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2638885, IEEE Transactions on Cognitive and Developmental Systems

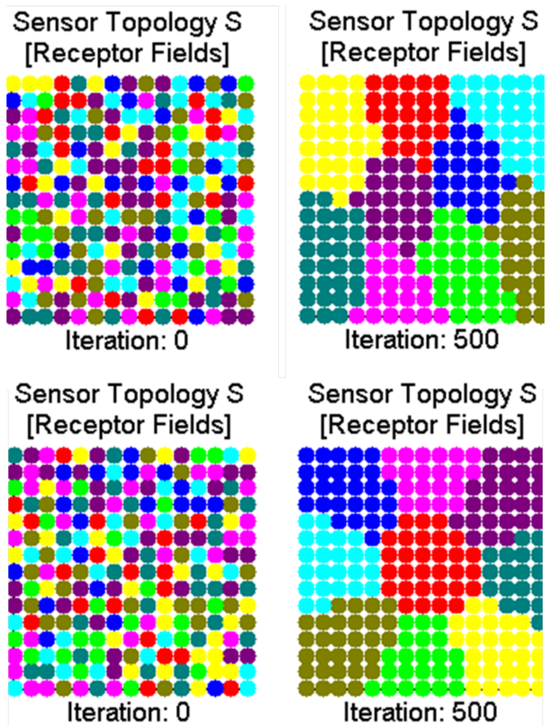

Fig. 6: Sensor RFs initialization and final organization after 500 iterations in one of the runs of ExpXY (Top) and ExpRZ (Bottom). Each color represents a receptive field which after training covers a continuous part of the considered visual area (vertical and horizontal axis represent sensors as pixels in an image $\mathbf{i}$) (Best seen in color).

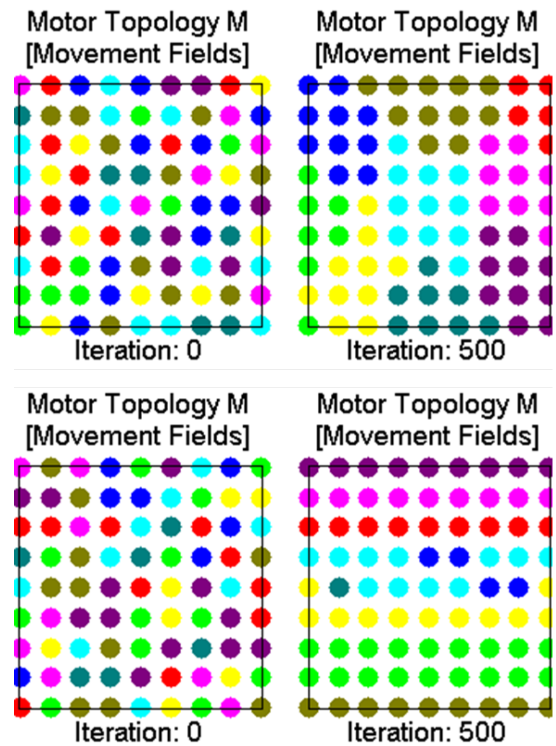

Fig. 7: Motor MFs initialization and final organization after 500 iterations in a run of ExpXY (Top - vertical axis represents up and down translations and horizontal axis represents left and right translations) and ExpRZ (Bottom - vertical axis represents rotations and horizontal axis represents zooms) (Bottom). Each color represents a movement field which after training covers a continuous part of the considered motor area (Best seen in color).

On the middle the location of the motor movement fields and its activation (gray shade) is shown. Finally, on the right, we can observe the motion flow map generated by the predictor, overlaid on the sensor receptive fields distribution. The arrows represent the main directions of flow of the resulting prediction $\mathbf{P}^k$. The predictor translates motor effects on the visual area, by weighting connections between the receptive fields and identifying areas of observation which will move from a receptive field (transmitter) to another (receiver). The arrows thus indicate the contribution of the transmitter in the formation of the target receptor field, with the weights proportional to the arrows gray level. Figure 9 displays the motion flow maps for many other actions.

The formation of the predicted image, step (iv), is illustrated in Figure 10. This is interpreted as the prediction 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 conclude 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.

E. Influence of Sensor Size

In this experiment, denoted ExpSize, we analyse the influence of the size of the sensor space. Maintaining the same motor structure ($n_m = 9$) and environment (Fig. 3) of the previous experiments, new models were trained with three different number of sensor receptive fields ($n_{s1} = 9$, $n_{s2} = 16$, $n_{s3} = 25$).

In Figure 11 we can observe the organized sensor topologies with the three different sensor sizes. Again, the reconstruction error RMSE was computed using a test set with the same size of the training set (8100 triplets). Examples of original and predicted images are shown: for a particular action and pre-action image, the prediction is computed and compared with the actual observed post-action image.

As expected, the quality of the reconstruction improves with the number of sensor receptive fields. Although the prediction error decreases with the number of available receptive fields, this also presents an increase on the number of empty receptive fields. In the case where 9 sensor receptive fields were considered, the model used all of them. However, when using a higher number of sensor receptive fields, some are not used for the adapted model (1 out of 16 RFs and 7 out of 25 RFs). All in all, a trade-off can be found in increasing the sensor complexity. The number of receptive fields should be increased only up to the point that prediction error decreases,

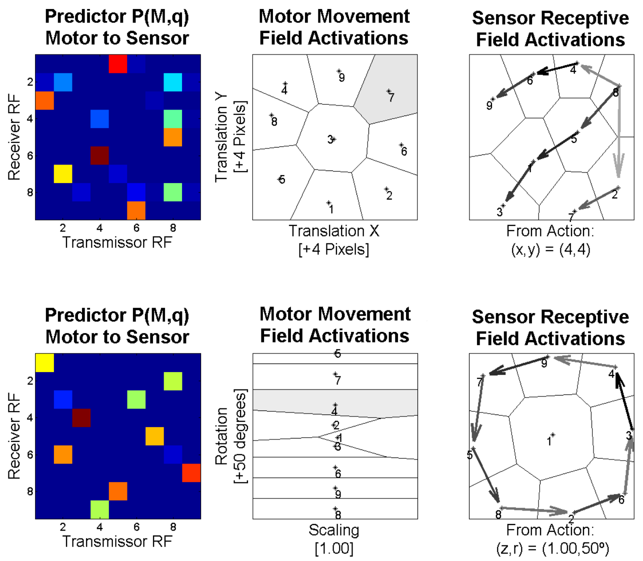

Fig. 8: (Left) Predictive structure $\mathbf{P}^k$. (Mid) Motor MFs activations corresponding to particular actions. (Right) Induced motion flow maps in the sensory space. (Top) Action $\mathbf{u} = (4,4)$ on the translation network (Bottom) Action $\mathbf{u} = (50°, 1.0)$ in the rotation/zoom network. The sensor RFs connections are represented by arrows with 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.

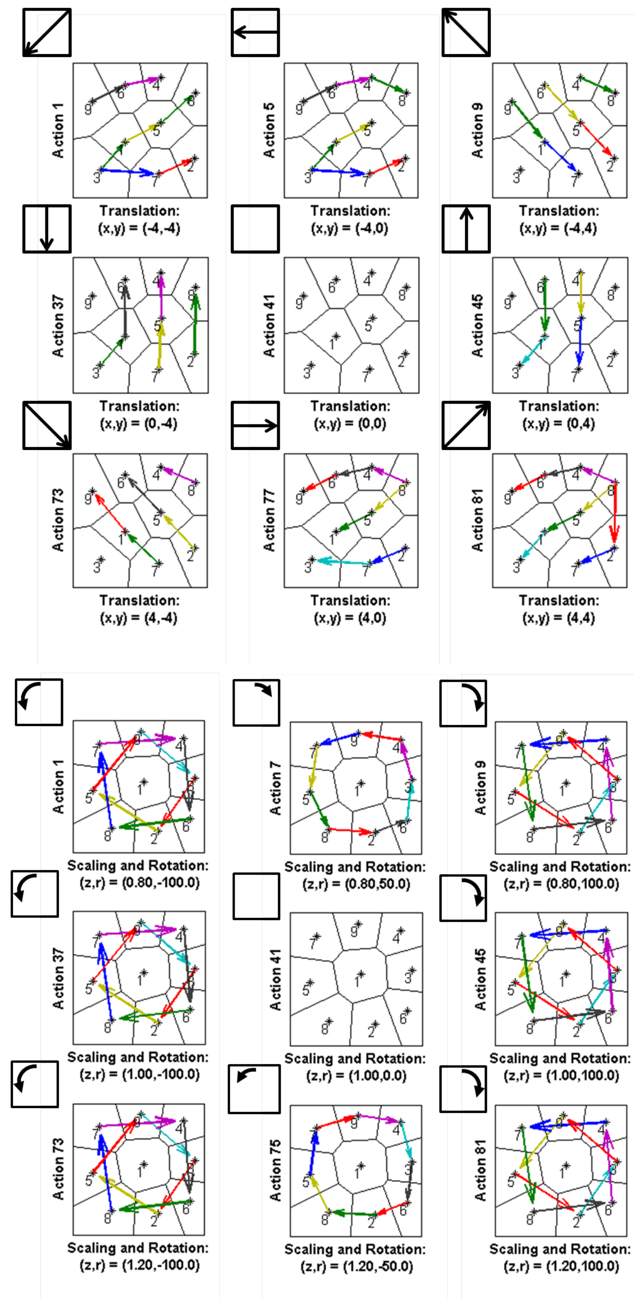

Fig. 9: Motion flow maps induced by predictive structure. Predictive structure influences sensor receptive fields after receiving a motor activation from action space ActXY (top) and action space ActRZ (bottom). On each example, the boxed arrow in the top left corner of each figure represents the direction and/or amplitude of the sensor translation or rotation action with respect to the environment.

otherwise unnecessary complexity and computational costs may be incurred.

F. Influence of Environment Type

As discussed in many works [19], [20], [21] the eye, retina and visual system evolved in many species in very distinctive manners, but still reaching highly efficient forms with specific ecological advantages. Three main characteristics can be enumerated which directly influence their structures: organism's nervous system, organism's motor capabilities and organism's perception of the environment.

Here we test the environment influence on the sensory structure developed by our model. Again we use 9 sensor receptive fields, 9 motor movement fields, and the ExpXY /ExpRZ experimental protocols, but now use four different environment images: 3 synthetic images (vertical stripes, diagonal stripes and dots) and 1 natural (textured picture of dry soil). In Figure 12, the 4 different environment and resulting sensor organizations, $\mathbf{S}$, are shown.

From the presented results we can conclude that the sensor structure organization depends on the environment. In this sense, our results validate the hypothesis in [22], that a retina does acquire knowledge, in its organization, about the natural scenes (environment). However, our results show that it is not only the environment that matters for the developed visual sensor topology. It also greatly depends on the motor repertoire of the agent. Even with very different environments, we can realize that only by changing the set of movements the agent can perform, a particular visual topology can emerge.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2638885, IEEE Transactions on Cognitive and Developmental Systems

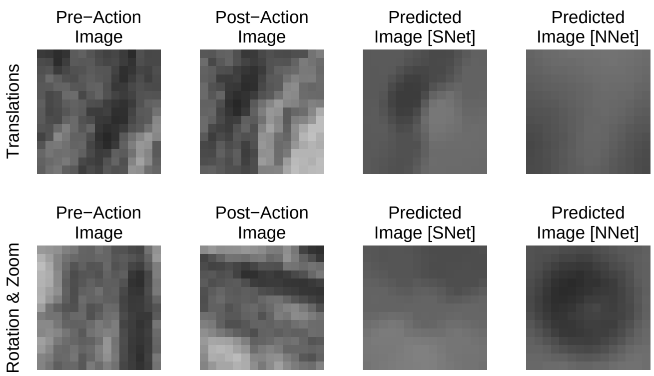

Fig. 10: Real and predicted image examples for the respective actions: (Top) Translation action example: $\mathbf{u} = (4, -4)$ and (Bottom) Rotation and zoom action example: $\mathbf{u} = (-75°, 1.20)$, using both SNet and linear NNet methods. As shown, reconstructions obtained by SNet optimization show a more coherent prediction of visual stimuli regarding the expected images (vertical and horizontal axis represent sensor pixels).

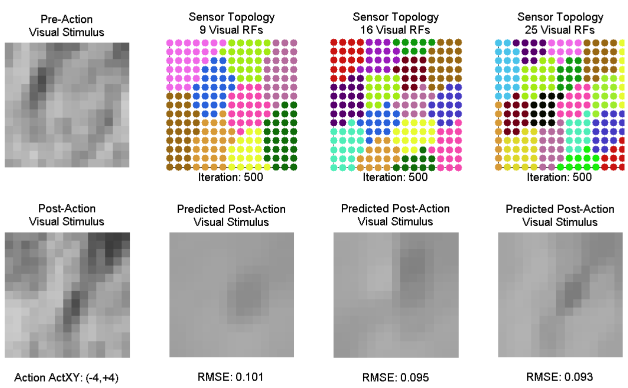

Fig. 11: Sensor Receptive Fields: Influence on prediction error and quality on image reconstruction (vertical and horizontal axis represent each sensor pixel).

As we can observe in Figure 12, translational movement in environments with stripes very clearly map the visual structure in the retina organization. However, the mapping is not so clear in the more random types of environments such as the artificial dots and the natural texture. Also, for rotation and zoom actions, the characteristics of the tested environments are not clearly reflected in the retina. There may be other environments where the retina may approximate better the visual characteristics but we did not actively search for those cases. In summary, we may conclude that the sensor organization depends not only in the environment characteristics but also in the agent's motor actions. However, it is still unclear how to make predictions on the sensor structure for different types of actions and environments.

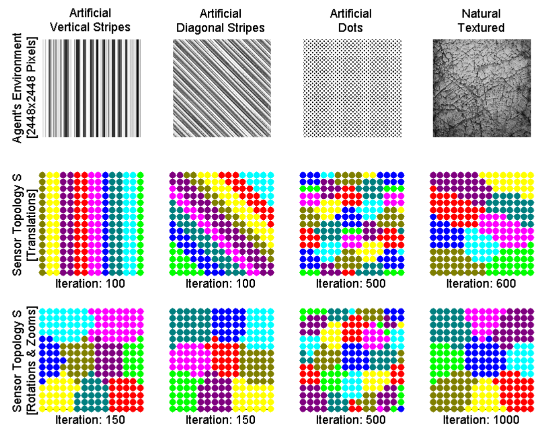

Fig. 12: Environment influence on visual sensor topology. Sequence of visual sensor topologies resulting from training the sensorimotor system using action spaces from ExpXY and ExpRZ, and four different environments: three artificial environments (vertical stripes, diagonal stripes and dots) and one natural environment (textured picture of dry soil). The number of iterations until convergence for each experience indicated is shown below.

G. Sensorimotor Development with Real Data

In this experiment, we used a Parrot AR.Drone2.0 aerial quadrotor to acquire images from a natural environment in Monsanto park in Lisbon. This drone is equipped with a fixed HD camera always pointing forward. During the flight a video was recorded at a rate of 30 frames per second, together with drone position variation (Δx,Δy) from GPS, orientation variation ($\Delta\theta$) and altitude variation (Δz). As such, the motor space is 4-dimensional. For the sake of simplicity, data from drone taking off or landing was excluded.

The data acquisition (image and actions) was performed while the drone followed a pre-planned trajectory, on constant altitude, where it had to pass over some specific GPS coordinates using its inner flight planner set through QRGround Flight Control, as shown in Figure 13.

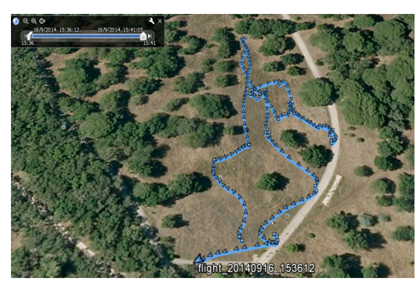

Fig. 13: Drone flight path in Monsanto, Lisboa.

The full data set recorded has 8340 samples, with a rate of 30 samples per second. However, with this rate, the variation between a pre-action and a post-action image was practically unnoticeable. This considered, the training samples were cut to 556 with a time difference between two

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2638885, IEEE Transactions on Cognitive and Developmental Systems

consecutive images of 0.5 seconds (2 samples per second). The retina was trained using 556 data triplets, $(\mathbf{i}_0, \mathbf{i}_1, \mathbf{q})$, with 95 different action identifiers. The original HD color images were converted to grayscale and reduced to 15x15 pixels through bilinear subsampling.

In this experiment a motor space with 4 degrees of freedom is considered. Differently form the simulated experiments, where actions were atomic and exact, here the motor space spans a continuous domain and must be quantized. Each degree of freedom was separately quantized in 4 bins, using the k-means clustering algorithm [23]). These were concatenated and then, to each unique combination of the concatenated vectors a specific action identifier $\mathbf{q}$ is assigned.

In Figure 14 three examples of visual stimuli prediction are shown, using two different complexities of sensor structure: one with 9 sensor receptive fields and another with 16. As

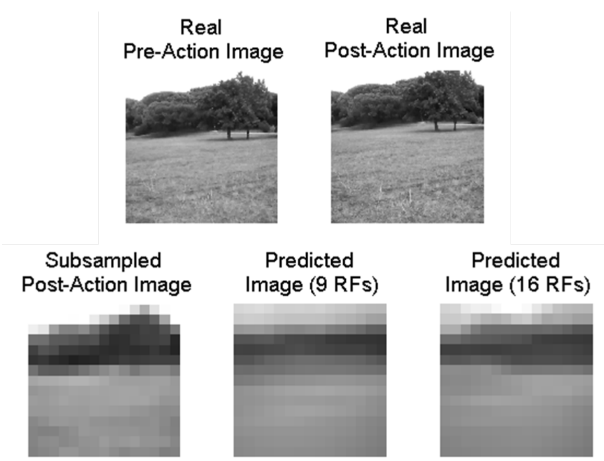

Fig. 14: Visual stimuli prediction using two different Sensor complexities (9 and 16 sensor receptive fields).

we can observe, the reconstruction is slightly better using the more complex retina. In Figure 15 are shown both sensor organization topologies and respective RMSE. Interestingly, the retina develops horizontal elongated receptive fields, reflecting both the structure of the environment (clear contrast between sky and land) and the structure of the motor system (dominated by motions inducing camera translations). The area with lower error corresponds to ground which occupies the bottom half of the field of view. Becasue the ground does not present significant texture, bigger receptive fields are developed in this region. Looking at the top half of the drone's field of view, it can be seen that a greater variability exists due to vegetation, originating a denser distribution of sensor receptive fields.

The presented sensorimotor model has a small dimension but was able demonstrate adaptation to challenging real-world scenario and stimulus prediction skills. However, if this model is to be used in a certain task, it may required a higher image resolution, which will demand a higher number of sensor receptive fields. For this purpose, it is essential to research more scalable and efficient mechanisms for sensorimotor optimization.

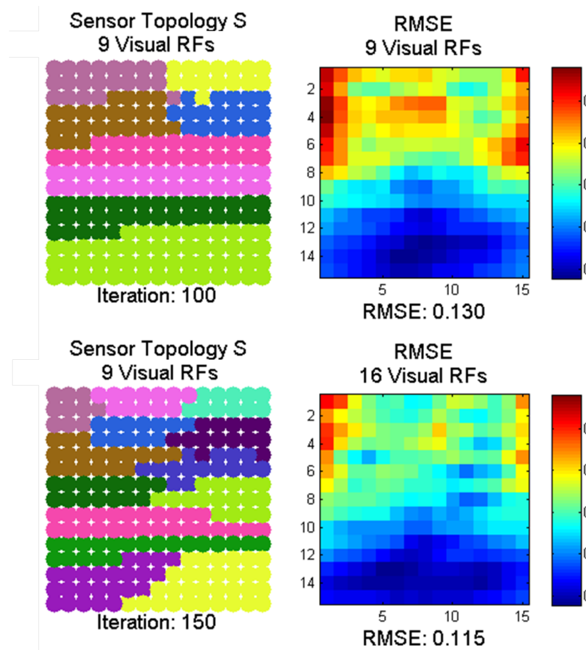

Fig. 15: Sensor organization topologies after training, and respective prediction error map (RMSE).

V. 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. The development of sensor 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 sensory and a motor system, as well as obtaining more reliable predictions of what an agent is expecting to see given its actions.

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 feed-forward neural networks and the Sensorimotor Network, suggests that the latter might prove useful in bringing computers a step closer to biological performance. At the same time, the sensorimotor approach presents a tight relationship between its structures and shows that by changing each sensor or motor configuration or even the agents environment, the system will successfully adapt and develop efficient topologies for visual stimuli prediction, even with real data (quadricopter) and different motor representations. This image processing capability makes such a system a good candidate for tasks such as anomaly detection or tracking.

Currently we are developing new approximate solvers for the proposed sensorimotor optimization problem. Up to now we have been able to tackle a small number of sensor and motor fields but new models will be able to address larger problems by an order of magnitude. Also of relevance for future work is the use Deep Learning techniques, which will allow sequential training of many layers and possibly address even larger problems. Finally, an online optimization algorithm would allow an easier adaptation of robots to dynamically changing 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 Project 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] J. Ruesch, "A computational approach on the co-development of visual sensorimotor structures," Ph.D. dissertation, Instituto Superior Tcnico, 2014.

[6] R. Santos, R. Ferreira, A. Cardoso, and A. Bernardino, "Sensorimotor networks vs neural networks for visual stimulus prediction," in *Development and Learning and Epigenetic Robotics (ICDL-Epirob), 2014 Joint IEEE International Conferences on*, Oct 2014, pp. 287–292.

[7] J. Ruesch, R. Ferreira, and A. Bernardino, "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.

[8] ——, "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.

[9] 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.

[10] 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.

[11] N. A. Mirza, C. L. Nehaniv, K. Dautenhahn, and R. te Boekhorst, "Anticipating Future Experience using Grounded Sensorimotor Informational Relationships," in *Proc. of Eleventh International Conference on the Simulation and Synthesis of Living Systems, Artificial Life XI, 2008.* , 2008, pp. 412 – 419.

[12] 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.

[13] 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.

[14] 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

[15] Y. Bengio, "Learning deep architectures for ai," *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[16] Y. Yamashita and Y. Tani, "Emergence of functional hierarchy in a multiple timescale neural network model: A humanoid robot experiment," *PLoS Computational Biology*, vol. 4, no. 11, pp. 1–18, 2008.

[17] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural Networks*, vol. 4, no. 2, pp. 251–257, 1991.

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

[19] R. Gregory, H. E. Ross, and N. Moray, "The curious eye of copilia," *Nature*, vol. 201, no. 4925, pp. 1166–1168, 1964.

[20] M. Land, "Movements of the retinae of jumping spiders (salticidae: Dendryphantinae) in response to visual stimuli," *Journal of experimental biology*, vol. 51, no. 2, pp. 471–493, 1969.

[21] J. Stone and P. Halasz, "Topography of the retina in the elephant loxodonta africana," *Brain, behavior and evolution*, vol. 34, no. 2, pp. 84–95, 1989.

[22] J. J. Atick and A. N. Redlich, "What does the retina know about natural scenes?" *Neural computation*, vol. 4, no. 2, pp. 196–210, 1992.

[23] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Applied statistics*, pp. 100–108, 1979.