Sensorimotor and Neural Networks for Visual Stimulus Prediction

Ricardo Manuel Raposo dos Santos
Institute for Systems and Robotics
Instituto Superior Técnico,
Lisbon, Portugal
Email: rsantos@isr.ist.utl.pt

***Abstract*—This work focuses on a recently developed special type of biologically inspired architecture, which here denoted as a Sensorimotor Network, able to co-develop sensorimotor structures directly from the 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 (visual and movement receptive fields) adapted to the robot and its environment. Here Sensorimotor Network is compared with a Multilayer Perceptron in the ability to create efficient predictors of visuomotor relationships. It is confirmed that the Sensorimotor Network is significantly more efficient in terms of required computations and is more precise (less prediction error) than a simple feedforward neural network in predicting self-induced visual stimuli. The sensorimotor approach is proved to be replicable in a modified neural network and capable of organizing the same meaningful structures. In addition, all sensorimotor structures are studied regarding their complexity, connectivity and visual perception influence. Finally, a sensorimotor model is trained using real data recorded during a quadricopter drone flight.**

***Index Terms*—Stimuli prediction, neural networks, sensorimotor network, visual receptive fields, motor 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 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 sensorimotor 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.

This thesis compares the model proposed in [4] with Multilayer Perceptron (MLP) with a single hidden layer. For sensorimotor prediction, it is shown that a network with a specific structure can attain significant advantages over fully connected networks. This work claims that co-developed structures yield better sensory predictions for the effects of actions, relatively to a more naive and straightforward approach which lacks a sensorimotor 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, a priori 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 [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 sensorimotor 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 two main principles are followed: 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.

In the 80s researchers started to realize neural networks potential when supported by the growing interest in human cognition for Artificial Intelligence applications and the rapid increase of computer processing power. At that time many works based on these networks were issued with special relevance to Fukushima's work at digit recognition [12] and Sejnowski's work at teaching a network to pronounce English written words [13]. Although, neural networks are still not good enough to compete with a brain, they already give computers the capability of learning by example and are effectively used in applications of object recognition, patterns recognition or data classification.

This thesis proposes the recently developed adaptive model of Sensorimotor Network [14] as a path to follow for better image processing and development of retina-like structures. The authors approach considers the interconnection between different areas of the brain (namely the visual and motor areas) and its adaptive properties that optimize the sensor, motor and predictive structures to the agents morphology and environment characteristics.This sensorimotor structures are presented as a biological equivalent of a Corollary Discharge Circuit (CDC), where motor signals emitted in the deep layers of the superior colliculus, are integrated and fed in the frontal eye field visual receptors [15].

III. Sensorimotor Prediction Architectures

An agent is considered 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, the light field is represented 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 sensory 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 two possible architectures are considered 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. Multilayer Perceptron

In this case a feed-forward linear network, equivalent to a Multillayer Perceptron (MLP), with n_s elements in its hidden layer emulate 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

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

which is represented in Figure **??**. 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. In Figure 1, the used feedforward neural network is represented.

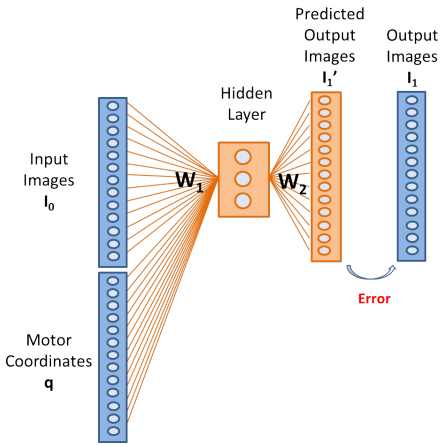

Fig. 1: Multilayer Perceptron: schematic diagram representing the total data triplets ($\mathbf{I}_0$,$\mathbf{I}_1$,$\mathbf{q}$) used to train the model (blue) and the trained parameters $(\mathbf{W}_1^*, \mathbf{W}_2^*)$ and stimuli prediction $\mathbf{I'}_1$ (orange). The hidden layer can have different activation functions (linear or non-linear).

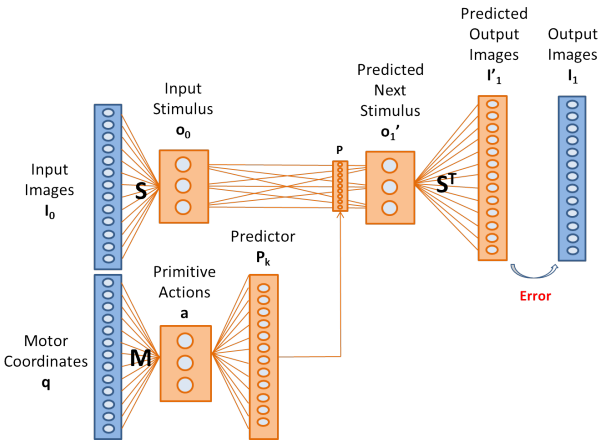

Fig. 2: Sensorimotor Network: schematic diagram representing data triplets ($\mathbf{I}_0$,$\mathbf{I}_1$,$\mathbf{q}$) used to train the model (blue) and the trained parameters $(\mathbf{P}^*, \mathbf{M}^*, \mathbf{S}^*)$ and stimuli prediction $\mathbf{I'}_1$ (orange).

B. Sensorimotor Network

The visual prediction system described in [4], models 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 2D 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 in 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], results in the optimization problem

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

and is represented in Figure 2. Unlike in the MLP 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.

C. Sensorimotor Neural Network

Additionally, the sensorimotor architecture is implemented in a neural network built using Neural Network toolbox from Matlab. After creating an unexisting multiplication block, Figure 3, by using matrix manipulation, the dot product function from Matlab and fixed weight matrices $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$, it was possible to train a modified neural network capable of evolving similar organized sensor and motor topologies as those proposed in [4].

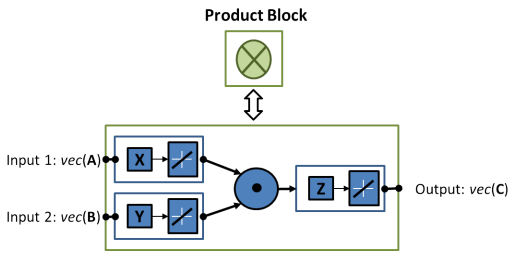

Fig. 3: Multiplication block using Matlab's Neural Network Toolbox blocks.

The full block diagram of the developed Sensorimotor Neural Network is shown in Figure 4. Training of such a model raises some issues. Matlab's Neural Network Toolbox has also some limitations when trying to impose some constraints in the weight matrices. Constraints like positivity (negative values being projected to 0) and normalization (applied to $\mathbf{S}$ and $\mathbf{M}$) have to be computed after each neural network training iteration. This implies that the sensorimotor neural network training has to be interrupted every time the projection and normalization is made. As in Sensorimotor network approach, ($\mathbf{P}$,$\mathbf{M}$,$\mathbf{S}$) are sequentially trained. Altogether, the developed neural network counts with 4 dynamic layers, whose weights are effectively trained, and 3 static layers from the added multiplication block.

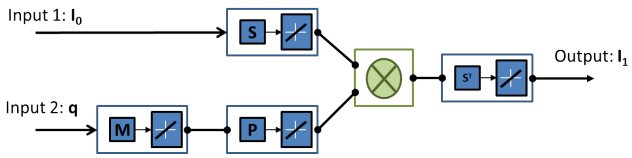

Fig. 4: Block diagram of Sensorimotor Network approach using Matlab's Neural Network Toolbox.

IV. Experimental Setup

A. Models Comparison

To compare the proposed biologically inspired Sensorimotor Network architecture, here on called SNet, with the Multilayer Perceptron, addressed as MLP, two experiments are designed. In the first experiment, ExpXY, the motor space spans actions leading to translations in the image plane, whereas in the second experiment, ExpRZ, actions leading to centered rotations and zooms in the image are used. 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, 10 runs for each training algorithm are performed. Each run is composed of a training batch of 8100 triplets, uniformly sampled from a discrete set of $N_m = 81$ canonical actions (100 triplets of each canonical action). A validation set with half the samples is used as stopping criteria and a test set with the same number of samples is used for prediction errors comparison. 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 experiments ExpXY and ExpRZ, 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. 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 Figure 5.

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 MLP $(\mathbf{W}_1, \mathbf{W}_2)$. 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

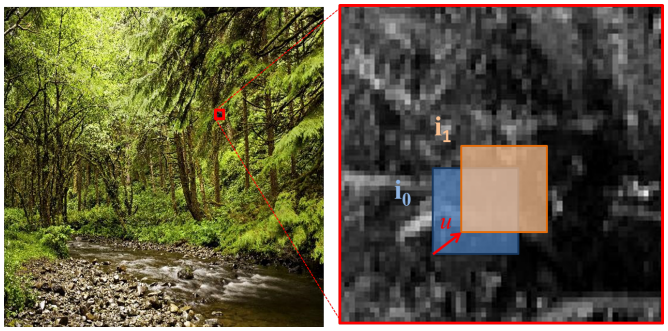

Fig. 5: Triplet acquisition process. In the left we show the full environment image. In the right we show a portion of the environment where the agent is placed to acquire the pre-action 15×15 pixel image, i_0, then transformed by action u, and acquire the post-action image, i_1 (Best seen in color).

a sensor structure composed by 9 receptive fields ($n_s = 9$), which is compared with a linear (*purelin*) and a non-linear (*tansig*) hidden layer of 9 neurons for the MLP model. The number of receptive/movement fields and hidden units can be chosen taking into account the resources available in the particular hardware used to deploy the system. In these experiments an identical number of sensor and motor fields is used but these numbers may be different. For instance, higher image resolution should be followed by higher number of visual receptive fields.

The optimization problem for the SNet showed in Eq. (3) is iteratively improved using a projected gradient descent method [16] within the sequential optimization of **P**, **M**, **S**, and the input triplets are considered in batches as in [4]. For SNet and MLP, while performing the optimization, the RMSE between the predicted and the expected images is computed,

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

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 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,

$$AIC = 2k - 2\log(L) \tag{5}$$

$$BIC = k\log(n) - 2\log(L) \tag{6}$$

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

$$L = \exp^{-\lambda \text{RMSE}^2} \tag{7}$$

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

B. Sensorimotor Properties

Here, additional experiments were performed to evaluate the sensor complexity, ExpSensor, and environment influence on an agents visual perception, ExpEnvironment. In addition, a following experience using a Parrot AR.Drone 2.0 was performed in order to train sensorimotor structures regarding real motor and sensor data.

In the first experiment, ExpSensor, using the described action space used in ExpXY with actions ranging $\mathbf{u} = \{-4:1:4\} \times \{-4:1:4\}$. Maintaining the same motor structure ($n_m = 9$) and static forest environment, the model was trained with three different number of visual receptive fields ($n_s1 = 9$,$n_s1 = 16$,$n_s3 = 25$) in the sensor structure.

In the experience checking the environment influence, ExpEnvironment, for different environments where used: 3 distinct images with artificial patterns and 1 image of forest mimicking a natural environment. Each environment was tested with two combinations of actions as in ExpXY and ExpRZ ($\mathbf{u} = \{-4:1:4\} \times \{-4:1:4\}$ and $\mathbf{u} = \{-100°:25°:100°\} \times \{0.80:0.05:1.20\}$, respectively). The sensor structure was composed by 9 visual receptive fields and the motor structure by 9 movement fields. Both ExpSensor and ExpEnvironment were trained using 100 data triplets per action for a total of 81 actions.

The last experiment was performed using visual and motor information acquired by a drone during its flight, which was used to train a sensorimotor structure with 9 fields in each motor and sensor layer. The used data is described with the results.

V. Results

In this section it is shown the results obtained from the optimization of the two models under comparison (Sensorimotor Network *vs* Multi-Layer Perceptron), using the methods and experimental setup described in the previous sections. The results from sensorimotor experiments and its training using real data are presented.

A. Statistical Models Comparison (ExpXY, ExpRZ)

After convergence of training on the 10 runs for each of sensorimotor and MLP methods (with the MLP hidden layer using two different activation function: linear and non-linear) several statistics are computed in order to evaluate and compare their performance. It was observed 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 MLP, in both experiments, 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 and Table II, 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.

ExpXY	Sensorimotor	MLP	MLP/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
ExpRZ	Sensorimotor	MLP	MLP/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 linear hidden layer MLP in both translation and rotation experiments. The presented values result from the average from all 10 runs. As observed sensorimotor approach uses less parameters, produces a bit less reconstruction error and has less loss of information. The differences are higher between the models in ExpRZ.

ExpXY	SNet	Non-Linear MLP	MLP/SNet
All Parameters	3483	5013	1,44
Parameters $\neq 0$	1140	5013	4,40
Parameters $\geq 10^{-3}$	803	4992	6,22
RMSE	0.1004	0.1241	1,24
AIC	2.654	10.026	3,78
BIC	10.628	45.115	4,24
ExpRZ	SNet	Non-Linear MLP	MLP/SNet
All Parameters	3483	5013	1,44
Parameters $\neq 0$	1053	5013	4,76
Parameters $\geq 10^{-3}$	743	4993	6,72
RMSE	0.0955	0.1233	1,29
AIC	2.442	10.026	4,11
BIC	9.817	45.115	4,60

TABLE II: Comparison between SNet and non-linear hidden layer MLP in both translation and rotation experiments. The presented values result from the average from all 10 runs. Non-Linear MLP uses about the same number of parameters and has a RMSE somewhat higher than Linear MLP.

In Figure 6 the reconstruction RMSE at each pixel of the retina is 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

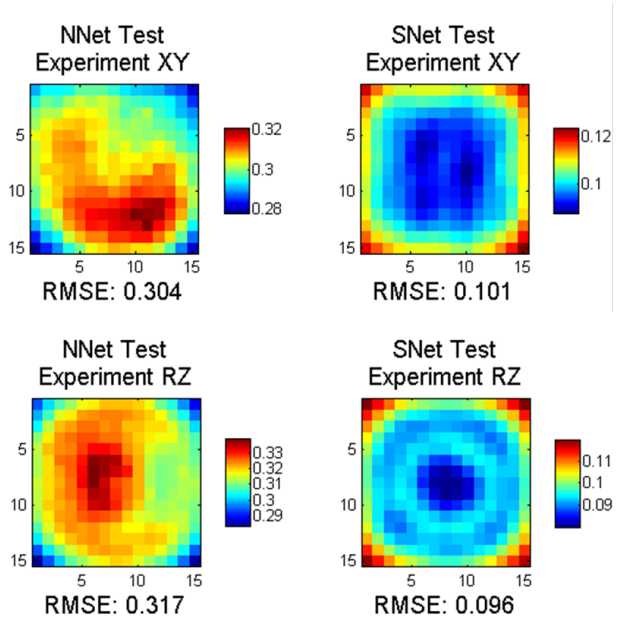

Fig. 6: Comparison between both methods regarding RMSE per pixel for reconstruction in a test set. (Top) Experiment XY run. (Bottom) Experiment RZ run (Best seen in color).

to the boundaries. Anyway, this fact is exacerbated in the multilayer perceptron on ExpRZ 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.

B. Sensorimotor Topology (ExpXY, ExpRZ)

Here the emergent properties [4] with respect to SNet organization (optimization problem in Eq. (3)) are revisited. 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 receptive fields (rows of **S**) organize into a regular structure (after 500 iterations) starting from a random initialization (see Fig. 7). 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 Figure 8 it can be observed the evolution of the motor fields (columns of **M**) for both experiments. ExpXY has its action space uniformly sampled by pixels, producing a near uniform organization of the motor fields. The performed zooms in ExpRZ 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 movement fields organization.

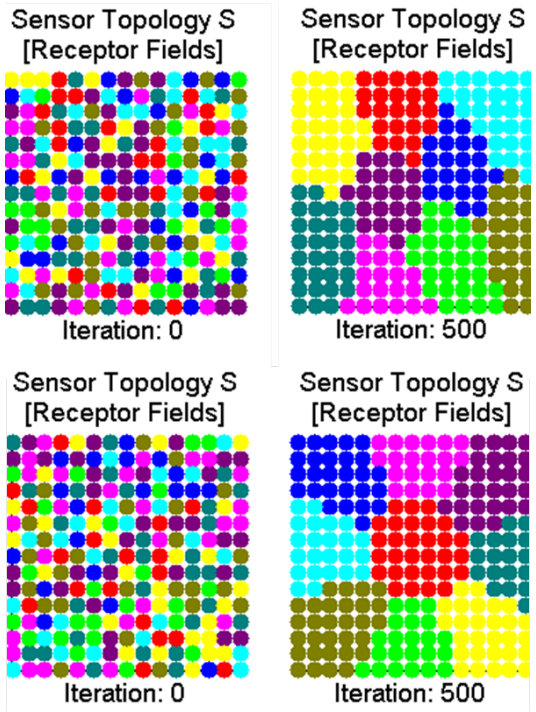

Fig. 7: Sensor RFs initialization and final organization after 500 iterations in one of the runs of ExpXY (Top) and ExpRZ (Bottom) (Best seen in color).

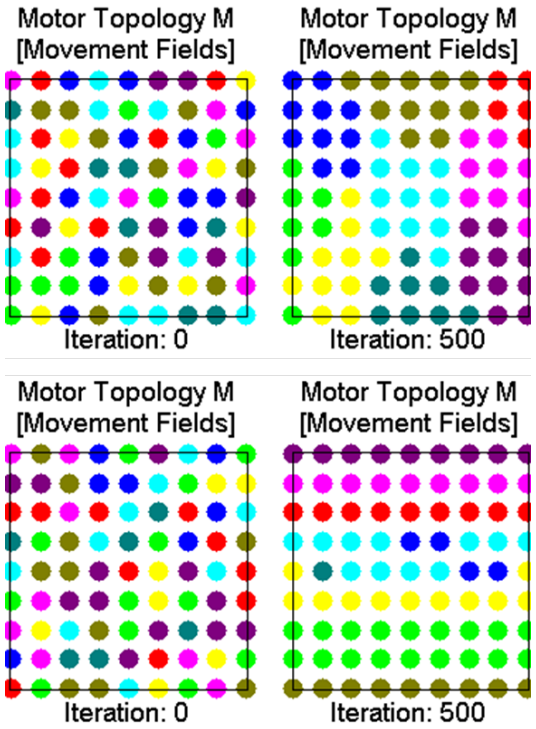

Fig. 8: Motor MRFs initialization and final organization after 500 iterations in a run of ExpXY (Top) and ExpRZ (Bottom) (Best seen in color).

C. Stimulus Predictions (ExpXY, ExpRZ)

After training the Sensorimotor Network, we can use it for making stimulus prediction of the agent's actions. Giving a certain planned motor action $\mathbf{q}$ it can be computed (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 $\mathbf{i'_1} = \mathbf{S^T o_1}$.

In Figure 9 steps (i), (ii) and (iii) for the translational action $\mathbf{u} = (4, 4)$ and the rotation/zoom action $\mathbf{u} = (50°, 1.0)$ are graphically illustrated. On the left, the resulting predictor $\mathbf{P}^k$ for the activated action is represented. On the middle the location of the motor movement fields and their activation (the shade of gray) for that specific action is shown. Finally, on the right it can observed the visual receptive fields distribution together with arrows representing the main directions of flow of the prediction result $\mathbf{P}^k$. The predictor translates motor effects on 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 weights proportional to the arrows gray level.

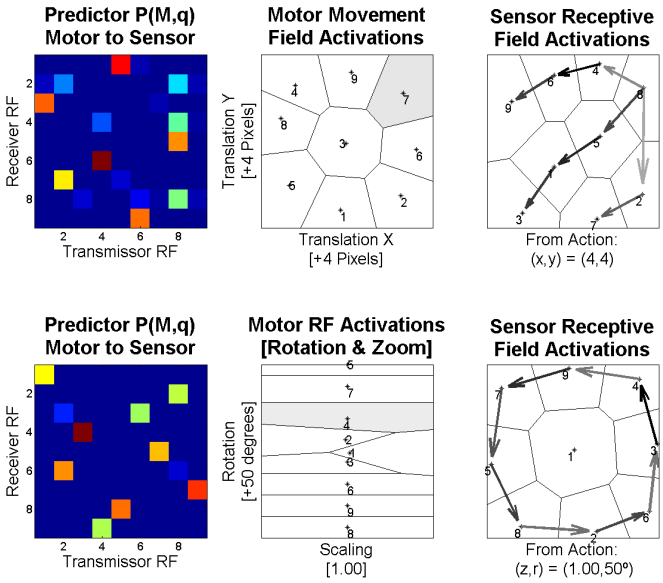

Fig. 9: (Left) Predictive structure $\mathbf{P}^k$. (Mid) Motor RF activations corresponding to particular actions. (Right) Induced prediction field in the sensory space. (Top) Action $\mathbf{u} = (4, 4)$ on the translation network (Bottom) Action $\mathbf{u} = (50°, 1.0)$ in the rotation/zoom network. 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.

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, it can be concluded 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 (visual receptive fields).

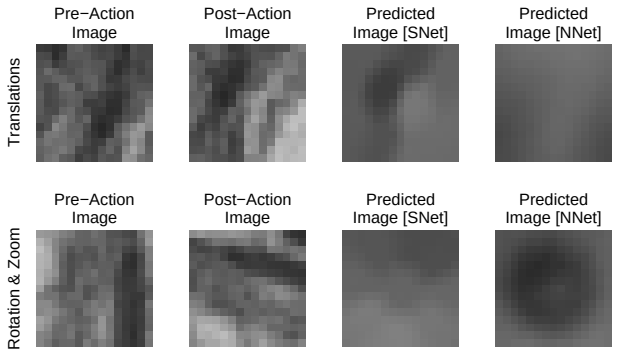

Fig. 10: Real and predicted image examples for the respective actions: (Top) Translation action example: $\mathbf{u} = (4, -4)$ and (Bottom) Rotation and zoom action example: $\mathbf{u} = (-75°, 1.20)$, using both SNet and MLP methods. As shown, reconstructions obtained by SNet optimization show a more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

D. Sensor Complexity Influence (ExpSensor)

Looking for the direct influence of sensor structure complexity on the stimuli prediction, some tests were made using experiment ExpSensor data set and model configuration. Three different models were trained, all in the same conditions, using the same action space, but with different number of available visual receptive fields in the sensory structure (9, 16 and 25). In Figure 11 it can be observed the organized sensor topologies from the three different sensor complexity models. Besides, the reconstruction error **RMSE** was computed using a test set with the same size of the training set (8100 triples). Below a reconstruction example is shown where for the same action and pre-action image, the visual stimuli prediction is computed and compared with the actual observed post-action visual stimulus.

As expected and observed, the quality of the reconstruction improves with the number of visual receptive fields. Although the prediction error decreases with the number of available receptive fields, this also presents an increasing relative number of empty receptive fields. In the case where 9 sensor receptive fields were considered, the model used all of them. However, when reaching higher number of visual receptive fields, some become unnecessary 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: on the one hand the prediction error decreases, but the amount of

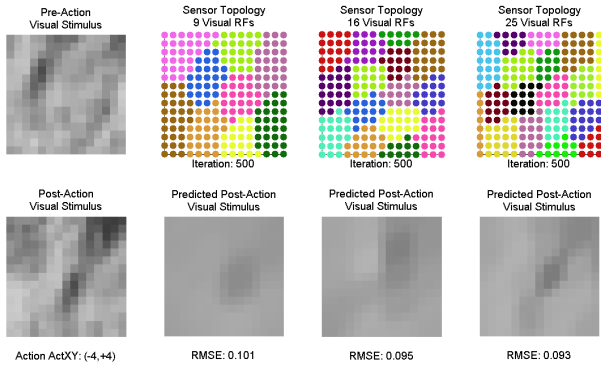

Fig. 11: Sensor Visual Receptive Fields: Influence on prediction error and quality on image reconstruction.

receptive fields consuming computational power, without any advantage for the prediction, increases.

E. Environment Influence (ExpEnvironment)

As proved in many works [17], [18], [19] the eyes, retinas and/or visual systems evolved in many species in very distinctive manners, but all highly efficient when vision appears as the most important sense for the organism. 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 the environment influence on the sensory structure within the sensorimotor system is tested. Using the action spaces mencioned for ExpEnvironment and its data sets four different environments were used for Sensorimotor Network training: 3 artifical (vertical stripes, diagonal stripes and dots) and 1 natural (textured picture of dry dirt). In Figure 12 there are represented the resulting sensor organizations, $\mathbf{S}$, for the 4 environments and the number of iterations until convergence.

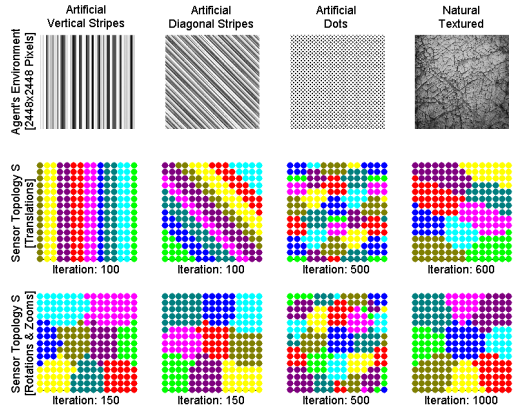

Fig. 12: Environment influence on visual sensor topology. Sequence of visual sensor topologies resulting from training the sensorimotor system using action spaces from ExpXY and ExpRZ, and four different environments: three artificial environments (vertical stripes, diagonal stripes and dots) and one natural environment (textured picture of dry dirt).

From the presented results it can be concluded that the sensor structure organization depends on the environment. Considering the question made in [20] and the tested sensorimotor system, it can be hypothesized that a retina does acquire knowledge, in its organization, about the natural scenes (environment). However, it is shown that the way the agent perceives its environment is the key factor for the resulting visual sensor topology. Even with very different topologies between environments, it can be observed that only by changing the set of movements the agent can perform, the way the same environment is perceived also changes.

Taking as an example, the artificial environment composed by vertical stripes, if an agent performs only translational movements parallel to the environment, the unique type of stimuli the agent will know corresponds to vertical stripes, then the most efficient retina it could develop should be one which translates the possible changes in the perceived stimuli (horizontal movement of the vertical stripes).

From another point of view, if the agent is only able to perform rotational and scaling movements, then the visual stimuli can change from vertical stripes to diagonal or even horizontal stripes. With such a variation of stimuli, it is expected that the retina topology should be different.

F. Sensorimotor Network with Real Data (ExpDrone)

A Parrot AR.Drone2.0 was used to acquire images from a natural environment in Monsanto park in Lisbon. This drone is equipped with a fixed front HD camera which during the experiment was always pointing to its movement direction. During the flight a video was recorded at a rate of 30 frames per second, together with drone position variations (Δx,Δy) from GPS, orientation variations ($\Delta\theta$) and absolute altitude (which corresponds to the state of the drone). For the sake of simplicity and since state is not explicitly modeled in this work, data from drone taking off or landing was removed and the altitude was admitted as constant.

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 locations defined by GPS coordinates using its inner flight planner set through QRGround Flight Control. Examples of acquired images, and respective bilinear subsampling to 15x15 pixels images for training, can be seen in Figure 13.

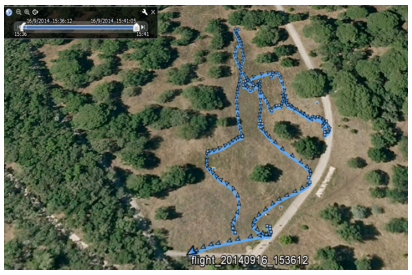

Fig. 13: Drone flight path in Monsanto, Lisboa.

The full data set recorded has 8340 samples, but with a rate of 30 recorded samples per second, 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 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.

Differently from the direct application of the Sensorimotor Network used in [14] where the action space discretizes a two dimensional motor space, in this experiment a motor space with 4 degrees of freedom is considered. Each degree of freedom was separately quantized in 4 bins, using k-means clustering algorithm [21]). 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 visual receptive fields and another with 16.

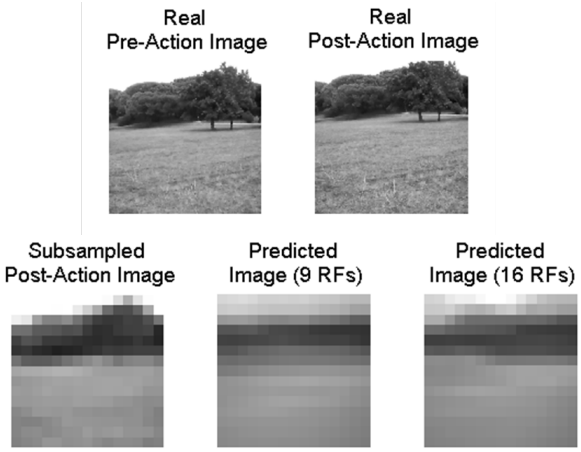

Fig. 14: Visual stimuli prediction using two different Sensor complexities (9 and 16 visual receptive fields).

As observable, and expected from previous results, the reconstruction is slightly better using the more complex retina. Above, in Figure 15 it is show both sensor organization topologies and respective RMSE. The area with lower error corresponds to ground which occupies the bottom half of the field of view with some deviations. During its flight, the ground suffers some vertical movements (bigger and more horizontal receptive fields). Looking at the top half of the drone's field of view, it can be seen that a greater variability exists, originating a denser distribution of visual receptive fields.

The tested sensorimotor is used with structures complex enough to successfully demonstrate its applicability and great predictive skills. However, if this model is to be used in a certain task, it can be required that training images become larger and the number of visual receptive fields and/or motor movement fields increase considerably. This would need a bigger time for training the model.

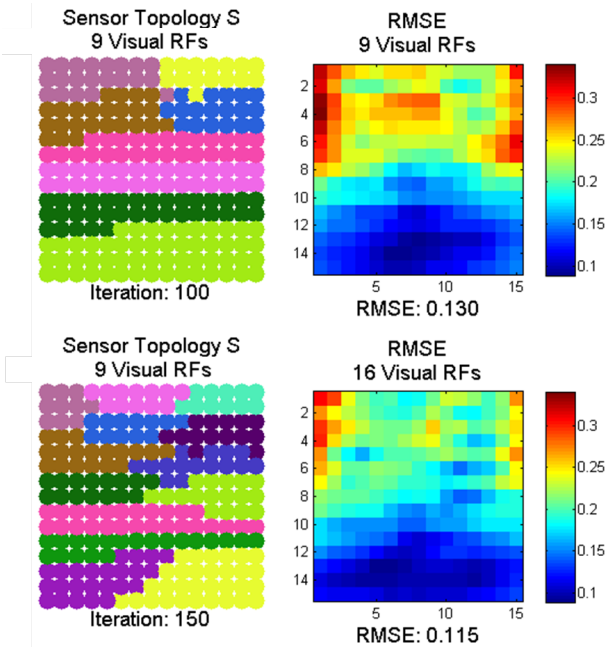

Fig. 15: Sensors organization topologies after training and respective prediction error, RMSE.

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 biologically inspired 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 when compared with a Multilayer Perceptron.

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 feed-forward neural networks and 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 to develop efficient topologies for visual stimuli prediction, even

with real data (as the one used in ExpDrone) with different motor representations. This image processing capability makes such a system a good candidate for tasks deployment such as anomaly detection or tracking.

Considering that it was possible to deploy a sensorimotor structure using modified neural networks, it could be important to follow the path of developing such system using a more state-of-the-art machine learning method such as Deep Learning which allows sequential training of many layers. 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] K. Fukushima, "Neocognitron: A hierarchical neural network capable of visual pattern recognition," *Neural networks*, vol. 1, no. 2, pp. 119–130, 1988.

[13] T. J. Sejnowski and C. R. Rosenberg, "Parallel networks that learn to pronounce english text," *Complex systems*, vol. 1, no. 1, pp. 145–168, 1987.

[14] J. Ruesch, "A computational approach on the co-development of visual sensorimotor structures," Ph.D. dissertation, Instituto Superior Tcnico, 2014.

[15] T. B. Crapse and M. A. Sommer, "Corollary discharge across the animal kingdom," *Nature Reviews Neuroscience*, vol. 9, no. 8, pp. 587–600, 2008.

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

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

[18] 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.

[19] 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.

[20] 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.

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