# An Integrated System based on Binocular Learned Receptive Fields for Saccade-vergence on Visually Salient Targets

Daniele Re[1], Agostino Gibaldi[1], Silvio P. Sabatini[1] and Michael W. Spratling[2]
[1]*Department of Informatics, Bioengineering, Robotics and System Engineering, University of Genoa, Genoa, Italy*
[2]*Department of Informatics, King's College London, London, U.K.*
*daniele.rejo@gmail.com, {agostino.gibaldi, silvio.sabatini}@unige.it, michael.spratling@kcl.ac.uk*

Keywords: Disparity, Binocular Vision, Stereopsis, Vergence, Saccade, Attention, Basis Function Networks, Neural Networks, Sensory-sensory Transformations, Sensory-motor Control, Learning, V1 Area, Receptive Field Learning.

Abstract: The human visual system uses saccadic and vergence eyes movements to foveate interesting objects with both eyes, and thus exploring the visual scene. To mimic this biological behavior in active vision, we proposed a bio-inspired integrated system able to learn a functional *sensory* representation of the environment, together with the *motor* commands for binocular eye coordination, directly by interacting with the environment itself. The proposed architecture, rather than sequentially combining different functionalities, is a robust integration of different modules that rely on a front-end of learned binocular receptive fields to specialize on different sub-tasks. The resulting modular architecture is able to detect salient targets in the scene and perform precise binocular saccadic and vergence movement on it. The performances of the proposed approach has been tested on the iCub Simulator, providing a quantitative evaluation of the computational potentiality of the learned sensory and motor resources.

## 1 INTRODUCTION

An intelligent perceptual system must be able to interact with its environment through sensing, processing and interpreting information about the external world at different levels of representation, and eventually solve complex problems in contingent situations. Conceiving such an intelligent system, a number of abilities are required, such as: to attain online resource allocation; to generate and execute complex plans; to deal with problems as they arise in real-time; and to reason with incomplete information and unpredictable events. To fulfill all these tasks, considerable computational resources are required, that might exceed those available. In active machine vision, it has been demonstrated that both perceptual and action processes can rely on the same computational resources (Antonelli et al., 2014; Ognibene and Baldassare, 2015), at an early level that mimics the localized, oriented and band-pass receptive fields available in the primary visual cortex (V1 area) of mammals (Daugman, 1985a). Notably, learning such resources (Olshausen et al., 1996; Olshausen and Field, 1997) has become a popular approach, for a number of advantages. In fact, it allows the emergence of spatial competences that can be exploited by the system to self-calibrate both to the working space and to the geometric features of its own body schema (Gibaldi et al., 2015c), and ultimately specialize task-dependent representations (Ballard et al., 1997). Considering visual exploration of the three-dimensional (3D) environment, this behaviour is composed of a sequential cascade of operations. First, the visual information impinging on the two retinas has to be encoded and interpreted, in order to gather the salient features in the visual scene. Next, a binocular coordination of the eyes is necessary to perform a saccadic movement to foveate a visual target in the 3D space. Finally, a vergence refinement precisely aligns the optical axes on the object of interest, allowing for a better interpretation of disparity information.

From this perspective, the control of such compounded operations in active vision, requires not just the implementation of different complementary modules, each capable of solving one of these *single* actions, but their joint integration in a structured framework that allows us to obtain the *complex* behaviours required for a natural interaction with the surrounding environment. Here we present an integrated framework for autonomous saccade-vergence control of a

Re D., Gibaldi A., P. Sabatini S. and W. Spratling M.
An Integrated System based on Binocular Learned Receptive Fields for Saccade-vergence on Visually Salient Targets.
DOI: 10.5220/0006124702040215
In *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2017)*, pages 204-215
ISBN: 978-989-758-227-1

binocular visual system. The proposed framework is able learn an efficient internal representation of 3D visual scene, both for the perceptual and motor space, in order to perform accurate binocular foveation towards salient visual targets. In the proposed approach, both the perceptual and motor capabilities are learned by a direct interaction with the working environment, in a concurrent process that closes the loop between action and perception at system level. Moreover, the distributed approach, used both for the perceptual and motor aspects of the framework, allows for a simple and straightforward communication among the different integrated modules, since they rely on similar neural codes for representing sensory and motor information.

The remaining of the paper is organized as follows: Section 2 reviews the state of the art; Section 3 presents the different modules and their integration in the proposed framework; the capabilities of the approach within the iCub simulator are evaluated in Section 4; in Section 5 we draw the conclusions.

## 2 STATE OF ART

*Encoding Visual Information* - In the early visual system, the sensory pathway is commonly considered a communication channel that performs an efficient coding of the sensory signals, *i.e.* it is able to represent the sensory information with the minimal amount of resources, while preserving the coded information. Over the last two decades, researchers have proposed different unsupervised learning algorithms to model a range of neural processes at the early sensory stages. Imposing a sparseness constraint, it is possible to learn basis functions that resemble V1 receptive fields (Olshausen et al., 1996; Olshausen and Field, 1997), forming an efficient image representations (Daugman, 1985b). The approach can be also extended to stereoscopic information (Hyvärinen and Hoyer, 2000; Okajima, 2004; Hyvärinen et al., 2009), exploiting the natural disparity distribution (Hunter and Hibbard, 2015) to obtain ideal disparity detectors (Hunter and Hibbard, 2016).

The monocular and binocular visual information encoded by this early sensory stage, can thus be exploited by subsequent processing stages with different decoding strategies, depending on the task at hand: the monocular responses can be interpreted as feature map, and used as input to an bottom-up attention module, whereas the binocular responses can be used to drive the disparity-vergence control.

*Attention Model* - Attention is considered the process of selecting and gating visual information, and has a fundamental role in perceiving the surrounding environment and driving the eye movements. In humans and primates, this process is mediated by two competitive mechanisms: a bottom-up interpretation of the visual information to obtain a saliency map of the visual features, and a top-down interpretation of the scene based on a prior knowledge about the scene.

From this perspective, it is worth considering that during the early development of the visual system, visual attention relies on bottom-up process mainly, since it is not yet supported by a sufficient cognitive development (see (Gerhardstein and Rovee-Collier, 2002), as review). Being our integrative model focused on modeling the early functionalities of the visual system, we adopt a bottom-up attentive behavior, which is more suited to that purpose. To this aim, visual search models can be designed by integrating different visual features from the image (orientation, color, direction of movement), on the top of which the most salient parts in a scene pop out. The seminal work of of Itti and Koch (Itti et al., 1998) paved the way to different approaches saliency-based attention (Houghton and Tipper, 1994; Ma and Zhang, 2003; Hu et al., 2004). In our model, we adopted the approach proposed by (Bruce and Tsotsos, 2005), for its deep biological inspiration. The authors related spatial visual saliency to regions characterized by large differences between the monocular response of simple cells within a local region, and the response of the cells with similar tuning in a surrounding region. Such an antagonist organization of the input mimics the lateral inhibition strategy present in area V1 (Rao and Ballard, 1999). The resulting saliency map is used to define the target position in visual space, in order to drive the ocular gaze through saccadic eyes movements.

*Saccadic Control* - In order to effectively direct the gaze toward an interesting visual target, the brain must transform the sensory coordinates of the stimulus into headcentric motor coordinates, so to generate an effective motor command in joint coordinates (Crawford and Guitton, 1997) How does the brain perform such sensorimotor transformations? In the past, several neural networks have been proposed to convert input information, *i.e.* eye position and retinal target position, into an output, *e.g.* the targets locations in head coordinates (Pouget and Sejnowski, 1997; Chinellato et al., 2011; Antonelli et al., 2014; Muhammad and Spratling, 2015). Typically, this mapping occurs by developing a distributed representation in a ”hidden layer” interposed between the input and output layers. In our work, we adopted a basis function network (Muhammad and Spratling, 2015) able to perform both sensory-sensory mapping (retinotopic-headcentered coordinates) and sensory-motor mapping (headcentered-joint coordinates).

*Disparity-Vergence Control* - During vergence eye movements, the eyes rotate in opposite direction in order to reduce and eventually nullify the binocular disparity of the fixated object. Disparity-vergence eye movements were first modeled using a simple feedback control system (Rashbass and Westheimer, 1961). Classical models of vergence control can generally be classified into three basic configurations: continuous feedback, feedback with pre-programmed control (Hung et al., 1986) and switched-channel with feedback (Pobuda and Erkelens, 1993). The primary limitation of these models is that they first require the computation of the disparity map for the extraction of the control signals, thus limiting the functionality of the vergence system within the range of where the system is able to solve the stereo correspondence problem. Subsequent approaches, gathering inspiration form the primates visual system, are based on a distributed representation of binocular information (Gibaldi et al., 2010; Wang and Shi, 2011; Lonini et al., 2013; Gibaldi et al., 2016), overcoming this limitation. On this basis, we exploited a neural network model that directly interprets the population response into a vergence control, to nullify the disparity in fovea (Gibaldi et al., 2010).

*System integration* - Here, we review the studies that are relevant to the proposed integrative model, with a particular care for those integrating active perceptual approach with bottom-up biologically inspired attention modules. Since our approach focuses on the interplay between oculomotor control and early visual processing, we will not take into account those works including higher cognitive processes (Borji et al., 2010), as they would deserve a dedicated consideration (Orquin and Loose, 2013). Bottom-up attentive module are important to select objects of interest (Serre et al., 2007) in order to direct the gaze across the scene (Wang and Shi, 2011). The former work, starting from a standard model of visual cortex (Serre et al., 2005), designed a 4-layers neural network able to learn a vocabulary of visual features from images and to use it for the recognition of real-world object-categories. The latter work, proposed a saccade model that integrates three driving factors of human attention reflected by eye movements: reference sensory responses, fovea periphery resolution discrepancy, and visual working memory. These capabilities are required also for the control of humanoid robots (Ruesch et al., 2008). In fact, the recent developments of artificial intelligence require humanoid robots able to cope with variable and unpredictable problems, similar to those tackled by the human perceptual system. On this common ground, efficient biologically inspired solutions are an effective approach (Pfeifer et al., 2007).

Specifically, in our integrative model, attention acts as a front-end module in acquiring the target of interest, for a subsequent oculomotor control, that is composed of a binocular coordinated saccade and a vergence movement. Specifically, we are interested in employing a computational approach derived by bio-inspired models of primates' visual cortex. Accordingly, while the saccadic control is based on a network of radial basis functions (Muhammad and Spratling, 2015), the latter relies on a substrate of binocular receptive fields (Gibaldi et al., 2010; Gibaldi et al., 2016).

# 3 THE INTEGRATED SYSTEM

The system we designed integrates different modules (see Fig. 1) to provide an active stereo head with autonomous exploration capabilities: 1) a front-end for the encoding of the visual information, 2) a bottom-up attentive model to obtain salient features in the visual scene, 3) a saccadic control module to fixate the object of interest, and 4) a vergence control module to refine the vergence posture. At the root of the perception process, we implemented a set of binocular basis functions (receptive fields), directly learned from the images captured by the cameras, that provide a distributed coding of monocular and binocular visual information. The information from each monocular channel is exploited by an attentive process in order to derive a bottom-up saliency map of the visual features. The relationship between the selected target in retinotopic coordinates and the required eye movement to binocularly foveate it, is learned through a basis function network, that eventually drives coordinated binocular saccadic movements in the 3D space. Finally, a closed-loop vergence control decodes the binocular disparity information to refine the binocular alignment on the object of interest. In the following, we will describe each single module and its implementation on the iCub Simulator.

## 3.1 Learning Binocular Receptive Fields

For binocular receptive field learning, we exploited a pre-existing algorithm proposed by (Olshausen et al., 1996; Olshausen and Field, 1997), which relies on an unsupervised strategy. Specifically, an image $I(\mathbf{x})$ can be locally represented in terms of a linear combination of basis functions $\phi_i(\mathbf{x})$, where $\mathbf{x} = (x,y)$ are the

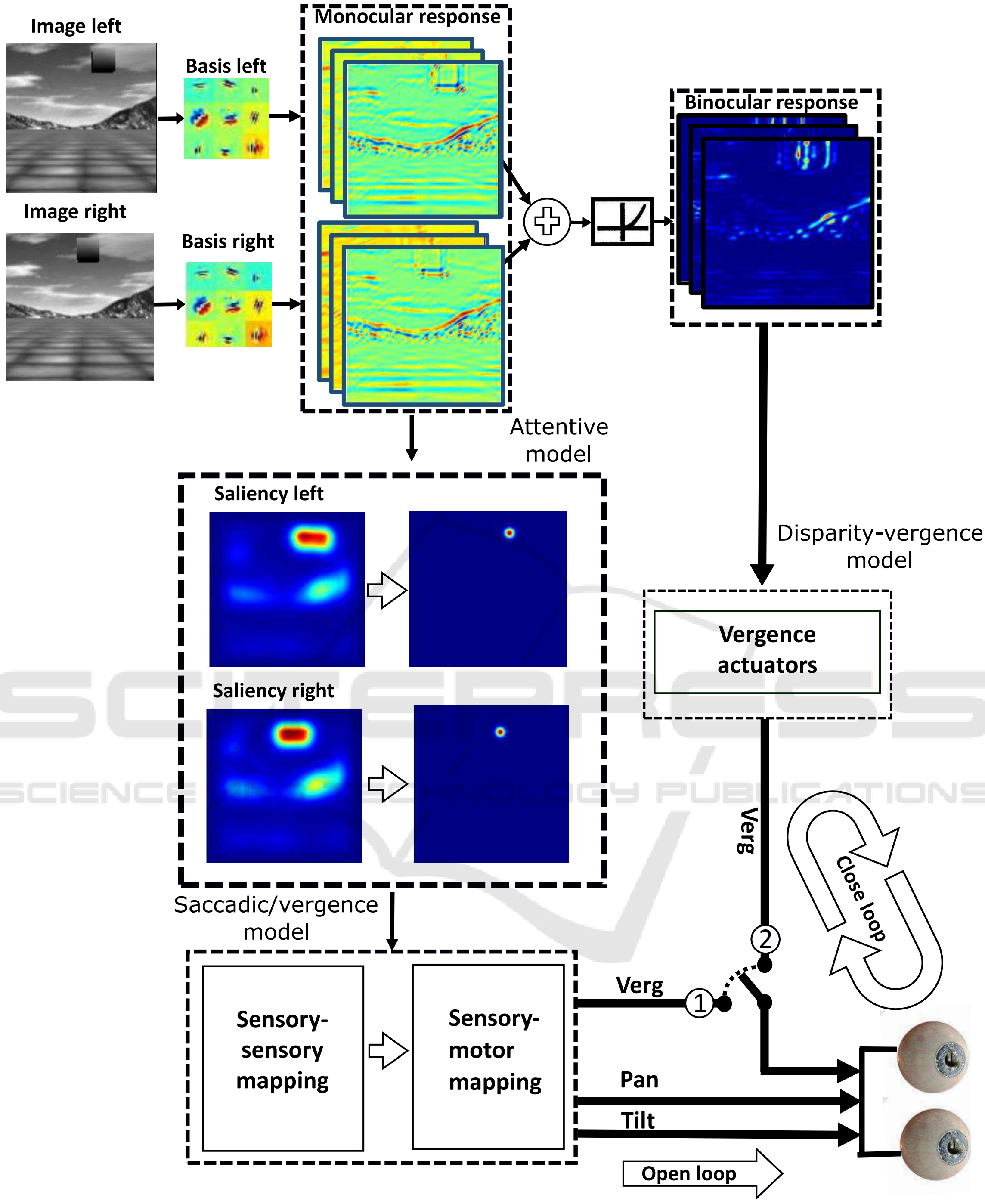


Figure 1: Schematic diagram of the integrated attentive-saccadic-vergence model. The monocular and binocular information obtained by filtering with the over-complete set of basis function is processed separately by the bottom-up attentive model and the disparity vergence control. The attentive model locates the most salient object on which to plan an open-loop binocular saccadic movement (verg, ①). At a subsequent stage, a closed-loop vergence refinement is guided by disparity information (verg, ②).

retinal coordinates. In order to learn the basis functions with the associated weights, the algorithm uses a set of patches extracted from natural scenes images, and seeks to maximize the sparseness of the encoded visual information. The basis functions that emerge, are a complete set that strongly resembles the receptive fields found in the primary visual cortex, and provide an efficient image representations (Daugman, 1985a).

In our implementation, we extended the approach to stereoscopic vision, in order to obtain binocular V1-like RFs. The image patches were taken from a set of twenty stereo images directly acquired from the iCub Simulator. A textured panel is placed at different depths in front of the iCub head, so as to obtain left and right patches with different binocular disparities. The monocular patches are then vertically concatenated to form a binocular patch $[p_k^L(\mathbf{x}), p_k^R(\mathbf{x})]^T$. According to (Lonini et al., 2013), these stereo patches can be approximated by:

$$\begin{bmatrix}\hat{p}_k^L(\mathbf{x})\\ \hat{p}_k^R(\mathbf{x})\end{bmatrix} = \sum_i^N a_{ik}\begin{bmatrix}\phi_i^L(\mathbf{x})\\ \phi_i^R(\mathbf{x})\end{bmatrix} \qquad (1)$$

where $N$ is the number of basis functions. In the implemented binocular approach, the resulting basis functions (*e.g.* Fig 2) are composed of a left, $\phi_i^L(\mathbf{x})$ and a right, $\phi_i^R(\mathbf{x})$ part. In order to characterize the properties of such functions, we can fit them as a bank of Gabor-like stereo receptive field pairs:

$$\phi_i^{L/R}(\mathbf{x}) \simeq \eta \cdot \exp\left(-\frac{(\mathbf{x}-\mathbf{x}_0^{L/R})^T(\mathbf{x}-\mathbf{x}_0^{L/R})}{2\sigma^2}\right)\cdot \cos(\mathbf{k_0^T}\mathbf{x}+\psi^{L/R}) \qquad (2)$$

where the following quantities are approximately equal in the left and right RFs: $\sigma$ is the variance of the Gaussian spatial support and defines the RF size, $\eta$ is a proper normalization constant, $\mathbf{k}_0 = (k_0\sin(\theta) - k_0\cos(\theta))^T$ is the spatial frequency of the RF, $k_0$ is the radial peak frequency orthogonal to the RF orientation $\theta$. The learned binocular RF profile is characterized by a difference between the phases ($\psi^L$ and $\psi^R$) of the monocular RFs and their positions ($\mathbf{x}_0^L$ and $\mathbf{x}_0^R$) on the image plane. Hence, the linear *monocular* response of a layer of simple cells is given by:

$$r_i^{L/R}(\mathbf{x}) = \int I^{L/R}(\mathbf{x}')\phi_i^{L/R}(\mathbf{x}'-\mathbf{x})d\mathbf{x}' \qquad (3)$$

with $i = 1,\ldots,N$. It is worth noting that $r(\mathbf{x})_i^{L/R}$ also define the feature map associated to the $i$-th feature. The response of a corresponding layer of *binocular* simple cells can be modeled as the cascade of the binocular linear response and a static non-linearity:

$$r_i^B(\mathbf{x}) = (r_i^L(\mathbf{x}) + r_i(\mathbf{x})^R)^2. \qquad (4)$$

By pooling the responses of simple cells with different monocular phase symmetries, but the same interocular phase and/or position shift, we obtain binocular complex cells with a specific disparity tuning, independent of the stimulus phase (Qian, 1994). Accordingly, the binocular cell is sensitive to a vector disparity $(\delta_H, \delta_V)$, oriented along the direction orthogonal to its spatial orientation $\theta$, which depends on the characteristics of the binocular receptive fields, and specifically on the frequency $\mathbf{k_0}$, the phase shift $\Delta\psi = \psi^L - \psi^R$ and the position shift $\mathbf{s} = \mathbf{x}_0^L - \mathbf{x}_0^R$:

$$r_i^B(\delta_H,\delta_V) = r_i^B(\mathbf{x};\mathbf{k_0},\mathbf{s},\Delta\psi) \propto \cos(1+\mathbf{k_0^T}(\delta-\mathbf{s})+\Delta\psi) \qquad (5)$$

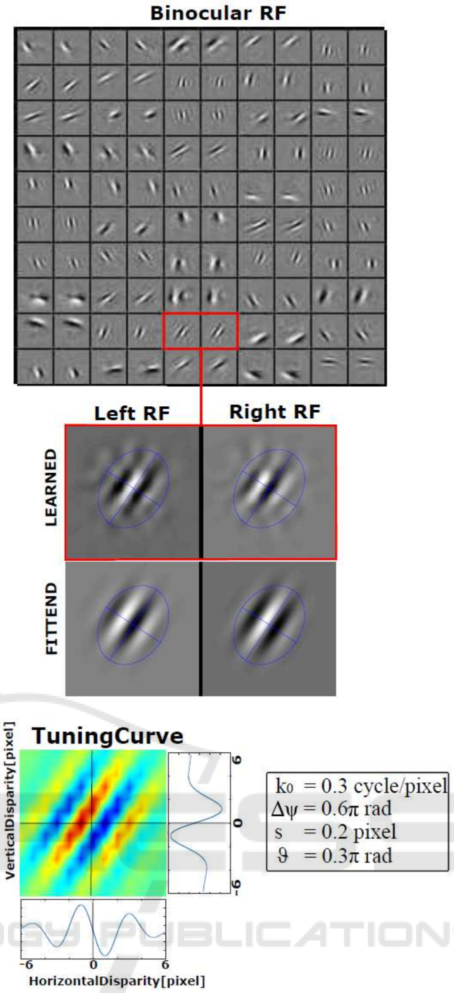


Figure 2: (Top) Examples of 50 learned basis functions, displayed pairing the left and right receptive field. Despite the similar size, orientation and frequency, each pair exhibits both a phase and a position shift in their profile (see Sec. 4). (Middle) An example of a learned binocular receptive field together with its fitting. The blue ellipses represent the size of the envelope and its horizontal and vertical meridians along the orientation of the receptive field. (Bottom) The 2D tuning curve of the binocular cell to horizontal and vertical retinal disparity, together with its horizontal and vertical cross sections. The tuning curve is derived as squared sum of the response of the two monocular receptive fields. (Table): the table shows the values of the spatial frequency($k_0$), phase shift($\Delta\psi$), location shift ($s$) and orientation($\theta$) of the above-mentioned basis.

that defines the tuning of the complex cell to a specific disparity value.

In the next stages, the monocular cell responses are fed to the attentive model to obtain the saccadic target, while the binocular responses are used to obtain the disparity-vergence control.

### 3.2 The Bottom-Up Attentive Module

This module is used to define interesting targets within the visual scene, to autonomously drive the saccadic eye control. Specifically, we exploited the *Attention based on Information Maximization* (AIM) algorithm (Bruce and Tsotsos, 2005). The authors define visual saliency as the relationship between the monocular response ($r^{L/R}$) of simple cells within a local region ($C_i$), and the response of the cells with similar tuning in a surrounding region ($S_i$). For each image location, the response value of each monocular basis can be interpreted as a local match between the frequency and orientation content of the image, and the tuning properties of that basis. Since visual features can be considered salient when they stand out the background (Bruce and Tsotsos, 2005), a classical center-surround function is used to process each feature map to obtain a saliency map $\mathcal{S}_i$ for each of the basis functions considered:

$$\mathcal{S}_i^{L/R}(\mathbf{x}) = \tfrac{1}{\sqrt{2\pi\sigma_V}} \textstyle\sum_{\mathbf{x}\in\Gamma} \mathrm{G}(\mathbf{x},\sigma_\mathrm{S}) \cdot \\ \exp\left[(r_{C_i}^{L/R}(\mathbf{x}) - r_{S_i}^{L/R}(\mathbf{x}))^2/(2\sigma_V^2)\right] \tag{6}$$

where and $r_{C_i}^{L/R}$ and $r_{S_i}^{L/R}$ are the monocular response for the local region and surround region for the *i*-th basis function, $\sigma_V$ is the variance of the Gaussian kernel used for spatial averaging over a neighborhood $\Gamma$ (Bruce and Tsotsos, 2005), $\frac{1}{\sqrt{2\pi\sigma_V}}$ is a normalizing factor, and $\mathrm{G}(\mathbf{x},\sigma_\mathrm{S})$ indicates the contribution of neighboring elements to the local estimate. The value of $\mathcal{S}_i^{L/R}(\mathbf{x})$ returns a saliency density where the unitary value suggests a redundancy between local and surround region, *i.e.* no saliency, and a value close to zero indicates a substantial content difference. The density information can be used to predict at each retinal location the saliency $l(\mathbf{x})$ of the monocular response over the whole set of visual features encoded:

$$l^{L/R}(\mathbf{x}) = \sum_{i=1}^{N} -\log(\mathcal{S}_i^{L/R}(\mathbf{x})) \tag{7}$$

thus returning large values for salient features in the image and *vice-versa*.

This attentive model is thus exploited to define salient locations within the visual scene as targets of gaze-shifts. In order to make the saccadic vergence control able to foveate on a generic salient object independently of its specific shape, a Mexican hat recurrent filtering is applied on the $l^{L/R}$ in order to reduce the saliency map to a blurred spot centered on the most salient object (see attentive block in Fig. 1), mimicking a winner-takes-all strategy (Itti et al., 1998). In the next section we will describe the implemented model for saccadic eye movements.

### 3.3 The Saccadic Module

In order to shift the gaze to a desired target, we implemented a sensory-sensory and a sensory-motor transformation. The goal is to transform visual information about the target location on the retinas and proprioceptive information about the positions of both eyes in the head, to obtain the necessary motor control to foveate the target with both eyes. To this purpose, we used the Predictive Coding/Bias Competition-Divisive Input Modulation (PC/BC-DIM) neural network (Muhammad and Spratling, 2015). PC/BC-DIM is a basis function network that performs a mapping between an input layer $\mathbf{i}$ and a reconstruction layer $\rho$ (see Fig 3). The range of possible mappings are encoded by connection weights ($\mathbf{W}$ and $\mathbf{V}$) and are mediated by a hidden layer of prediction nodes $\pi$ that encodes the distinct causes that can underlie the input:

$$\begin{cases} \rho = \mathbf{V}\pi \\ \varepsilon = \mathbf{i} \oslash (c_2 + \rho) \\ \pi \leftarrow (c_1 + \pi) \otimes \mathbf{W}\varepsilon. \end{cases} \tag{8}$$

Where $\varepsilon$ is the error between the input and the network's reconstruction of the input ($\rho$); $c_1$ and $c_2$ are constants; and $\oslash$ and $\otimes$ indicate element-wise division and multiplication, respectively. Both $\mathbf{i}$ and $\rho$ are partitioned into four parts representing the visual target position in retinotopic coordinates, the eye pan value, the eye tilt, and the head-centered location of the target.

The model employs the connection weights $\mathbf{V}$ and $\mathbf{W}$ as a "dictionary" containing all the possible combinations among the stimulus location in retinal coordinates, the eye position (pan and tilt values), and the head-centered bearing of the visual target. Their values are determined by a training process in which a stationary visual target is presented to the iCub Simulator while the cameras move, generating distinct combinations of eye pan/tilt and retinal inputs. After an exhaustive number of eye movements are realized for a specific position of the target, the process is repeated for a different stimulus position, thus defining a new head-centered position. After the training process, the model is able to return the pan and tilt values

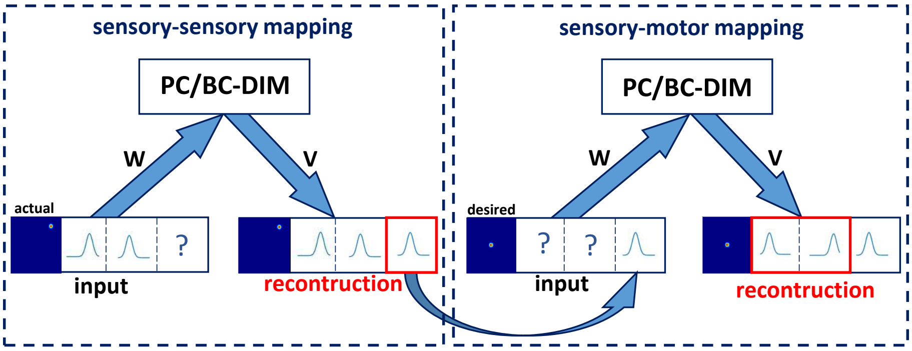


Figure 3: The figure shows the PC/BC-DIM model being used to calculate the eye pan and tilt values necessary to foveate the target (for a single eye). In the first step, the network is provided with the saliency map showing the visual target in retinotopic coordinates ($1^{st}$ partition) and pan and tilt eye position values ($2^{nd}$ and $3^{rd}$ partitions). The network calculates the head-centered coordinates of the visual target ($4^{th}$ partition of the reconstruction array). In the second step, the network takes as new input the head-centered coordinates from the previous step and a reference signal showing the desired position of the target at the center of the retina. The network calculates, in the reconstruction array, the pan and tilt values necessary to foveate the target.

necessary to move the fovea onto the desired target position, as described in Fig. 3.

This process of planning eye movements was performed for the left and right eye separately. The iCub Simulator has a single actuator for conjugate tilt movement, a single actuator for conjugate pan movement, and a separate control for vergence. These values were determined by combining the separate left and right eye pan and tilt values calculated by PC/BC-DIM as follows:

$$\begin{cases} \mathrm{Pan} = (\mathrm{Pan}_{\,\mathrm{L}} + \mathrm{Pan}_{\,\mathrm{R}})/2 \\ \mathrm{Tilt} = (\mathrm{Tilt}_{\,\mathrm{L}} + \mathrm{Tilt}_{\,\mathrm{R}})/2 \\ \mathrm{Vergence} = \mathrm{Pan}_{\,\mathrm{L}} - \mathrm{Pan}_{\,\mathrm{R}}. \end{cases} \tag{9}$$

It is worth considering that the saccadic module works in a ballistic open-loop manner: after the command there is no visual feedback to correct the resulting vergence movement in order to fixate the target in depth. A visually-driven closed-loop refinement is thus required to correctly align the eyes on the object of interest. To this aim, we used the disparity-vergence module described in the following section.

### 3.4 Disparity-Vergence Module

To obtain an effective disparity-vergence control, we used the disparity tuning properties of the binocular complex cells described in Section 3.1. The desired vergence control is an odd symmetric signal that evokes a convergent movement for crossed disparity and a divergent movement for uncrossed disparity. In this way, the vergence control can be applied in closed-loop until the disparity on the target is null, *i.e.* until the complex cell population response is maximum (Gibaldi et al., 2015b)).

The response of a vergence neuron $r^V$, that drives the convergent and divergent movements of the robot eyes, is obtained through a weighted sum of the binocular responses over a spatial neighborhood $\Omega$:

$$r^V = \sum_{\mathbf{x}\in\Omega} \sum_{i}^{N} w_i r_i^B(\mathbf{x}) \tag{10}$$

The connection weights $w_i$ are learned by minimizing the following cost function:

$$\underset{w_i}{\operatorname{argmin}} \quad \left|\left|\Sigma_{i=1}^{N} r_i^B(\delta_H,0) w_i - v_H\right|\right|^2 + \tag{11}$$
$$+\lambda \left|\left|\Sigma_{i=1}^{N} r_i^B(0,\delta_V)(w_i-1)\right|\right|^2$$

where $v_H$ is the desired control profile, $\delta_H$ is the horizontal disparity and $\delta_V$ is vertical disparity, whereas $r_i^B(\delta_H)$ and $r_i^B(\delta_V)$ are the binocular tuning curves for horizontal and vertical disparity, respectively, and $\lambda > 0$ is a factor that balances the relevance of the second term over the first.

### 3.5 Implementation and Learning

The integrated system has been implemented within the iCub Simulator (Tikhanoff et al., 2008), which is an open-source computer simulator for the humanoid robot iCub, developed to be a test-bed for robotic control algorithms. Specifically, we used this environment since the iCub stereo head has the necessary characteristics for binocular active vision (Beira et al., 2006).

*Binocular Receptive Fields* - The iCub robot gazes at a textured panel with different vergence angles in order to produce a disparity approximately in the range $[-2,2]$ pixels. The images are captured from the left and right cameras at a resolution of 128 × 128 pixels, covering ≈ 24° of visual field. From those images, we randomly extracted sets of 300 stereoscopic image patches, of size 15×15 pixel, *i.e.* covering ≈ 3° of visual field. The stereoscopic patches are then are fed to the algorithm described in section 3.1, in order to learn the set of basis functions. The procedure is iterated for ≈ 300 sets of patches, before the receptive field learning converges to a stable solution.

*Attentive Model* - Each of the 150 learned basis is used to generate a feature maps of the same size of the input image (128 × 128). Each feature map is transformed in a saliency map through a comparison between a center (single pixels) and surround (39 × 39 pixels) values in a local region of the feature map (see Eq. 7). Thereafter, an overall likelihood for all coefficients corresponding to a single location is given by the product of the likelihoods associated with each individual basis type (see Eq. 7). The final output return a map, again of the same size of the input image (128 × 128), with high coefficients associated to the less expected values. In order to select the highest salient regions only, the map is filtered by an iterative *Difference of Gaussian* function (30 × 30, $\sigma_C = 4.5$, $\sigma_S = 6$ pixels), in order to obtain the target for the subsequent saccadic control.

*Learning Oculo-Motor Transformation* - In order to learn the **W** and **V** matrices required by the the PC/BC-DIM network, the visual space is sequentially explored. A visual target (a box of 0.05 $m^3$) is placed on a 3 × 3 grid of 3D points ranging from -0.2 m to 0.2 m at steps of 0.2 m along the *x*-axis, and from -0.2 m to 0.2 m at steps of 0.2 m along the *y*-axis. The grid is placed at a distance of 0.5 m from the robot head (*z*-axis). For each position of the visual target, the eyes perform a set of movements toward the box, starting from a grid of 31 × 41 points in a range $[-20°, 20°]$ pan and $[-15°, 15°]$ tilt, at steps of 1°. For each iteration, the images (128 × 128 pizels) is filtered with a 91 Gaussian filter bank ($\sigma = 7$ pixel) that covers all the image locations. The obtained 1 × 91 vector is concatenated with a 1 × 11 array for pan movement, a 1 × 7 array for tilt movement and a value for the headcentered position of the box. Once all the ocular movements are performed for that specific box position, the location is changed and a new learning set is implemented adding a new vector element corresponding to the new headcentric location. This procedure allows us to obtain the **W** and **V** matrices with the learned combination from which the PC/BC-DIM network can obtain headcentric information (sensory-sensory mapping) or joints information (sensory-motor mapping). During the saccadic control, the attentive output will be used as retinal input to the saccadic model.

*Vergence Control* - Subsequently, the obtained basis functions are used to learn the weights for the disparity-vergence control (as described in section 3.4). The disparity tuning curves, used to compute the weights *w*, as in Eq.11, have been obtained by random dot stereograms with retinal disparity varying between $[-6, 6]$ pixels, both for its horizontal and vertical component. The obtained control is able to produce the correct vergence movement while the stimulus disparity is approximately in a range three times larger than the one used to learn the basis function (*i.e.* $[-2, 2]$ pixels).

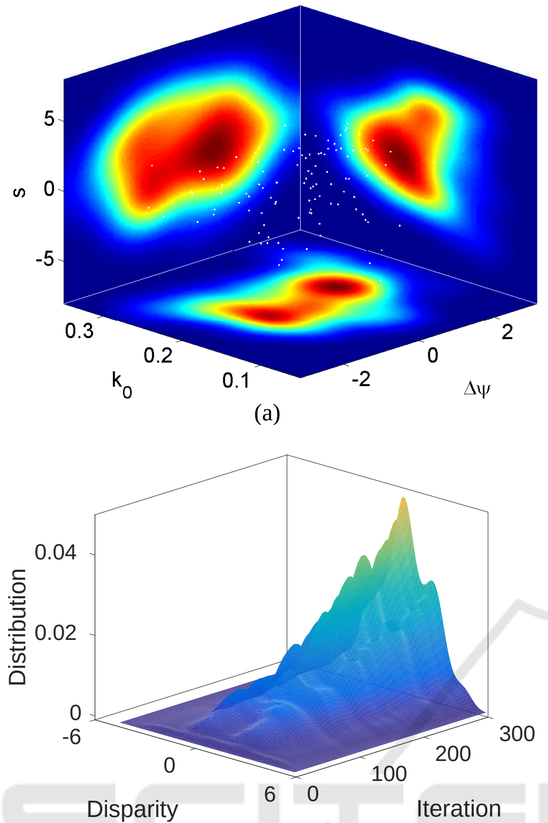
## 4 RESULTS

The following results show the accuracy and robustness of the integrated system starting from a reduced set of high informative resources. The implemented algorithm is indeed effective in achieving its main goal of moving the fixation point of a simulated stereo head towards salient locations at different depths.

**Testing the basis functions** - As a preliminary assessment of our approach, we evaluated if the learned resources are able to provide an efficient coding of binocular information. To this purpose, we fit the learned monocular basis functions with Gabor functions (see Fig. 2) in order to characterize the resources by the hybrid position-phase-based model (Ohzawa et al., 1990). Fig. 4a shows the distribution of the fitted parameters of frequency tuning $k_o$, phase $\Delta\psi$ and position shift *s*. It is interesting to notice how the binocular computational resources properly tile the parameter space, with a distribution that qualitatively resembles the actual distribution of V1 complex cells of area V1 (Prince et al., 2002; Gibaldi et al., 2015a).

In order to characterize the effectiveness of the sensory coding performed by the learned binocular resources, we used the Fisher information (Ralf and Bethge, 2010). To this purpose, we derived the disparity tuning curves in response to random dot stereograms in which the disparity is varied in the range of $[-6, 6]$ pixels Fig. 4b shows the trend of Fisher information over the iterations. It is evident how at the first iteration, *i.e.* when the basis functions are randomly initialized, the population carries no information about retinal disparity. Along the learning, the binocular information coding improves at each iteration over the disparity range, and, more particularly, it is peaked at zero disparity which is specifically informative for guiding vergence behavior.

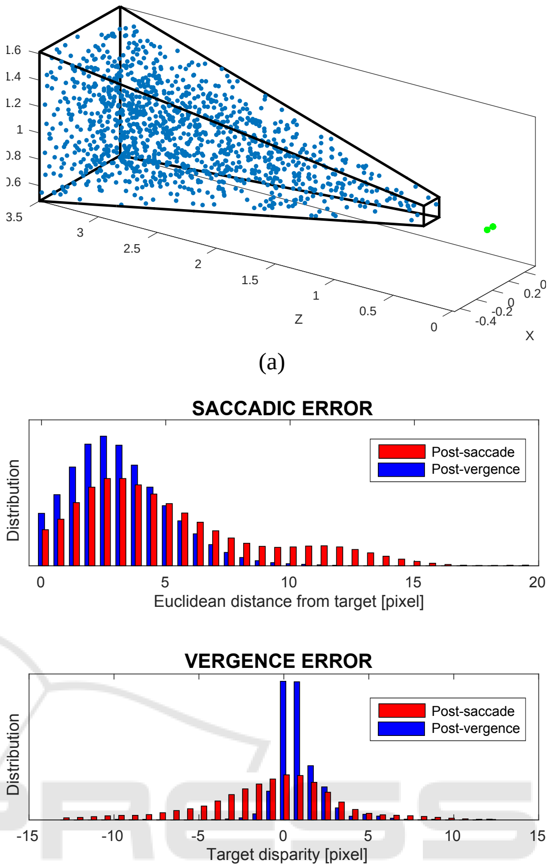
**Testing the integrated model** - In order to test



(a)

(b)

Figure 4: (a) Representation of learned binocular cells with respect to frequency $\mathbf{k}_0$, phase difference $\Delta\psi$ and position shift $\mathbf{s}$ (white dots), together with the joint distributions. (b) Fisher information carried by the binocular cells (Ralf and Bethge, 2010) for each disparity value; plotted against the iterations of the basis functions learning process.



(a)

(b)

Figure 5: (a) Random locations (500) in which the target is placed during the test. Specifically, the positions fall in the visual field of the robot, within a frustum defined by a range of $[-15+15]$ degrees of pan and $[-10+10]$ degree of tilt covering a depth ranging from 0.5 to 3.5m. (b) Distribution of the saccadic (top) and vergence error (bottom) at target location, after the 3D saccadic movement (red) and after applying the disparity-vergence control (blue).

the integrated behaviour of the system, we used the iCub Simulator to implement a simple environment suited to the task at hand: a “salient” object is placed at random positions against a textured background (see Fig. 5a). The system firstly uses the bottom-up attentive model to find the salient object. The target is brought in the fovea through a coordinate saccadic-vergence movement implemented by the PC/BC-DIM model. Then, the disparity-vergence control is iterated until the binocular disparity of the target reaches a threshold value close to zero. Once a proper fixation is reached, the salient object is displaced to a new random position. The target object is a square box of 0.038 m, presented at 500 random positions within a portion of space represented by the frustum shown in Fig.5a. For each target position the iCub-simulator executes an open-loop saccadic movement, followed by a closed loop vergence movement. The performance of the integrated system, has been evaluated by computing the saccadic error and the vergence error. In order to evidence the positive effect of the vergence control on the fixation posture, we computed these two quantities after the binocular saccade (Post-saccade) and after the vergence refinement (Post-vergence). The saccadic error is computed as the mean distance between the center of mass of the object’s image and the “foveas” of the cameras (see Fig. 5b, top), and quantifies the accuracy of the performed saccadic movement in foveating the target (mean error $< 3.5$ pixels). The vergence error is computed as the residual retinal disparity (absolute value) over a foveal sub-window corresponding to the target area (see Fig. 5b, bottom). After the 3D saccadic control, the residual disparity is $\approx 3.2$ pixels and it drops to $\approx 1.2$ pixels after the action of the vergence control. This result demonstrates the efficacy of the implemented algorithm to properly fixate the visual target in depth. Moreover, the vergence control, bringing the object of interest into the foveas, is also helpful to reduce the post saccadic error (see Fig. 5b, top).

For a demonstration of the overall capabilities of the proposed model, see the video *https://www.youtube.com/watch?v=EFsEu25-nR4&feature=youtu.be*. The video represents a sequential series of tasks performed by the iCub robot, simulated within the iCub Simulator:

1. *Receptive fields learning*: the proposed algorithm for binocular receptive field learning is fed with stereo images acquired from the iCub head. The vergence angle is randomly changed in order to obtain stereo images with variable binocular disparity, and thus to obtain receptive fields with different disparity sensitivities.
2. *Disparity vergence control*: the performance of the vergence control is shown with a step stimulus (Hung et al., 1986; Gibaldi et al., 2010; Gibaldi et al., 2016). A frontoparallel textured plane suddenly changes depth, generating binocular disparities in the foveal area of the stereo images, and consequently triggering the closed-loop vergence control to nullify the binocular disparity.
3. *The integrated model*: the proposed approach is tested with a sequence of attention, saccadic movement and vergence refinement. The bottom-up attentive module firstly selects the most salient image region within the scene, then a binocular saccade is performed to bring this region in the foveae of both the eyes, finally the closed-loop vergence control refines the eye position on the selected area, in order to nullify the binocular disparity.

## 5 CONCLUSIONS

In this paper, we proposed an integrated bio-inspired architecture able to learn functional sensory and motor competences directly from the interaction with the 3D environment. The visual front-end of learned V1-like computational resources provides an efficient coding of the binocular visual information, instrumental to different complementary tasks. The flexibility and adaptability of the distributed coding allows us to exploit the population response at different levels of complexity, from disparity-vergence control in closed-loop, to visual saliency on which to learn, plan and perform open-loop saccadic and vergence movements in the 3D environment. The resulting system's performance goes well beyond those obtained by the previous work on saccade (Muhammad and Spratling, 2015) and vergence (Gibaldi et al., 2010; Gibaldi et al., 2016) control considered in isolation. Advantages have been observed in terms of both accuracy and generalization capability.

Summarizing, the proposed bio-inspired approach, rather than sequentially combining different functionalities, defines an integrated and coherent architecture where each module relies on the same source of information and applies it to specialized sub-tasks. This would allow us not just to solve the single separate tasks, but also to develop complex behaviours for an active natural interaction of a robot agent with the environment.

A future extension of the present work will be dedicated to include in the network a top-down module for visual attention, in order to provide the robot with a higher level behavior, possibly endowed by cognitive capabilities. Moreover, the proposed approach will be tested on a real robot system (*e.g.* the iCub stereo head (Beira et al., 2006)).

## REFERENCES

Antonelli, M., Gibaldi, A., Beuth, F., et al. (2014). A hierarchical system for a distributed representation of the peripersonal space of a humanoid robot. *Autonomous Mental Development, IEEE Transactions on*, 6(4):259–273.

Ballard, D. H., Hayhoe, M. M., Pook, P. K., and Rao, R. P. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, 20(04):723–742.

Beira, R., Lopes, M., Praça, M., Santos-Victor, J., Bernardino, A., Metta, G., Becchi, F., and Saltarén, R. (2006). Design of the Robot-Cub (iCub) head. In *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.*, pages 94–100. IEEE.

Borji, A., Ahmadabadi, M. N., Araabi, B. N., and Hamidi, M. (2010). Online learning of task-driven object-based visual attention control. *Image and Vision Computing*, 28(7):1130–1145.

Bruce, N. and Tsotsos, J. (2005). Saliency based on information maximization. In *Advances in neural information processing systems*, pages 155–162.

Chinellato, E., Antonelli, M., Grzyb, B. J., and Del Pobil, A. P. (2011). Implicit sensorimotor mapping of the peripersonal space by gazing and reaching. *IEEE Trans. on Autonomous Mental Development*, 3:43–53.

Crawford, J. D. and Guitton, D. (1997). Visual-motor transformations required for accurate and kinematically correct saccades. *Journal of Neurophysiology*, 78(3):1447–1467.

Daugman, J. G. (1985a). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *JOSA A*, 2(7):1160–1169.

Daugman, J. G. (1985b). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *JOSA A*, 2(7):1160–1169.

Gerhardstein, P. and Rovee-Collier, C. (2002). The development of visual search in infants and very young children. *Journal of Experimental Child Psychology*, 81(2):194–215.

Gibaldi, A., Canessa, A., and Sabatini, S. (2015a). Vergence control learning through real V1 disparity tuning curves. In *Neural Engineering (NER), 2015 7th International IEEE/EMBS Conference*, pages 332–335.

Gibaldi, A., Canessa, A., Solari, F., and Sabatini, S. (2015b). Autonomous learning of disparity–vergence behavior through distributed coding and population reward: Basic mechanisms and real-world conditioning on a robot stereo head. *RAS*, 71:23–34.

Gibaldi, A., Chessa, M., Canessa, A., Sabatini, S., and Solari, F. (2010). A cortical model for binocular vergence control without explicit calculation of disparity. *Neurocomputing*, 73(7):1065–1073.

Gibaldi, A., Sabatini, S. P., Argentieri, S., and Ji, Z. (2015c). Emerging spatial competences: From machine perception to sensorimotor intelligence. *Robotics and Autonomous Systems*, (71):1–2.

Gibaldi, A., Vanegas, M., Canessa, A., and Sabatini, S. P. (2016). A portable bio-inspired architecture for efficient robotic vergence control. *International Journal of Computer Vision*, pages 1–22.

Houghton, G. and Tipper, S. P. (1994). A model of inhibitory mechanisms in selective attention.

Hu, Y., Xie, X., Ma, W.-Y., Chia, L.-T., and Rajan, D. (2004). Salient region detection using weighted feature maps based on the human visual attention model. In *Pacific-Rim Conference on Multimedia*, pages 993–1000. Springer.

Hung, G. K., Semmlow, J. L., and Ciufferda, K. J. (1986). A dual-mode dynamic model of the vergence eye movement system. *IEEE Transactions on Biomedical Engineering*, (11):1021–1028.

Hunter, D. W. and Hibbard, P. B. (2015). Distribution of independent components of binocular natural images. *Journal of vision*, 15(13):6–6.

Hunter, D. W. and Hibbard, P. B. (2016). Ideal binocular disparity detectors learned using independent subspace analysis on binocular natural image pairs. *PloS one*, 11(3):e0150117.

Hyvärinen, A. and Hoyer, P. (2000). Emergence of phase- and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neural computation*, 12(7):1705–1720.

Hyvärinen, A., Hurri, J., and Hoyer, P. O. (2009). *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision.*, volume 39. Springer Science & Business Media.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (11):1254–1259.

Lonini, L., Zhao, Y., Chandrashekhariah, P., Shi, B. E., and Triesch, J. (2013). Autonomous learning of active multi-scale binocular vision. In *Development and Learning and Epigenetic Robotics (ICDL), 2013 IEEE Third Joint International Conference on*, pages 1–6.

Ma, Y.-F. and Zhang, H.-J. (2003). Contrast-based image attention analysis by using fuzzy growing. In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 374–381. ACM.

Muhammad, W. and Spratling, M. (2015). A neural model of binocular saccade planning and vergence control. *Adaptive Behavior*, 23(5):265–282.

Ognibene, D. and Baldassare, G. (2015). Ecological active vision: Four bioinspired principles to integrate bottom-up and adaptive top-down attention tested with a simple camera-arm robot. *Autonomous Mental Development, IEEE Transactions on*, 7(1):3–25.

Ohzawa, I., DeAngelis, G., and Freeman, R. (1990). Stereoscopic depth discrimination in the visual cortex: neurons ideally suited as disparity detectors. *Science*, 249(4972):1037–1041.

Okajima, K. (2004). Binocular disparity encoding cells generated through an infomax based learning algorithm. *Neural Networks*, 17(7):953–962.

Olshausen, B. A. et al. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609.

Olshausen, B. A. and Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision research*, 37(23):3311–3325.

Orquin, J. L. and Loose, S. M. (2013). Attention and choice: A review on eye movements in decision making. *Acta psychologica*, 144(1):190–206.

Pfeifer, R., Lungarella, M., and Iida, F. (2007). Self-organization, embodiment, and biologically inspired robotics. *science*, 318(5853):1088–1093.

Pobuda, M. and Erkelens, C. J. (1993). The relationship between absolute disparity and ocular vergence. *Biological Cybernetics*, 68(3):221–228.

Pouget, A. and Sejnowski, T. J. (1997). Spatial transformations in the parietal cortex using basis functions. *Journal of cognitive neuroscience*, 9(2):222–237.

Prince, S., Pointon, A., Cumming, B., and Parker, A. (2002). Quantitative analysis of the responses of V1 neurons to horizontal disparity in dynamic random-dot stereograms. *Journal of Neurophysiology*, 87(1):191–208.

Qian, N. (1994). Computing stereo disparity and motion with known binocular cell properties. *Neural Computation*, 6(3):390–404.

Ralf, H. and Bethge, M. (2010). Evaluating neuronal codes for inference using fisher information. In *Advances in neural information processing systems*.

Rao, R. and Ballard, D. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.*, 2(1):79–87.

Rashbass, C. and Westheimer, G. (1961). Disjunctive eye movements. *The Journal of Physiology*, 159(2):339.

Ruesch, J., Lopes, M., Bernardino, A., Hornstein, J., Santos-Victor, J., and Pfeifer, R. (2008). Multimodal saliency-based bottom-up attention a framework for the humanoid robot icub. In *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*, pages 962–967. IEEE.

Serre, T., Kouh, M., Cadieu, C., Knoblich, U., Kreiman, G., and Poggio, T. (2005). A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex. Technical report, DTIC Document.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., and Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE transactions on pattern analysis and machine intelligence*, 29(3):411–426.

Tikhanoff, V., Cangelosi, A., Fitzpatrick, P., et al. (2008). An open-source simulator for cognitive robotics research: the prototype of the iCub humanoid robot simulator. In *Proc. of the 8th workshop on performance metrics for intelligent systems*, pages 57–61. ACM.

Wang, Y. and Shi, B. E. (2011). Improved binocular vergence control via a neural network that maximizes an internally defined reward. *IEEE Transactions on Autonomous Mental Development*, 3(3):247–256.