Self-Organization of Visual Sensor Topologies Based on Spatiotemporal Cross-Correlation

Jonas Ruesch, Ricardo Ferreira, and Alexandre Bernardino

Instituto Superior Técnico, 1049-001 Lisbon, Portugal,
jruesch,ricardo,alex@isr.ist.utl.pt

Abstract. In living organisms, the morphology of sensory organs and the behavior of a sensor's host are strongly tied together. For visual organs, this interrelationship is heavily influenced by the spatial topology of the sensor and how it is moved with respect to an organism's environment. Here we present a computational approach to the organization of spatial layouts of visual sensors according to given sensor-environment interaction patterns. We propose that prediction and spatiotemporal correlation are key principles for the development of visual sensors well-adapted to an agent's interaction with its environment. This proposition is first motivated by studying the interdependency of morphology and behavior of a number of visual systems in nature. Subsequently, we encode the characteristics observed in living organisms by formulating an optimization problem which maximizes the average spatiotemporal correlation between actual and predicted stimuli. We demonstrate that the proposed formulation leads to spatial self-organization of visual receptive fields, and leads to different sensor topologies according to different sensor displacement patterns. The obtained results demonstrate the explanatory power of our approach with respect to i) the development of spatially coherent light receptive fields on a visual sensor surface, and ii) the particular topological organization of receptive fields depending on sensorimotor activity.

Keywords: visual sensor topology, self-organization, sensorimotor coupling

1 Introduction

By simply observing the active behavior and visual organs of different animal species, important hints can be obtained on how an organism constructs visual percepts. Primates use a sophisticated oculomotor system to sequentially move and stabilize their eyes with relation to different target locations [1]. Most airborne insects on the other hand, have their eyes rigidly attached to their body or head; instead of focusing on particular target locations, these animals analyze how the projection of the environment translates on their sensors during flight [2]. In general, three interrelated aspects contribute to how biological vision systems record raw visual stimuli: i) the characteristics of the environment in which an animal is living, ii) the way a sensor is moved with respect to the environment, and iii) the physical and morphological design of a visual organ.

In this work, we consider i) to be a general environment and we investigate a possible principle how ii) influences iii).

A closer look at the morphology of biological visual sensors reveals profound differences between different organisms. While all visual organs found in nature record visual stimuli through a number of light sensitive receptors – and hence always record a spatially discretized stimulus – the spatial density distribution of visual receptors varies greatly between species. Studies measuring the distribution of retinal ganglion cells in camera-type eyes, or the ommatidia distribution in compound eyes, suggest that receptor distributions are directly tied to an animal's behavior and environment. Most prominently, primates and other mammalians with binocular vision feature a fovea – a small, high-resolution area in the center of their retina – and a radially close to logarithmically decreasing receptor density. In [3], it is pointed out that such a log-polar-like receptor distribution corresponds to a mapping function which transforms image rotations and dilations (zoom) into simple coordinate shifts in the log-polar coordinate system. Thus, if an eye featuring such a receptor distribution is focusing on an object and that object is rotated or scaled, the projected image is merely shifted along the log-polar coordinate axes. It was argued that this property results in an advantage for the human visual cortex, as it could achieve image invariance for these transformations at a low computational cost by simply shifting the image. Similar to ganglion cell distributions found in camera-type eyes, the density of ommatidia in arthropods varies significantly over the spatial extension of their compound eye. Many flying insects for example have about a two times higher spatial resolution in the frontal visual eye field than compared to the lateral part [4]. A possible advantage of such a distribution is discussed in [5]. There, it is demonstrated that high density of light recording receptors in frontal and caudal regions, and decreasing density in lateral regions, leads to a uniform translation of projected stimuli on the eye during straight locomotion and can facilitate visual distance estimation.

Motivated by observations related to the relationship of behavior and morphology in natural visual systems, we explore in this paper the hypothesis that visual organs develop such as to simplify neural circuitry for predicting on average experienced stimulus flow patterns. We first propose a criterion based on spatiotemporal cross-correlation to evaluate such a receptor-to-receptor flow property, and we subsequently use the introduced criterion as a cost function to synthesize visual sensor topologies on a given sensor surface using a given set of stimulus transformations. The obtained results suggest that the introduced criterion is able to capture important properties of the relationship between the spatial layout of a visual sensor and the way the sensor is moved with respect to the environment.

1.1 Related Work

In an inventive work [6], Clippingdale and Wilson present a numerical experiment motivated by the spatial organization of visual sensors in nature. Using an abstract setup where visual receptors are represented as a set of points on a disk,

an appealing principle is motivated on how to capture the relationship between form and behavior. In line with our observations for natural visual systems, the basic idea is a rule capable of generating sensor layouts which simplify stimulus transformation patterns under a given behavior: assuming the given points are transformed by a set of sensor displacement actions, the relative position of each point is updated such as to reduce the overall motion-prediction error between points. Interestingly, this update rule leads to foveal point distributions when considering stimulus transformations plausible e.g. for the mammalian visual system. Furthermore, using different action probability distributions for horizontal and vertical translations, elliptic (visual streak-like) point layouts can be obtained. For an illustration see Figure 10 in [6]. Formally, Clippingdale and Wilson proved the following: a set of points randomly distributed on a disk converges to a stable configuration given: i) points are conjointly transformed by rotations, dilations and translations which are applied according to a given probability distribution; and ii) after a transformation action is applied, each point is moved towards transformed points which are lying closest to the point under consideration. It was shown, the final point distribution is the configuration where each point has on average the smallest possible distance to the next closest transformed point under the given action probability distribution. This approach is based on two important assumptions: visual receptors have no spatial extension (i.e. are points), and the error between original and transformed receptors can be measured as an Euclidean distance between spatial locations of receptors. The first assumption is clearly an abstraction of a real visual sensor. The second assumption can be further divided into two requirements: the spatial layout of the visual sensor is known to the algorithm, and the prediction error of visual stimuli is directly related to spatial distance. While it is arguable if an agent can have complete knowledge of the spatial layout of its sensor, the assumption that the prediction error is equivalent to spatial distance is unlikely to hold for spatially extended visual receptors of different sizes.

Related to the question of how the distance measure underlying the optimization proposed by Clippingdale and Wilson could be translated to real visual sensors, the authors of this paper investigated in previous work how the interrelationship between form and behavior could be quantified for sensors with spatially extended receptors and unknown topologies [7]. Based on the complexity of the model required to predict stimulus changes, a measure was introduced which evaluates the coupling between sensor displacements and sensor topologies. It has been shown that a given sensor topology implicitly defines actions for which future sensory stimuli can be predicted with less parameters. In this work we use a similar strategy to optimize the coupling between a sensor's topology and executed motor actions.

1.2 Contribution

We develop a computational method for synthesizing visual sensor topologies according to on average experienced stimulus transformations. To establish a

relation between a sensor's spatial layout and experienced stimulus transformations, we adopt the basic principle proposed in [6]. Though, instead of considering point-like sensor elements, we simulate a realistic visual sensor which records stimuli through receptors where each receptor integrates luminance according to a receptive field. Different from [6], we impose that the algorithm has no access to information about the topological layout of the sensor being organized. This means, the organization of the sensor layout has to be achieved solely by observing the activation of an orderless array of visual receptors. Hence, the implementation of a rule similar to the one proposed in [6] becomes considerably more challenging. In particular, the Euclidean distance measure between transformed and original points has to be replaced with a measure related to how activation is transported between visual receptors when the recorded stimulus changes. We will address this issue by introducing a criterion based on spatiotemporal cross-correlation of receptor activation. This criterion allows us then to implement an optimization which organizes the layout of visual receptors depending on sensorimotor activity. At the same time, we also required the algorithm to find a suitable shape for the receptive fields (RFs) of the spatially extended receptors. We show that spatially coherent RFs can evolve driven only by the low spatial frequency of natural images. By rewarding spatial correlation within RFs, smoothly overlapping clusters organize on the sensor surface without any further constraint on the spatial shape of a receptor's integration area. In practice, receptors can be initialized with a randomly chosen luminance integration function and eventually develop into compact receptive fields.

The following steps summarize the approach followed in this paper:

1. A system with a given sensor surface, a given motor space and a predefined number of visual receptive fields is considered.
2. Each visual receptive field is described as a discretized, randomly initialized function according to which visual input is integrated from the sensor surface.
3. By maximizing spatial correlation of visual stimuli recorded through receptive fields, the development of spatially coherent visual receptors is achieved.
4. By extending spatial correlation to spatiotemporal correlation between visual stimuli of transformed and original receptors, sensor topologies dependent on the agent's motor activity are developed.

2 Approach

An artificial agent with a given sensor surface $\mathcal{I} \subset \mathbb{R}^2$ and a given number of motion degrees of freedom is considered. The sensor surface records a projection of the environment given as a function $i_s : \mathcal{I} \to \mathbb{R}$ defining a luminance value for each point on the surface when the agent is in state s. For numerical purposes, i is sampled at N spatial locations $\mathbf{x}_n$ as a discrete grayscale image $\mathbf{i} = [i(\mathbf{x}_1)\ i(\mathbf{x}_2)\ \ldots\ i(\mathbf{x}_N)]^\top$. The topology of the visual sensor is composed of M visual receptors, where M is a parameter of the proposed method and is much smaller than N. Each visual receptor m integrates a visual stimulus through a receptive field (RF). The RF is described as a vector of weights $\mathbf{r}_m$ defining

how much each entry in $\mathbf{i}$ contributes to receptor m. Note that $\mathbf{r}_m$ is allowed to encode any receptive field function and no spatial coherence is assumed. By assembling weight vectors $\mathbf{r}_m$ for all M visual receptors as the rows of a matrix $\mathbf{R}$, a stimulus recorded by the agent in state s can be written as $\mathbf{R}\mathbf{i}_s$.

After observing state s, the agent can choose to take an action a from a discrete set of actions $\mathcal{A}$ representative of the agent's behavior. This action induces a change in the observed grayscale image from $\mathbf{i}_s^-$ to $\mathbf{i}_s^+$; here we assume that this change is predictable.[1] As the agent explores its environment, we collect before and after images for each particular action a in the matrices $(\mathbf{I}_a^-, \mathbf{I}_a^+)$, where samples are arranged in columns. For the whole set of actions $\mathcal{A}$, these matrices are collected in a dataset $\mathcal{D} = \{(\mathbf{I}_a^-, \mathbf{I}_a^+), a \in \mathcal{A}\}$.

With the introduced terminology, we now proceed to develop an optimization problem which evolves the sensor topology $\mathbf{R}$ such that the previously described properties are induced: i) spatially coherent receptive fields are formed, and ii) the topological layout of the sensor reflects stimulus translations induced by the behavior of the host. We propose to find an optimal $\mathbf{R}$ as the solution to an optimization problem:

$$\mathbf{R}^* = \underset{\mathbf{R}\in\mathcal{R}}{\operatorname{argmax}}[F(\mathcal{D}, \mathbf{R}) - G(\mathbf{R})], \tag{1}$$

where F denotes a function evaluating the spatiotemporal cross-correlation of a set of samples $(\mathbf{I}_a^-, \mathbf{I}_a^+)$, and G represents a cost for growing receptive fields. The constraint set $\mathcal{R}$ is chosen as $\mathcal{R} = \{\mathbf{R} : \mathbf{R} \geq \mathbf{0},\ \mathbf{R}^\top \mathbf{1} = \mathbf{1}\}$, such as to guarantee that the visual receptive fields occupy the whole sensor surface and luminance cannot be subtracted. In the remainder of this section, we unroll the complete definition of this optimization problem by developing F and G.

Consider first an immobile agent with a single null action leading to a reduced data set $\bar{\mathcal{D}} = \{\mathbf{I}^-\}$ of stimuli recorded in different states s. In this case, we consider a reasonable sensor topology $\mathbf{R}$ to be one which leads to high correlation within a batch of recorded stimuli $\mathbf{R}\mathbf{i}_s$. The rational behind this is that bigger differences between simultaneous receptive field activations indicate that the agent is able to pick-up more information from the images $\mathbf{i}_s$, in an information theoretic sense. Furthermore, correlation must be normalized with respect to the size of a receptive field such that different sized receptive fields are comparable. Implementing these two requests, we propose a first version of F for an immobile agent to be a size normalized correlation between stimuli $\mathbf{i}_s$ like:

$$\bar{F}(\bar{\mathcal{D}}, \mathbf{R}) = \sum_{s=1}^{S} \left(\hat{\mathbf{R}}\mathbf{i}_s^-\right)^\top \left(\hat{\mathbf{R}}\mathbf{i}_s^-\right), \qquad \hat{\mathbf{R}} = \frac{\mathbf{R}}{\sqrt{\mathbf{R}\mathbf{1}\mathbf{1}^\top}}, \tag{2}$$

where in $\hat{\mathbf{R}}$ the division and square root operators are applied element wise.

In a second step, an active agent and a full data set $\mathcal{D} = \{(\mathbf{I}_a^-, \mathbf{I}_a^+)\}$ is considered. To establish a temporal relationship between receptive fields, we

[1] See also [8], Appendix A for the constraints posed on such actions and how this situation relates to a physical agent acting in a 3-dimensional world.

now adapt $\bar{F}$ to compute correlation between pre- and post-action stimuli. We remind the reader that it is a priori unknown how to temporally relate receptive fields and how stimuli change under an action a. This is naturally solved by considering a prediction operator which describes a mapping of receptors for a given action, allowing for comparison of stimuli at different points in time. In [9] Crapse and Sommer provide an excellent review of the ubiquity of stimulus prediction in living organisms and [8] gives an argument for the use of linear prediction. Thus, assuming that for an action a we can predict a visual stimulus as $\mathbf{R}\mathbf{i}_a^+ = \mathbf{P}^a_{(\mathbf{R})}\mathbf{R}\mathbf{i}_a^-$ we revise $\bar{F}$ like

$$F\left(\mathcal{D}, \mathbf{R}\right) = \sum_{a\in\mathcal{A}}\sum_{s=1}^{S}\left(\hat{\mathbf{R}}\mathbf{i}^+_{s,a}\right)^\top\left(\mathbf{P}^a_{(\mathbf{R})}\hat{\mathbf{R}}\mathbf{i}^-_{s,a}\right), \qquad \hat{\mathbf{R}} = \frac{\mathbf{R}}{\sqrt{\mathbf{R}\mathbf{1}\mathbf{1}^\top}}, \tag{3}$$

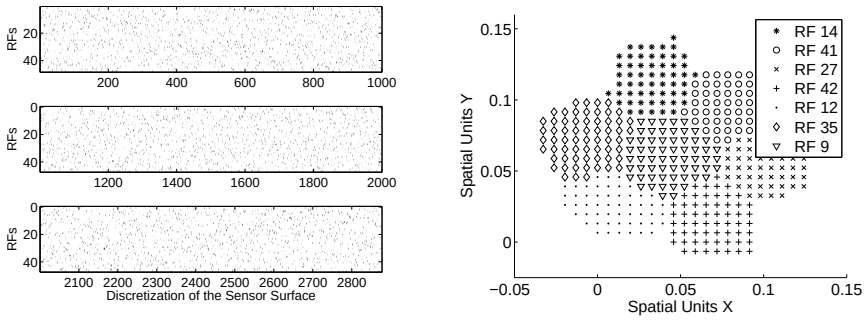
where a prediction operator $\mathbf{P}^a_{(\mathbf{R})}$ is learnt from a batch of samples $(\mathbf{I}_a^-, \mathbf{I}_a^+)$. We request $\mathbf{P}^a_{(\mathbf{R})} \geq \mathbf{0}$ and propose $\mathbf{P}^a_{(\mathbf{R})}$ to be the solution to a positive least squares problem. As demonstrated in [7] this yields a predictor reflecting the complexity of stimulus flow patterns under actions.

Finally $G\left(\mathbf{R}\right)$ is chosen in such a way as to impose a cost on the growth of receptive fields. Choosing $G(\mathbf{R}) = \omega\|\mathbf{R}\|_2^2$ provides control over the smoothness of the receptive field boundaries. For $\omega = 0$ solutions with hard receptive field boundaries are obtained.

3 Method

We consider the sensor surface to be a disk, discretized at $N = 2877$ locations in a grid-like layout, and being organized into $M = 48$ receptive fields. The environment is given as a plane textured by a very high resolution image depicting a real world scene. A state s consists of a position of the sensor surface with respect to this plane. In this paper we assume the sensor surface to be parallel to the plane and each location records luminance over the covered area into discrete grayscale images $\mathbf{i}$. This sensor interacts with the environment through four types of actions, translations in x- and y-directions, rotations and changes in distance to the plane (zoom). An action set $\mathcal{A}$ is obtained by sampling a particular action probability distribution representative of the agent's behavior. For the results presented in this paper each behavior is represented with 60 samples as shown in Fig. 2. For each action a a pair of samples is obtained by positioning the agent in a random state on the environment and taking the chosen action a. This process is repeated 68 ($> M$) times for each a, acquiring the dataset $\mathcal{D} = \{(\mathbf{I}_a^-, \mathbf{I}_a^+)\}$.

To find $\mathbf{R}^*$ we iteratively improve the optimization problem given in Eq. (1) using a projected gradient descent method [10]. At each iteration we learn predictors $\mathbf{P}^a_{(\mathbf{R})}$ that best satisfy $\mathbf{R}\mathbf{i}_a^+ = \mathbf{P}^a_{(\mathbf{R})}\mathbf{R}\mathbf{i}_a^-$ in a positive least squares sense using the optimization method known from [11]. Note that, even though $\mathbf{P}^a_{(\mathbf{R})}$ cannot be obtained as a closed form solution, the gradient needed to iterate Eq. (1)



(a) Converged Matrix $\mathbf{R}^*$ (b) Topology of $\mathbf{R}^*$ on the Sensor Surface

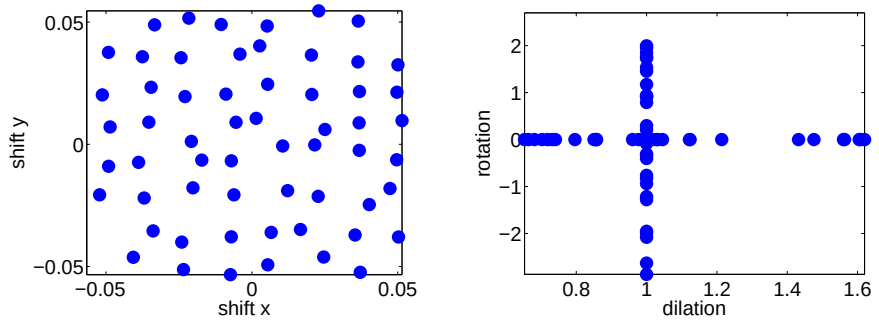
Fig. 1. Emergent clustering of receptive fields (RFs). Left: A converged but topologically orderless matrix $\mathbf{R}$ as seen by the algorithm; each entry specifies the contribution of a location on the sensor surface to a receptive field (RF); the sensor surface is discretized into 2877 pixels (x-axis), and the matrix $\mathbf{R}$ codes for 48 RFs. Right: The sensor surface and the coverage of 7 selected RFs at spatial locations where their contribution is predominant; this view reveals the implicitly present topological clustering in $\mathbf{R}$.

can still be found in closed form by applying the implicit function theorem to the Karush-Kuhn-Tucker optimality conditions of the positive least squares optimization problem [12]. While it is no problem to find a solution for $\mathbf{R}$ with an online method, convergence is much slower, we therefore choose here the batch approach for practical reasons. However, we note that under different circumstances an online implementation might be preferable, e.g. for a purely biologically inspired implementation in a robot with stronger memory constraints and a longer exploration phase.

The experiments presented in Sect. 4 were initialized as follows: the topology of the sensor $\mathbf{R}$ was randomly initialized according to a uniform distribution between zero and one, and then projected to obey the constraints $\mathcal{R}$. The cost for growing receptive fields was kept at a constant level $\omega = 0.3$. It is important to note that with a randomized initialization, nothing prevents the adaptation process from converging to a locally optimal solution. From a biological point of view, we accept these solutions as possible branches of evolutionary development.

4 Results

To demonstrate the correlation principle introduced in Eq.(2) we start by showing the results for an immobile agent. This example, although discarding any meaningful behavior, shows a crucial capability of the proposed method namely the requested property i) the development of spatially coherent light receptive fields on a visual sensor surface. Figure 1 highlights the discovery of topological order from the orderless sampling of the underlying image.



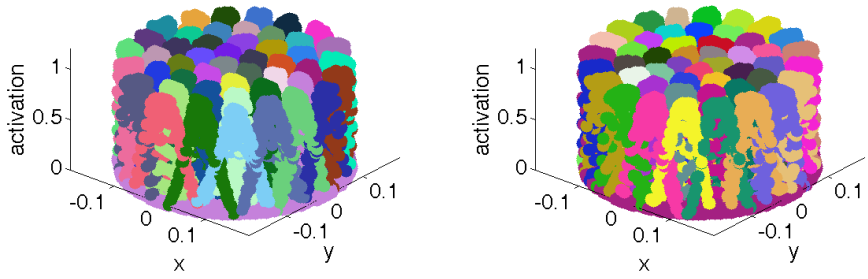
(a) Shift (b) Rotation and Zoom

Fig. 2. Two different behaviors represented as action distributions. Left: uniform 2-dimensional translations in a given range covering 10 times the distance between discrete sampling locations on the sensor surface in each direction. Shift units are normalized with respect to the environment. Right: independent zoom and rotation actions distributed uniformly on each axis. Rotations are given in radians and dilations are given as a scale factor. Both operate with respect to the center of the sensor surface. Zoom actions range from 0.6 to 1.66 and rotations cover $-\pi$ to π.

As external observers we have the privilege of knowing the spatial locations where the sensor surface was sampled and as such we are able to plot the topological ordering of receptive fields on the sensor surface as shown in Fig. 1(b). In the two dimensional visualization we choose to show at each discrete sensor surface location the predominant receptor. The clustering property of the receptive field elements is clearly demonstrated. Since in this case no action is taken, this clustering is a sole consequence of the interaction between the correlation based cost function and the low frequency characteristic of the observed environment. Note that the agent does not have access to the sampling locations of the sensor surface and is thus unaware of the final topological ordering. The proposed algorithm operates solely on matrix $\mathbf{R}$ which is absent of any topological meaning even in the final converged state, as shown in Fig. 1(a).

For active agents we will now consider two different behaviors as shown in Fig. 2(a) and Fig. 2(b). The first consists of a uniform action probability distribution of 2-dimensional translations over the sensor surface in a given range. This scenario relates to translational unbiased oculomotor control causing random stimulus displacements. The second behavior is composed of independent zoom and rotation actions distributed uniformly on each axis. This mimics the behavior of an object manipulating agent where the oculomotor system stabilizes the sensor on target, mechanically compensating for image translations but not image rotations or scaling. These setups demonstrate that the agent's behavior induces different topologies of receptive fields on the sensor surface.

In Fig. 2 the converged layouts for the two considered action distributions are shown. The nature of the two converged topologies exhibits macroscopic



(a) Shift (b) Rotation and Zoom

Fig. 3. Sensor topologies obtained under behaviors visualized in Fig. 2(a) and Fig. 2(b).

differences: in the translation only case we can identify a tendency for hexagonal tiling structures over the entire sensor surface (apart from boundary effects), whereas in the rotation and zoom case the receptors organize radially in clear circular rings. Unlike in Fig. 1, the 3-dimensional perspective shows the smooth overlapping between receptive field elements.

To better comprehend the resulting sensor layouts, we refer back to the work of Clippingdale and Wilson [6], where the fitness of a layout relates directly to the distance between predicted and original point locations. In our case, just as in [6] a perfect sensor layout is one where receptors exactly map one onto another for every considered action resulting in $\mathbf{P}^a_{(R)}$ matrices where each row contains exactly one non-zero entry. Any deviation from this case leads to an increase in prediction error and lowers correlation. This fact allows us to replace the Euclidean distance as used by Clippingdale and Wilson by one based solely on correlation between sensory readings disregarding any knowledge about the sensor topology.

5 Conclusion and Outlook

This paper explored how the behavior of an artificial agent can shape the topology of a visual sensor. We proposed that a well suited sensor is one which simplifies stimulus flow patterns – and hence stimulus prediction – under a given set of actions. We showed that this quality is captured by spatiotemporal cross-correlation and can be used to self-organize visual sensor topologies on a given surface. The method proposed in this work simultaneously develops spatially coherent receptive fields and organizes their layout according to an executed behavior.

Recognizing the mutual coupling of morphology and active behavior in organisms evolved in nature, we believe that in artificial agents physical structure and actuation should eventually emerge through a co-developmental process.

Working in this direction, we will investigate in future contributions the reciprocal influence of physical form on behavior in order to deduce suitable actions from a given sensor topology.

Acknowledgments. This work was supported by the European Commission proj. FP7-ICT-248366 RoboSoM, by the Portuguese Government – Fundação para a Ciência e Tecnologia (FCT) proj. PEst-OE/EEI/LA0009/2011, proj. DCCAL PTDC/EEA-CRO/105413/2008, and FCT grant SFRH/BD/44649/2008.

References

1. Hayhoe, M., Ballard, D.: Eye movements in natural behavior. Trends in Cognitive Sciences **9** (2005) 188–194
2. Egelhaaf, M., Kern, R., Krapp, H.G., Kretzberg, J., Kurtz, R., Warzecha, A.K.: Neural enconding of behaviourally relevant visual-motion information in the fly. Trends in Neurosciences **25** (2002) 96–102
3. Schwartz, E.L.: Computational anatomy and functional architecture of striate cortex: A spatial mapping approach to perceptual coding. Vision Research **20**(1) (1980) 645–669
4. Petrowitz, R., Dahmen, H., Egelhaaf, M., Krapp, H.G.: Arrangement of optical axes and spatial resolution in the compound eye of the female blowfly *Calliphora*. J. Comp. Physiology A **186** (2000) 737–746
5. Lichtensteiger, L., Eggenberger, P.: Evolving the morphology of a compound eye on a robot. In: Proc. 3rd Europ. Worksh. on Adv. Mobile Robots. (1999) 127–134
6. Clippingdale, S.M., Wilson, R.: Self-similar neural networks based on a kohonen learning rule. Neural Networks **9**(5) (1996) 747–763
7. Ruesch, J., Ferreira, R., Bernardino, A.: A measure of good motor actions for active visual perception. In: Proc. Int. Conf. Dev. and Learning ICDL. (2011)
8. Ruesch, J., Ferreira, R., Bernardino, A.: Predicting visual stimuli from self-induced actions: an adaptive model of a corollary discharge circuit. IEEE Transactions on Autonomous Mental Development, submitted.
9. Crapse, T.B., Sommer, M.A.: Corollary discharge across the animal kingdom. Nat. Rev. Neuroscience **9** (2008) 587–600
10. Absil, P.A., Mahoney, R., Sepulchre, R.: Optimization Algorithms on Matrix Manifolds. Princeton University Press (2008)
11. Barzilai, J., Borwein, J.: Two-point step size gradient methods. IMA Journal of Numerical Analysis **8** (1988) 141–148
12. Bertsekas, D.P.: Constrained Optimization and Lagrange Multiplier Methods. Athena Scientific (1996)