Please cite as:
S.P. Sabatini, F. Solari, A. Canessa, M. Chessa, A. Gibaldi. Early Perception-Action Cycles in Binocular Vision: Visuomotor Paradigms and Cortical-like Architectures.
In: Developing and Applying Biologically-Inspired Vision Systems: Interdisciplinary Concepts. Marc Pomplun and Jun Suzuki (Eds.), IGI Global, 2013.

Early Perception-Action Cycles in Binocular Vision: Visuomotor Paradigms and Cortical-like Architectures

Silvio P. Sabatini, Fabio Solari, Andrea Canessa, Manuela Chessa and Agostino Gibaldi
DIBE – University of Genoa, Italy

ABSTRACT

Pushed by wide neurophysiological evidences of modulatory effects of motor and premotor signals on the visual receptive fields across several cortical areas, there is a growing attention for moving the active vision paradigm from systems in which just the effects of action influence the perception, to systems here the acting itself, and even its planning, operate in parallel with perception, thus really closing the loops and taking full advantage of a concurrent/anticipatory perception-action processing. In this context, cortical-like architectures for both vergence control and depth perception (in the 3D peripersonal space) that incorporate adaptive tuning mechanisms of the disparity detectors are presented. The proposed approach points out the advantages and the flexibility of distributed and hierarchical cortical-like architectures against solutions based on a conventional systemic coupling of sensing and motor components, which in general poses integration problems since too heterogeneous and complex processes must be coupled.

INTRODUCTION

There is a growing attention for moving the active vision perspective from systems in which just the effects of action influence the perception, to systems where the acting itself, and even its planning, operate in parallel with perception, thus really closing the loops and taking full advantage of a concurrent/anticipatory perception/action processing. From this perspective, the motor system of a humanoid should be an integral part of its perceptual machinery (e.g., see Int. Journal of Humanoid Research Special issue on the "Active Vision of Humanoids", 2010). Traditionally, however, in robot vision systems, perception-action loops close at a "system level" (by decoupling *de facto* the vision modules from those dedicated for motor control and motor planning), and the exploitation of the computational effects of the eye movements on the visual processes are very seldom in artificial artifacts.
The limit of the approach was that solving specific high-level tasks usually requires sensory-motor shortcuts at the system level, and specific knowledge-based rules or heuristic algorithms have to be included to establish behavioural consistency relationships among the extracted perceptual features and the desired actions. The risk is to abandon distributed representations of multiple solutions to prematurely construct integrated description of cognitive entities and commit the system to a particular behaviour.
Conversely, our claim is that early/complex interactions between vision and motor control are crucial in determining the effective performance of an active binocular vision system with a minimal amount of resources and coping with uncertainties and inaccuracies of real systems.

There is ample evidence for the pivotal role of programmed eye-movements in the computations that are performed in the process of seeing (as opposite to "looking at"). Yet, how to profitably integrate these accumulating evidences with the computational theories of stereo vision has not been fully exploited, as rectified images are still calculated in current humanoid active-disparity vision modules, relying on the encoders data, only. The complexity of integrating efficiently and with flexibility the different aspects of binocular active vision indeed prevented till now a full validation of the visuomotor approaches to 3D perception in real world situations. We believe that the advantages of binocular visuomotor strategies could be fully understood only if one jointly analyzes and models the problem of neural computation of stereo information, and if one takes into account the limited accuracy of the motor system. Unfortunately, models in this joint field are few (Theimer & Mallot, 1994; Hansard & Horaud, 2008; Read & Cumming, 2006) and rarely address all the computational issues.

In this work we defend a visuomotor approach to 3D perception by proposing the instantiation of visuomotor optimization principles concurrently with the design of distributed neural models/architectures that can efficiently embody them. Specifically, we present two case-studies that show how large-scale networks of V1-like binocular cells can provide a flexible medium on which to base coding/decoding adaptation mechanisms related to sensorimotor schema: at the coding level, the position of the eyes in the orbits can adapt the disparity tuning to minimize the necessary resources, while preserving reliable estimates (i.e., adjustable tuning mechanisms based on the posture of the eyes to improve depth vision (Chessa, Sabatini & Solari, 2009)). At the decoding level, read-out mechanisms of the disparity population code can specialize to gain vergence control servos over a wider range of disparities than that would be possible through an explicit calculation of the disparity map (Gibaldi, Canessa, Chessa, Solari & Sabatini, 2010).

BACKGROUND

Purposive (active) vision is an important source of information and provides a number of cues about the 3D layout of objects in a scene that could be used for planning and controlling goal-directed behaviors. Although computer and robot vision are today making technological progress, the current active vision solutions are increasingly driven by applications like robotic manipulation or surveillance and are still far from reaching a real purposive behavior. More precisely, research in robot vision is more geared towards active "looking at" scenes and objects rather that "seeing", gathering the visual data and having to act in a limited time and space. According to the sensorimotor theory of perception (O'Regan & Noë, 2001; Noë, 2004), "seeing" is something we do rather than a sequence of hierarchical interpretative processes. From this perspective, the experience of "seeing" in not necessarily generated, but it expresses itself in the behavior.

Following these premises, we can study the topic of seeing using an embodied artificial intelligence.

Perception-Action Synergies in Purposive Vision

An intelligent perceptual agent must be able to interact with its environment through sensing, processing and transforming information about the surrounding world into different levels of representation, and eventually solve complex problems in a real dynamic world. The idea of constructing such an intelligent system has been the ultimate goal of the Artificial Intelligence (AI) and has led this Community to investigate a number of robot capabilities, such as: to generate and execute complex plans; to perform online resource allocation; to deal with problems as they arise in real-time (reaction); and to reason with incomplete information and unpredictable events. In view of the inherent difficulty of the problem, while robots are now commonplace in today's manufacturing industry, these devices are typically characterized by a lack of the ability to adapt to a dynamic environment. In short, they lack intelligent sensing and action. One of the main reasons of this failure must be ascribed to the traditional AI view an intelligent system as a set of independent cognitive modules: perception, planning, learning, and execution (see Fig. 1A). Within this view, complex problems are solved by the execution of individual modules that can exchange information through well-defined interfaces. The assumption of independent processes brought up enormous difficulties because of the unexpected interactions of those parts (Costa, Rillo, Barros & Bianchi, 1998).

Figure 1. The modular view of an intelligent system: "Complex problems are solved by the execution of individual modules that can exchange information through well defined interfaces" (Costa, Rillo, Barros & Bianchi, 1998) (A). The synergic view of an intelligent system: "Perception should not be considered

as a self-contained module, but as an entity containing other intelligent capabilities, i.e., planning, reasoning and learning, all of which cooperate to solve specific tasks" (Costa, Rillo, Barros & Bianchi, 1998) (B).

The embodiment concept, which describes the ability of the system to acquire knowledge of its world from the consistent coupling of its own action and perception, has rooted in the scientific Community and has characterized many significant advances in Cognitive Vision research of these years.

During the last decades, there has been a growing interest in the use of active control of image formation to simplify and accelerate scene understanding. Since the pioneering ideas of Bajcsy (1988) and Aloimonos, Weiss & Bandopadhay (1987), the approach has been extended by several groups (Ballard, 1991; Eklundh & Pahlavan, 1992; Brown, 1990) demonstrating how multiple simple behaviors (e.g., saccadic, vergence, vestibulo-ocular reflex and neck motion) may be used for controlling visual perception. Different computational paradigms have been introduced through the years, by developing the initial concept of "animate vision" in which the visual calculations are embedded in a sensory-motor repertoire that reduces degrees of freedom and yields to a number of computational advantages.
In more general terms, an important research direction emerged, with the aim of designing a robotic agent that could autonomously acquire visual skills from its interactions with an uncommitted environment in order to achieve some set of goals. Learning new visual skills in a dynamic, task-driven fashion so as to complete an a priori unknown visual task is known as the purposive vision paradigm. Purposive vision is essentially orthogonal to the reconstructionist approach of computer vision (Marr, 1982).

The goal of reconstructive or recovery vision (Dean, Allen & Aloimonos, 1995) is to derive indeed, from one or more images of a scene, an accurate three-dimensional description of the objects in the world and quantitatively recover its properties from image cues such as shading, contours, motion, stereo, color, etc. Thus recovery emphasizes the study of task-independent visual abilities carried out by a passive observer through a hierarchy of bottom-up processes (from primal sketch, to 21/2-D, and then to scene interpretation) (Marr, 1982). By contrary, the purposive vision paradigm does not consist in generating a complete, detailed, symbolic 3D model of the surrounding environment, rather it emphasizes the fact that vision is task-oriented and that the performing agent should focus its attention only on the parts of the environment that are relevant to its task. Therefore, purposive vision stresses the dependency between action and perception: selecting actions becomes the inherent goal of the visual sensing process. This paradigm often leads to the breakdown of the visual task into several sub-problems that are managed by a supervision module that ultimately selects the suitable reactions.

According to the purposive approach, vision should not be considered as a self-contained module, but as an entity containing other intelligent capabilities, i.e., planning, reasoning and learning, all of which cooperate to solve specific tasks. Intelligence should not be divided into isolated cognitive modules, but decomposed in terms of behaviors (Polpitiya, Ghosh, Martin & Dayawansa, 2004). Thus a synergic approach replaces the modular (see Fig. 1B). Each ring corresponds to a perception/action-loop that can be associated to specific behaviors decomposed with increasing degrees of complexity.

Purposive vision is very close to active vision, and these concepts are sometimes used interchangeably. Just like purposive vision, active vision criticizes the passive point of view of the reconstructionist approach, and it argues that visual perception is an exploratory activity. However, active vision is essentially interested in experimental setups where the position of the visual sensors can be governed by the effectors. This approach is evidently inspired by human vision, for which muscles can orientate the head, the eyes and the pupils, and the visual sensor is regarded as a component in a larger system, able to make either deliberate actions to simplify the visual sensing process, or to react to visual events. By learning to control such effectors, the agent can acquire better information and resolve ambiguities in the visual data, for example by acquiring images of a scene from different viewpoints. Some ill-posed

problems in computer vision become well-posed by employing active vision (Bajcsy, 1988; Krotkov, Henriksen & Kories, 1990; Ballard & Ozcandarli, 1988; Brown, 1990).
From this perspective, (1) general perceptual problems that are ill posed and nonlinear for a passive observer become well posed and linear, (2) computational resources are directed to important areas ("attentive" mode (Clark & Ferrier, 1988)), therefore improving the efficiency of the visual systems.

In this respect, the concept of "purposive vision" is much more general and powerful than that of "active vision" since it does not essentially confine interactivity to the positioning of visual sensors. Though, to a large extent, current models of vision do not fully exploit its potentialities. Indeed, models of vision generally assume action and perception are still sequential processes: it is the effect of action that influences the perception (e.g., eye movements serve to select a scene for perception) and not the acting itself, even less its planning (but see Hamker (2005a) and Hamker (2005b)).

A visuomotor approach to 3D perception

In absence of motion parallax, and disposing of a binocular stereo vision system, 3D information can be gathered without introducing active components (except for covering a larger field of view). In such conditions, stereopsis (i.e., position parallax) is usually thought as a static problem, since the disparity map obtained by a fixed-geometry stereo camera pair, with (nearly) parallel axes, is sufficient to reconstruct the 3D spatial layout of the observed scene, and camera movements are often regarded as unnecessary complicating factor. Things are dramatically different if we consider a binocular foveated system with a vergent stereo geometry. In such conditions, the perceptual process intrinsically gains a motor dimension as the 3D information is collected dynamically with respect to the fixation point. The eye rotations, although insufficient *per se* to provide depth cues (but see Santini & Rucci (2007)), strongly constraint the stereo vision processing and visual information is gathered through a continuous interaction with the environment. This is especially true for visual exploration of the peripersonal space when large values of vergence occur, and disparity geometry strongly depends on the viewing geometry.
In general, the importance of the visuomotor aspects of 3D visual perception can be understood not just to develop algorithms that enable robots to cope with changes in the environment, but, and more importantly, to acquire task-independent skills as a living being. Even while the problem of controlling on a visual basis the vergence of a stereo camera system has specific and rather straightforward solutions, the joint treatment of the vergence control and of 3D perception still represents a challenging cognitive problem. The zero-disparity condition in the fixation point solves indeed the vergence task, but nullifies the visually-based information for the 3D position of the fixated target point. Only the residual disparities elsewhere in the visual field are cues for stereopsis. The momentarily existing (and continuously changing) fixation point, i.e. where the system verges, becomes a reference that can be parameterized by the relative orientations of the eyes. Moreover, how the system verge in terms of the full three dimensional rotational position of the cameras has an impact for the accuracy of stereopsis. Changing camera positions influences the local shape of the zero disparity surface near the fixation point (i.e., the surface horopter), and optimal motor control provides several advantages for vision, such as optimal use of the range over which the disparity detectors operate, or a "corrective" warping of the images to adjust the slant of the observed surface when it deviates from the stereotypical frontoparallel case. Although these facts have been understood in the psychological literature for over a long time (Howard & Rogers, 2002), it is only recently that computational and theoretical approaches attempted an engineering formulation of these concepts in order to provide operative guidelines to quantitatively analyze their visual (and motor) advantages, and optimal design criteria for active artificial vision systems (Schreiber, Crawford, Fetter & Tweed, 2001; Schor, Maxwell, McCandless & Graf, 2002; Read & Cumming, 2004; Hansard & Horaud, 2008; Hansard & Horaud, 2010). From a behavioural point of view, the empirical derivation of the horopter has been used to simulate the optimally stereo-viewed surface in foveal vision (Schreiber, Tweed & Schor, 2006; Schreiber, Hillis, Filippini, Schor & Banks, 2008),

and complementary non-visual information exploited to estimate absolute distances when the system is engaged in reaching tasks towards non foveally viewed targets (Greenvald & Knill, 2009; Blohm, Khan, Ren, Schreiber & Crawford, 2008). From a computational point of view, a debate on the role of vertical disparities to calibrate depth perception pointed out that the vertical disparity is not simply tolerated, but it is actively detected and used in perception (Read & Cumming, 2006; Read, Phillipson & Glennerster, 2009; Serrano-Pedraza & Read, 2009; Serrano-Pedraza, Phillipson, & Read, 2010) and to guide vergence behaviour (Yang, FitzGibbon & Miles 2003; Sheliga & Miles, 2003; Gibaldi, Canessa, Chessa, Solari & Sabatini, 2010, but see Rambold and Miles, (2008)). As a whole, the 2D vector disparity pattern should be recovered when incorporating eye movements since limiting the search of the binocular correspondences on the epipolar lines predicted by the current position of the eyes would be vulnerable to any inaccuracies in the sensed eye position. Exploiting mutual dependencies between the disparity patterns and the epipolar geometry e.g., by adapting the computational resources or the processing to the fixation constraint, can be a viable solution to the inaccuracy of eye position sensing (cf. the approach proposed in Chessa, Canessa, Gibaldi, Solari and Sabatini (2009)). In general, to overcome the difficulties in obtaining an accurate estimation of epipolar geometry, Monaco, Bovik and Cormack (2009) demonstrated a symbiotic relation between foveation and uncalibrated active vision systems to minimize the number of points per epipolar space and thus improve the efficiency of the search for stereo matches.

Phase-based stereo vision processing

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 $I^L[\mathbf{x}+\delta(\mathbf{x})]$ (Sanger, 1988). 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 amplitude vanishes and the phase becomes unreliable (Fleet, Jepson & Jenkin, 1991). This property of the phase signal yields good predictions of binocular disparity by

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

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:

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

As a consequence of the linear phase model, the instantaneous frequency is generally constant and close to the tuning frequency of the filter $(\phi_x \approx k_0)$, 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 (Fleet, Jepson & Jenkin, 1991).

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 have both horizontal and vertical components. 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 detected as a phase shift $\mathbf{k}^T\boldsymbol{\delta}(\mathbf{x})$ in the local spectrum, where $\mathbf{k}(\mathbf{x})$ is the local (i.e., instantaneous) frequency vector defined as the phase gradient:

$$\mathbf{k}(\mathbf{x}) = \nabla\phi(\mathbf{x}) = \left(\frac{\partial\phi(x,y)}{\partial x}, \frac{\partial\phi(x,y)}{\partial y}\right)^T \qquad \text{Eq. 1}$$

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. 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 structure is intrinsically 1D, with a dominant orientation θ_s (let us think of an oriented edge or of an oriented grating with frequency $\mathbf{k}_s = (k_s \sin\theta_s, k_s \cos\theta_s)^T$, as extreme cases), the aperture problem (Morgan & Castet, 1997) restricts detectable disparity to the direction orthogonal to the edge (i.e., to the direction of the dominant frequency vector $\mathbf{k}_s$):

$$\boldsymbol{\delta}_{\theta_s}(\mathbf{x}) = \frac{\mathbf{k}_s}{k_s}\frac{\lfloor\Delta\phi_{\theta_s}(\mathbf{x})\rfloor_{2\pi}}{k(\mathbf{x})} \approx \frac{\mathbf{k}_s}{k_s}\frac{\lfloor\Delta\phi_{\theta_s}(\mathbf{x})\rfloor_{2\pi}}{k_s}, \qquad \text{Eq. 2}$$

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 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. 2 becomes:

$$\boldsymbol{\delta}_{\theta_s}(\mathbf{x}) \approx \frac{\mathbf{k}_0}{k_0}\frac{\lfloor\Delta\phi_{\theta_s}(\mathbf{x})\rfloor_{2\pi}}{k_0},$$

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 is the one orthogonal to the filter's orientation θ, i.e., the projection in the direction of the filter's frequency:

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

Again, $\mathbf{k}(\mathbf{x})$can be derived by Eq. 1 or approximated by the peak frequency of the Gabor filter 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 (Theimer & Mallot, 1994; Sabatini, Gastaldi, Solari, Pauwels, Van Hulle, Diaz, Ros, Pugeault & Krueger, 2010), 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. 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. 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$\boldsymbol{\delta}(\mathbf{x})$:

$$\boldsymbol{\delta}(\mathbf{x}) = \arg\min_{\boldsymbol{\delta}(\mathbf{x})} \sum_\theta c_\theta(\mathbf{x})\left(\delta_\theta(\mathbf{x}) - \frac{\mathbf{k}_0^T}{k_0}\boldsymbol{\delta}(\mathbf{x})\right)^2,$$

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.

CORTICAL ARCHITECTURES FOR 3D ACTIVE MEASUREMENTS IN THE PERIPERSONAL SPACE

Distributed Representation of Binocular Disparity

The phase-based disparity estimation approach presented in the previous section implies, for each spatial orientation channel θ (and for any given scale), explicit measurements of the local phase difference $\Delta\phi$ in the image pairs, from which we obtain the direct measure of the binocular disparity componentδ_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.

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 (RFs) of V1 simple cells (Sanger, 1988; Fleet, 1994; Fleet, Wagner & Heeger, 1996a; Qian, 1994; Ohzawa Freeman & DeAngelis, 1990; Prince, Cumming & Parker, 2002). 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 $\mathbf{x}$ and oriented along θ, can be written as:

$$ {}^{\theta}_{\Delta\psi} r_{s,\psi_0}(\mathbf{x}) = {}^{\theta} r_{l,\psi^L}(\mathbf{x}) + {}^{\theta} r_{r,\psi^R}(\mathbf{x}) $$

where:

$$ \begin{aligned} {}^{\theta} r_{l,\psi^L}(\mathbf{x}) &= I^L * h^L\left(\mathbf{x};\theta,\psi_0+\psi^L\right) \\ {}^{\theta} r_{r,\psi^R}(\mathbf{x}) &= I^R * h^R\left(\mathbf{x};\theta,\psi_0+\psi^R\right) \end{aligned} \qquad \text{Eq. 3} $$

and:

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

is a real-valued RF, 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, and Θ is the rotation matrix defined by:

$$ \Theta = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} $$

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. 2A) and their response is defined as:

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

For any fixed orientation, if we characterize a “quadrature pair” of simple cells by a complex-valued RF:

$$ h(\mathbf{x}) \equiv h_C(\mathbf{x}) + jh_S(\mathbf{x}) = g(\mathbf{x};\psi) $$

where:

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

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

$$\begin{aligned} {}^{\theta}_{\Delta\psi} r_c(\mathbf{x}) &= \left|{}^{\theta}_{\Delta\psi} r_{s,0}(\mathbf{x}) + {}^{\theta}_{\Delta\psi} r_{s,\pi/2}(\mathbf{x})\right|^2 = \left|Q^L(\mathbf{x})e^{j\psi^L} + Q^R(\mathbf{x})e^{j\psi^R}\right|^2 = \\ &= \left|e^{j\psi^L}\left(Q^L(\mathbf{x}) + Q^R(\mathbf{x})e^{j\Delta\psi}\right)\right|^2 = \left|Q^L(\mathbf{x}) + Q^R(\mathbf{x})e^{j\Delta\psi}\right|^2 \end{aligned} \qquad \text{Eq. 4}$$

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. 4 formally establishes the equivalence between phase-based techniques and energy-based models (Qian & Mikaelian, 2000). Indeed, the maximum of r_c responses is obtained when the two phasors Q^L and Q^R are aligned in the complex plane (see Fig. 2B), that is when $\Delta\psi$ compensates for the different Fourier phases of the right and left image patches within the cell's RF (cf. Sanger (1988)).

Notwithstanding the formal equivalence between phase-based techniques and energy-based models, the latter 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.

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, gL and gR, respectively (A). The responses of the two "quadrature pair" for the left and the right eye are characterized by a phase shift $\Delta\phi$ proportional to the disparity (LEFT). The maximum response of a complex cell are obtained when the two phasors are aligned. This is obtained when the RF phase shift $\Delta\psi$ equals $\Delta\phi$ (RIGHT). (B).The population of binocular receptive fields for each retinal location (C).

On the ground of these computational principles, a large-scale cortical network has been designed to encode disparity information from a stereo image pair. For each pixel, the network employs a population of simple and complex cells sensitive to $N_O \times N_P$ 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. 2C). For each simple cell we can control the orientation θ of the binocular RF $h(\mathbf{x})$ 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 $\delta_{pref} = \Delta\psi_\theta / k_0$, along the direction orthogonal to θ. The spatial frequency k_0 and the spatial envelope are fixed on the basis of optimal design criteria described in Sabatini, Gastaldi, Solari, Pauwels, Van Hulle, Diaz, Ros, Pugeault & Krueger (2010). The complex cell inherits the spatial properties of the simple cells, and its response $r_c(\mathbf{x})$ is given by Eq. 4.

For each orientation, the population is, in this way, capable of providing reliable disparity estimates in the range between $\pm\Delta$, where $\Delta = \Delta\psi_{\max}/k_0$ can be defined as the maximum detectable disparity of the population.

Embedding fixation constraints in binocular energy-based models of depth perception

When we look around in a cluttered environment, or when we inspect a small object, the eyes coordinate to make the lines of sight to intersect in the target dynamically, thus ensuring binocular fusion and accurate scanning of the object. In general (but in particular for relatively large vergence angles), binocular coordinated movements of the eyes affect the two-dimensional (horizontal and vertical) disparity pattern in the peripheral part of the image of the fixated object, as well as in the background,. Consequently, the strategy adopted to move the eyes can directly influence the perception of depth, the visual behavior, and eventually the 3D spatial awareness of the world around us.

The epipolar lines, defined as the loci of all the possible matching points for every retinal location, represent the most convenient/immediate way to characterize the disparity patterns experienced by a binocular vergent system (see Fig. 3A). When we look straight ahead at infinity (i.e., with parallel optical axes) all the epipolar lines are horizontal. Conversely, whenever the gaze changes and the vergence increases the epipolar lines move and become more and more tilted. This movement causes an increase of the observed disparities, and, as a consequence, the vision system has to cope with larger search zones within which the stereo correspondences have to be found. From this perspective, having a general design strategy for the oculomotor system behavior that minimizes the motion of the epipolar lines, would reduce the search zone and thus the computational cost of finding visual correspondences (see Fig. 3C). On the way around, the predictable components of disparity, depending on the relative position of the eyes, may be used as *priors* to optimally allocate the computational resources to ease the recovery of the (unpredictable) components of disparity, which are dependent on the structure of the scene, only. This suggests the possibility of a *mutual calibration* of the vision and the oculomotor system to compensate disparity components due to the epipolar geometry.

The eye with its three degrees of freedom could, in principle, assume an infinite number of torsional postures for any gaze direction. Though, Listing’s Law states that the eye assumes only those orientations that can be reached from the primary position by a single rotation about an axis in a plane called Listing’s plane (Tweed & Vilis, 1990). This corresponds, according to Donders, to a single possible torsional position for each combination of the azimuth and the elevation angles. During binocular convergence, Listing’s Law is still valid, but the plane is rotated temporally both for the left and the right eye. These convergent-dependent changes of orientation of Listing’s plane have been referred to as the binocular extension of Listing’s Law or L2 (Mok, Ro, Cadera, Crawford & Vilis, 1992). From a functional point of view, Listing's Law can be understood as a limiter of the degrees of freedom of the oculomotor system (when the behavioral situation demands or permits it), which constraints torsional components to reduce rotation (Listing's Law, or LL) (Tweed & Vilis, 1990), or to reduce the cyclovergence and restricts the motion of the epipolar lines (Binocular LL, or L2) (Mok, Ro, Cadera, Crawford & Vilis, 1992), thus permitting stereo matching to work with smaller search zones.

Although, from a conceptual point of view, the oculomotor parameterization of active stereopsis is a well-established issue (Jenkin, 1996; Hansard & Horaud, 2008), mapping the oculomotor constraints into the neural population coding and decoding strategies is still an open problem. In this section, we describe how it is possible to exploit such oculomotor information in the specification of the architectural parameters/resources of the distributed representation of disparity information.

To this end, we have calculated natural scene disparity statistics for a binocular vergent system engaged in natural viewing in a peripersonal workspace. For comparison, both the Helmholtz (i.e., tilt-pan, zero torsion) and the more sophisticated oculomotor models based on LL and its binocular extension L2 have been considered (see Fig. 3B).

Figure 3. Given a image point on the left retina, all its possible corresponding points on the right one lie on line: the epipolar line (heavy black line). The epipolar lines change their orientation depending on the relative orientation of the two eyes (A). In the three represented cases the eyes fixate the same point in the world but in different ways, following an Helmholtz (or Tilt-Pan), a Listing and a L2 system. The surface is always orthogonal to the gaze line. The blue and red lines represent the projections of the horizontal and vertical meridians for the left and the right eye on the surface. In case of perfect alignment the two lines superimpose becoming purple (B). For the three systems in (B) the epipolar lines for a grid of 3x3 retinal points of the left eye are depicted for different gaze direction. The reference point on the left eye are represented by an open circle, while the epipolar lines are represented by blue lines. The value of version and elevation angle varies in a range [-40° 40°]. The epipolar lines tilt and move depending on the version and the elevation of the gaze direction, describing an area: the correspondence search zone. It is possible to see how the spread of the search zone is linked to the geometry of the system (C).

Data acquisition - For the simulations shown in the following, we first captured 3D data from a real-world scene by using a 3D laser scanner (Konica Minolta Vivid 910), with the optimal 3D measurement operating range from 0.6 m to 1.2 m, which is appropriate for analyzing the disparity information experienced by an active observer in his/her peripersonal space. The system allows also capturing the color textures at a resolution of 640 × 480 pixels. Each scan contained up to 307, 200 points within a variable field of view, which was adjusted with respect to the size of the object to be scanned. For this work we considered cluttered desks with a collection of hundred real-world objects. The whole scene, as well as the single objects were scanned, registered and merged together to obtain full models of more than 13,000,000 of points each (see Fig. 4A). Off-line registrations of data guarantee an accuracy of about 0.1 mm. A full 360-degree view of the scene is acquired to minimize the occlusion problems that occur when one simulates changes in the vantage point of the virtual observer.

Simulated fixations in the acquired peripersonal scenes – The real-world environment, captured by the 3D laser scanner, is then "explored" through an active vision simulator. Such simulator has been implemented in C++, using OpenGL libraries and the Coin3D toolkit (http://www.coin3d.org/) developed for effective 3D graphics rendering. This system is capable of handling the commonly used 3D modeling formats (e.g., VRML), and thus the data acquired by the 3D laser scanner. To obtain the toe-in stereoscopic visualization of the scene, useful to mimic an active stereo vision system rather than to make humans perceive depth, we have modified the SoCamera node of the Coin3D toolkit. Moreover, the developed tool allows us to access the buffers (see Fig. 4B) used for the 3D rendering of the scenes. The 3D data and the textures are loaded in the active vision simulator, then the left and right projections, the horizontal and the vertical ground truth disparity maps, are obtained, for each possible fixation point. More details on the simulator are reported in (Chessa, Sabatini & Solari, 2011).
The developed tool has been used to create a database of real-world range data and stereo image pairs for a variety of fixations (see Fig. 4C), in order to guide modeling and for algorithmic and behavioral benchmarks in real-world but fully measured environments.

Figure 4. Example of real-world scene acquired by the laser scanner. Together with the 3D data the system is able to attach the real color texture to the scanned objects (A). Two outputs of the active vision system simulator: the Z buffer (B) and the left and right image pairs (C). The disparity δ can be divided into two components: one, unpredictable, due to the scene, called residual disparity δ_s, and one, predictable, due the geometry of the adopted vision system, called epipolar disparity δ_e (D). The mean vector disparity patterns and the standard ellipses,

averaged over all the fixation, are depicted for a Helmholtz, a Listing and a L2 system. Here only a grid of 7x7 retinal points are shown for the sake of clarity (E).

Statistical analysis - For a given eye posture we computed the distribution of the horizontal and vertical disparities for all the objects whose images fall within an angle of ±22.5° in both retinas. The other parameters used were: a resolution of 601 × 601 pixels, a focal length of 10 mm, and an interocular distance of 6 cm. We repeated the calculation for 100 different vantage points, corresponding to different positions and orientations of the cyclopean visual axis, and for a set of fixation points. The fixation points varied in the range of 0◦ ÷360° for the azimuth angle, and in the range of 0◦ ÷31.82° for the polar angle. More precisely, the fixation points were obtained by backprojecting a 11×11 grid of equally spaced points of the cyclopean retina on the closest visible surface of the scene. Under the same experimental conditions, the disparity patterns were calculated for two different eye movement paradigms (Listing and Tilt-Pan). Figures 14 and 15 demonstrate that large vertical disparities can occur in the peripheral field of view, especially for tertiary eye positions. The mean vector disparity patterns, together with their standard ellipses (measuring the joint dispersion of the bivariate distribution) are shown in Fig. 4E. It is worth noting that, though, as expected, the mean disparity patterns calculated for each fixation are characterized by significant differences (not shown), these are attenuated averaging over all the fixations we considered.

From the analysis of the simulation data, it is worth noting that the mean value of the disparities changes with the fixation point, thus it is possible to distinguish two disparity components: the first ($\boldsymbol{\delta}_s$), unpredictable, due to the structure of the 3D scene and the second ($\boldsymbol{\delta}_e$), more predictable, due to the geometry of the binocular system (see Fig. 4D):

$$\boldsymbol{\delta} = \boldsymbol{\delta}_s + \boldsymbol{\delta}_e \,.$$

Adjustable energy models - The component of the disparity due to the epipolar geometry can be embedded in the distributed representation of disparity information with the position shift mechanisms (Fleet et al., 1996).

The position-shift model assumes that there is a population of simple cells whose left and right RFs are always identical in shape, but can be centered at different spatial locations. Accordingly, we can consider a family of binocular energy neurons whose right monocular receptive field is shifted by a set of offsets $\mathbf{d} = \boldsymbol{\delta}_e$ with respect to the center of the corresponding monocular receptive field in the left retina. Formally, the response of a simple cell with RF center in **x** and oriented along □**,** can be written as:

$${}^{\theta}r_{s,\psi_0}(x;d) \;=\; {}^{\theta}r_{L,\psi_0}(x) + {}^{\theta}r_{R,\psi_0}(x-d)$$

where ${}^{\theta}r_{L,\psi_0}(x)$, ${}^{\theta}r_{R,\psi_0}(x)$ are expressed as in Eq. 3. Still, we can define the response of a binocular energy complex cell as the squared sum of a quadrature pair of simple cells:

$${}^{\theta}r_c(x;d) \;=\; {}^{\theta}r^2_{s,0}(x;d) + {}^{\theta}r^2_{s,\pi/2}(x;d) \,.$$

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

$$\delta^{\theta}_{pref} = d^{\theta} \,.$$

By using jointly the position-shift and the phase-shift disparity encoding mechanisms, the spatial relationships between target points on the left eye and the mean of the corresponding points for the right eye are embedded into a hybrid energy-based model where phase-shifts and position-shifts play a different role: position-shifts are used to compensate the global components of the averaged disparity pattern over all the fixations, whereas phase-shifts are used to estimate the residual 2D disparity.

Figure 5. The subplots represent a grid of 7x7 retinal points for a value of the gaze corresponding to zero elevation and zero version angles. For different points of the left retina (open circles) we estimated the corresponding point for the right retina (color dots) by a population of disparity detectors. The disparity relates to point image projections of randomly positioned planes in the peripersonal space, when the fixation point is in the primary position. The red dots represent the mean of the disparities, whereas the black dots represent the true mean disparity. Distribution of the estimated disparities between the left and the right retina are depicted without and with the compensation of the predictable components of the disparity pattern (A). Disparity estimation by embedding fixation constraints into the binocular energy model for a stereo pair representing a fronto-parallel plane and an indoor scenario obtained by the active vision simulator. (TOP) Ground truth horizontal and vertical disparity maps. (MIDDLE) Estimation of the disparity by using the distributed architecture without embedding any fixation constraint. (BOTTOM) Estimation of the disparity by using the distributed architecture by embedding the fixation constraints: a position shift derived from the mean values of disparities averaged over all the fixation points. The results are obtained by using 43×43 pixels receptive fields, tuned to a disparity range from −8 to 8 pixels.

Fig. 5A shows the distribution of the estimated disparities between the left and the right retinas without and with a compensation of the predictable components of the disparity pattern. For different points of the left retina (open circles) we estimated the corresponding point for the right retina (colored dots) by a population of disparity detectors. The disparity relates to point image projections of randomly oriented surfaces in the peripersonal space. The red dots represent the mean of the estimated disparities, whereas the black dots represent the true mean disparity. It is worth noting that, by embedding the mean values computed with respect to the reference situation, the mean values of the estimated disparities (red dots) become closer to the true values of disparities (black dots). Fig. 5B shows the estimation of the 2D disparity without global components compensation and by embedding these components into the model, for a frontoparallel plane and for an indoor scenario obtained by the active vision simulator. It is worth noting that the reliability of the disparity representation is improved, by embedding the component due to the epipolar geometry of the system.

	# cells	AVG	STD	%
Venus stereo pair				
without phases shifts redistribution	33	0.84	0.63	91
with phases shifts redistribution	17	0.72	0.56	91
Tsukuba stereo pair				
without phases shifts redistribution	33	0.36	0.37	56
with phases shifts redistribution	17	0.28	0.20	91

Table 1. Average error (AVG), standard deviation of the error (STD) and density (expressed as the percentage of estimated values with respect to the total number of pixels in the image) for the Venus and the Tsukuba stereo pairs, without and with the redistribution of the cells. The redistribution has been performed by taking into account the known sign of the true disparity values. The first column shows the number of cells (for each spatial orientation) necessary to obtain the estimation of the disparity.

Moreover, the activity of the population of neurons in the distributed representation can be adapted by changing the distribution of the units, and by this minimize the necessary resources while preserving reliable estimates. To this aim, we have also tested if a prior knowledge of a particular feature (e.g the sign of the disparity value of the disparity, or the range of the values) can be used to redistribute the sensitivity coverage of the cells' population and its density, by properly choosing the phase-shifts, while keeping fixed the other parameters. We show the compared results, obtained by using two stereo image pairs (Venus and Tsukuba) for which the ground truth is available (Scharstein & Szeliski, 2002). We have redistributed the cells of the population, accordingly to the (known) sign of the disparity and we have compared the results with the ground truth disparity maps. Table 1 shows how the same (and in certain cases better) reliability is obtained by halving the units of the population.

In conclusion, the position shifts mechanism can be seen as a pre-wired design strategy that takes into account an initial adaptation of the system with respect to a "typical" viewing condition.

Reading-out the disparity population code to specialize vergence control servos

Previous vergence models that are based on a population of disparity detectors, require first the computation of the disparity map for the extraction of the control signals (Theimer & Mallot, 1994; Patel, Ogmen & Jiang, 1996), thus limiting the functionality of the vergence system within the range of disparities in which the system is able to fuse the left and right images. Making a parallel with the biological system, this means that vergence eye movements would be reliable only inside the Panum's area, where they are not necessary. Though each neuron of Medial Superior Temporal area (MST) sensitive to retinal

disparity, has been found to encode only some limited aspects of the motor response for vergence eye movements, the activity of the whole population directly correlates with the magnitude, direction, and time course of the initial vergence motor response (Takemura, Inoue, Quaia & Miles, 2001; Takemura, Kawano, Quaia & Miles, 2006). Mimicking the behaviour of the cells of the MST area, we present a model that, by combining the response of a population of complex cells, does not take a decision on the disparity values (disparity map), but extracts disparity-vergence responses that allows us to nullify the disparity in the fovea, even when the stimulus disparities are far beyond the fusible range. Furthermore, on the basis of the Dual Mode theory (Semmlow, Hung & Ciuffreda, 1986; Hung, Semmlow & Ciuffreda, 1986), vergence eye movements are not controlled by a simple continuous feedback system (Krishnan & Stark, 1977), but they exhibit dual-mode slow and fast responses. Accordingly, the model provides two distinct vergence control mechanisms: a "fast" mode enabled in the presence of large disparities, and a "slow" mode enabled in the presence of small disparities. An additional signal provides the switch between the two modes, according to the stimulus disparities.

The population used is made of complex cells that are, by construction, tuned to oriented disparities, i.e. jointly tuned to horizontal δ_H and vertical disparities δ_V. By recovering the full disparity vector, the disparity detectability range would still be limited to $\pm\Delta$, and the horizontal component of the full disparity vector will then used for the control of horizontal 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 solving the aperture problem and thus without explicit computation of the disparity map. To this end, we can consider that to drive horizontal vergence, the meaningful feature is δ_H only, thus 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 RF is used as a degree of freedom to extend the sensitivity range of the cells' population to horizontal disparity stimuli. 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:

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

Fig. 4C shows the horizontal disparity tuning curves obtained by the population for different orientations of the receptive fields (blue lines). 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 COM decoding strategy, the estimated horizontal disparity δ_H^{est} is obtained by:

$$\delta_H^{est} = \frac{\sum_{j=1}^{N_O}\sum_{i=1}^{N_P} \frac{\Delta\psi}{2\pi k_0 \cos\theta} r_c^{ij}}{\sum_{i=1}^{N_P} r_c^{ij}}$$

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 estimate of the disparity can be considered correct when the stimulus disparity is within $[-\Delta,\Delta]$. By analyzing the tuning curves of the population 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 N_O different estimates of the disparity:

$$\delta_{H,\theta_j}^{est} = \frac{\sum_{i=1}^{N_P} \frac{\Delta\psi}{2\pi k_0 \cos\theta_j} r_c^{ij}}{\sum_{i=1}^{N_P} r_c^{ij}}$$

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. $\delta_H^{\theta=0} \to \infty$). 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 taken into account.

Control signal extraction –

A desired feature of the disparity tuning curves for vergence is an odd symmetry with a linear segment passing smoothly through zero disparity (see Fig. 6D), which defines critical servo range over which changes 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. Given a stimulus with an horizontal disparity δ_H, we want to combine the population responses in order to extract a vergence control proportional to the disparity to be reduced, regardless of a possible non-zero vertical disparity δ_V. To this end, starting from the 2D (multichannel) responses of the population of binocular cells, we exploit the responses at different orientations to design linear servos that work outside the reliability range of disparity estimation. Yet, to cope with the attendant sensitivity to vertical disparity, which is an undesirable effect that alters the control action, a limiting factor of the influence of vertical disparity must be introduced. The desired disparity vergence response curves r_v^k is approximated by a weighted combination of the population cell responses, where disparity tuning curves act as basis functions:

$$r_v^k = \frac{\sum_{j=1}^{N_O}\sum_{i=1}^{N_P} w_{ij}^k r_c^{ij}}{\sum_{j=1}^{N_O}\sum_{i=1}^{N_P} r_c^{ij}}$$

Figure 6. Simplified scheme of the neural circuitry involved in the control of vergence eye movements. The left and right images are processed by a population of disparity detectors, inspired by complex cells of area V1. The population produces a distributed representation of the retinal disparity, and through convolution with weighting kernels it is decoded in order to obtain a family of vergence cells that are able to provide a direct vergence motor response, that is sent to the ocular plant. Since the task is to drive vergence eye movements so as to improve the fixation and the estimation of disparity, the information is gathered only from the central (parafoveal) portion of the visual field (A). The υ_H^K *target curves to be approximated by the LMS minimization in order to obtain the* r_{NE}, r_{FA}, r_{TN}, r_{TF}, r_{T0} *signals (B). Disparity-tuning curves of the cells of the whole population stimulated by horizontal disparities varying between* $[-\Delta, \Delta]$. *The frequency of the curves decreases as the orientation of the RFs approaches to the vertical one (C). The effective LONG (solid green line) and SHORT (dashed red line) signals computed by the model, stimulated with a random dot stereograms. 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 (D). The effective* r_{T0} *signal that is able to act like a switch between the two controls. When it is below a defined threshold* TH, *the disparities are large, and it enables the long control (grey area, green line), otherwise the short control is active (white area, blue line) (E).*

The normalization term is introduced to make the vergence response independent of the image contrast. The weights w_{ij}^k are obtained through a recursive LMS algorithm that minimizes the following functional:

$$E\left(\mathbf{w}^k\right) = \left\|\sum_{j=1}^{N_O}\sum_{i=1}^{N_P} r_c^{ij}\left(\delta_H\right) w_{ij}^k - \upsilon_H^K\right\|^2 + \lambda\left\|\sum_{j=1}^{N_O}\sum_{i=1}^{N_P} r_c^{ij}\left(\delta_V\right) w_{ij}^k - 1\right\|^2$$

where $\lambda > 0$ balances the relevance of the second term over the first. In our simulations we fixed $\lambda = 1$. The weights w_{ij}^k are obtained through a recursive LMS algorithm. More precisely, given the profile of the desired vergence curve $r_v^k\left(\delta_H\right)$, such curve is approximated by a weighted sum of the tuning curves for horizontal disparity $r_c^{ij}\left(\delta_H;\theta,\Delta\psi\right)$. To gain the insensitivity to vertical disparity we add a constraint term in the minimization to ensure that the sum of the vertical disparity tuning curves $r_c^{ij}\left(\delta_V;\theta,\Delta\psi\right)$, weighted with the same $\mathbf{w}^k$, is approximately constant.

Dual-Mode Behaviour - In analogy to a common classification (Poggio, 1995), we distinguish five categories of r_v cells: near (r_{NE}) and far (r_{FA}) dedicated to coarse vergence, tuned near (r_{TN}), tuned far (r_{TF}) and tuned zero (r_{T0}) for fine vergence. More precisely, r_{NE} and r_{TN} drive convergence movements, while r_{FA} and r_{TF} divergence movements, in a push-pull system (see Fig. 6A-B). In practice the fast-coarse control is given by $LONG = r_{NE} - r_{FA}$, while the slow-fine is given by $SHORT = r_{TN} - r_{TF}$. The *SHORT* control signal is designed to proportionally generate, in a small range of disparities, the vergence to be achieved, and allows a precise and stable fixation (see Fig. 6D). Out of its range of linearity, the short signal decreases and it loses efficiency to the point where it changes sign, thus generating a vergence movement opposite to the desired one. Instead for small disparities the *LONG* control signal yields overactive vergence signal that makes 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 (see Fig. 6E). 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. 6D-E). On the contrary, for large stimulus disparities, r_{T0} is below the threshold and it enables the *LONG* control (see grey regions in Fig. 6D-E).

Figure 7. Eyes fixating on a plane perpendicular to the binocular gaze line, with a RDS texture attached. The gaze direction, for a primary fixation, is defined by an Azimuth = 0° and an Elevation = 0° (A), and by an Azimuth = 40° and an Elevation = 40° for a tertiary fixation (B). In the first case the horizontal disparity pattern δ_x is quasi-constant and the vertical one δ_y has a value close to zero (C), while in the second case the horizontal is no more constant and the vertical has a not negligible value (D). The disparity-vergence response in case of a constant disparity pattern with zero vertical component, is close to the desired one for both the LONG (green line) and the SHORT (blue line) controls (E). Indeed, in case of a tertiary fixation, thanks to the reduced sensitivity to vertical disparity, the vergence control is able to be effective with more complicate disparity patterns (F). Time course of the fixation point (red line) respect to the depth of the stimulus (blue line), in case of a step, a diverging ramp and a sinusoid for the primary (G) and tertiary (H) position of the gaze line.

Results - We tested the proposed model in a virtual environment in which the eyes, characterized by random azimuth and elevation angles, look at a plane with a random dot texture. The plane is at a distance Z with respect to the cyclopic position, and perpendicular to the binocular line of sight (see Fig 7A-B). In the first experiment, at the beginning of each trial, 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. In the second and in the third experiments, on the contrary, the plane is free to move along the gaze line as a ramp or a sinusoid, and the fixation point has to follow it in depth (see Fig. 7G-H). Considering that the gaze direction is allowed to span the entire environment, and it is not constrained to null azimuth and elevation, the vertical disparity components are not negligible and it may strongly affect the population response, as can be seen from the disparity patterns (see Fig 7C-D).
We considered the gaze direction for a primary fixation (Azimuth = 0° and an Elevation = 0°), and for a tertiary fixation (Azimuth = 40° and an Elevation = 40°). Even if in tertiary position, the vertical disparity pattern has a not negligible value, the vergence control is able to work properly for disparities in a range three times larger than the control gathered from the estimation of disparity (see Fig 7E-F).

FUTURE RESEARCH DIRECTIONS

Because humans and primates outperform the best machine vision systems by almost any measure, building a system that emulates cortical visual processing has always been an attractive idea. However, for the most part, the use of visual neuroscience in computer vision has been limited to a justification of Gabor filters for early vision tasks. Perhaps, an interesting exception relates to the research on the basic processing mechanisms of binocular vision where relevant breakthroughs in recent years have been reached by a synergic dialogue across disciplines (mixing "engineering" and "experimental" approaches). This dialogue primed experiments and scientific debates, which eventually converged to the understating/formulation of the basic principles of neuronal coding of motion and stereopis, as well as to powerful bio-inspired algorithms for depth and optic flow (Ohzawa, DeAngelis & Freeman, 1997; Qian & Mikaelian, 2000; Fleet & Jepson, 1993; Fleet, Wagner & Heeger, 1996; Fleet & Jepson, 1990; Heeger, 1988; Simoncelli & Heeger, 1998; Gautama & Van Hulle, 2001; Chen & Qian, 2004; Jepson, Fleet & Jenkin, 1991; Sabatini & Solari, 2004; Perrone, 2004).

In this framework, phase-based computational paradigms became very popular for the reliability of the extracted stereomotion features and for the interesting correlations with the properties of simple and complex cells in primary visual cortex. A general fresh view of complex cells as phase-insensitive units, but effective encoders of differential phase properties, has been recently proposed (Sabatini, Gastaldi, Solari, Pauwels, Van Hulle, Diaz, Ros, Pugeault & Krueger, 2010). On this basis, consolidated direct phase-based measure techniques imposed themselves (e.g., Sanger (1988) and Fleet, Jepson, and Jenkin (1991)), to which correspond distributed coding approaches (Fleet, Wagner & Heeger, 1996; Chen & Qian, 2004) based on populations of binocular energy units (Ohzawa, Freeman & DeAngelis, 1990; Ohzawa, DeAngelis & Freeman,1997; Qian, 1994) in which the output from receptive fields in both eyes is linearly combined by V1 simple cells and this sum is then passed through an output nonlinearity. The equivalence between phase-based techniques and energy-based models has been formally demonstrated (Qian & Mikaelian, 2000). Yet, the latter 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.

Binocular energy units are now consolidated models of complex cells in area V1 as demonstrated by the numerous recent works that propose architectural variants to enrich their functionality or that adopt them to describe complex perceptual behaviors (Read, 2002; Read, Parker & Cumming, 2002; Tanabe & Cumming, 2008; Bridge & Cumming, 2008; Haefner & Cumming, 2008; Read & Cumming, 2006; Serrano-Pedraza & Read, 2009; Nishimoto, Ishida & Ohzawa, 2006; Sanada & Ohzawa, 2006; Miura, Sugita, Matsuura, Inaba, Kawano & Miles, 2008).

Similarly, in the ICT community there are several examples of neuromorphic (i.e., distributed) approaches profitably used to challenge conventional solutions to computer vision problems, by introducing sophisticated interpretation of biologically plausible operations (e.g., Tsang and Shi (2007), Tsang and Shi (2009), Bayerl and Neumann (2007); see also Franz and Triesh (2007), Wang and Shi (2009), Solgi and Weng (2008)).

Although the performances of these models were promising, they have never been largely employed in real-world applications. This is mainly due to their high computational cost. The specific design approach followed to implement the distributed architecture presented in this chapter demonstrated that it is possible to implement 'neuromorphic' solutions that are characterized by an affordable computational cost, to be efficiently employed in closed-loop robotic applications. Pilot GPU-based implementations of the distributed architecture for the computation of 2D disparity, using the Nvidia CUDA Library yielded encouraging results. Therefore, cortical-like architectures as bio-inspired structural paradigms to solve computer vision tasks, can represent a viable solution for the next-generation robot vision systems, which are capable to calibrate and adapt autonomously through the interaction with the environment. The distributed character of processing and representation ensures the necessary entry points for closing perception-action cycles from the very initial processing stages.

From a broader sensorimotor perspective, the challenge is to extend the problem of a visuomotor awareness of the 3D peripersonal space to other body parts, e.g. head and arms, so possibly using multisensory feedback, to extract information useful to build representations of the 3D space which are coherent and stable with respect to time, towards an egocentric and heterogeneous multimodal representation of space.

CONCLUSION

In this chapter, we present two case-studies that show how large-scale networks of V1-like binocular cells can provide a flexible medium on which to base coding/decoding adaptation mechanisms related to sensorimotor schema: at the coding level, the position of the eyes in the orbits can adapt the disparity tuning to minimize the necessary resources, while preserving reliable estimates (i.e., adjustable tuning mechanisms based on the posture of the eyes to improve depth vision (Chessa, Canessa, Gibaldi, Solari & Sabatini, 2009). At the decoding level, specialized read-out mechanisms can be obtained for directly extracting disparity-vergence responses without explicit calculation of the disparity map, to gain (1) linear servos with fast reaction and precision, and (2) wide working range with a reduced amount of resources (Gibaldi, Canessa, Chessa, Solari & Sabatini, 2010).

The extraction of binocular features relies upon a full (i.e., amplitude, orientation and phase) harmonic representation of the visual signal, operated by a set of “simple cell” units (S-cells). Such representation allows us to reconsider and analyze the flexibility and robustness of the multi-channel perceptual coding, adopted by the early stages of the mammalian visual cortex, for the “atomic” components of early vision. Oriented disparity tuning emerges in layers of binocular energy “complex cell” units (C-cells) which gathers S-cells outputs according to specific architectural schemes.

From a methodological point of view, the proposed approach points out the advantages and the flexibility of distributed and hierarchical cortical-like architectures against solutions based on a conventional systemic coupling of sensing and motor components, which in general poses integration problems since too heterogeneous and complex processes must be coupled. Through the distributed coding, indeed, it is possible to avoid a sequentialization of sensorial and motor processes, that is certainly desirable for the development of cognitive abilities at a pre-interpretative (i.e., sub-symbolic) level, e.g., when a system must learn (binocular) eye coordination, handling the inaccuracies of the motor system, and calibrate the active measurements of the space around it. The design strategy of these active visual cortical networks jointly involves three concurrent aspects: (i) *signal processing*, by defining the proper descriptive elements of the visual signal (in the Gibsonian sense) and the operators to measure them (cf. the Plenoptic Function (Adelson & Bergen, 1991), (ii) the *geometry of the system* and its kinematics, which directly relates to the embodiment concept, and (iii) the *connectionism paradigms*, that define neuromorphic architectural solutions for information processing and representation. The connectionism paradigm (i.e., hierarchical, distributed computing) is crucial to guarantee accessibility and interaction of the information at different levels of coding and decoding, by postponing decisions as much as possible.

REFERENCES

Adelson, E.H. & Bergen, J.R. (1991). *Computational Models of Visual Processing*, Cambridge, MA, MIT Press.

Aloimonos, J.Y., Weiss, I. & Bandopadhay, A. (1987). Active vision. *International Journal on Computer Vision,* 1(4), 333-356.

Bajcsy, R. (1988) Active perception. *IEEE Proceedings.* Vol. 76(8), 996-1006.

Ballard, D. (1991). Animate vision. *Artificial intelligence*, 48(1), 1-27.

Ballard, D.H. & Ozcandarli, A. (1988, Dec). *Eye fixation and early vision: Kinematic Depth*. Paper presented at the meeting of the IEEE 2nd International Conference on Computer Vision, Tarpon Springs, Fla.

Bayerl, P. & Neumann, H. (2007). A fast biologically inspired algorithm for recurrent motion estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 29(2), 246-260.

Blohm, G., Khan, A.Z., Ren, L., Schreiber, K.M. & Crawford, J.D. (2008). Depth estimation from retinal disparity requires eye and head orientation signals. *Journal of Vision*, 8(16), 3, 1-23.

Bridge, H. & Cumming, B.C. (2008). Representation of binocular surfaces by cortical neurons. *Curr. Opin. Neurobiol.,* 18(4), 425-30.

Brown, C. (1990). Prediction and cooperation in gaze control. *Biological Cybernetics*, 63(1), 61-70.

Chen, Y. & Qian, M. (2004). A coarse to fine energy model with both phase-shift and position-shift receptive field mechanisms. *Neural Computation,*16, 1545-1577.

Chessa, M., Sabatini, SP, & Solari, F. (2009). A fast joint bioinspired algorithm for optic flow and two-dimensional disparity estimation. In Piater, Justus, *7th Int. Conference on Computer Vision Systems (ICVS'09)*, *Lecture Notes in Computer Science*, Vol. 5815, pp. 13-15.

Chessa, M., Solari, F. & Sabatini, S.P. (2011). Virtual Reality to Simulate Visual Tasks for Robotic Systems. In Jae-Jin Kim, *Virtual Reality* (pp. 71-92), InTech.

Clark, J.J & Ferrier, N.J. (1988). *Modal control of attentive vision system*. In Proceedings of the International Conference on Computer Vision (pp. 514–523).

Costa, A.H.R., Rillo, C., Barros, L.N.D. & Bianchi, R.A.C. (1998). Integrating purposive vision with deliberative and reactive planning: engineering support for robotic applications. *J. Braz. Comp. Soc., 4(3).*

Dean, T., Allen, J. & Aloimonos, Y. (1995). Artificial intelligence: theory and practice. Redwood City, CA: Benjamin/Cummings Publishing Co.

Eklundh, J.O. & Pahlavan, K. (1992, April). *Eye and head-eye system.* SPIE Applications of AI X: Machine Vision and Robotics, Orlando, Fla.

Fleet, D. (1994). Disparity from local weighted phase-correlation. In Proc. of the *IEEE Int. Conf. on Systems,Man and Cybernetics, Vol. 1,* (pp. 48–54).

Fleet, D.J. & Jepson, A.D. (1990). Computation of component image velocity from local phase information. *International Journal of Computer Vision,* 5(1), 77-104.

Fleet, D.J. & Jepson, A.D. (1993). Stability of phase information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 15(12), 1253-1268.

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

Fleet, D.J., Wagner, H. & Heeger, D.J. (1996). Neural encoding of binocular disparity: Energy models, position shifts and phase shifts. *Vision Research*, 36(12), 1839–1857.

Franz, A. & Triesch, J. (2007). Emergence of disparity tuning during the development of vergence eye movements. *In International Conference on Development and Learning.*

Gautama, T. & Van Hulle, M.M. (2001). Function of center-surround antagonism for motion in visual area MT/V5: a modeling study. *Vision Research,*41(28), 3917-3930.

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

Greenwald, H.S. & Knill, D.C. (2009). Cue integration outside central fixation: A study of grasping in depth. *Journal of Vision*, 9(2), 11, 1-16.

Haefner, R.M. & Cumming, B.G. (2009). An improved estimator of variance explained in the presence of noise. *Advances in Neural Information Processing Systems,* 21, 585-592.

Hamker, F.H. (2005a). The Reentry Hypothesis: The Putative Interaction of the Frontal Eye Field, Ventrolateral Prefrontal Cortex, and Areas V4, IT for Attention and Eye Movement. *Cerebral Cortex,* 15, 431-447.

Hamker, F.H. (2005b). A computational model of visual stability and change detection during eye movements in real world scenes. *Visual Cognition,* 12, 1161-1176.

Hansard, M. & Horaud, R. (2008). Cyclopean geometry of binocular vision. *J. Opt. Soc. Amer.*, 25, 2357–2369.

Hansard, M. & Horaud, R. (2010). Cyclorotation models for eyes and cameras. *IEEE Transactions on System, Man, and Cybernetics–Part B: Cybernetics,* 40, 151–161.

Heeger, D. (1988). Optical flow using spatiotemporal filters. *International Journal of Computer Vision,*1, 270-302.

Howard, I. P. & Rogers, B. J. (2002) *Seeing in depth: Volume 2, Depth perception*. Ontario, Canada: I Porteous Publishing.

Hung, G.K., Semmlow, J.L. & Ciuffreda, K.J. (1986). A dual-mode dynamic model of the vergence eye movement system. *IEEE Trans. Biomed. Eng.*, 36(11), 1021–1028.

Jenkin, M.R.M. (1996). Stereopsis near the horoptor. *Proc. 4th ICARCV*.

Jepson, A.D., Fleet, D.J. & Jenkin, R.M. (1991). Phase-based disparity measurement. *CVGIP: Image Understanding,*53, 198-210.

Krishnan, V.V. & Stark, L.A. (1977). A heuristic model for the human vergence eye movement system. *IEEE Trans. Biomed. Eng.*, 24, 44–49.

Krotkow, E., Henriksen, K. & Kories, R. (1990). Stereo ranging from verging cameras. *IEEE Transaction on PAMI*, 12(12), 1200-1205.

Marr, D. (1982) Vision. New York, NY: W.H. Freeman and Company.

Miura, K., Sugita, Y., Matsuura, K., Inaba, N., Kawano, K. & Miles, F.A. (2008) The initial disparity vergence elicited with single and dual grating stimuli in monkeys: evidence for disparity energy sensing and nonlinear interactions. *J. Neurophysiol,* 100(5), 2907-18.

Mok, D., Ro, A., Cadera, W., Crawford, J.D. & Vilis, T. (1990) Rotation of listing's plane during vergence. *Vision Research*, 32, 2055–2064.

Monaco, J.P., Bovik, A.C. & Cormack, L.K. (2009). Active, foveated, uncalibrated stereovision. *Int. J. of Computer Vision*, 85(2), 192-207.

Morgan, M.J. & Castet, E. (1997). The aperture problem in stereopsis. *Vision Research*, 37, 2737–2744.

Nishimoto, S., Ishida, T. & Ohzawa, I. (2006). Receptive Field Properties of Neurons in the Early Visual Cortex Revealed by Local Spectral Reverse Correlation. *Journal of Neuroscience,* 26(12), 3269–3280.

Noë, A. (2004). *Action in perception*. Cambridge, MA: MIT Press.

O'Regan, K. & Noë, A. (2001). A sensorimotor account of vision anvisual consciousness . *Behavioral and Brain Sciences,* 24(5), 883–917.

Ohzawa, I., DeAngelis, G.C. & Freeman, R.D. (1997). Encoding of binocular disparity by complex cells in the cat's visual cortex. *The Journal of Neurophysiology.*77, 2879-2909.

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

Perrone, J.A. (2004). A visual motion sensor based on the properties of V1 and MT neurons. *Vision Research,*44, 1733-1755.

Poggio, G.F. (1995). Mechanism of stereopsis in monkey visual cortex. *Cerebral Cortex*, 5, 193–204.

Polpitiya, A.D. & Ghosh, B.K. (2002, May). *Modelling and control of eye-movement with muscolotendon dynamics*. Paper presented at the meeting of the American Control Conf., Anchorage, AK.

Prince, S.J.D., Cumming, B.G. & Parker, A.J. (2002). Range and mechanism of encoding of horizontal disparity in macaque v1. *Journal of Neurophysiology*, 87, 209–221.

Qian, N. & Mikaelian, S. (2000) Relationship between phase and energy methods for disparity computation. *Neural Computation,* 12, 303-316.

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

Rambold, H.A., & Miles, F. (2008). A human vergence eye movements to oblique disparity stimuli: evidence for an anisotropy favoring horizontal disparities. *Vision Research*, 48, 2006–2019.

Read, J.C.A. & Cumming, B.G. (2004). Understanding the cortical specialization for horizontal disparity. *Neural Computation,* 16(10), 1983-2020.

Read, J.C.A. & Cumming, B.G. (2006). Does depth perception require vertical disparity detectors? *Journal of Vision,* 6(12), 1323-1355.

Read, J.C.A. & Cumming, B.G. (2006). Does depth perception require vertical disparity detectors? *Journal of Vision,* 6(12), 1323-1355.

Read, J.C.A. (2002). A Bayesian approach to the stereo correspondence problem. *Neural Computation,* 14, 1371-1392.

Read, J.C.A., Parker, A.J. & Cumming, B.G. (2002). A simple model accounts for the response of disparity-tuned V1 neurons to anti-correlated images. *Visual Neuroscience,* 19, 735-753.

Read, J.C.A., Phillipson, G.P. & Glennerster, A. (2009). Latitude and longitude vertical disparities. *Journal of Vision*, 9(13), 1-37.

Sabatini, S.P. & Solari, F. (2004). Emergence of motion-in-depth selectivity in the visual cortex through linear combination of binocular energy complex cells with different ocular dominance. *Neurocomputing,* 58–60, 865-872.

Sabatini, S.P., Gastaldi, G., Solari, F., Pauwels, K., Van Hulle, M.M., Diaz, J., Ros, E., Pugeault, N. & Krueger, N. (2010). A Compact Harmonic Code for Early Vision based on Anisotropic Frequency Channels. *Computer Vision and Image Understanding*, 114, 681-699.

Sanada, T.M. & Ohzawa, I. (2006). Encoding of three-dimensional surface slant in cat visual areas 17 and 18. *Journal of Neurophysiology,* 95, 2768-2786.

Sanger, T.D. (1988). Stereo disparity computation using Gabor filters. *Biol. Cybern.*, 59, 405–418.

Santini, F. & Rucci, M. (2007) Active estimation of distance in a robotic system that replicates human eye movement. *Robotics and Autonomous Systems,* 55(2), 107-121.

Scharstein, D. & Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1/2/3), 7-42.

Schor, C.M., Maxwell, J.S., McCandless, J. & Graf, E. (2002) Adaptive control of vergence in humans. *Ann. N.Y. Acad. Sci.*, 956, 297–305.

Schreiber, K., Crawford, J.D., Fetter, M. & Tweed, D. (2001) The motor side of depth vision. *Nature,* 410, 819–822.

Schreiber, K.M., Hillis, J.M., Filippini, H.R., Schor, C.M. & Banks, M.S. (2008). The surface of the empirical horopter. Journal of Vision, 8(3), 1–20.

Schreiber, K.M., Tweed, D.B. & Schor, C.M. (2006). The extended horopter: Quantifying retinal correspondence across changes of 3d eye position. *Journal of Vision*, 6, 64–74.

Semmlow, J.L., Hung, G. K. & Ciuffreda, K. J. (1986). Quantitative assessment of disparity vergence components. *Invest. Ophthalmol. Vision Science*, 27, 558–564.

Serrano-Pedraza, I. & Read, J.C.A. (2009). Stereo vision requires an explicit encoding of vertical disparity. *Journal of Vision*, 9(4), 1-13.

Serrano-Pedraza, I., Phillipson, G.P. & Read, J.C.A. (2010). A specialization for vertical disparity discontinuities. *Journal of Vision*, 10(3), 1-25.

Sheliga, B.M. & Miles, F.A. (2003). Perception can influence the vergence responses associated with open-loop gaze shifts in 3-d. *Journal of Vision*, 3, 654–676.

Simoncelli, E.P. & Heeger, D.J. (1998). A model of neuronal responses in visual area MT. *Vision Research,*38(5), 743-761.

Solgi, M. & Weng, J. (2008, Nov). *Developmental stereo: topographic iconic-abstract map from top-down connection.* In International Neural Network Society. Symposia Series New developments in Neural Networks, Auckland, New Zealand-

Takemura, A., Inoue, Y., Kawano, K., Quaia, C. & Miles, F. A. (2001). Single-unit activity in cortical area MST associated with disparity vergence eye movements: Evidence for population coding. *Journal of Neurophysiology*, 85, 2245–2266.

Takemura, A., Kawano, K., Quaia, C. & Miles, F.A. (2006). Population coding of vergence eye movements in cortical area MST. In L. Harris and M. Jenkin, *Levels of Perception*.

Tanabe, S. & Cumming, B.G. (2008). Mechanisms underlying the transformation of disparity signals from V1 to V2 in the macaque. *J. Neurosci.,* 28(44), 11304-14.

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

Tsang, E.K.C. & Shi, B.E. (2007). Estimating disparity with confidence from energy neurons. *NIPS2007.*

Tsang, E.K.C. & Shi, B.E. (2009). Disparity Estimation by Pooling Evidence From Energy Neurons. *IEEE Transactions on Neural Networks,* 20(11), 1772-1782.

Tweed, D. & Vilis, T. (1990). Geometric relations of eye position and velocity vectors during saccades. *Vision Research,* 30(1), 111-127.

Wang, Y. & Shi, B.E. (2009). Autonomous development of vergence control driven by disparity energy neuron populations. *Neural Comp,* 22, 1-22.

Yang, D.S., FitzGibbon, E.J. & Miles, F.A. (2003). Short-latency disparityvergence eye movements in humans: sensitivity

ADDITIONAL READING SECTION

Daugman, J.G. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by twodimensional visual cortical filters. *Journal of the Optical Society of America,* 2(7), 1160-1169.

Adelson, E. H. & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America,* 2(2), 284-299.

Anzai, A., Ohzawa, I. & Freeman, R.D. (2001). Joint-encoding of motion and depth by visual cortical neurons: neural basis of the Pulfrich effect. *Nature Neuroscience,* 4, 513-518.

De Valois, R. & De Valois, K. (1988). *Spatial vision.* Oxford University Press.

Emerson, R.C., Bergen, J.R. & Adelson, E.H. (1992). Directionally selective complex cells and the computation of motion energy in cat visual cortex. *Vision Research,* 32, 203-218.

Heeger, D.J. (1987). Model for the extraction of image flow. *Journal of Optical Society of America,*4, 1455-1471.

Li, Z. & Atick, J.J. (1994). Towards a theory of striate cortex. *Neural Computation,* 6, 127-146.

Li, Z. (1996). A theory of the visual motion coding in the primary visual cortex. Neural Computation, 8(4), 705-730.

Priebe, N.J., Lisberger, S.G. & Movshon, J.A. (2006). Tuning for spatiotemporal frequency and speed in directionally selective neurons of macaque striate cortex. *Journal of Neuroscience,*26, 2941-2950.

Raiguel, S.E., Van Hulle, M.M., Xiao, D.K., Marcar, V.L. & Orban, G.A. (1995). Shape and spatial distribution of receptive fields and antagonistic surrounds in area MT (V5) of the macaque. *The European Journal of Neuroscience,* 7, 2064-2082.

Rust, N.C., Mante, V., Simoncelli, E.P. & Movshon, J.A. (2006). How MT cells analyze the motion of visual patterns. Nature Neuroscience,9, 1421-1431.

Simoncelli, E.P. (1993). *Distributed analysis and representation of visual motion.* Unpublished doctoral dissertation, Massachusetts Institute of Technology, Cambridge MA.

Zhu, Y. & Qian, N. (1996). Binocular receptive field profiles, disparity tuning and characteristic disparity. *Neural Computation.*

KEY TERMS & DEFINITIONS

Perception-action loop, Active vision, Population coding, Binocular disparity, Binocular energy model, Vergence eye movements, Eye movements.

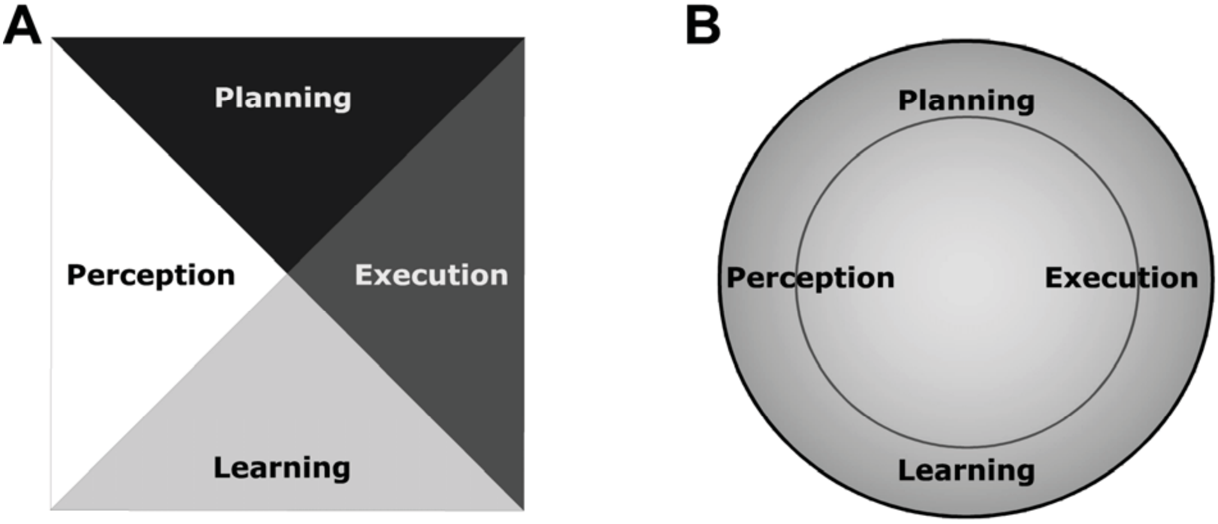


Figure 1

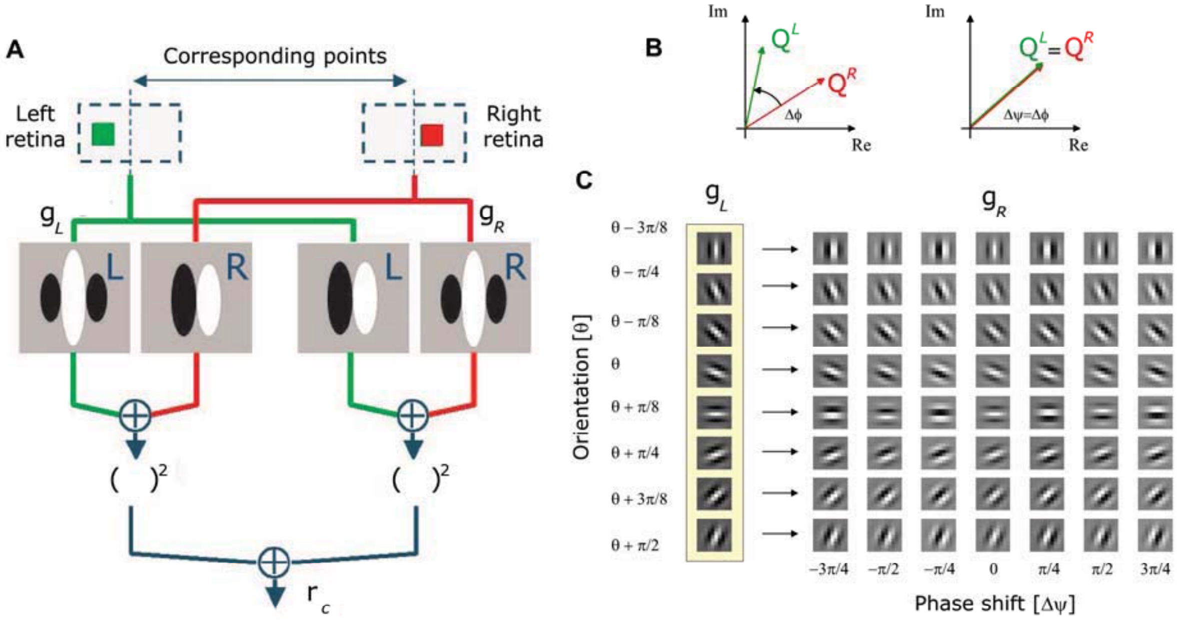


Figure 2

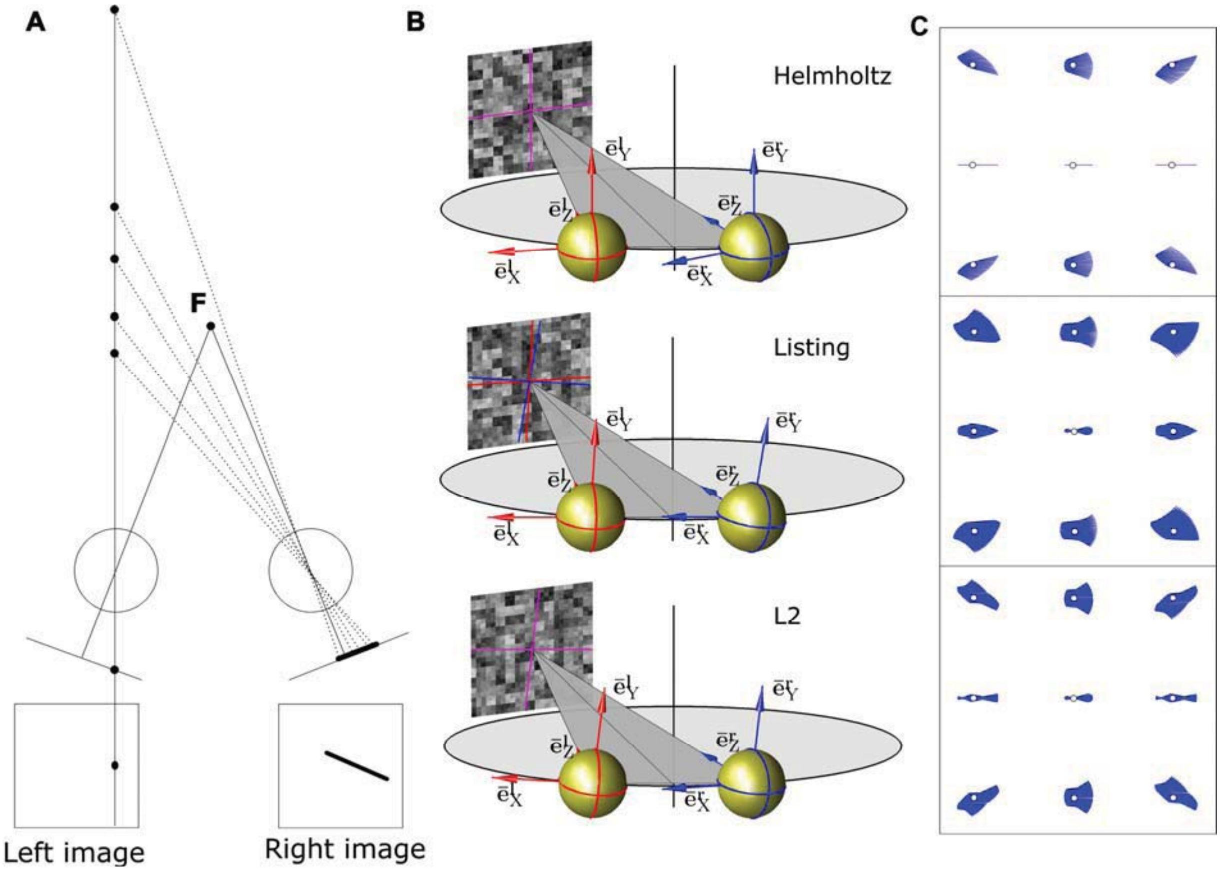
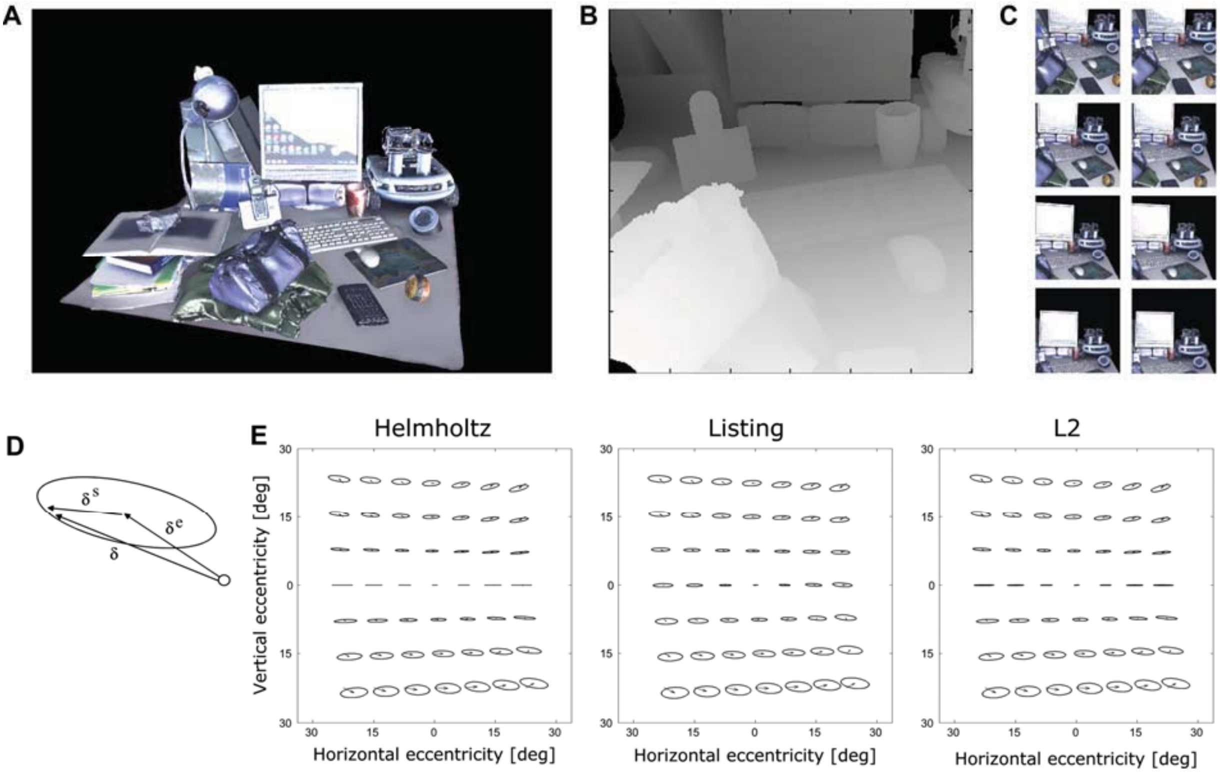

A
F
Left image
Right image
B
Helmholtz
Listing
L2
$\bar{e}^l_Y$
$\bar{e}^l_Z$
$\bar{e}^l_X$
$\bar{e}^r_Y$
$\bar{e}^r_Z$
$\bar{e}^r_X$
C

Figure 3


A
B
C
D
δ^s
δ^e
δ
E
Helmholtz
Listing
L2
Vertical eccentricity [deg]
30
15
0
15
30
Horizontal eccentricity [deg]
30
15
0
15
30

\
Figure 4

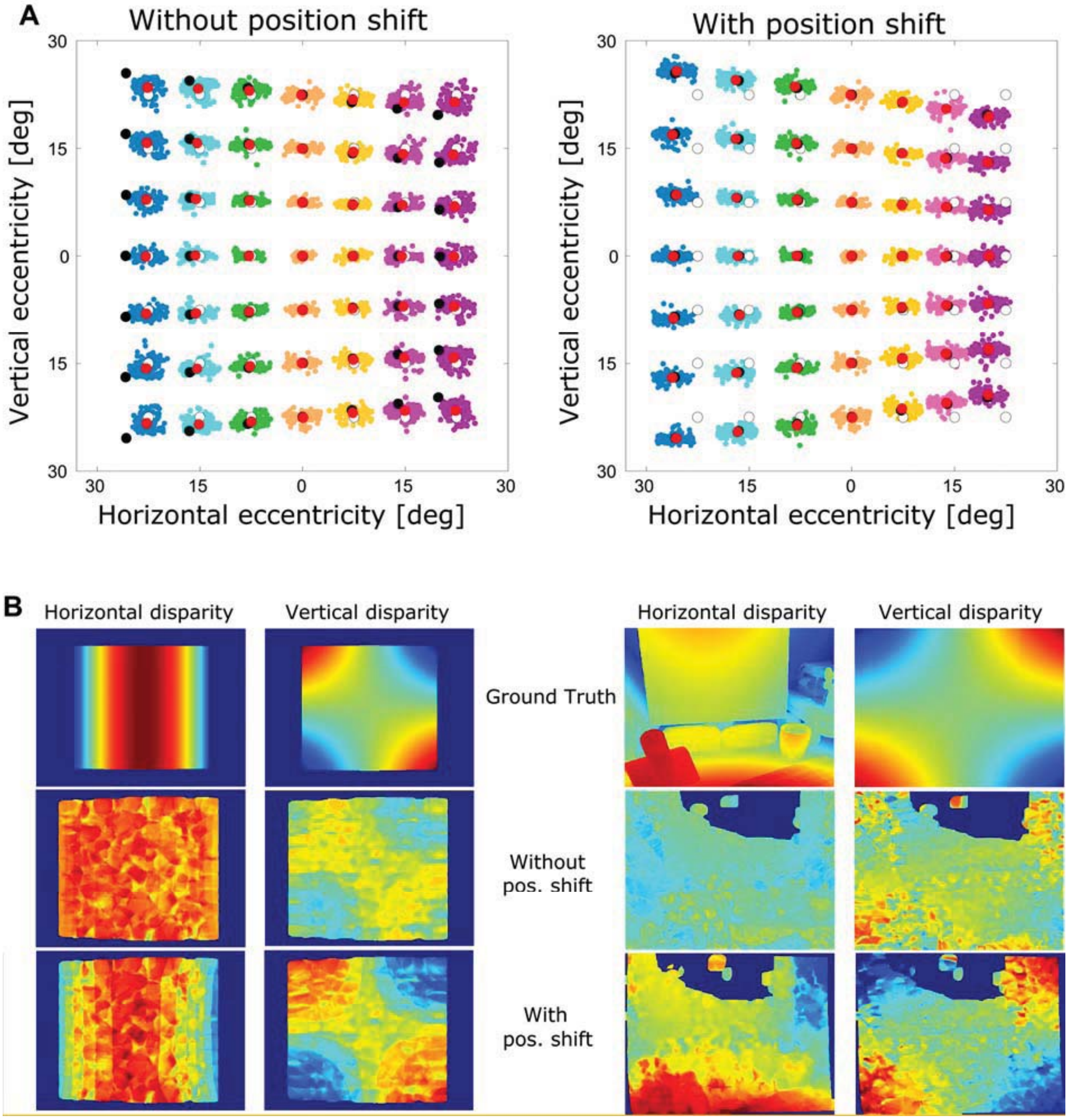

A
Without position shift
With position shift
Vertical eccentricity [deg]
Horizontal eccentricity [deg]
30
15
0
15
30
B
Horizontal disparity
Vertical disparity
Ground Truth
Without
pos. shift
With
pos. shift

Figure 5

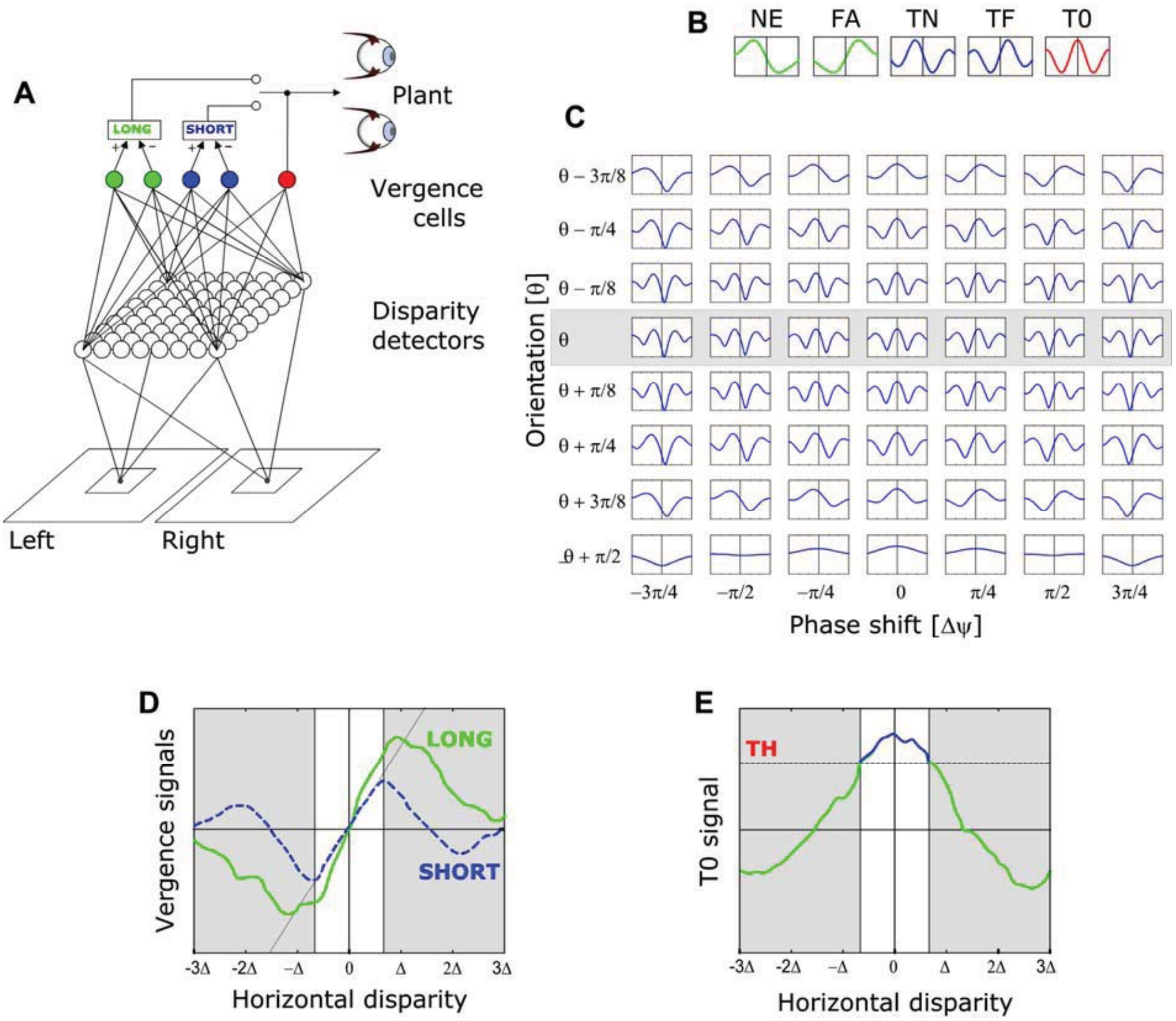

A
LONG
SHORT
+ −
+ −
Plant
Vergence cells
Disparity detectors
Left
Right
B
NE
FA
TN
TF
T0
C
θ − 3π/8
θ − π/4
θ − π/8
θ
θ + π/8
θ + π/4
θ + 3π/8
θ + π/2
Orientation [θ]
−3π/4
−π/2
−π/4
0
π/4
π/2
3π/4
Phase shift [Δψ]
D
Vergence signals
LONG
SHORT
-3Δ
-2Δ
−Δ
0
Δ
2Δ
3Δ
Horizontal disparity
E
T0 signal
TH
-3Δ
-2Δ
−Δ
0
Δ
2Δ
3Δ
Horizontal disparity

Figure 6

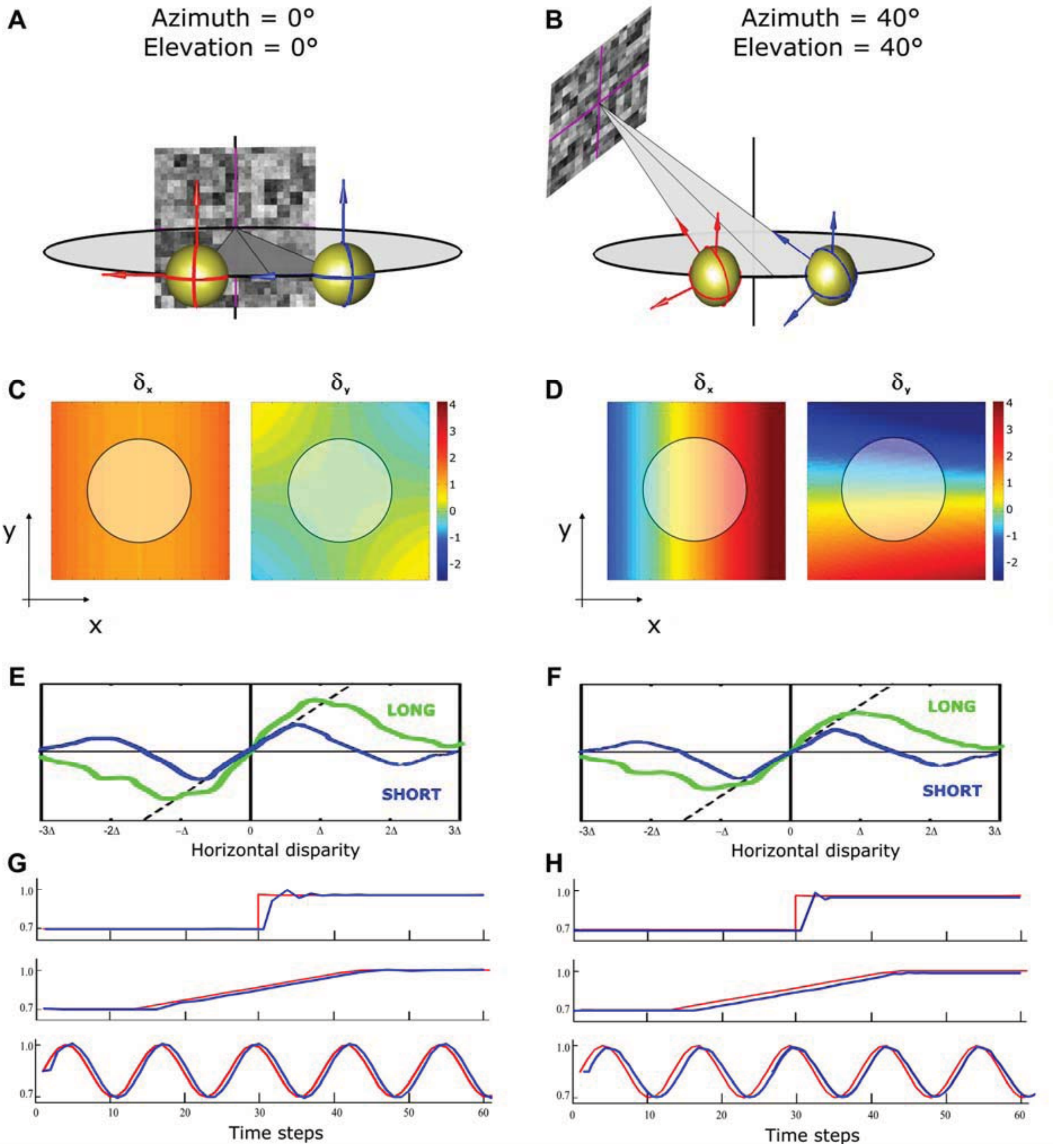

A
Azimuth = 0°
Elevation = 0°
B
Azimuth = 40°
Elevation = 40°
C
δx
δy
y
x
4
3
2
1
0
-1
-2
D
δx
δy
y
x
E
LONG
SHORT
-3Δ
-2Δ
-Δ
0
Δ
2Δ
3Δ
Horizontal disparity
F
LONG
SHORT
Horizontal disparity
G
1.0
0.7
0
10
20
30
40
50
60
Time steps
H
Time steps

Figure 7