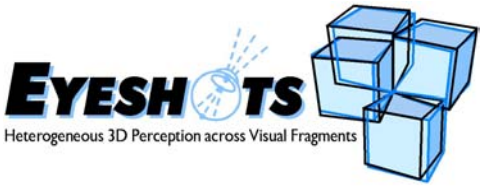


Project no.: FP7-ICT-217077
Project full title: Heterogeneous 3-D Perception across Visual Fragments
Project Acronym: EYESHOTS
Deliverable no: D2.1
Title of the deliverable: Convolutional network for vergence control.

Date of Delivery:	09 September 2009	
Organization name of lead contractor for this deliverable:	K.U.Leuven	
Author(s):	A. Gibaldi, N. Chumerin, M.Chessa, K. Pauwels, F. Solari, S.P. Sabatini, M. Van Hulle	
Participant(s):	K.U.Leuven, UG	
Workpackage contributing to the deliverable:	WP2	
Nature:	Report	
Version:	2.1	
Total number of pages:	44	
Responsible person:	Marc Van Hulle	
Revised by:	F. Hamker	
Start date of project:	1 March 2008	**Duration:** 36 months

Project Co-funded by the European Commission within the Seventh Framework Programme		
Dissemination Level		
PU	Public	**X**
PP	Restricted to other program participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Abstract:

Computational models for the control of horizontal vergence, based on a distributed representation of disparity, are proposed and analyzed. The models directly extract the linear servos from the post-processed response of a population of disparity tuned complex cells, without explicit calculation of the disparity map. The disparity-vergence curves have been either designed on the basis of a desired behavior, or learned by examples. Training and evaluation of the networks are discussed. The resulting vergence controls yield to stable fixation and has small response time to a wide range of disparities.

Contents

1 Executive summary

One of the objective of Workpackage 2 is to develop a convolutional network-based vergence control from a population of disparity-based feature. To this end, we investigated the specialization of disparity detectors at different levels in a hierarchical network architecture to see the effect of learning specific coding and decoding strategies for active vergence control and depth vision. The extraction of binocular features occurs through a cortical-like population network, developed by partner UG. The network provides a harmonic (*i.e.*, amplitude and phase) representation of the visual signal, operated by a set of "simple cell" units (S-cells). At the level of S-cells, the "totipotency" of the representation contains all the necessary basic components to differentiate into several classes of visual descriptors. Stereo and - in perspective - stereomotion percepts emerge in layers of disparity energy "complex cell" units (C-cells) that gather S-cells outputs according to specific architectural schemes. These computations can be supported by neuromorphic architectural resources organized as hierarchical arrays of interacting nodes. On this basis, convolutional network paradigms and learning processes have been introduced to develop a high degree variability of the cell's responses towards the specialization of disparity detectors for the control of vergence. The desired linear servos have been either designed on the basis of the disparity-vergence curves observed in the Medial Superior Temporal cortical area, or learned by examples. The selected learning paradigm is inspired by LeNet5 [1], since it is expected to have a good performance being such a network optimized at every level of the hierarchy. To this end, the LeNet architecture has been extended to increase its flexibility and including new functionalities. Specifically, differently from most of the conventional vergence control models [2, 3, 4, 5, 6], based on the minimization of the horizontal disparity, we propose to avoid implicit computation of the disparity map and extract the vergence control signal directly from the population responses over the "foveal" region. A neural network paradigm has been chosen for this type of conversion/extraction procedure. An increasing complexity strategy in the learning process is adopted: starting from the simplest one-unit architecture we increase the number of units/layers until an acceptable level of generalization error is reached. In order to learn the vergence control, we developed and used a simulator to create the training datasets. Each sample in the training dataset contains the stereo image pairs, the actual vergence angle, the actual gaze orientation, and the desired (for this particular case) vergence angle ("ground truth"). Using these datasets and the simulated environment it has been possible to train and evaluate the proposed neural network based vergence controller. We conclude that:

1. The vergence can be controlled using convolutional networks arranged in a closed loop, for different orientations of the gaze.

2. A strategy for reading-out binocular energy population codes for short-latency disparity-vergence eye movements can be devised. Specific features are: (i) wide working range with a reduced number of resource (single scale), (ii) linear servos with fast reactions and precision.

In general, we can take full advantage of the flexibility and adaptability of distributed computing to specialize disparity detectors for vergence control and depth vision.

On this ground, further generalization of the network paradigm will be explored, also with the aim of including (i) dynamic (*i.e.*, spatiotemporal) disparity tuning, and (ii) attentional

signals (based on object properties) that might guide intentional exploration of the selected object.

The results described in this deliverable have been partially presented at ESANN'09, ICVS'09, and submitted to ISABEL'09.

2 Introduction

Experimental evidences show that, although depth perception and vergence eye movements are based on the activity of complex cells of the primary visual cortex, the brain adopts specific and separate mechanisms to combine binocular information and carry out the two distinct tasks. Vergence control models that are based on a distributed population of disparity detectors, usually require first the computation of the disparity map, thus limiting the functionality of the vergence system inside the sensitivity range of the population of cells specialized for depth perception. For the control of vergence larger disparities have to be discriminated while keeping a good accuracy around the fixation point for allowing finer refinement and achieving stable fixations. Thus, alternative strategies might be employed. In this work, we developed models that combine the population responses without taking a decision, but extracting, directly from the population responses, a disparity-vergence response that allows us to nullify the disparity in the fovea, even if the stimulus presented is far beyond the disparity sensitivity range. The disparity-vergence response is obtained by a weighted combination of the population response. First, the weights were computed in order to obtain desired set disparity-vergence responses on which to base a 'dual-mode' vergence control mechanism; then the weights were directly learned from examples of the desired vergence behaviour. We tested the proposed model in a virtual environment achieving stable fixation and small response time to a wide range of disparities. The vergence movements produced are able bring and to keep the fixation point both on a steady and on a moving stimulus. Section 3 and Section 4, respectively, report on the basic population network of disparity detectors and the proposed 'dual-mode' strategy for binocular vergence, devised by UG. Section 5 reports on the two networks (linear and convolutional) developed by K.U.Leuven to learn disparity-vergence behaviours on the basis of the population responses.

3 Distributed representation of binocular disparity

3.1 Computational theory

3.1.1 Multichannel band-pass representation of the visual signal

An efficient (internal) representation is necessary to guarantee all potential visual information can be made available for higher level analysis. At an early level, feature detection occurs through initial local *quantitative* measurements of basic image properties (*e.g.*, edge, bar, orientation, movement, binocular disparity, colour) referable to spatial differential structure of the image luminance and its temporal evolution (cf. linear cortical cell responses). Later stages in vision can make use of these initial measurements by combining them in various ways, to

come up with categorical *qualitative* descriptors, in which information is used in a non-local way to formulate more global spatial and temporal predictions. The receptive fields of the cells in the primary visual cortex have been interpreted as fuzzy differential operators (or local *jets* [7]) that provide regularized partial derivatives of the image luminance in the neighborhood of a given point $\mathbf{x} = (x, y)$, along different directions and at several levels of resolution, simultaneously. Given the 2D nature of the visual signal, the spatial direction of the derivative (*i.e.*, the orientation of the corresponding local filter) is an important "parameter". Within a local jet, the directionally biased receptive fields are represented by a set of similar filter profiles that merely differ in orientation.

Alternatively, considering the space/spatial-frequency duality [8], the local jets can be described through a set of independent spatial-frequency channels, which are selectively sensitive to a different limited range of spatial frequencies. These spatial-frequency channels are equally apt as the spatial ones. From this perspective, it is formally possible to derive, on a local basis, a complete harmonic representation (phase, energy/amplitude, and orientation, for any frequency channel) of any visual stimulus, by defining the associated analytic signal in a combined space-frequency domain through filtering operations with complex-valued band-pass kernels. Formally, due to the impossibility of a direct definition of the analytic signal in two dimensions, a 2D spatial frequency filtering would require an association between spatial frequency and orientation channels. Accordingly, for each orientation channel θ, an image $I(\mathbf{x})$ is filtered with a complex-valued filter:

$$f_A^\theta(\mathbf{x}) = f^\theta(\mathbf{x}) - j f_\mathcal{H}^\theta(\mathbf{x}) \tag{1}$$

where $f_\mathcal{H}^\theta(\mathbf{x})$ is the Hilbert transform of $f^\theta(\mathbf{x})$ with respect to the axis orthogonal to the filter's orientation. This results in a complex-valued *analytic image*:

$$Q_A^\theta(\mathbf{x}) = I * f_A^\theta(\mathbf{x}) = C_\theta(\mathbf{x}) + j S_\theta(\mathbf{x}) \ , \tag{2}$$

where $C_\theta(\mathbf{x})$ and $S_\theta(\mathbf{x})$ denote the responses of the quadrature filter pair. For each spatial location, the amplitude $\rho_\theta = \sqrt{C_\theta^2 + S_\theta^2}$ and the phase $\phi_\theta = \arctan(S_\theta / C_\theta)$ envelopes measure the harmonic information content in a limited range of frequencies and orientations to which the channel is tuned.

In the harmonic space, it is in general an important requirement to have both the spatial width of the filters and the spatial frequency bandwidth small, so that good localization and good approximation of the harmonic information is realized simultaneously. Gabor functions reaching the maximal joint resolution in space and spatial frequency domains are specifically suitable for this purpose and are extensively used in computational vision [8]. Different band-pass filters have been proposed as an alternative to Gabor functions, on the basis of specific properties of the basis functions [9, 10, 11, 12, 13, 14, 15, 16], or according to theoretical and practical considerations of the whole space-frequency transform [17, 18, 19, 20, 21, 22]. A detailed comparison of the different filters evades the scope of this report and numerous comparative reviews can be already found in the literature (*e.g.*, see [23] [24] [25]).

We have considered a discrete set of oriented Gabor filters with different angles θ. To generate a filter with orientation θ (measured from the positive horizontal axis), we can rotate a vertically oriented filter by $\theta - 90^\circ$ with respect to the filter's center (positive angle means counterclockwise rotation):

$$g(\mathbf{x},\theta,\psi) = \eta \cdot \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{x_\theta^2}{2\sigma_x^2} - \frac{y_\theta^2}{2\sigma_y^2}\right) \mathrm{cis}(k_0 x_\theta + \psi) \tag{3}$$

with

$$\begin{cases} x_\theta = x\cos(\theta - 90^\circ) + y\sin(\theta - 90^\circ) \\ y_\theta = -x\sin(\theta - 90^\circ) + y\cos(\theta - 90^\circ) \end{cases}$$

k_0 denotes the *radial peak frequency*, ψ relates to the filter symmetry, and σ's relates to the spatial filter extension, and $cis(\circ)$ is intended to be $\cos(\circ) + j\sin(\circ)$. The parameter η is a proper normalization constant (*e.g.*, chosen to the unitary maximum condition or to the unitary energy condition). Equivalently, the set of Gabor filters can be defined by a quadratic form as:

$$g(\mathbf{x},\theta,\psi) = \eta \cdot \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x}\right) \mathrm{cis}(\mathbf{k}_0^T\mathbf{x} + \psi) \tag{4}$$

where $\mathbf{k}_0 = (k_0\sin\theta, -k_0\cos\theta)^T$ is the oriented spatial frequency vector[1], and the matrix $\mathbf{A}$ can be derived from a diagonal matrix $\mathbf{D}$ (corresponding to a vertically oriented Gabor filter) by multiplication with the rotation matrix $\mathbf{\Theta}$:

$$\mathbf{A} = \mathbf{\Theta}^T\mathbf{D}\mathbf{\Theta} = \begin{pmatrix} \sin\theta & \cos\theta \\ -\cos\theta & \sin\theta \end{pmatrix} \begin{pmatrix} \sigma_x^{-2} & 0 \\ 0 & \sigma_y^{-2} \end{pmatrix} \begin{pmatrix} \sin\theta & -\cos\theta \\ \cos\theta & \sin\theta \end{pmatrix}. \tag{5}$$

It is worth noting that the peak radial frequency k_0 and the width σ_x of the Gaussian envelope in the Gabor function are linked by the constant relative bandwidth factor β (in octave)[2] as:

$$\sigma_x = \frac{1}{k_0}\left(\frac{2^\beta + 1}{2^\beta - 1}\right). \tag{6}$$

Typically, β is chosen around 1 ($\beta \in [0.8, 1.2]$). The relative bandwidth constancy yields self-similar filters across the scales: filters with different radial peak frequencies, but identical orientation angle are simply geometrically scaled version of each other. The aspect ratio σ_x/σ_y normally takes values between 0.25 and 1 and, together with the radial peak frequency, defines the orientation bandwidth of the filter[3]. In the following, to bind the orientation bandwidth of

[1]The orientation of the Gabor filter in space and the orientation of the bandpass channel in the frequency domain are related by $\theta = \arg(\mathbf{k}_0) + \frac{\pi}{2}$.

[2]The relative bandwidth of a Gabor filter is defined as

$$\beta = \log_2\left(\frac{k_0 + \Delta k/2}{k_0 - \Delta k/2}\right) = \log_2\left(\frac{k_0\sigma_x + 1}{k_0\sigma_x - 1}\right)$$

when one chooses the cut-off frequency at one-standard-deviation of the amplitude spectrum of the Gabor function ($1/\sigma_x$) to define the absolute bandwidth Δk.

[3]The orientation bandwidth is the angle between two lines that pass through the frequency origin and are tangent to the one-standard-deviation contour of the amplitude spectrum of the Gabor function. It is given by

$$B_\theta = \arctan\left(\frac{2^\beta - 1}{2^\beta + 1}\right).$$

the filter to the presence of the sinusoidal term only, we fix the aspect ratio to 1 (*i.e.*, $\sigma_x = \sigma_y = \sigma$).

The values of all the design parameters have been chosen to have a good coverage of the space-frequency domain, to guarantee a uniform orientation coverage and to keep the spatial support to a minimum, in order to cut down the computational cost. Therefore, we determined the smallest filter on the basis of the highest allowable frequency without aliasing, and we doubled the sampling when the model analysis requires a higher precision in the filter's profile (or, from a different perspective, a larger spatial support in pixels). [Note: this design strategy reveals itself particularly effective for economic multi-scale analysis through pyramidal techniques [26]. Yet, for all the simulations conducted in this work we considered a single scale, only]. Accordingly, we fixed the maximum radial peak frequency (k_0) by considering the Nyquist condition and a constant relative bandwidth β around one octave, that allows us to cover the frequency domain without loss of information. The result was a minimal 11×11 filter mask capable of resolving sub-pixel phase differences. To satisfy the quadrature requirement all the even symmetric filters have been "corrected" to cancel the DC sensitivity. The filters have been expressed as sums of x-y separable functions to implement separate 1D convolutions instead of 2D convolutions in a similar way that [27], with a consequent further drop of the computational burden. For a detailed description of the filters used, see the Appendix.

3.1.2 Phase-based disparity detection

Depth perception derives from the differences in the positions of corresponding points in the stereo image pair projected on the two retinas of a binocular system. When the camera axes are parallel, on the basis of a local approximation of the Fourier Shift Theorem, the phase-based stereopsis defines the disparity $\delta(\mathbf{x})$ as the one-dimensional (1D) shift necessary to align, along the direction of the horizontal epipolar lines, the phase values of bandpass filtered versions of the stereo image pair $I^R(\mathbf{x})$ and $I^L[\mathbf{x} + \delta(\mathbf{x})]$ [28]. In general, this type of local measurement of the phase results stable, and a quasilinear behaviour of the phase vs. space is observed over relatively large spatial extents, except around singular points where the amplitudes $\rho(\mathbf{x})$ vanishes and the phase becomes unreliable [29]. This property of the phase signal yields good predictions of binocular disparity by

$$\delta(\mathbf{x}) = \frac{\lfloor \phi^L(\mathbf{x}) - \phi^R(\mathbf{x}) \rfloor_{2\pi}}{k(\mathbf{x})} = \frac{\lfloor \Delta\phi(\mathbf{x}) \rfloor_{2\pi}}{k(\mathbf{x})}, \tag{7}$$

where ϕ^L and ϕ^R are the local phase in the left and right image, respectively, and $k(\mathbf{x})$ is the average instantaneous frequency of the bandpass signal, measured by using the phase derivative ϕ_x from the left and right filter outputs:

$$k(\mathbf{x}) = \frac{\phi_x^L(\mathbf{x}) + \phi_x^R(\mathbf{x})}{2}. \tag{8}$$

As a consequence of the linear phase model, the instantaneous frequency is generally constant and close to the tuning frequency of the filter ($\phi_x \simeq k_0$), except near singularities where abrupt frequency changes occur as a function of spatial position. Therefore, a disparity estimate at a point $\mathbf{x}$ is accepted only if $|\phi_x - k_0| < k_0\mu$, where μ is a proper threshold [29].

Equivalently, the principal part of the interocular phase difference necessary to estimate the binocular disparity can be obtained directly, without explicit manipulation of the left and right phase and thereby without incurring the 'wrapping' effects on the resulting disparity map [30] (see also [31, 32]):

$$\begin{aligned} \lfloor \Delta\phi \rfloor_{2\pi} &= \arg(Q^L Q^{*R}) && (9)\\ &= \operatorname{atan2}\left(\operatorname{Im}(Q^L Q^{*R}), \operatorname{Re}(Q^L Q^{*R})\right) && (10)\\ &= \operatorname{atan2}\left(C^R S^L - C^L S^R, C^L C^R + S^L S^R\right) && (11) \end{aligned}$$

where $Q^L = Q^L(\mathbf{x}) = I^L * g(\mathbf{x}, 0°, \psi)$, $Q^R = Q^R(\mathbf{x}) = I^R * g(\mathbf{x}, 0°, \psi)$ and Q^* denotes complex conjugate of Q.

When the camera axes are moving freely, as it occurs in a binocular active vision system, stereopsis cannot longer be considered a 1D problem and the disparities can be both *horizontal* and *vertical*. Therefore, the 1D phase difference approach must be extended to the 2D case.

Still relying upon the local approximation of the Fourier Shift Theorem, the 2D local vector disparity $\boldsymbol{\delta}(\mathbf{x})$ between the left and right images can be related/detected as a phase shift $\mathbf{k}^T(\mathbf{x})\boldsymbol{\delta}(\mathbf{x})$ in the local spectrum, where $\mathbf{k}(\mathbf{x})$ is the local (*i.e.*, instantaneous) frequency vector defined as the phase gradient:

$$\mathbf{k}(\mathbf{x}) = \nabla\phi(\mathbf{x}) = \left(\frac{\partial\phi(x,y)}{\partial x}, \frac{\partial\phi(x,y)}{\partial y}\right)^T \tag{12}$$

with

$$\phi(\mathbf{x}) = \frac{\phi^L(\mathbf{x}) + \phi^R(\mathbf{x})}{2}.$$

Given the 1D character of both the local phase and the instantaneous frequency, their measures strictly depend on the choice of one reference orientation axis, thus preventing the determination of the full disparity vector by a punctual single-channel measurement. We will see that only the projected disparity component on the direction orthogonal to the dominant local orientation of the filtered image can be detected.

Let us distinguish two cases. When the image (stimulus) structure is intrinsically 1D, with a dominant orientation θ_s (let us think of an oriented edge or of an oriented grating with frequency vector $\mathbf{k}_s = (k_s \sin\theta_s, k_s \cos\theta_s)^T$, as extreme cases), the aperture problem [33] restricts detectable disparity to the direction orthogonal to the edge (*i.e.*, to the direction of the dominant frequency vector $\mathbf{k}_s$):

$$\boldsymbol{\delta}_{\theta_s}(\mathbf{x}) = \frac{\mathbf{k}_s}{k_s}\frac{\lfloor \Delta\phi_{\theta_s}(\mathbf{x}) \rfloor_{2\pi}}{k(\mathbf{x})} \simeq \frac{\mathbf{k}_s}{k_s}\frac{\lfloor \Delta\phi_{\theta_s}(\mathbf{x}) \rfloor_{2\pi}}{k_s} \tag{13}$$

where $k(\mathbf{x})$ is the magnitude of the instantaneous frequency. That is, only the projection $\boldsymbol{\delta}_{\theta_s}$ of the disparity $\boldsymbol{\delta}$ onto the direction of the stimulus frequency $\mathbf{k}_s$ is observed. A spatial disparity in a direction orthogonal to $\mathbf{k}_s$ cannot be measured. For an intrinsic 1D image structure, indeed, the spectrum energy is confined within a very narrow bandwidth and it is gathered by the bandwidth (Δk, B_θ) of a single activated channel. This is a realistic assumption for a relatively large

number of orientation channels. Moreover, in this condition, when the dominant frequency of the stimulus $\mathbf{k}_s$ is unknown, it can be approximated by k_0, and thus Eq. (13) becomes:

$$\boldsymbol{\delta}_{\theta_s}(\mathbf{x}) \sim \frac{\mathbf{k}_0}{k_0} \frac{\lfloor \Delta\phi_{\theta_s}(\mathbf{x}) \rfloor_{2\pi}}{k_0}. \tag{14}$$

When the image structure is intrinsically 2D (let us think of a rich texture or a white noise, as an extreme case), the visual signal has local frequency components in more than one direction and the dominant direction is given by the orientation of the Gabor filter. Similarly, the only detectable disparity by a band-pass oriented channel ($\Delta k, B_\theta$) is the one orthogonal to the filter's orientation θ, *i.e.*, the projection in the direction of the filter's frequency:

$$\boldsymbol{\delta}_{\theta}(\mathbf{x}) = \frac{\mathbf{k}_0}{k_0} \frac{\lfloor \Delta\phi_{\theta}(\mathbf{x}) \rfloor_{2\pi}}{k(\mathbf{x})}. \tag{15}$$

Again, $k(\mathbf{x})$ can be derived by Eq. (12) or approximated by the peak frequency of the Gabor filter $\mathbf{k}_0$.

By considering the whole set of oriented filters, we can derive the projected disparities in the directions of all the frequency components of the multi-channel band-pass representation, and obtain the full disparity vector by intersection of constraints [3], thus solving the aperture problem. Without measurement errors, the vector disparity determined by each orientation channel consists of projection $\boldsymbol{\delta}_\theta(\mathbf{x})$ in $\mathbf{k}_0$-direction and unknown orthogonal component (see Fig. 1). The full disparity vector $\boldsymbol{\delta}(\mathbf{x})$ can be recovered from at least two projections $\boldsymbol{\delta}_\theta(\mathbf{x})$, which are not linearly dependent. The end points of the vectors $\boldsymbol{\delta}_\theta(\mathbf{x})$ for fixed $\mathbf{k}_0$ are located on a circle through the origin and the end point of $\boldsymbol{\delta}_\theta(\mathbf{x})$. Taking into account measurement errors of $\Delta\phi_\theta$ and , the redundancy of more than two projections can be used to minimize the mean square error for $\boldsymbol{\delta}(\mathbf{x})$:

$$\boldsymbol{\delta}(\mathbf{x}) = \underset{\boldsymbol{\delta}(\mathbf{x})}{\operatorname{argmin}} \sum_\theta c_\theta(\mathbf{x}) \left(\delta_\theta(\mathbf{x}) - \frac{\mathbf{k}_0^T}{k_0} \boldsymbol{\delta}(\mathbf{x}) \right)^2. \tag{16}$$

where the coefficient $c_\theta(\mathbf{x}) = 1$ when the component disparity along direction θ for pixel $\mathbf{x}$ is a *valid* (*i.e.*reliable) component on the basis of a confidence measure, and is null otherwise. In this way, the influence of erroneous filter responses is reduced.

3.2 Distributed models

The phase-based disparity estimation approach presented in section 3.1 implies *explicit* measurements, for each spatial orientation channel θ (and for any given scale) of the local phase difference $\Delta\phi$ in the image pairs, from which we obtain the *direct* measure of the binocular disparity component δ_θ. Similarly, we can consider a distributed approach in which the binocular disparity δ is never measured but implicitly coded by the population activity of cells that act as "disparity detectors" - over a proper range of disparity values. Such models are inspired by the experimental evidences on how the brain and, specifically, the primary visual cortex (V1), implements early mechanisms for stereopsis. Using such a distributed code it is possible to achieve a very flexible and robust representation of binocular disparity for each spatial position in the retinal image.

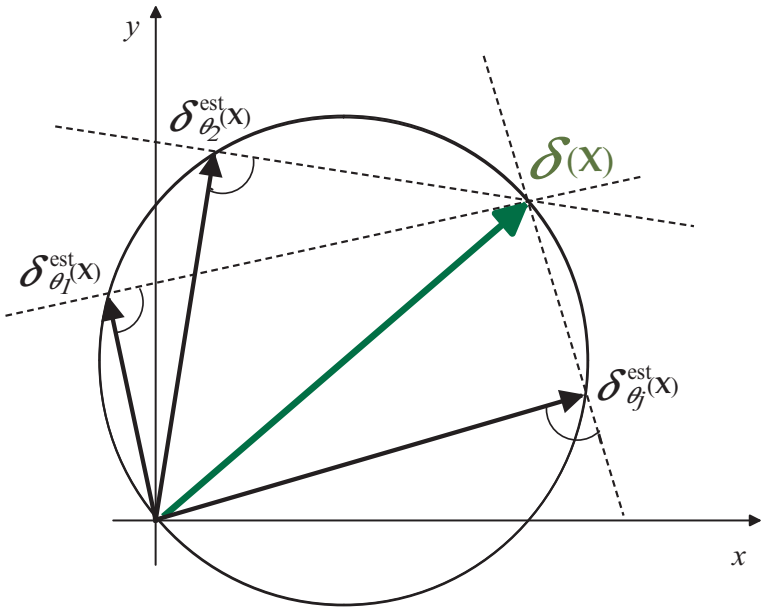


Figure 1: Recovery of the 2D disparity vector. By construction, the end points of all the (correct) estimates δ_θ^{est} of the disparity component with respect to the orientation θ are located on a circle through the origin. The true full disparity is the longest vector whose end point lies on the circle.

3.2.1 Phase-shift and binocular energy models

An abundance of neurophysiological evidences report that the cortical cells' sensitivity to binocular disparity is related to interocular phase shifts in the Gabor-like receptive fields of V1 simple cells ([28][34][35][36][37][38]). It is worth noting that models based on a difference in the position of the left and right RFs (position-shift models) or hybrid approaches have been proposed (we will discuss the consequences of this model extentions in the Section). The phase-shift model posits that the center of the left and right eye RFs coincides, but the arrangements of the RF subregions are different. Formally, the response of a simple cell with RF center in $\boldsymbol{x}$ and oriented along θ, can be written as:

$$ {}^{\theta}_{\Delta\psi}r_{s,\psi_0}(\mathbf{x}) = I^L * h^L(\mathbf{x};\theta,\psi_0+\psi^L) + I^R * h^R(\mathbf{x};\theta,\psi_0+\psi^R) \tag{17}$$

where

$$ h(\mathbf{x}) = h(\mathbf{x};\theta,\psi) = \eta \exp\left(-\frac{1}{2\sigma^2}\mathbf{x}^T\mathbf{x}\right)\cos(\mathbf{k}_0^T\mathbf{x}+\psi) \tag{18}$$

is a real-valued RF (cf Eq. (4)), ψ_0 is a "central" value of the phase of the RF, and ψ^L and ψ^R are the phases that characterize the binocular RF profile.

In order to make the disparity tuning independent of the monocular local Fourier phase of the images (but only on the interocular phase difference), binocular energy complex cells play the role. Such "energy units" are defined as the squared sum of a quadrature pair of simple cells (see Fig. 2) and their response is defined as:

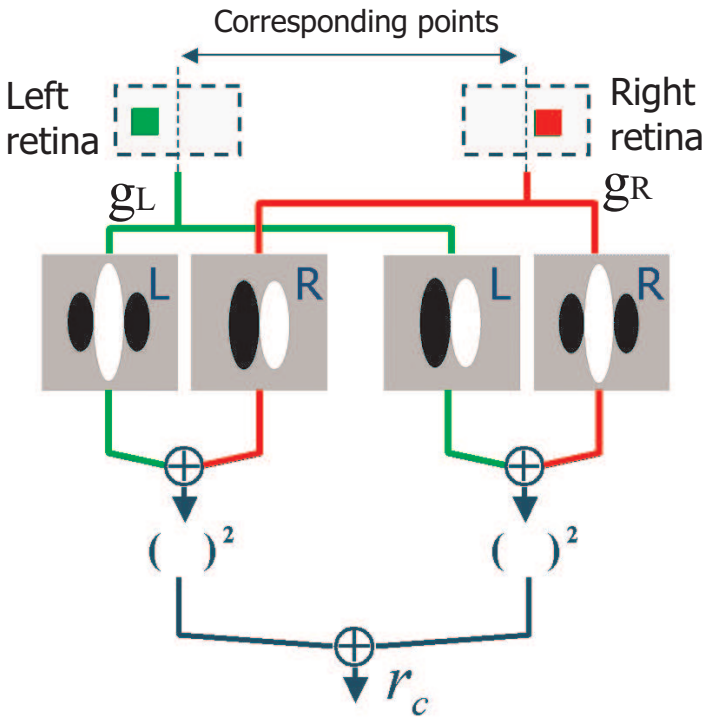


Figure 2: The complex cell response is constructed as the squared sum of a quadrature pair of simple cells. The green and red pathways relate to the monocular "quadrature pair" of simple cell RFs, g^L and g^R, respectively.

$$\overset{\theta}{}_{\Delta\psi} r_c(\mathbf{x}) = \overset{\theta}{}_{\Delta\psi}\, r^2_{s,0}(\mathbf{x}) + \overset{\theta}{}_{\Delta\psi}\, r^2_{s,\pi/2}(\mathbf{x}) \tag{19}$$

Linking phase-based and energy-based models For any fixed orientation, if we characterize a "quadrature pair" of simple cells by a complex-valued RF (cf Eq. (4)):

$$\mathrm{h}(\mathbf{x}) \stackrel{\triangle}{=} h_C(\mathbf{x}) + j h_S(\mathbf{x}) = g(\mathbf{x};\psi) \tag{20}$$

then we can write the expression of the response of the "quadrature pair" as:

$$\begin{aligned} Q(\mathbf{x}) &= I^L * g^L(\mathbf{x}) + I^R * g^R(\mathbf{x}) = I^L * g(\mathbf{x}) e^{j\psi^L} + I^R * g(\mathbf{x}) e^{j\psi^R} = \\ &= Q^L(\mathbf{x}) e^{j\psi^L} + Q^R(\mathbf{x}) e^{j\psi^R}. \end{aligned}$$

The response of a complex "energy" cell is then

$$\begin{aligned} \overset{\theta}{}_{\Delta\psi} r_c(\mathbf{x}) &= \left| \overset{\theta}{}_{\Delta\psi} r_{s,0}(\mathbf{x}) + \overset{\theta}{}_{\Delta\psi}\, r_{s,\pi/2}(\mathbf{x}) \right|^2 = \left| Q^L(\mathbf{x}) e^{j\psi^L} + Q^R(\mathbf{x}) e^{j\psi^R} \right|^2 = \\ &= \left| e^{j\psi^L} \left(Q^L(\mathbf{x}) + Q^R(\mathbf{x}) e^{j\Delta\psi} \right) \right|^2 = \left| Q^L(\mathbf{x}) + Q^R(\mathbf{x}) e^{j\Delta\psi} \right|^2 \end{aligned} \tag{21}$$

where $\Delta\psi = \psi^L - \psi^R$. Therefore, complex cells' responses depend on $\Delta\psi$ only, instead of on ψ^L and ψ^R individually.

Eq. (21) formally establishes the equivalence between phase-based techniques and energy-based models [39]. Indeed, the maximum of r_c responses is obtained when the two phasors Q^L and Q^R are aligned in the complex plane, that is when $\Delta\psi$ compensates for the different Fourier phases of the right and left image patches within the cell's RF (cf. [28]).

Notwithstanding the formal equivalence between phase-based techniques and energy-based models, the latters prove themselves more robust to noise and more flexible, since they can intrinsically embed adaptive mechanisms both at coding and decoding levels of the population code. From algebraic and trigonometric manipulation we can derive the tuning curve of the complex cell:

$$\overset{\theta}{_{\Delta\psi}}r_c(\mathbf{x}) = |Q^L(\mathbf{x})|^2 + 2|Q^L(\mathbf{x})Q^{*R}(\mathbf{x})|\cos(\delta^\theta k_0 - \Delta\psi) + |Q^R(\mathbf{x})|^2. \tag{22}$$

Accordingly, the stimulus disparity, along direction θ, to which the cell is tuned is:

$$\delta^\theta_{pref}(\mathbf{x}) = \frac{\lfloor \Delta\psi(\mathbf{x}) \rfloor_{2\pi}}{k_0}. \tag{23}$$

Including position shift: hybrid models The position-shift model posits that there is a population of energy neurons with different receptive field position shifts. Accordingly we can consider a family of binocular energy neurons whose right monocular subfield is shifted by a set of distances d compared to the retinal position of the left monocular subfield. Usually position-shift are used in combination with phase-shift models to overcome the restriction on the maximum disparity detectability stemmed by the fact that the phase shifts are unique only between $-\pi$ and π. These hybrid models posit that there is a population of binocular energy neurons with different RF positions and different RF phase shifts. In the following we will restrict our analysis to phase-shift model only, and we will deserve a model extention for future work.

3.2.2 Characterization of the population of disparity detectors

Coding Disparity information is extracted from a sequence of stereo image pairs by using a distributed cortical architecture that resorts to a population of simple and complex cells. The population is composed of cells sensitive to $N_p \times N_o$ vector disparities $\boldsymbol{\delta} = (\delta_H, \delta_V)$ with N_p magnitude values distributed in the range $[-\Delta, \Delta]$ pixels and along N_o orientations uniformly distributed between 0 and π (see Fig. 3). For each simple cell we can control the ocular dominance of the binocular receptive field $h(\mathbf{x})$, its orientation θ with respect to the horizontal axis and the interocular phase shift $\Delta\psi$ along the rotated axis, which confers to the cell its specific tuning to a disparity $\delta^\theta_{pref} = \Delta\psi_\theta / k_0$, along the direction orthogonal to θ. The spatial frequency k_0 and the spatial envelope are fixed on the basis of the design criteria described in Section 3.1. The complex cell inherits the spatial properties of the simple cells, and its response $r_c^{ij}(\mathbf{x})$ is given by Eq. (21): For each orientation, the population is, in this way, capable of providing reliable disparity estimates in the range between $-\Delta$ and Δ, where $\Delta = \Delta\psi_{max}/k_0$ can be defined as the maximum detectable disparity of the population.

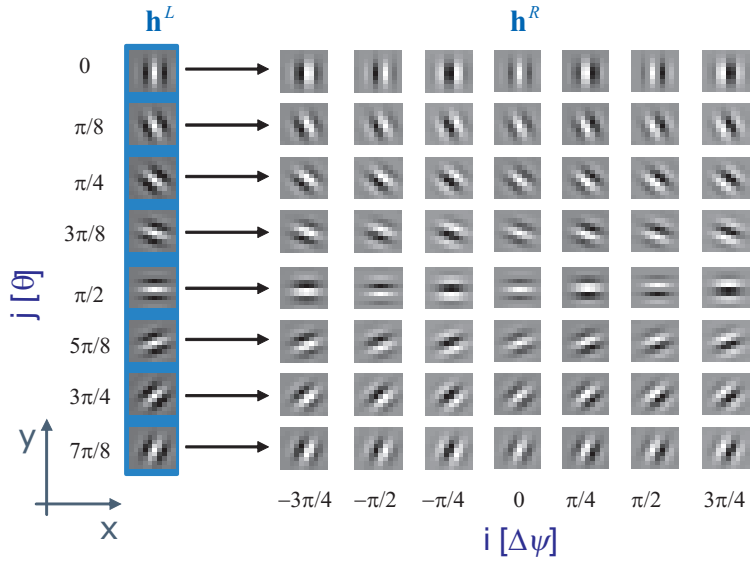


Figure 3: The population of binocular receptive fields for each retinal location.

Fig. 4 shows examples of tuning curves obtained from the population network stimulated with δ_H only, compared to the variety of tuning curves for δ_H, observed experimentally in V1 cortical cells [38].

Decoding Once the disparity along each spatial orientation have been coded by the population activity, it is necessary to read out this information, to obtain a reliable estimate. The decoding strategy, the number of the cells in the population and their distribution are jointly related. To decode the population by a winners-take-all strategy, a large number of cells along each spatial orientation would be necessary, thus increasing the computational cost and the memory occupancy of the approach. To obtain precise feature estimation, while keeping the number of cells as low as possible, thus an affordable computational cost, a *weighted sum* (*i.e.*, a center of gravity) of the responses for each orientation is calculated. The *component disparity* $\delta^{est}_{\theta_j}$ is obtained by:

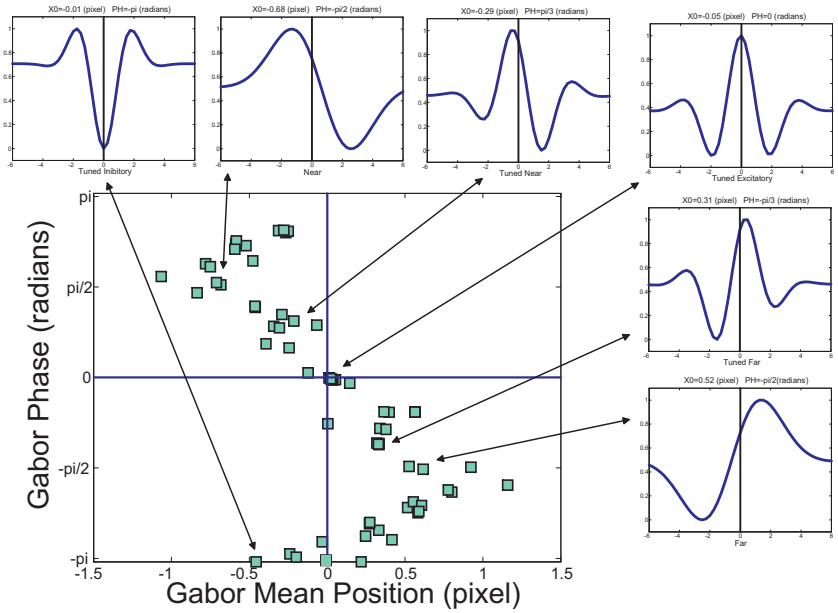
$$\delta^{est}_{\theta_j} = \frac{\sum_{i=1}^{N_p} \frac{\Delta\psi_i}{k_0 \cos\theta_j} r_c^{ij}}{\sum_{i=1}^{N_p} r_c^{ij}} \tag{24}$$

Other decoding methods [40], such as the *maximum likelihood* estimator, have been considered, but the center of gravity of the population activity is the best compromise between simplicity, low computational cost and accuracy of the estimates.

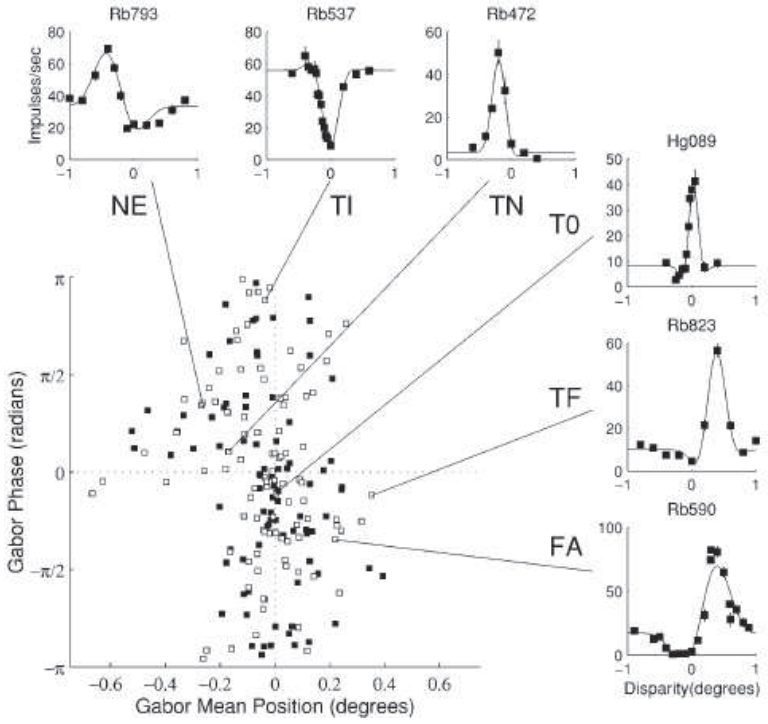
Confidence values, based on local energy, are used to provide a reliability measure for each disparity estimate.

To decode the full (horizontal and vertical) disparity we can still rely on the intersection of constraints (channel interaction) introduced in Section 3.1.2 that combine the population estimates for each orientation channel.

Summarizing, on the basis of these principles, a cortical-like architecture for disparity es-



(a)



(b)

Figure 4: (a) Distribution of the tuning curves obtained from the population network. The distribution has been obtained for $N_p = 7$ and $N_o = 8$. (b) The distribution observed for real V1 cortical cells [38]. The insets represent examples of disparity tuning curves fitted by Gabor function. The model cells' distribution and the tuning profiles closely resemble the experimental ones.

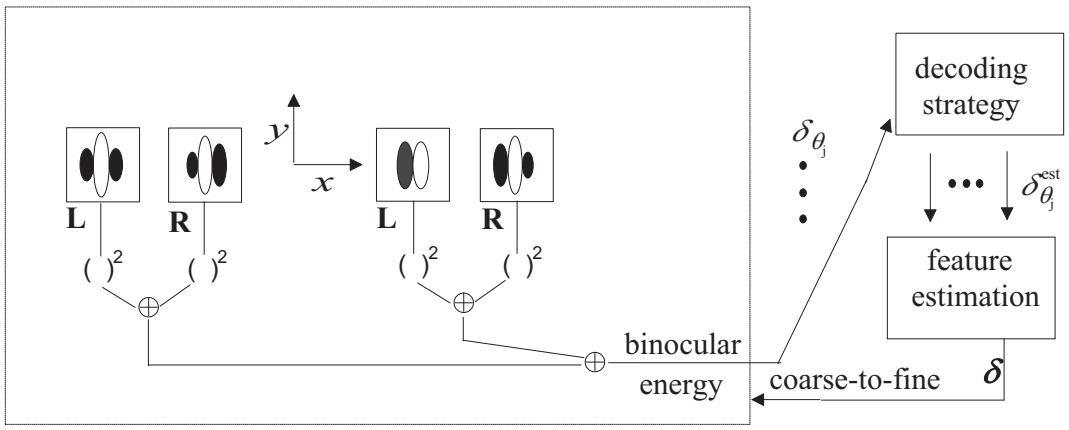


Figure 5: Basic scheme of the neuromorphic architecture for the computation of the 2D disparity.

timation can be devised [41]. The overall scheme of the proposed architecture is shown in Fig. (5). Three distinct levels of processing can be distinguished: (1) the distributed coding of disparity across different orientation channels, (2) the decoding stage for each channel, and (3) the estimation of the full disparity through channel interaction. If one wants to consider several scales, coarse-to-fine strategies can be straightforwardly embodied, *e.g.*, by including in the scheme a refinement loop as re-entrant connections in the filtering stage (see [41] [42]).

Toward a generalized architecture for active stereopsis In active stereopsis, besides handling horizontal and vertical disparities, we have to explicitly consider vergence mechanisms in the processing loop. From this perspective, in the next Section, we address the problem of the refinement of vergence, which does not necessarily implies first a refinement of the estimation of the disparity map. Indeed, experimental evidences (see *e.g.*, [43] [44] [45]) pointed out that mechanisms guiding eye movements are in general different from those supporting depth perception. We will see that, by specializing disparity detectors for vergence control, we can obtain linear servos with fast reaction and precision that work over a wide range of disparities with a reduced number of resources single scale).

4 Strategies for vergence without explicit calculation of disparity

4.1 Reading binocular energy population codes for short-latency disparity-vergence eye movements

As described in Section 3, the population of complex cells are, by construction, tuned to oriented disparities, *i.e.*, jointly tuned to horizontal (δ_H) and vertical disparities (δ_V). In general, indeed, the retinal disparity is a two-dimensional (2D) feature and the full decoding of the population response would require the proper solution of the aperture problem [33]. This can be achieved, by example, through the intersection of the constraints provided by the different

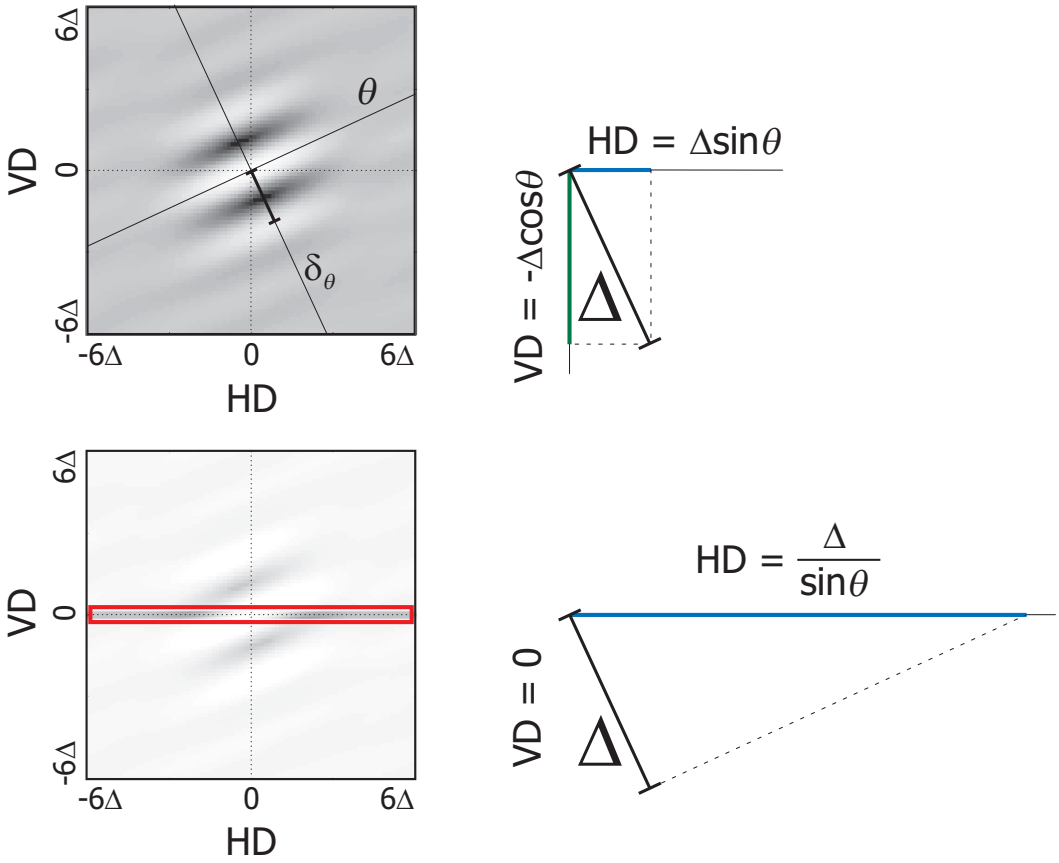


Figure 6: Each complex cell is, by construction, tuned to an oriented disparity δ_θ, *i.e.*, each cell is jointly tuned to horizontal (HD) and vertical (VD) disparities. (Top): For each oriented disparity, its contribution to the HD and VD is calculated by projections on the horizontal and vertical lines. (Bottom): By assuming VD$= 0$, the orientation of the RF is used as a degree of freedom to extend the sensitivity range of the cell to horizontal disparity stimuli (HD).

orientation channels (cf. [3]). If one proceedes in such a way, that is by recovering the full disparity vector, the disparity detectability range would still be limited to $\pm\Delta$, and the horizontal (vertical) component of the full disparity vector will then used for the control of horizontal (vertical) vergence. Unless one uses computationally expensive multiscale techniques for widening the disparity detectability range, this approach would considerably limit the working range of the vergence control. As for the control of vergence, larger disparities have to be discriminated while keeping a good accuracy around the fixation point for allowing finer refinement and achieving stable fixations, alternative strategies might be employed to gain effective vergence signals directly from the complex cell population responses, without explicit computation of the disparity map. To this end, we can map the 2D disparity feature space into the 1D space of the projected horizontal disparities, where the orientation θ plays the role of a parameter. More precisely, by assuming $\delta_V = 0$, the dimensionality of the problem of disparity estimation reduces to one, and the orientation of the receptive field is used as a degree of freedom to extend the sensitivity range of the cells' population to horizontal disparity stimuli (see Fig. 6).

In this way, each orientation channel has a sensitivity for the horizontal disparity that can be obtained by the projection of the oriented phase difference on the (horizontal) epipolar line in the following way:

$$\delta_H^\theta = \frac{\Delta\psi}{2\pi k_0 cos\theta} \tag{25}$$

Fig.7a shows the horizontal disparity tuning curves obtained of the population for different orientations of the receptive fields. To decode the horizontal disparity at a specific image point, the whole activity of the population of cells, with receptive fields centered in that location, is considered. By using a center-of-mass decoding strategy, the estimated horizontal disparity δ_H^{est} is obtained by:

$$\delta_H^{est} = \frac{\sum_{i=1}^{N_p}\sum_{j=1}^{N_o}\frac{\Delta\psi_i}{2\pi k_0\cos\theta_j}r_c^{ij}}{\sum_{i=1}^{N_p}\sum_{j=1}^{N_o}r_c^{ij}} \tag{26}$$

where r_c^{ij} denotes the response of the complex cell characterized by the i-th phase difference and by the j-th orientation. The dashed line plots in Fig.7b-c show the resulting disparity curves obtained by population decoding. The estimate of the disparity can be considered correct when the stimulus disparity is within $\pm\Delta$.

By analyzing the tuning curves of the population (see Fig.7a) we observe that the peak sensitivity of cells that belong to a single orientation channel is uniformly distributed in a range that increases with the orientation angle θ of the receptive field, as the horizontal projection of the frequency of the Gabor function declines to zero. Thus, applying the center of mass decoding strategy, separately for each orientation, we can obtain j different estimates of the disparity:

$$\delta_{H,\theta_j}^{est} = \frac{\sum_{j=1}^{N_p}\frac{\Delta\psi_i}{2\pi k_0\cos\theta_j}r_c^{ij}}{\sum_{i=1}^{N_p}r_c^{ij}} \tag{27}$$

It is worthy to note that the increase of the sensitivity range, as the orientation of the receptive fields deviates from the vertical, comes at the price of a reduced reliability and accuracy of the measure (as an extreme case, horizontal receptive fields are unable to detect horizontal disparities, *i.e.*, $\delta_H^{\theta=0} \rightarrow \infty$). In any case, the estimate of the disparity can be considered correct in a range around $[-\Delta, \Delta]$, only.

Moreover, since the 1D tuning curves of the population were obtained under the assumption of horizontal disparity only, when the vertical disparity in the images differs from zero, the correctness of estimate of the actual component of the horizontal disparity has to be verified. We observe that (see Fig.7b and Fig.7c, top row), the disparity estimated by the whole population is unaffected by non null vertical disparities, as well as the estimate obtained by the orientation $\theta = 0$ (vertically oriented cells are indeed , by definition, sensitive to horizontal disparity only). On the contrary, the estimated disparity obtained for $\theta \neq 0$ shows a dependence on vertical disparity, that increases with θ (see Fig.7c, middle and bottom row), and leads to a systematic error response.

4.1.1 Control signal extraction

A desired feature of disparity-vergence curves is an odd symmetry with a linear segment passing smoothly through zero disparity, which defines critical servo ranges over which changes

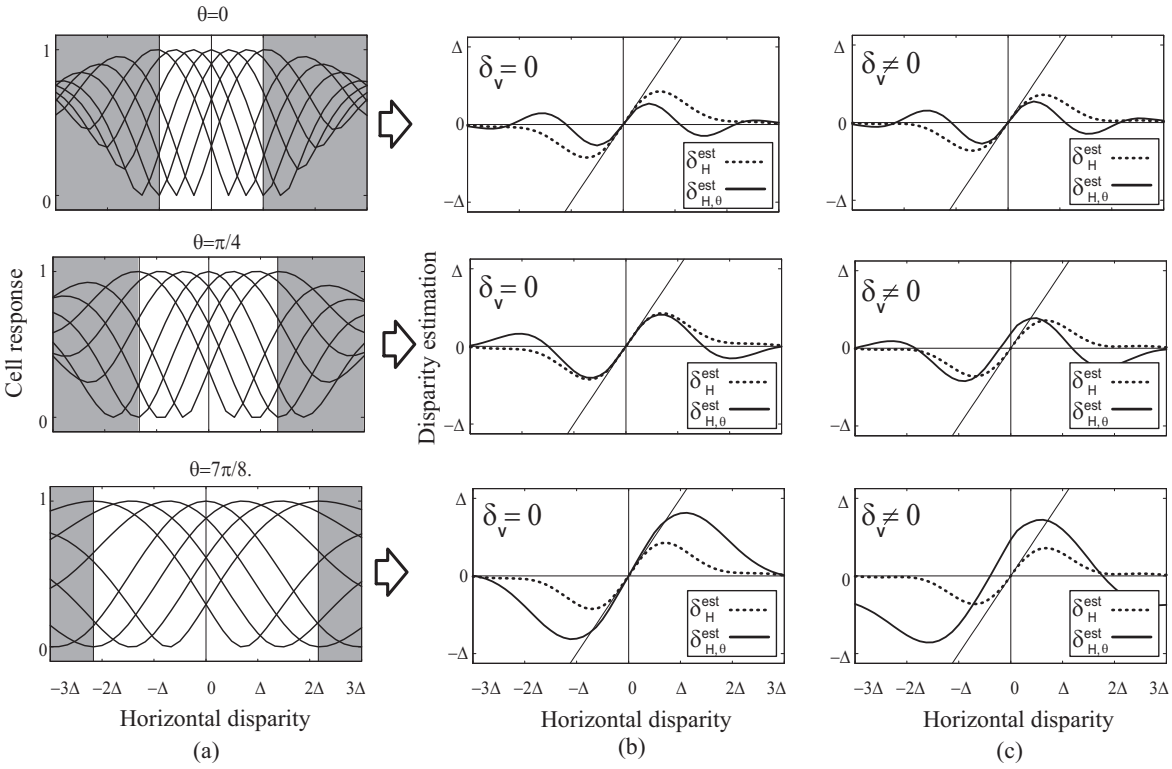


Figure 7: (a) Disparity tuning curves of complex cells at different orientations. (b) Estimated horizontal disparity using single orientation channels in presence of horizontal disparity only ($\delta_V = 0$). (c) Estimated horizontal disparity using single orientation channels in presence of a fixed amount of vertical disparity ($\delta_V \neq 0$). Dashed line plots refer to the horizontal disparity estimates obtained by combining all the orientation channels.

in the stimulus horizontal disparity elicit roughly proportional changes in the amount of horizontal vergence eye movements, $\Delta\alpha = p\delta_H$, where α is the vergence angle. Starting from the estimated disparity curves shown in Fig.7b, we can exploit the responses at different orientations to design linear servos that work outside the reliability range of disparity estimation. Yet, we have to cope with the attendant sensitivity to vertical disparity, which is an undesirable effect that alters the control action. Hence, given a stimulus with horizontal and vertical disparity δ_H and δ_V, we want to combine the population responses in order to extract a vergence control proportional to the δ_H to be reduced, regardless of any possible δ_V. We demonstrate that such disparity vergence response can be approximated by proper weighting of the population cell responses where disparity tuning curves act as basis functions [46]. Due to these considerations, the population responses are combined with two very specific goals: (1) to obtain signals proportional to horizontal disparities, (2) to make these signals be insensitive to the presence of vertical disparities. The disparity vergence response curves r_v^k are obtained by a weighted sum of the complex cell responses (see Fig.9):

$$r_v^k = \sum_{i=1}^{N_p} \sum_{j=1}^{N_o} w_{ij}^k r_c^{ij} \tag{28}$$

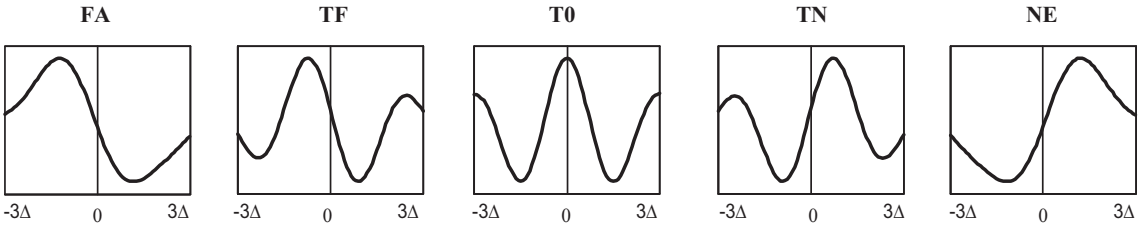


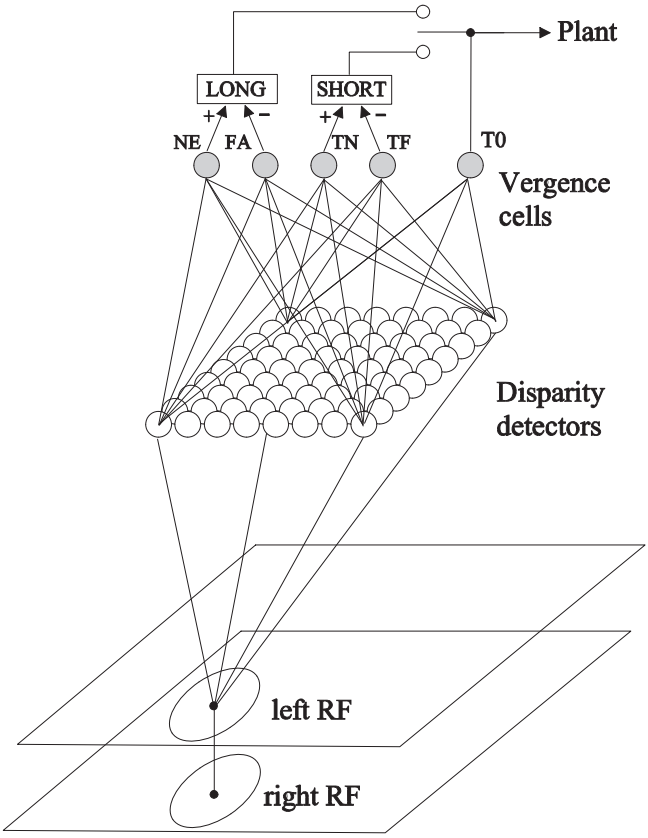
Figure 8: The v_H^k target curves to be approximated by the LMS minimization. Each of them is designed to have a tuning to disparities of different magnitude.

where the index k denotes the different kind of the desired vergence response curves. Referring to a common classification [47] we divide the V1 cells in five categories: near (*NE*) and far (*FA*) dedicated to coarse stereopsis, and tuned near (*TN*), tuned far (*TF*) and tuned zero (*T0*) for fine stereopsis. The weights w_{ij}^k are obtained through a recursive LMS algorithm. From the control point of view, we assume that small values of vertical disparities do not affect the disparity-vergence curves. Moreover, to mildly constraint the solution of the problem and, in the meantime to ensure a good control stability, we pose the VD independence constraint for $\mathrm{HD} \simeq 0$, only. Under this assumption, we can design the disparity-vergence curves that define the visual servos by considering the tuning curves obtained separately for VD=0 and HD=0 (*i.e.*, the orthogonal cross-section of the oriented 2D disparity tuning curves of the binocular energy model). More precisely, the profile of the desired vergence curve v_H^k (see Fig.8) is approximated by a weighted sum of the tuning curves for horizontal disparity $r_c(\delta_H; \theta, \Delta\psi)$.

To gain the insensitivity to vertical disparity we add a constraint term in the minimization formula. This term ensures that the sum of the vertical disparity tuning curves $r^c(\delta_V; \theta, \Delta\psi)$, weighted with the same $\mathbf{w}^k$, approximates v_V^k. To overcome the difficulties of approximating a constant with a combination of a limited number of periodic basis functions, we impose v_V^k to have a profile that is mildly constant as the one that can be obtained by summing the tuning curves all together ($v_V^k = \sum_{i=1}^{N_p} \sum_{j=1}^{N_o} r_c^{ij}(\delta_V)$). Hence, the weights $\mathbf{w}^k$ are obtained by minimizing the following functional:

$$E(\mathbf{w}^k) = \left\| \sum_{i=1}^{N_p} \sum_{j=1}^{N_o} r_c^{ij}(\delta_H) w_{ij}^k - v_H^k \right\|^2 + \lambda \left\| \sum_{i=1}^{N_p} \sum_{j=1}^{N_o} r_c^{ij}(\delta_V)(w_{ij}^k - 1) \right\|^2 \tag{29}$$

where $\lambda > 0$ balances the relevance of the second term over the first. In our simulations we fixed $\lambda = 1$ in order to give the same relevance to both δ_H and δ_V. To test the functionality of the model, at this stage, we used the same kind of stimuli adopted to compute the disparity tuning curves of the cells, so that we expect the disparity vergence tuning curve to be the same we drew from the minimization. The stimuli have a disparity varying in the same range used for the tuning curves, and the control computed has the same shape of the desired curves (Fig.9b). A drawback that arises is that if the image contrast is lowered, disparity vergence tuning curves hold the same shape, but their gain is consequently lowered, with the effect that the speed of the vergence movements is modulated by the image contrast. The estimated disparity does not show this effect because the center of mass decoding strategy means to take a decision on the disparity value, regardless to the contrast of the stimulus (cf. [35]). By analogy with the



(a)

Figure 9: Extraction of the vergence control signals: each location of the left and right image is filtered with a population of diparity detectors whose response r_c is combined with five different families of weights w^k, in order to extract five signals $r^{FA}, r^{TF}, r^{T0}, r^{TN}$ and r^{NE}, tuned to disparities of different magnitudes. These signals are combined in a differential way, io order to extract the LONG and SHORT controls, used to drive the vergence eye movements, while the r^{T0} works as a switch between them.

formula used to decode the disparity, we can introduce the same normalization term to let the system work in the proper way independently of the image contrast:

$$r_v^k = \frac{\sum_{i=1}^{N_p} \sum_{j=1}^{N_o} w_{ij}^k r_c^{ij}}{\sum_{i=1}^{N_p} \sum_{j=1}^{N_o} r_c^{ij}} \tag{30}$$

4.1.2 Signal Choice

With reference to the five categories of the disparity-vergence curves, it is plausible to think that the first two generate the fast and coarse component and the others the slow and fine component of the vergence movements. In practice the fast-coarse control is given by LONG$= r^{NE} - r^{FA}$, while the slow-fine is given by SHORT$= r^{TN} - r^{TF}$ (see Fig.9). The SHORT control signal is

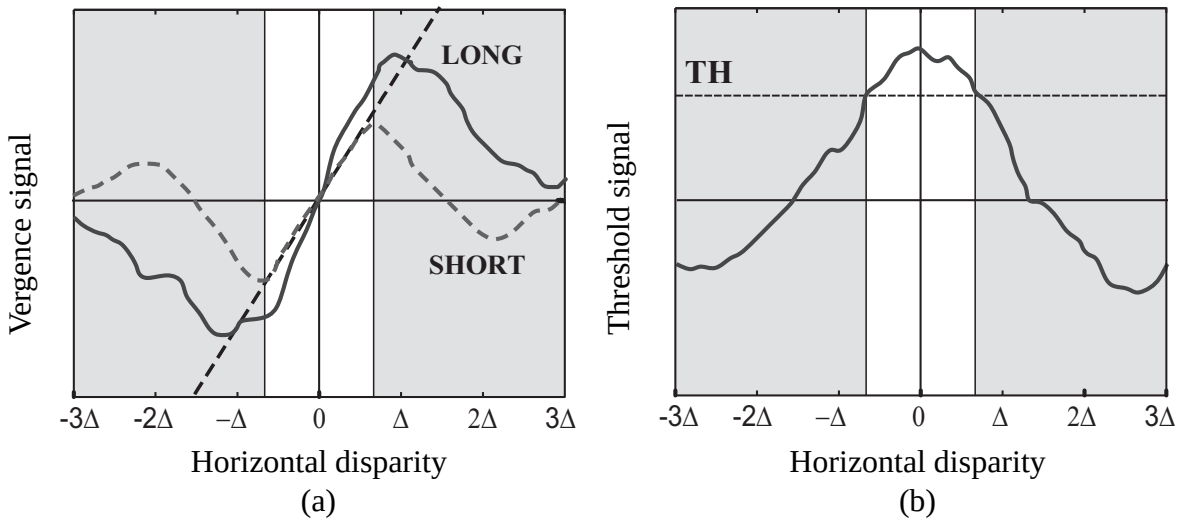


Figure 10: The effective LONG, SHORT (a), and T0 (b) signals computed by the model stimulated with a random dot stereograms (RDS). The SHORT control is able to work in a linear and precise manner for small disparities, while the LONG one works in a coarse but effective way for larger disparities. Since the $T0$ signal is high for small disparities, it is able to act like a switch between the two controls.

designed to proportionally generate, in a small range of disparities, the vergence to be achieved, and allows a precise and stable fixation (Fig.9b). Out of its range of linearity, the SHORT signal decreases and loses efficiency to the point where it changes sign, thus generating a vergence movement opposite to the desired one. On the contrary for small disparities the LONG control signal yields overactive vergence signal that make the system to oscillate, whereas for larger disparities it provides a rapid and effective signal.

The role of the r^{T0} signal, is to act as a switch between the SHORT and the LONG controls. When the binocular disparities are small, r^{T0} is above a proper threshold *TH*, and it enables the SHORT control (see white regions in Fig.9b). On the contrary, for large stimulus disparities, r^{T0} is below the threshold and it enables the LONG control (see grey regions in Fig.9b).

A straightforward but meaningful effect that arises from calculating the SHORT and the LONG controls in a differential way is a strong robustness to noise. If we add a Gaussian white noise to the population response, both the decoding of the disparity and the computation of the r_v^k signals, would be affected. Since the weights $\mathbf{w}^k$ are normalized, it is easy to demonstrate that the noise terms on r^{NE} and r^{FA} cancel each other while differentiating to compute the LONG control, and so it happens for the SHORT one. Simulation results evidenced that, when one adopts the differential SHORT and LONG control signals, the S/N ratio is $\sim 6\mathrm{dB}$ higher than the input S/N ratio for the complex cell responses.

4.2 Effects of vertical disparity

The optimized control we want to obtain from the proposed technique is a control of the horizontal vergence that would be able to yield the same movement for a given horizontal disparity

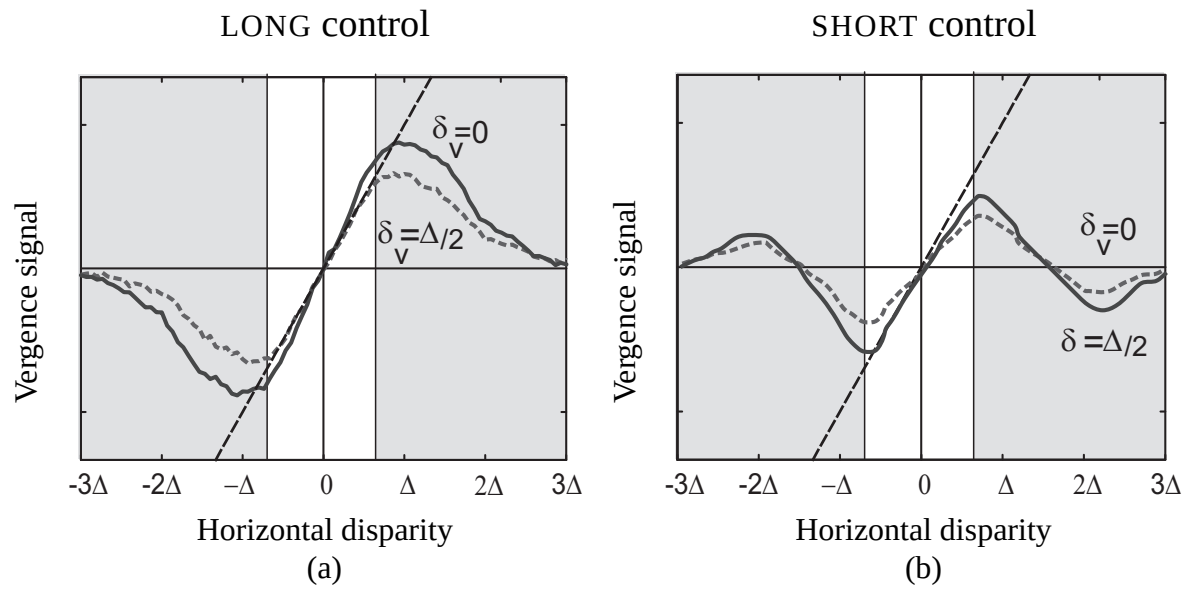


Figure 11: The effective LONG (a), and SHORT (b) signals computed by the model stimulated with a random dot stereograms (RDS) in the absence (solid line) and in presence (dashed line) of a vertical disparity pedestal.

δ_H, without suffering any effect from the vertical disparity δ_V. Indeed if the δ_V constraint is not taken into account in the minimization process used to obtain the weights w (see Eq.29), the resulting control shows a strong dependence on vertical disparity, as it appears evident in the disparity-vergence tuning curves shown in Fig.7 right column. The control loses the zero crossing and its odd symmetry, which are instrumental features to ensure that at the steady state, the eyes fixate on the closest surface along the axis of fixation, not before, nor beyond.

Although, the regularization term we introduced in Eq.29 has the sake of forcing the control to be insensitive to the vertical disparity, simulations with RDSs showed that the behaviour is different from the expected one.

The problem of this approach is due to the fact that the minimization is computed by considering, for every complex cell, its response r_c^{ij} to the horizontal and the vertical disparity only, for the first and the second term of the functional, respectively. As a matter of fact, in the functional in Eq.29, what we consider are the cross-sections for $\delta_H = 0$ and $\delta_V = 0$ of the two-dimensional (2D) tuning profile that characterizes each complex cell. Though, the 2D disparity tuning profile of a binocular energy unit can be oriented by any angle, depending on the orientation channel we consider, and it is separable for $\theta = 0$ and $\theta = \pi/2$ only.

Hence, the problem arises if a vertical disparity pedestal is added to the stimulus, the section of the 2D profile one should consider, is the one at the imposed δ_V. Otherwise, the more the filter is tilted from the vertical and the more the δ_V is, the most the tuning curves change, producing the effect of making the decoding for vergence unreliable.

Thus if the δ_V is set to 0, the control is working in the conditions it is designed for, and its effectiveness is the highest. In Fig.14a we show the evolution in time of the actual horizontal disparity, when the vergence control to correct active. The value of each plot at the first time step is the initial horizontal disparity step we imposed. The system is able to cope with

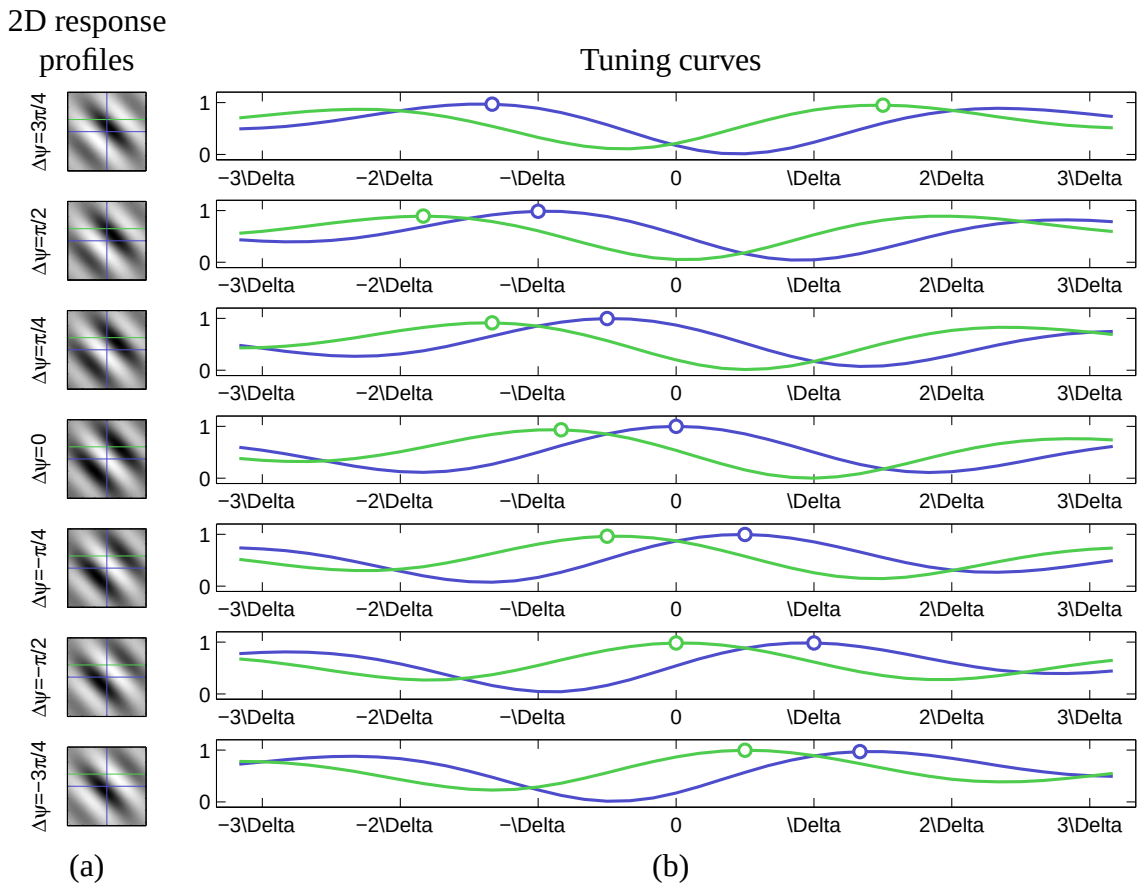


Figure 12: (a) The profile of the response of the complex cell defined by $\theta = \pi/4$, tested with a RDS with δ_H and δ_V ranging from -3Δ and 3Δ. (b) Tuning curves for the same complex cells, taken at different fixed vertical disparity, the blue one is for $\delta_V = 0$ and the green one is for $\delta_V = \Delta$. The empty circles highlight the position of the peak for each curve.

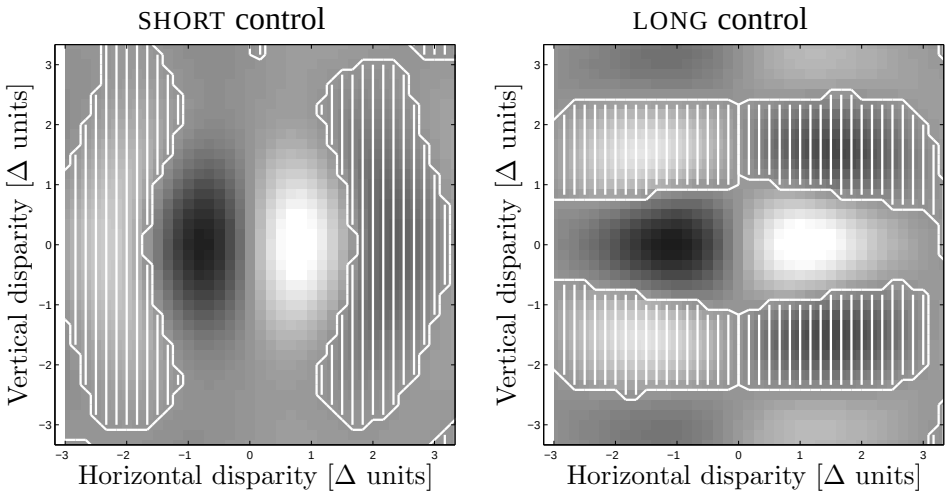


Figure 13: 2D vergence tuning profiles for the SHORT and LONG control mode obtained as weighted sums of the 2D disparity tuning profiles of the complex cells. The weights are derived by Eq.29 and by Eq.30

disparity values ranging from -3Δ to $3\Delta$and the control of vergence reduces to zero the stimulus disparity. Moreover is highlighted when the system relies upon the SHORT control (filled circles) and the LONG control (open circles). As expected, for small disparities the working mode is the former, while for larger disparities is the latter, depending on the threshold of the $T0$ signal (see Fig.10b). At the same way if vertical disparity is small (see Fig.14b), the tuning of the population responses is almost unaffected by δ_V, and the only visible effect is a slight slow down of the vergence control. Increasing the value of δ_V (see Fig.14c-d), besides a more consistent slow down of the control, another drawback is the reducing of the range of δ_H the control is able to cope with. This effect is particularly evident on the LONG mode, because it resort mainly on the cells whose orientation tuning largely deviates from the vertical, thus being more sensitive to δ_V. Fig.13 shows the SHORT and LONG controls obtained as weighted sums of the 2D disparity tuning profiles of the complex cells, and in particular haw the two control are slowed down by δ_V. Moreover the areas where the controls are unreliable is highlighted with white lines.

4.3 Results

4.3.1 Test with Random Dot Stereograms

We tested the proposed model with synthetic stimuli consisting of random dot stereograms (RDS) in which the stereo image pairs are shifted horizontally. Specifically, we applied horizontal disparity steps varying from -3Δ to 3Δ. The model works in a perception-action loop in which the vergence movements are simulated reducing step by step the disparity between the left and right images by an amount proportional to the vergence control, computed both through the estimation of the disparity δ_H^{est} (see 26) and through the vergence signals r_v^k (see 30) Fig.16 shows the percentage of vergence movement accomplished by the two mechanisms

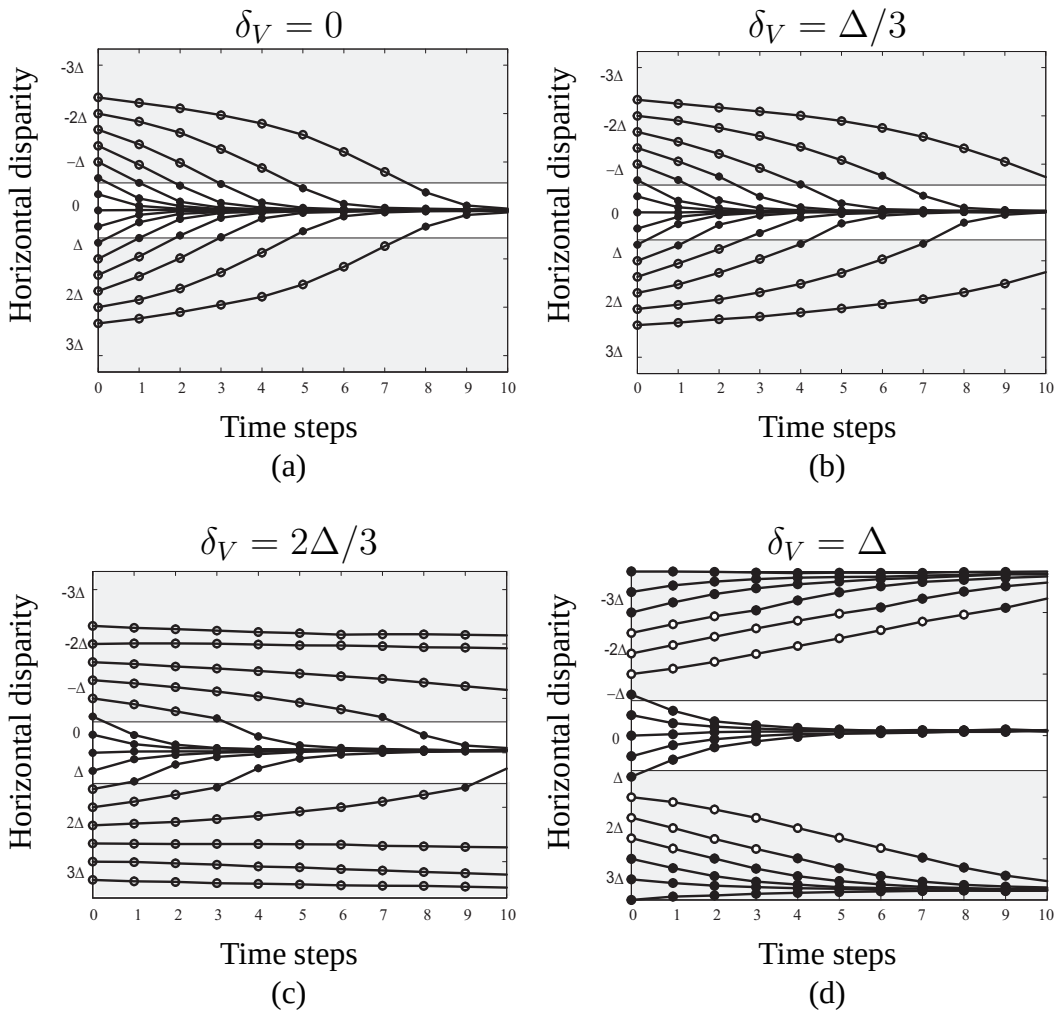


Figure 14: Evolution in time of the vergence control testes with a RDS. Fixed a vertical disparity pedestal, varing from 0 to Δ, each trace represent the evolution of the vergence strating from a different value of horizontal disparity. It is clear that considering a small vertical disparity (b), its effect on the horizontal vergence is negligible. Increasing it above a certain value (c) and (d), the vergence control is slowed down and its range of effectiveness is reduced. Filled and opend circles denote the action of the SHORT and LONG controls, respectively.

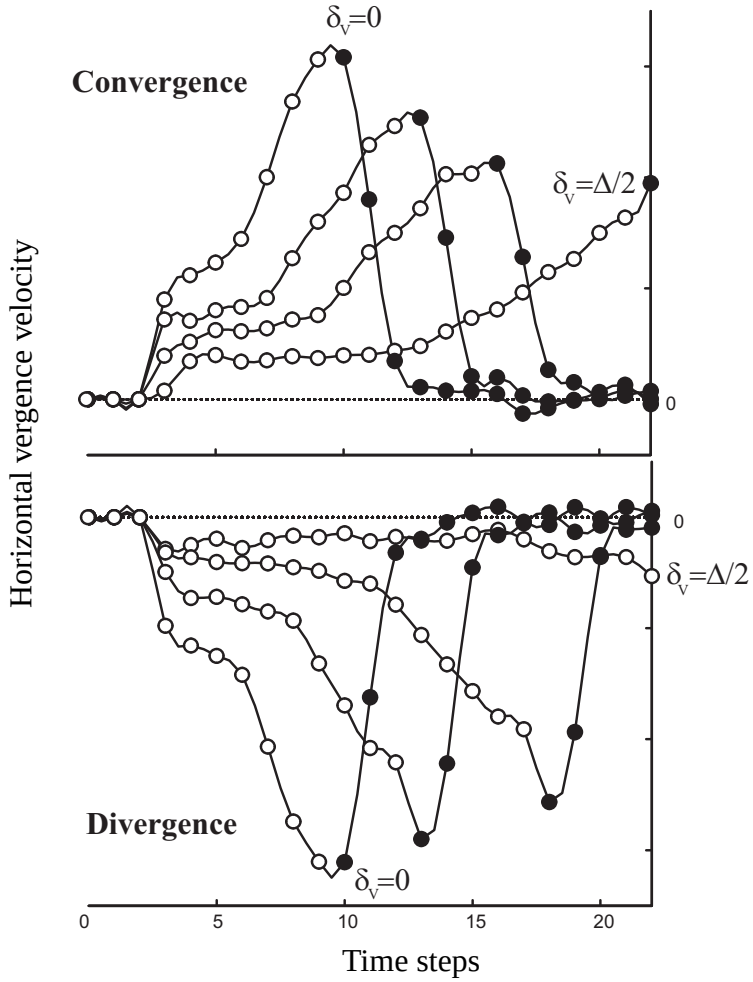


Figure 15: Horizontal vergence velocity (deg/timestep) in presence of a vertical disparity pedestal of increasing magnitude.

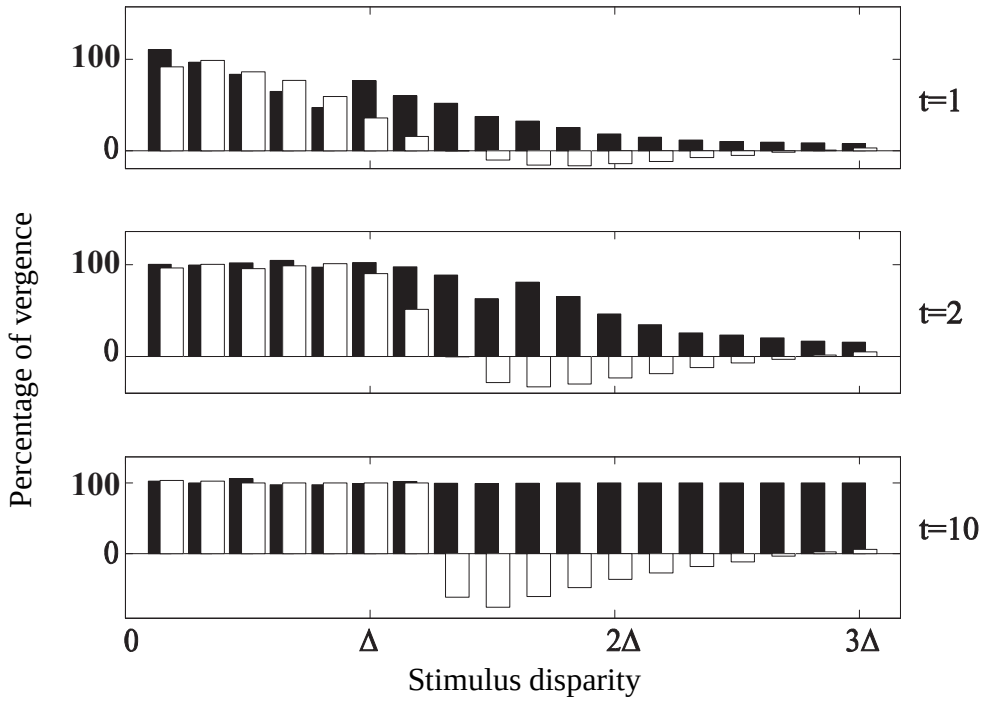
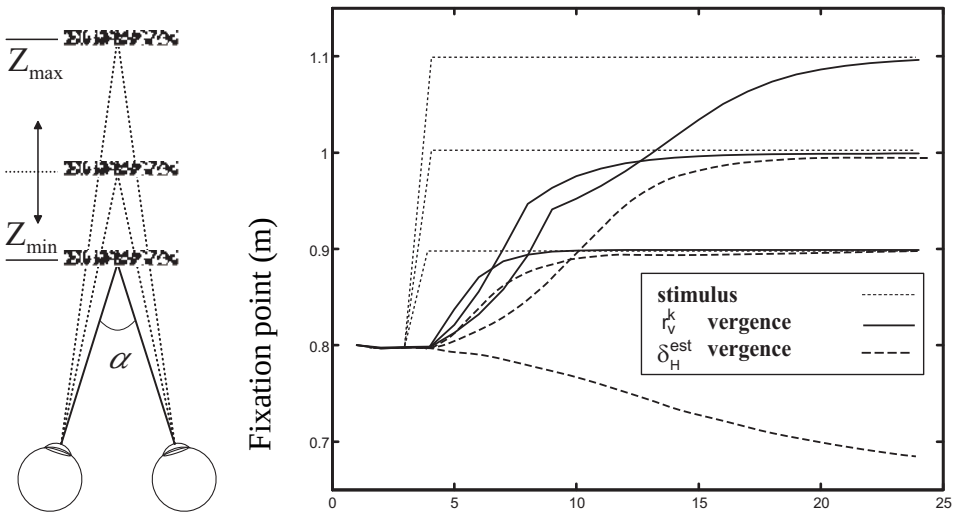


Figure 16: Comparison of percentage of vergence achieved by the model using the estimated disparity δ_H^{est} (white bars), and the r_v^k signals to control the vergence. The stimulus used is a RDS with a disparity step in the range from -3Δ to 3Δ. Only the positive axis is represented because the response is symmetric around zero disparity. The graphs represent the status of the system after 1, 2 and 10 time steps. At each step the r_v^k signals are able to reach the target in the whole range tested, while δ_H^{est} yields a wrong control for disparities larger than Δ.



(a) (b)

Figure 17: (a) Simulated experimental setup, consisting of the eyes looking at a plane characterized by a RDS pattern, and perpendicular to the binocular line of sight. (b) Behaviour of the vergence control using r_v^k vs δ_H^{est} in case of a diverging step. The r_v^k control (solid line) is able to reach the depth of the plane (dotted line) in all the cases presented, while the δ_H^{est} control (dashed line) produce a wrong movement for a depth step above a certain threshold.

for different time steps. Because of the behaviour of the two mechanisms is symmetric with respect to zero disparity, we show the positive semiaxis only. A percentage value higher than 100 indicates an overshoot of the movement, whereas a value lower than zero indicates a movement in the opposite (*i.e.*, wrong) direction. After the first time step (Fig.16 top row), if the stimulus disparity is within Δ, the behaviour is slightly better for δ_H^{est} (white bars), whereas outside this range it produces a vergence movement that is the opposite of the one requested. The r_v^k signals (black bars) produce almost the same movement of δ_H^{est} for small disparity steps, but they are able to achieve slow but effective vergence movements up to the limit of the tested range. At the second time step (Fig.16 middle row), for disparity steps smaller than Δ, both the mechanisms reach the target, and for higher disparities the behaviour is similar to the previous time step. After 10 time steps (Fig.16 bottom row), we observed that δ_H^{est} was able to work in the proper way only for disparities within Δ, whereas r_v^k was able to reach the target in all the tested range.

4.3.2 Test with a frontoparallel plane

Considering a virtual environment in which the eyes, characterized by null version and elevation angle, and by a vergence angle α, look at a plane with a random dot texture (Fig.17a). The plane is at a depth Z with respect to the cyclopic position, and perpendicular to the binocular line of sight. The interocular distance is $b = 70mm$, the nodal length is $f_0 = 17mm$, and the stimulus is projected onto the retinal plane, with a size of $6mm$, thus considering a field of view of almost 20 degree. At the first time step the plane and the fixation point are at the

same Z, then the plane is moved to a new depth, and the vergence angle starts to change step by step, until the fixation point reaches the depth of the plane. Considering the position of the eyes, the vergence variation is applied symmetrically: $\Delta\alpha_R = -\Delta\alpha_L = -\arctan(\frac{r}{2f_0})$, where r is computed by considering the weighted average of the vergence responses r_v^k or of the estimated disparities δ_H^{est}. The area where the average is computed, is a neighborhood of the fovea of 5°, and its size is based on physiological experiments [48] that show that it is the maximum extent of the retina where the disparity stimulus is integrated to drive vergence eye movements in humans.

The first test considers a fixed frontoparallel plane at a given distance, while the eyes are fixating on the surface of the plane. The plane steps back and forth by an amount that varies from trial to trial. Fig.17b shows that a control based on the disparity computation produces the correct change of the vergence angle (dotted lines), when the size of the step is restrained. On the other hand, the implemented model is able to produce a faster change of the fixation point (solid lines), and, even for larger depth steps, the model is able to ensure a reliable vergence control. Moreover, once the fixation point has reached the plane in depth, the disparity in the fovea is approximately zero and the system is able to ensure a stable fixation.

The second test considers a frontoparallel plane whose position in depth varies continuously in time as a ramp and a sinusoid. The slope of the ramp is varied from $0.5\ cm$ per time step to a pure step (Fig.18a). While for small values only the SHORT control is enabled, in the other cases the initial part of the vergence is produced by the LONG one, and the interplay between the two controls is very similar to the one observed in the transient and sustained components of the physiological responses [5]. In support of this hypothesis, in case of both a divergent and a convergent ramp, the simulated vergence movements are qualitatively very similar to the results obtained in physiological experiments. In the same way, the frequency of the sinusoid that controls the depth of the plane was varied between 7 and 38 time steps, and again the simulated results (Fig.18b) are qualitatively similar to the experimental data [5]. Increasing the frequency, it is evident a transition from a slow and smooth tracking of the plane, due to the SHORT control, to a combination of the LONG and SHORT controls. When the frequency becomes too high, the system is no more able to follow the stimulus in depth.

5 Network Paradigms for vergence control

In this section we present a modular architecture (see subsection 5.1) and two networks for learning vergence control: a linear (subsection 5.2) and a convolutional one (subsection 5.3). Training and evaluation are also discussed.

5.1 Vergence control model

For the vergence control paradigm modeling we used the setup shown on Fig. 19. This setup consists of the vergence simulator module, the disparity detector population module (described in Section 3.2) and the vergence control network module. The vergence simulator consists of a robotic head model and a ray-tracing engine. The robotic head model has the same parameters as in subsection 4.3.2 (the baseline $b = 70\ \mathrm{mm}$, the focal length $f_0 = 17\ \mathrm{mm}$ and field of view

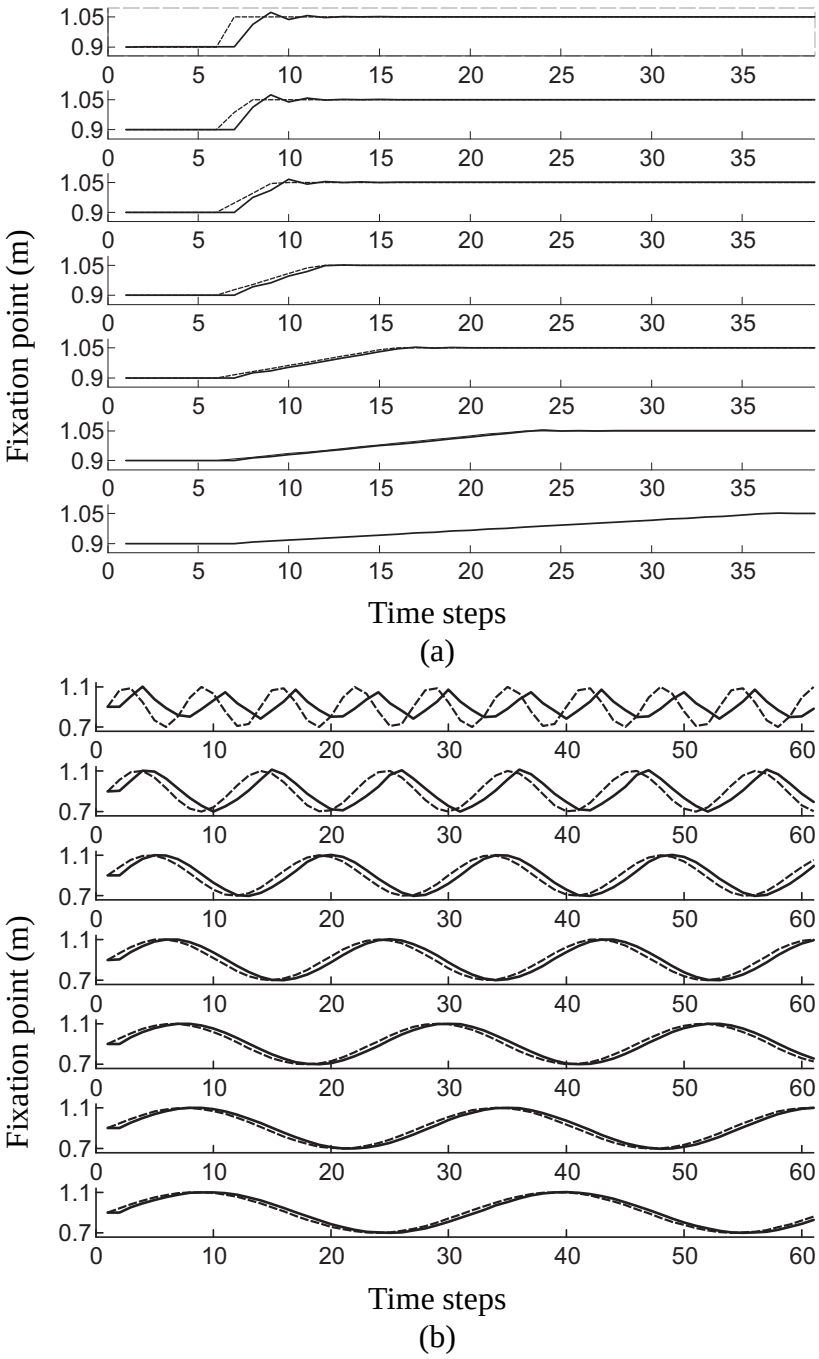


Figure 18: (a) Vergence response in time to diverging ramps with different slopes, and (b) to sinusoids characterized by different periods. The dotted line represents the depth of the stimulus and the solid one is the depth of the fixation point.

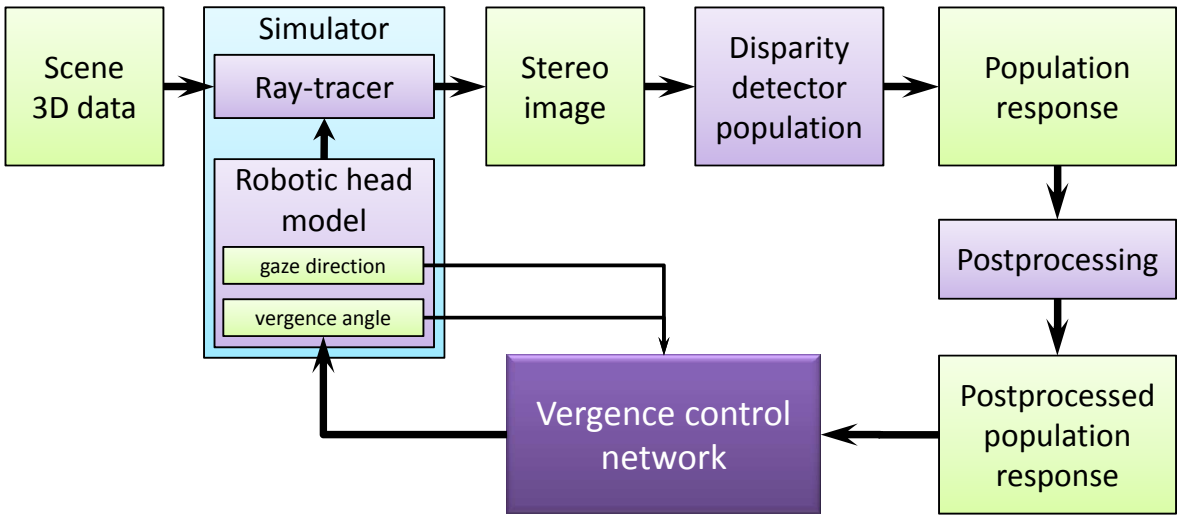


Figure 19: The scheme of the experimental setup for vergence control model training.

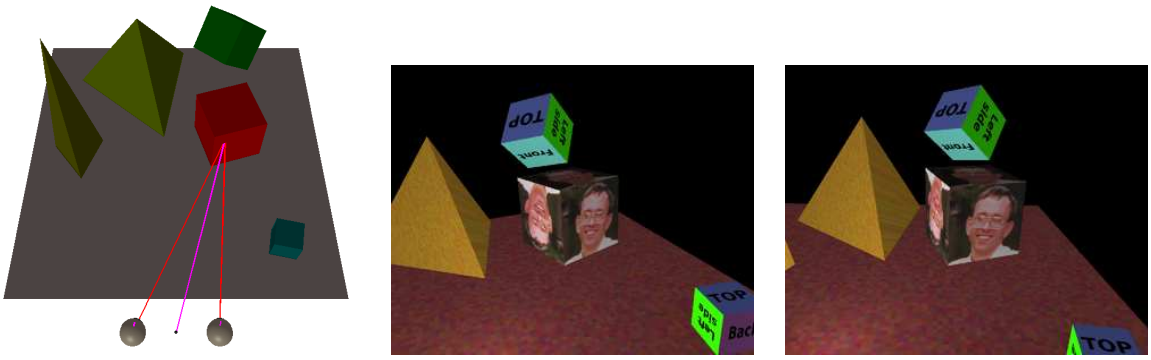
$\approx 20°$) and can be controlled externally by the gaze direction (version) and the vergence angle. Using the robotic head model and scene information the ray-tracing engine renders left and right views (see Fig. 20), which then are fed to the disparity detector population module. In order to speed up the simulations we decided to work with a single-scale architecture of the disparity detector population and use images of low-resolution (41×41).

The response of the population through the postprocessing module reaches the vergence control network (VC-net). The actual gaze direction and actual vergence are used as an additional input to the VC-net. The output of the VC-net is a parameter controlling the actual vergence, which in this work is the vergence angle. Obviously, given a gaze direction, which in our case is fixed, there is only one value of the vergence angle, which brings the fixation point onto the surface of the attended object. The VC-net is expected to approximate this value as close as possible given the input data.

The postprocessing of the population response was different for the two considered VC-nets. In the case of the linear network, the postprocessing was defined as a 2D pooling over first two (spatial) dimensions of the population response (see subsection 5.2).

On the one hand, the pooling operation reduces the amount of data to process, but on the other hand, it has a major drawback as it discards the spatial information about the disparity encoded in the population response. The simulations shows that, in the simplified case, this is still acceptable, but not in general case (see subsection 5.2).

That is why, in the general when the linear network has an unpredictable systematic error, we do not do any pooling directly, but let the convolutional network to do this in the first two layers (see subsection 5.3). In this case, it is convenient to consider the postprocessing as an identical operator.



(a) (b) (c)

Figure 20: An example of a synthetic scene layout (a) used by the vergence simulator to render corresponding left (b) and right (c) eye views.

5.1.1 Vergence control database

The vergence simulator was used also for the creation of a *vergence database*, which has been used for training and testing of the vergence control network. The database contains a set of synthetic scenes and a set of samples. Each synthetic scene consists of several simple (plane triangle, cube, tetrahedra etc.) textured objects placed into room-like environment (see Fig. 20). All the textures (real-world images, checkerboard-like images, random noise etc.) we used, were corrupted by 5% Gaussian noise in order to obtain a better response from the population. Samples of the database consist of the *gaze direction*, the *actual vergence angle*, the *stereo pair* (left and right eyes images), the *population response* for the stereo pair and the *desired vergence angle*. The actual vergence angle is a distorted (with Gaussian noise) version of the desired one.

With the database it is easy to prepare training pairs. As the input data vector is constructed from the gaze direction, the actual vergence angle and postprocessed population response are computed. The output consists of only one parameter: the desired vergence angle.

5.2 Linear servo network

The first attempt in the modeling of a network for vergence control was done with the simplest possible network that reproduces the results from section 4 using learning from examples taken from the vergence database.

5.2.1 Vergence angle vs. distance to the fixation point

Given a robotic head baseline b and a gaze direction vector $\mathbf{g} = (g_x, g_y, g_z)^T$, ($\|\mathbf{g}\| = 1$) it is possible to infer the distance d to the fixation point (from the middle of the head's baseline) using the vergence angle α:

$$d = \frac{b}{2}\left(s + \sqrt{s^2 + 1}\right), \text{ where } s = \frac{1}{\tan\alpha\sqrt{1 - g_x^2}} \tag{31}$$

and *vice versa*:

$$\begin{aligned} \alpha &= \arccos\left(\frac{\mathbf{v}_l^T \mathbf{v}_r}{\|\mathbf{v}_l\| \cdot \|\mathbf{v}_r\|}\right), \text{ where} \\ \mathbf{v}_r &= d \cdot \mathbf{g} + (b/2, 0, 0)^T, \text{ and} \\ \mathbf{v}_l &= d \cdot \mathbf{g} - (b/2, 0, 0)^T. \end{aligned} \tag{32}$$

From the equations 31 and 32, one can see that by considering a fixed gaze direction and fixed baseline, the vergence angle is equivalent to the distance to the fixation point (nevertheless they have a nonlinear relationship). We used the deviation of the actual distance to the fixation point from the desired one as an additional measure of vergence performance. From our point of view, this measure is more natural compared to the deviation of the vergence angle.

5.2.2 Postprocessing of population response

As it has been mentioned above, we have defined the postprocessing of the population response ($r_c = \{r_c^{ij}\}_{ij}$, where r_c is a four dimensional array $n_r \times n_c \times N_o \times N_p$, n_r is the number of rows, n_c is the number of columns, N_p is the number of phase shifts and N_o is the number of orientations) for the linear network as two-dimensional spatial pooling over the first two dimensions of r_c with a two-dimensional Gaussian kernel G_σ:

$$P_{ij} = G_\sigma * r_c^{ij}, \tag{33}$$

where r_c^{ij} is (two-dimensional) population response map for i-th orientation and j-th phase shift. The kernel G_σ has the same size $n_r \times n_c$ as the size of a population response map r_c^{ij}, so the result of the convolution is a scalar value P_{ij}.

This step drastically reduces the amount of data to process. After pooling, the network has to process only a two-dimensional ($N_o \times N_p$) pooled population response instead of four-dimensional ($n_r \times n_c \times N_o \times N_p$) array.

5.2.3 Training

We consider two cases for the experiment: a *simplified* and a *general* case. The simplified case is shown on Fig. 17a: the gaze direction of the robotic head is orthogonal to its baseline and the stimulus is a frontoparallel plane, thus, also orthogonal to the gaze direction. The stimulus is allowed to move only in depth. In the general case, all restrictions on the orientation of the gaze, as well on the stimulus position, type and orientation, are dropped. One of the examples is shown on Fig. 20 with the only difference in the resolution of the rendered images (for the simulation we used much lower resolution).

For each experiment case we have prepared several vergence databases with the number of synthetic scenes ranged from 100 to 1000 and the number of samples from 200 to 4000.

The input vector for the linear VC-net was constructed as a concatenation of the pooled population response (56 values), the gaze direction (2 values) and the actual vergence (1 value), so its dimensionality is 59. The output is a prediction of the vergence angle, which is a scalar

value. Due to the linearity of the network, there was no reason to introduce any hidden layers, so the linear VC-net consisted of only one linear unit. This simplest possible vergence control network has only 60 parameters (including bias), which can be learned either directly (using robust linear regression) or iteratively (using gradient descent) from the training database. Not surprisingly, both training approaches produced almost identical solutions on the same training data in the simplified case.

5.2.4 Evaluation and results

For the evaluation of the VC-net, we have adopted the methodology described in subsection 4.3.2 with the next differences:

- the stimuli are allowed to move in the direction of the gaze (not only in Z direction),
- the rendered stimuli cover 60-100% of the image area (allowing for depth discontinuities),
- in the general case the stimuli can be not only 2D plane rectangles but also 3D primitives (cubes or tetrahedrons),
- in the general case the stimuli can have arbitrary position and orientation inside the workspace.

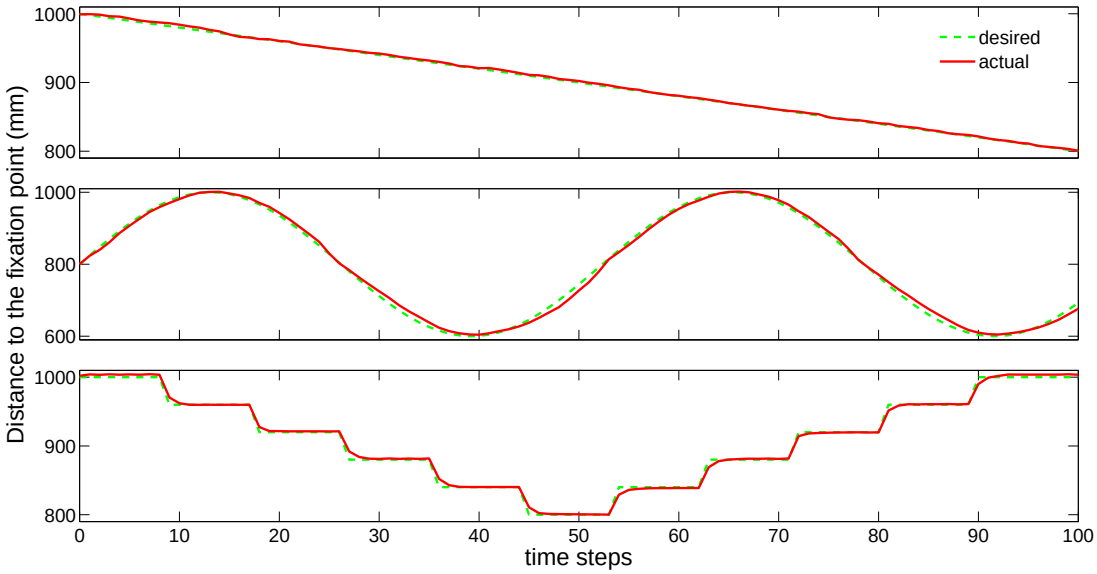
The first item is very important for the general case, when the gaze direction is not necessarily parallel to the Z-axis.

Three standard tests (ramp, sinusoid and staircase) were carried out for the simplified as well as the general case. The typical results of the performance, measured in terms of distance to the fixation point, are shown on Fig. 21.

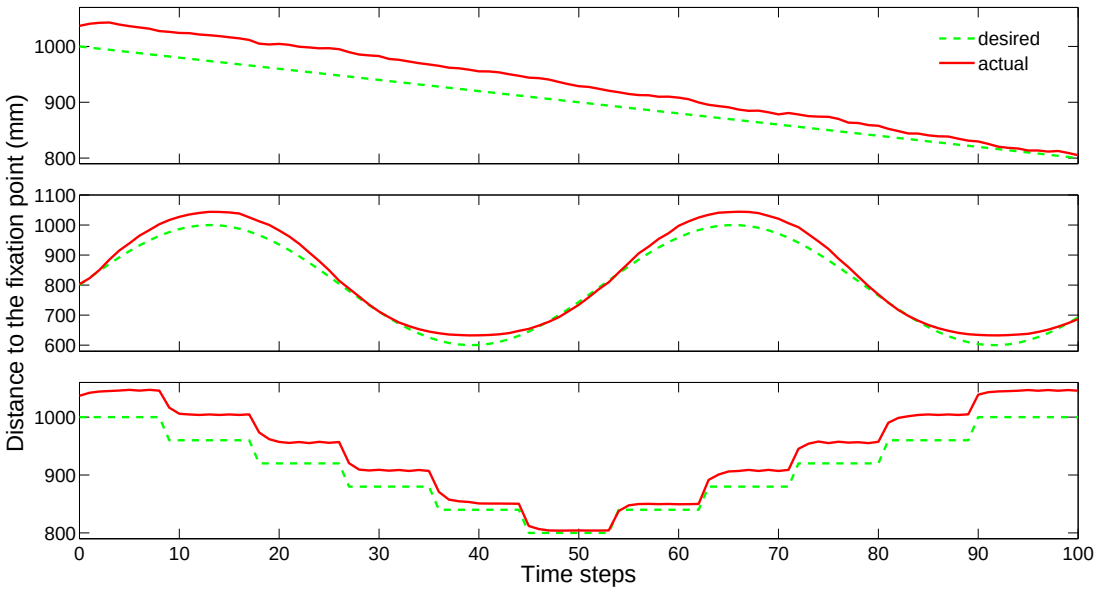
The results of the evaluation of the linear VC-network show, that in the simplified setup, it can produce an accurate and robust vergence control using spatially pooled population responses. The relative error (distance-based as well as angular measure) was always less then 1% in all tests of the simplified scenario (see Fig. 21a).

But in the general case this approach has unpredictable systematic error, which in our tests was up to 7% (see Fig. 21b). The large magnitude of the vergence error of the linear network in the general case can be explained, from our point of view, by the presence of the vertical disparity asymmetric patterns (due to the arbitrary orientation and position of the stimuli) and by the disparity discontinuities (caused by the limited size of the stimuli). In the simplified scenario, the vertical disparity information is symmetrically spread over the spatial dimensions of the population response, and is discarded in the preprocessing stage by spatial pooling. This does not happen in the general case, so the pooled population response is biased by the residual vertical disparity, and linear network, in turn produces a biased vergence control signal.

This situation motivated us to investigate a more complex paradigm for the vergence control, which should be able to recognize particular patterns in the population responses in the general case, and produce a proper vergence control signal. For this purpose we have chosen a convolutional network [49, 50, 51], as it has proved to be one of the best paradigms for object recognition.



(a) Simplified scenario



(b) General case scenario

Figure 21: A typical examples of the depth-based performance plots for a linear vergence control network in the simplified (a) and general case (b) scenarios.

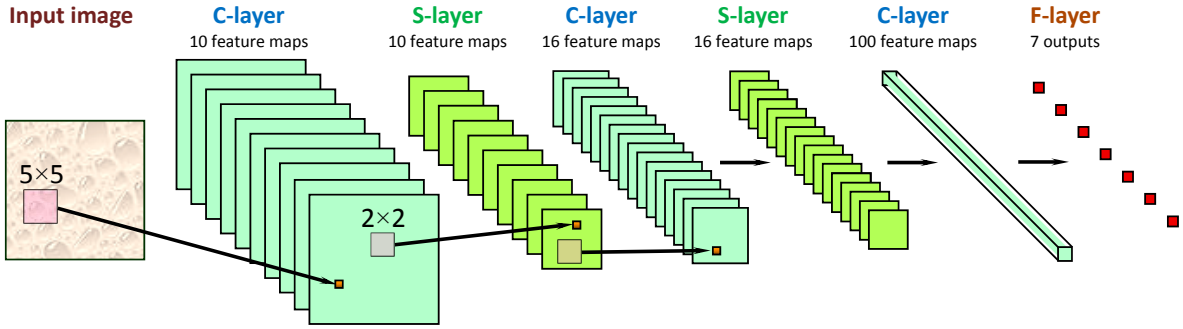


Figure 22: An example of typical convolutional network. The architecture (number of layers and/or feature maps) of the network can differ, depending on the complexity of the task.

5.3 Convolutional network

The first *convolutional network* (CN) appeared in the work of Fukushima in [49] and was called Neocognitron. The basic architectural ideas behind the CN (*local receptive fields*, *shared weights*, and spatial or temporal *subsampling*) allow such networks to achieve some degree of shift and deformation invariance and at the same time reduce the number of training parameters.

Since 1989, Yann LeCun and co-workers have introduced in [1] a series of convolutional networks with the general name *LeNet*, which contrary to the Neocognitron use supervised training. In this case, the big advantage is that the whole network is optimized for the given task, making this approach usable for real-world applications. LeNet have been successfully applied to character recognition nonlinear-dimensionality reduction of image-sets [52] and even to obstacle avoidance in an autonomous robot [53].

A typical convolutional network is a feed-forward network of layers of three types: *convolutional* (C-layer), *subsampling* (S-layer) and *fully-connected* (F-layer). The C-layers and S-layers usually come in pairs and are interleaved, and F-layers come at the end (see Fig. 22). The output of a C-layer is organized as a set of *feature maps*. Each feature map contains the output of a set of neurons with local receptive fields. All neurons in the feature map share the same weights, so the feature map is responsible for a particular local visual feature which is encoded in the weights of these neurons. The computation of a feature map starts with a 2D convolution of the input with a fixed kernel defined by the neuron's weights. A feature map can have inputs from several feature maps of the previous layer. In order to condense the extracted features and make them more invariant with respect to spatial deformations, the C-layer is typically followed by an S-layer which does a local averaging and subsampling. Each neuron in a F-layer just does summation of bias with all weighted inputs and then propagates the sum through a nonlinear transfer function (RBF or sigmoid).

The network is trained in a supervised manner using backpropagation. For the efficient training of large CNs, LeCun and colleagues proposed a number of tricks and a modification of the Levenberg-Marquardt algorithm [54].

5.3.1 Extended convolutional network

For the modeling of CN-based vergence control we have developed our own version of the convolution network in MATLAB. This network can be considered as an extension of LeCun's LeNet because it has the next features:

- any directed acyclic graph can be used for connecting the layers of the network;
- the network can have any number of arbitrarily sized input and output layers;
- the neuron's receptive field (RF) can have an arbitrary stride (step of local RF tiling), which means that in the S-layer, RFs can overlap and in the C-layer the stride can differ from 1;
- any layer or feature map of the network can be switched from trainable to nontrainable (and vice versa) mode even during training;
- new layer type: M-layer.

The M-layer works similarly to the C-layer with the only difference in the subsampling operation $s(\mathbf{x}, a) = a \sum_i x_i$ is replaced by a softmax-like M-operation:

$$m(\mathbf{x}, a) = \frac{\sum_i x_i e^{a x_i}}{\sum_i e^{a x_i}}, \tag{34}$$

where the receptive field is denoted by $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. This function $m(\mathbf{x}, a)$ has been chosen because its properties:

- $m(\mathbf{x}, a) \approx \max\{x_i\}_i$, if a is a large positive (*e.g.* $a = 100$);
- $m(\mathbf{x}, a) \approx \min\{x_i\}_i$, if a is a large negative (*e.g.* $a = -100$);
- $m(\mathbf{x}, a) = \dfrac{\sum_{i=1}^{n} x_i}{n}$, if $a = 0$.

5.3.2 Convolution network design

The idea behind the use of the convolutional network as a vergence controller consist in an assumption that this powerful network, after proper training, will be able to recognize disparity patterns directly from the population responses, and convert them into a proper vergence signal. The architecture of the convolutional network, used for our experiments, is CSFF and is depicted on Fig. 23. The main challenge in this approach was the amount of data: the population response consists of 56 (7×8) maps of resolution 41×41 (rendered image resolution), so the input of the network has 94136 ($41 \times 41 \times 8 \times 7$) components. In order to be able to train the network with such high dimensional input data, we had to reduce the number of the training parameters. The first (convolutional) layer was set as fixed (nontrainable) with Gaussian kernels of size 19×19 with standard deviation 6. The second (subsampling) layer has also 56 feature maps size of which was set to 3×3.

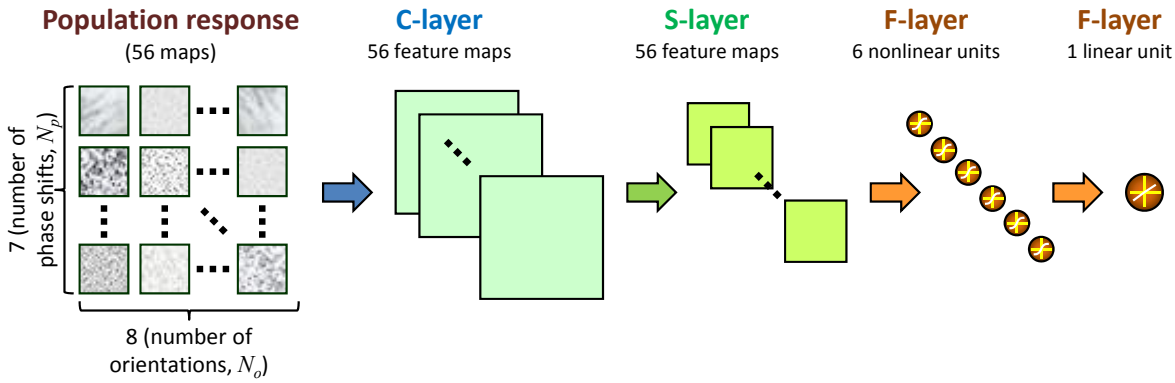


Figure 23: Convolutional network (and its input) used for the vergence control.

5.3.3 Evaluation and results

For the evaluation of the convolutional network we have used exactly the same tests as for the linear network (see subsection 5.2). The performance was very similar in both scenarios (see Fig. 24). The average relative error (in distance-based measure) for both scenarios is less than 1%. Comparing the performances of two the networks, it is possible to say that the convolutional one has a more pronounced inertia with respect to the linear one, but it still is able to handle the general case tasks with an acceptable accuracy and robustness. But, on the other hand, vergence control based on the convolutional network is much more computationally expensive than linear-based.

6 Conclusions

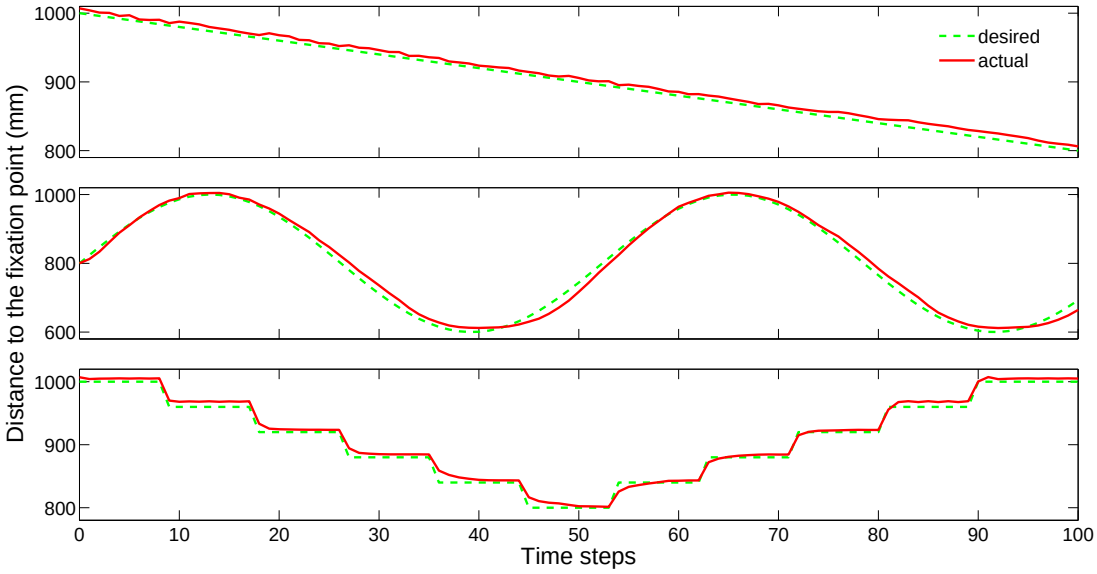
Most of the conventional vergence control models, are based on the minimization of the horizontal disparity.

Conversely, we proposed to avoid the explicit computation of the disparity map and extract the vergence control signal directly from the population response, over the “foveal” region, of a cortical-like network organized as a hierarchical arrays of binocular complex cells. With the specific design approach followed to implement the distributed architecture, we demonstrated that we can take full advantage of the flexibility and adaptability of distributed computing to specialize disparity detectors for vergence control and depth vision.

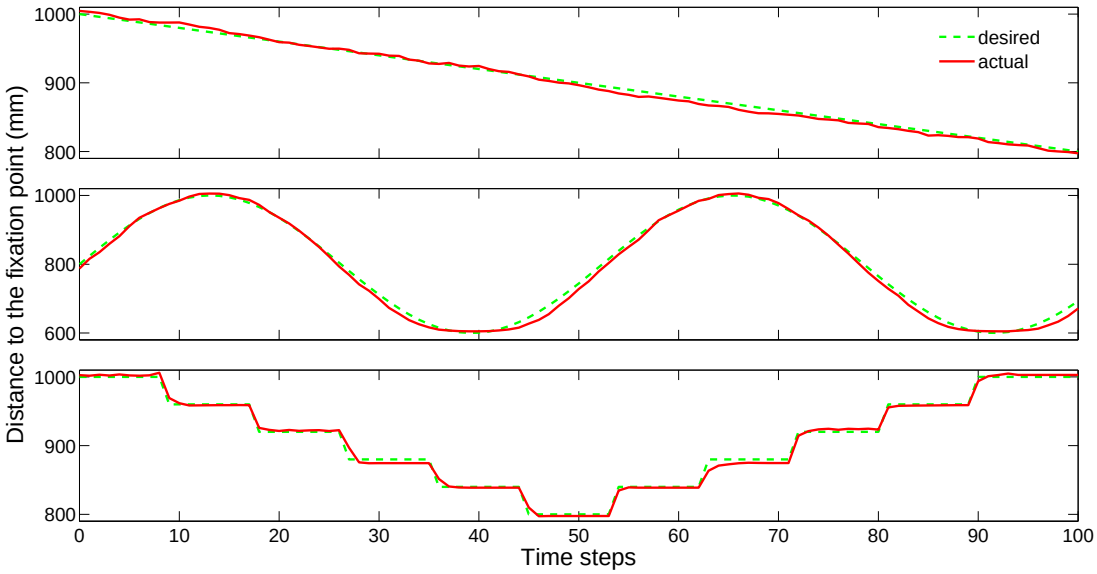
On this ground, a neural network paradigm has been proposed for building disparity-vergence control. Specifically, an increasing complexity strategy in the learning process is adopted: starting from the simplest one-unit architecture we increased the number of units/layers until an acceptable level of generalization error is reached.

Although the model only resorts to a population of neurons in a single scale, we demonstrated that, using a convolutional network, accurate and fast vergence control can be achieved in a closed loop, for different orientations of the gaze.

In this direction of the development of the vergence control networks, our next steps of investigation are the following:



(a) Simplified scenario



(b) General case scenario

Figure 24: A typical examples of the depth-based performance plots for a convolutional vergence control network in the simplified (a) and general case (b) scenarios.

- we can allow the first (C-)layer of the convolutional network to be trained (in a supervised or unsupervised manner);
- we can replace the disparity detector population by additional non-trainable layers of the convolutional network;
- we can include dynamic (*i.e.*, spatiotemporal) disparity tuning and attentional signals (based on object properties) that might guide intentional exploration of the selected object.

References

[1] Y. LeCun, L. Bottou, G. Orr, and K. Muller. Efficient backprop. In G. Orr and Muller K., editors, *Neural Networks: Tricks of the trade*. Springer, 1998.

[2] J. Horng, J. Semmlow, G.K. Hung, and K. Ciuffreda. Initial component in disparity vergence: A model-based study. *IEEE Trans. Biomed. Eng.*, 45:249–257, 1998.

[3] W. M. Theimer and H. A. Mallot. Phase-based vergence control and depth reconstruction using active vision. *CVGIP, Image understanding*, 60(3):343–358, 1994.

[4] H. Ogmen S. S. Patel and B. C. Jiang. Neural network model of short-term horizontal disparity vergence dynamics. *Vision Research*, 37(10):1383–1399, 1996.

[5] J. L. Semmlow G. K. Hung and K. J. Ciuffreda. A dual-mode dynamic model of the vergence eye movement system. *Trans. on Biomedical Engineering*, 36(11):1021–1028, 1986.

[6] V.V. Krishnan and L. A Stark. A heuristic model for the human vergence eye movement system. *IEEE Trans. Biomed. Eng.*, 24:44–49, 1977.

[7] J.J. Koenderink and A.J. van Doorn. Representation of local geometry in the visual system. *Biol. Cybern.*, 55:367–375, 1987.

[8] J.G. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *J. Opt. Soc. Amer. A*, A/2:1160–1169, 1985.

[9] R.A. Young. The Gaussian derivative theory of spatial vision: Analysis of cortical cell receptive field line-weighting profiles. Technical Report GMR-4920, General Motors Research, 1985.

[10] A.B. Watson. The cortex transform: rapid computation of simulated neural images. *Computer Vision, Graphics, and Image Processing*, 39:311–327, 1987.

[11] M.J. Hawken and A.J. Parker. Spatial properties of neurons in the monkey striate cortex. *Proc. Roy. Soc. Lond. B*, 231:251–288, 1987.

[12] D.J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Amer. A*, 4:2379–2394, 1987.

[13] J.B. Martens. The Hermite transform - Theory. *IEEE Trans. Acoust., Speech, Signal Processing*, 38:1595–1606, 1990.

[14] D.G. Stork and H.R. Wilson. Do Gabor functions provide appropriate decriptions of visual cortical receptive fields? *J. Opt. Soc. Amer. A*, 7(8):1362–1373, 1990.

[15] J. Yang. Do Gabor functions provide appropriate decriptions of visual cortical receptive fields?: comment. *J. Opt. Soc. Amer. A*, 9(2):334–336, 1992.

[16] S.A. Klein and B. Beutter. Minimizing and maximizing the joint space-spatial frequency uncertainty of Gabor-like functions: comment. *J. Opt. Soc. Amer. A*, 9(2):337–340, 1992.

[17] S.G. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 11(7):674–693, 1989.

[18] T. Reed and H. Wechsler. Segmentation of textured images and gestalt organization using spatial/spatialfrequency representations. *IEEE Trans. Pattern Analysis Mach. Intell.*, 12:1–12, 1990.

[19] W.T. Freeman and E.H. Adelson. The design and use of steerable filters. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13:891–906, 1991.

[20] P. Perona. Steerable-scalable kernels for edge detection and junction analysis. *Image Vis. Comput.*, 10:663–672, 1992.

[21] E.P. Simoncelli, W.T. Freeman, E.H. Adelson, and D.J. Heeger. Shiftable multiscale transforms. *IEEE Trans. on Information Theory*, 38(2):587–607, 1992.

[22] M. Felsberg and G. Sommer. The monogenic scale-space: A unifying approach to phase-based image processing in scale-space. *Journal of Mathematical Imaging and Vision*, 21:5–26, 2004.

[23] L.D. Jacobson and H. Wechsler. Joint spatial/spatial-frequency representation. *Signal Processing*, 14:37–68, 1988.

[24] H. Wechsler. *Computational Vision*. Academic Press, 1990.

[25] R. Navarro, A. Tabernero, and G. Cristobal. Image representation with gabor wavelets and its applications. In P. W. Hawkes, editor, *Advances in Imaging and Electron Physics*, pages 1–84. Academic Press, San Diego CA, 1996.

[26] E.H. Adelson, C.H. Anderson, J.R. Bergen, P.J. Burt, and J.M. Ogden. Pyramid methods in image processing. *RCA Engineer*, 29(6):33–41, 1984.

[27] O. Nestares, R. Navarro, J. Portilla, and A. Tabernero. Efficient spatial-domain implementation of a multiscale image representation based on Gabor functions. *Journal of Electronic Imaging*, 7(1):166–173, 1998.

[28] T.D. Sanger. Stereo disparity computation using Gabor filters. *Biol. Cybern.*, 59:405–418, 1988.

[29] D.J. Fleet, A.D. Jepson, and M.R.M. Jenkin. Phase-based disparity measurement. *CVGIP: Image Understanding*, 53(2):198–210, 1991.

[30] F. Solari, S.P. Sabatini, and G.M. Bisio. Fast technique for phase-based disparity estimation with no explicit calculation of phase. *Elect. Letters*, 37(23):1382–1383, 2001.

[31] A.D. Jepson and M.R.M. Jenkin. The fast computation of disparity from phase differences. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR'89)*, pages 398–403, 1989.

[32] Carl-Johan Westelius. Preattentive gaze control for robot vision. Lic. Thesis LiU-Tek-Lic-1992:14, ISY, Linköping University, SE-581 83 Linköping, Sweden, June 1992. Thesis No. 322, ISBN 91-7870-961-X.

[33] M. J. Morgan and E. Castet. The aperture problem in stereopsis. *Vis Res.*, 37:2737–2744, 1997.

[34] D Fleet. Disparity from local weighted phase-correlation. In *Proc. of the IEEE Int. Conf. on Systems,Man and Cybernetics*, volume 1, pages 48–54, 1994.

[35] H. Wagner D.J. Fleet and D.J. Heeger. *Modelling binocular neurons in the primary visual cortex.* Jenkin, M. and Harris, L., Cambridge University Press, 1996.

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

[37] Freeman R. D. I. Ohzawa and G. C. DeAngelis. Stereoscopic depth discrimination in the visual cortex: Neurons ideally suited as disparity detectors. *Science*, 249:1037–1041, 1990.

[38] B.G. Cumming S.J.D. Prince and A. J. Parker. Range and mechanism of encoding of horizontal disparity in macaque v1. *J. Neurophysiol.*, 87:209–221, 2002.

[39] N. Qian and S. Mikaelian. Relationship between phase and energy methods for disparity computation. *Neural Comp.*, 12:279–292, 2000.

[40] P. Dayan A. Pouget and R. S. Howard. Computation and inference with population codes. *Annu. Rev. Neurosci*, 26:381–410, 2003.

[41] M. Chessa, S.P. Sabatini, and F. Solari. A fast joint bioinspired algorithm for optic flow and two-dimensional disparity estimation. In *Proc. International Conference on Computer Vision Systems (ICVS'09)*, Liege, Belgium, October 2009.

[42] Sabatini S.P., G. Gastaldi, F. Solari, J. Diaz, E. Ros, K. Pauwels, M. Van Hulle, N. Pugealult, and N. Krueger. Compact and accurate early vision processing in the harmonic space. In *Proc. VISAPP'07*, 8-11 March, 2007, Barcelona, Spain, 2007.

[43] C. Busettini G. S. Masson and F. A. Miles. Vergence eye movements in response to binocular disparity without depth perception. *Nature*, 389:283–286, 1997.

[44] B.G. Cumming and A.J. Parker. Responses of primary visual cortical neurons to binocular disparity without depth perception. *Nature*, 389:280–283, 1997.

[45] D.A. Wismeijer, R. van Ee, and C.J. Erkelens. Depth cues, rather than perceived depth, govern vergence. *Exp. Brain Research*, 184:61–70, 2008.

[46] A. Gibaldi, M. Chessa, A. Canessa, S.P. Sabatini, and F. Solari. A neural model for binocular vergence control without explicit calculation of disparity. *Neurocomputing*, In press.

[47] G. F. Poggio. Mechanism of stereopsis in monkey visual cortex. *Cerebral Cortex*, 5:193–204, 1995.

[48] I. P. Howard R. S. Allison and X. Fang. The stimulus integration area for horizontal vergence. *Exp. Brain Res.*, 156:305–313, 2004.

[49] K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, 1980.

[50] R. Linsker. From basic network principles to neural architecture: Emergence of orientation columns. *Proceedings of the National Academy of Sciences*, 83(22):8779–8783, 1986.

[51] Y. LeCun, B. Boser, JS Denker, D. Henderson, RE Howard, W. Hubbard, and LD Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[52] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *Proc. Computer Vision and Pattern Recognition Conference (CVPR'06)*. IEEE Press, 2006.

[53] Y. LeCun, U. Muller, J. Ben, E. Cosatto, and B. Flepp. Off-road obstacle avoidance through end-to-end learning. *Advances in neural information processing systems*, 18, 2006.

[54] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. volume 86, pages 2278–2324, 1998.