2011 11th IEEE-RAS International Conference on Humanoid Robots
Bled, Slovenia, October 26-28, 2011

A neuromorphic control module for real-time vergence eye movements on the iCub robot head

Agostino Gibaldi, Andrea Canessa, Manuela Chessa, Silvio P. Sabatini, Fabio Solari
PSPC Lab, Department of Biophysical and Electronic Engineering,
University of Genoa, Via allOpera Pia 11/A - 16145 Genova, Italy
http://www.pspc.dibe.unige.it/Members/

***Abstract*—We implemented a cortical model of vergence eye movements on a humanoid robot head (iCub). The proposed control strategy resorts on a computational substrate of modeled V1 complex cells that provides a distributed representation of binocular disparity information. The model includes a normalization stage that allows for a vergence control independent of the texture of the object and of luminance changes. The disparity information is exploited to provide a signal able to nullify the binocular disparity in a foveal region.**

Experimental tests demonstrated that the robot head executes effective vergence movements, on a visual stimulus posed within the peripersonal space, that falls in the perifoveal field of view. The algorithm works in real-time, and provides a stable control, robust with respect to the mechanical inaccuracies that affect the mechanical system. The eye posture resulting from the vergence movement, is instrumental for an active vision system to optimally perceive the 3D structure of the environment from disparity.

The similarity of the vergence speed profile and trajectories of the fixation point with respect to psychophysical data, allows us to use the system for a validation of the neurophysiological models for triggering vergence eye movements.

I. Introduction

Stereoscopic vision is a fundamental feature for active vision systems. The ability of a robot system to interact with the surrounding environment is a direct consequence of its correct sensing of the 3D space. Particularly at close short distances, foveating the object of interest with both cameras is necessary to reconstruct the spatial layout of the observed scene, for tasks such as manipulation and navigation.

In Humans and primates, the retinal binocular disparity is the primary cue the brain uses to achieve depth perception and to control the movement of the eyes in order to actively get a better perception of the scene, on the basis of the characteristics of the scene itself. The given stimulus disparity is equally effective to provide depth perception and to trigger the correct vergence eye movements [1]. Neurons in primary visual cortex (V1) have a strong sensitivity to disparity, that can be described by a tuning curve, that is the cell's response as a function of the stimulus disparity. An optimal stimulus to stimulate the neurons, and thus to derive their disparity tuning curves, is the random dot stereogram (RDS), drawn with a defined amount of disparity that varies in the sensitivity range of the cells. When stimulated with anticorrelated random dot stereograms [1] [2] (*i.e.* where the contrast of the left image is opposite respect to the right one), V1 neurons exhibit inverted tuning curves (*i.e.* vertically flipped), as predicted by the binocular energy model. While with such a stimulus the perception of depth fails, a vergence movement opposite to the desired one is triggered. This suggests that disparity-vergence responses might follow a faster reactive stream that directly involves V1 cells without resorting to a high level interpretation of depth.

The bio-inspired vergence models previously developed, require first the computation of the disparity map for the extraction of the control signals [3] [4], thus limiting the functionality of the vergence system within the range of disparity detectors, 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 within a limited range known as Panum fusional area, where they are not useful.

Neurophysiological studies in the Medial Superior Temporal area (MST), sensitive to retinal disparity, reported that MST neurons have been found to encode some aspects of the motor response for vergence eye movements, and the activity of the whole population directly correlates with the magnitude, direction, and time course of the initial vergence motor response [2]. By mimicking the behaviour of the cells of the MST area, we developed a model that, by combining the response of a population of complex cells, does not take a decision on the disparity values (*i.e.* does not compute the 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.

Recent attempts present in literature, focus on how the sensory representation is able to guide vergence movements in a behaving organism. The model described by [5], resorts on the response of a population of complex cells tuned to zero disparity ($\Delta\psi = 0$) to earn the emergence of disparity tuning in a neural network as a plausible substrate for guiding vergence movements. In this approach, the image used for the training is a one-dimensional stimulus, thus the problem is reduced to horizontal disparity only. A subsequent model [6] proposes to use directly the output of three populations of complex cells, tuned to different disparity values, to guide vergence movements. The approach shows a good effectiveness in providing the correct vergence on a moving stimulus, but the complexity of the problem is tightly limited. In fact, first, the stimulus considered is a vertical bright bar on a dark background, and second, the stereo head is calibrated

for lens distortions and camera misalignments, with vergence movements simulated on a LCD screen placed in front of the head. This means that the stimulus used is very far from a real stimulus, and the retinal images produced have zero distortion and zero vertical disparity.

In this paper, we implemented a neuromorphic control for vergence eye movements based on the binocular energy model of V1 complex cells [7]. The distributed representation of disparity allows us to interpret directly the information to gather the correct vergence control on real-word stimuli, like highly textured objects and complex 3D shapes. The control is designed to cope with the vertical disparity deriving both from the vergent geometry, and from the misalignment of the optical axes, due to the imprecision of the real mechanical system. To overcome instability effects resulting from unpredictable and changing lighting condition of the environment, we implemented a divisive normalization circuit [8]. Furthermore, a single scale approach yields a good accuracy of the control with minimal amount of resources, and thus produces the real time behaviour needed for robot control.

The capabilities of the model for the control of vergence were tested both with "controlled" stimuli to have a quantitative evaluation and a direct comparison with psychophysical data, and in a real word situation, to verify behaviour with complex stimuli in the peripersonal space of the robot.

The paper is organized as follows. In Sec. II we present the resources considered, based on the binocular energy model, how to specialize them for the vergence control, and how to obtain a control robust to different textures and to luminance changes. In Sec. III we present how we implemented the algorithm in real time, and how to used it on the iCub robot head, at gaze directions different from the primary position (*i.e.* straight-ahead). Furthermore in Sec. IV we present the results and discuss the effectiveness of the approach comparatively with the psychophysical data. Finally in Sec. V we draw the conclusions, and we present future developments.

II. A CORTICAL MODEL OF VERGENCE EYE MOVEMENTS

Depth perception and vergence eye movements derive from the differences in the positions of corresponding points in the stereo image pair projected on the two retinas of a binocular system, that is the retinal disparity $\boldsymbol{\delta}(\mathbf{x})$, defined by a *horizontal* δ_H and a *vertical* δ_V component. Neurophysiological evidences report that the cortical cells' sensitivity to binocular disparity is related to interocular phase shifts in the Gabor-like receptive fields (RF) of V1 simple cells [9][10][8].

A distributed cortical architecture, based on V1 binocular energy model, is capable of providing reliable encoding of the information of disparity. The response of the architecture can be used to estimate the disparity map [11], from which to derive the proper vergence control, like in previous vergence models [3][4]. The derived control is effective when the disparity is within the theoretical reliability range $[-\Delta, \Delta]$, specified by the parameters of the receptive fields. The drawback of this approach is to limit the capability of the vergence control within a range where the system is already able to produce a correct perception of the 3D structure of the environment, consequently where vergence movements are not crucial.

We implemented an alternative strategy [7] to gather an effective vergence control directly from the population responses, *i.e.* without explicit computation of the disparity map, so to have a system able to cope with larger disparities than $[-\Delta, \Delta]$, and to achieve stable fixations. Since the meaningful information for vergence comes from the perifoveal part of the image only, we used response in a neighborhood placed in the center of the image.

A. Binocular energy model

Relying upon the local approximation of the Fourier Shift Theorem, the 2D local vector disparity $\boldsymbol{\delta}(\mathbf{x})$ between the left $I^L(\mathbf{x})$ and right $I^R(\mathbf{x})$ images can be 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 average instantaneous frequency of the band-pass signal measured using the phase derivative [12]. This property of the phase $\phi^{L/R}(\mathbf{x})$ yields good estimate of binocular disparity by:

$$\delta(\mathbf{x}) = \frac{\lfloor \phi^L(\mathbf{x}) - \phi^R(\mathbf{x}) \rfloor_{2\pi}}{k(\mathbf{x})} = \frac{\lfloor \Delta\phi(\mathbf{x}) \rfloor_{2\pi}}{k(\mathbf{x})}, \qquad (1)$$

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. Formally, the response of left and right monocular RFs centered in $\boldsymbol{x}$ and oriented along θ, are:

$$r_{L/R}(\mathbf{x}; \theta, \psi^{L/R}) = I^{L/R} * h^{L/R}(\mathbf{x}; \theta, \psi^{L/R}) \qquad (2)$$

where $h(\mathbf{x})$ is a complex-valued Gabor filter, defined by:

$$h(\mathbf{x}) \triangleq h(\mathbf{x}; \theta, \psi) = \eta e^{\left(-\frac{1}{2\sigma^2}\mathbf{x}_\theta^T \mathbf{x}_\theta\right)} e^{\jmath(k_0 \mathbf{x}_\theta + \psi)} \qquad (3)$$

where $\mathbf{x}_\theta$ is the rotated coordinate system by an angle θ, k_0 is the filter radial peak frequency, and ψ is the phase value that characterizes the binocular RF profile.

Thus, the response of a complex cell of V1 area is:

$$r_c(\mathbf{x}; \theta, \Delta\psi) = \left| r_L(\mathbf{x}; \theta, \psi^L) + r_R(\mathbf{x}; \theta, \psi^R) \right|^2 \qquad (4)$$

where $\Delta\psi = \psi^L - \psi^R$ defines the disparity along the direction θ, to which the cell is tuned, that is $\delta^\theta_{pref} = \lfloor \Delta\psi \rfloor_{2\pi} / k_0$.

Since the phase shifts are unique only between $-\pi$ and π, we can define the theoretical value of the maximum encoded disparity $\pm\Delta$ by the maximum phase shift that can be used to design the RFs: $\pm\Delta = \delta^\theta_{pref}|_{\Delta\psi = \pm\pi} = \pm\pi / k_0$.

On the basis of the phase-shift energy model [8] it is possible to construct a population of complex cells in order to encode the disparity information of a stereo image pair. 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 and along N_o orientations uniformly distributed between 0 and π.

The distributed cortical architecture is, in this way, capable of providing reliable encoding of the information of disparity, while its module is smaller than Δ.

B. Vergence control

The spatial neighborhood Ω, considered to compute the vergence control is defined by a Gaussian profile centered in fovea and with standard deviation of $1.5°$, from which:

$$r_{v_H} = \sum_{\mathbf{x}\in\Omega} \sum_{i=1}^{N_p} \sum_{j=1}^{N_o} w_{ij} r_c^{ij}(\mathbf{x}). \tag{5}$$

Since the desired horizontal vergence control must be sensitive to the horizontal component of the vector disparity δ_H and insensitive to the vertical component δ_V, the weights w_{ij} are obtained with a functional that considers both the features:

$$\begin{aligned} E(\mathbf{w}) \quad &= \left\| \sum_{i=1}^{N_p} \sum_{j=1}^{N_o} r_c^{ij}(\delta_H) w_{ij} - \upsilon_H \right\|^2 + \\ &+ \lambda \left\| \sum_{i=1}^{N_p} \sum_{j=1}^{N_o} r_c^{ij}(\delta_V)(w_{ij} - 1) \right\|^2 \end{aligned} \tag{6}$$

where $r_c^{ij}(\delta_H)$ and $r_c^{ij}(\delta_V)$ are the disparity-tuning curves of the population of complex cells, υ_H is the desired behaviour to δ_H, and $\lambda > 0$ balances the relevance of the first term (sensitivity to δ_H) over the second (sensitivity to δ_V).

In general, the insensitivity to δ_V is an important feature. Such an insensitivity allows the control to work in the case of a disparity pattern with a meaningful vertical component, *i.e.* a slanted surface, or an oblique gaze direction. Working with a real robot head, we have to cope with mechanical errors. Misalignment of the optical sensors or of the optical axes, produce unpredictable values of δ_V in the stereo image, that affect the effectiveness of the control. Thus, in designing the weights w_{ij}, the factor λ can be modulated in order to provide a vergence control with a stronger insensitivity to vertical disparity. The resulting system provides a vergence control that is effective in a range of horizontal disparities about three times the theoretical range, *i.e.* $[-3\Delta, 3\Delta]$, and is insensitive to the vertical component of disparity, in a range of $[-\Delta, \Delta]$.

C. Luminance and texture invariance

The vergence control gathered in this way is able to provide the correct behaviour in a range of both horizontal and vertical disparities that is larger than the one supported by the computation of the disparity map. A further problem that arises in real word situations is that the complex cell response has a quadratic dependence on the energy of the binocular image. In fact, from the Fourier transform of the monocular image $\tilde{I}^{L/R}(\omega)$, assuming for the sake of simplicity that locally $\tilde{I}^L \approx \tilde{I}^R$, the response of the complex cell (see Eq. 4) becomes:

$$r_c(\mathbf{x}) \approx \frac{16\pi^4 \left|\tilde{I}\right|^2}{\sigma^4} \left(1 + e^{-\frac{|\delta|^2}{\sigma^2}} + 2e^{-\frac{|\delta|^2}{2\sigma^2}} \cos(k_0^T \delta - \Delta\psi)\right)$$

Since the vergence control is computed by a linear summation of the population response (see Eq.5), the module of the Fourier transform $|\tilde{I}|^2$, acts as a multiplicative gain on the control itself. Basically, considering as reference a mean value of $|\tilde{I}|^2$, if the images have low luminance of with a low power spectrum in the bandwidth of the RFs, it results in a slowdown of the control. On the contrary, a high value of $|\tilde{I}|^2$ produces a wider movement than the required, with overshoot and oscillations of the fixation point around the correct position in depth. The desired feature of an ideal vergence signal is a sensitivity to the stimulus disparity only, regardless the luminance of the environment and the texture of the observed object. In order to remove the dependence of the control signal to the energy of the image, it is possible to implement a divisive normalization stage [8]. Such an extension of the binocular energy model was introduced to explain the response saturation to interocular contrast of the complex cell response [8], but yields interesting effects on the amplitude on the population response to natural binocular images. According to this modification, a cell's selectivity is attributed to the energy stage made by a linear summation and of a squaring, as in Eq. 4, and its nonlinear behavior is attributed to division, the normalization stage. The response of each complex cell is divided by a normalization factor E_{bin}, obtained pooling the activity of the complex cells over phases and orientations:

$$\begin{aligned} E_{bin}(\mathbf{x}) \quad &= \quad \int_0^{\pi} \int_{-\pi}^{\pi} r_c(\mathbf{x}; \theta, \Delta\psi)\, d\Delta\psi\, d\theta = \\ &= \quad \frac{32\pi^5}{\sigma^4} \left(1 + e^{-\frac{|\delta|^2}{\sigma^2}}\right) |I|^2 \end{aligned} \tag{7}$$

The normalizing signal, proportional to the local Fourier energy of the stimulus $|\tilde{I}|^2$, has the effect to rescale the cell responses with respect to the stimulus luminance, thus preserving the dependence on the stimulus disparity δ:

$$\begin{aligned} \hat{r}_c(\mathbf{x}) \quad &= \quad r_c(\mathbf{x}) / E_{bin}(\mathbf{x}) = \\ &= \quad 1/2\pi \left(1 + \cos(\Delta\psi - \delta k) \operatorname{sech}\left(\frac{|\delta|^2}{2\sigma^2}\right)\right). \end{aligned} \tag{8}$$

Since locally we can consider $\tilde{I}^L \approx \tilde{I}^R$, from an operative point of view the normalization factor is computed for each retinal location as a summation of $E_{bin}(\mathbf{x})$ over a neighborhood, and weighted by a Gaussian function defined by the same variance σ^2 used to design the RFs.

Thereby the vergence control, being derived from a linear summation of the response of the disparity detectors, is not affected by the stimulus energy too. In case of real images the divisive normalization allows us to remove the dependence on the texture and the luminance of the binocular image, and to maintain an effective and stable vergence control sensitive to changes of the stimulus disparity only.

III. Implementation on the iCub head

A. Materials

The presented vergence algorithm is based on a model of the cortical mechanisms of V1 area, and it is worthy to evaluate its effectiveness on a platform that is designed to resemble the human head. The iCub robot system, engineered to serve as a research tool for embodied cognition, visuomotor coordination, and development [13], is an ideal platform for

the validation of the algorithm. What is interesting for the proposed model, is the baseline of $70mm$, *i.e.* similar to the baseline of a human being. This allows the system, working in the peripersonal space, to experience binocular images with disparities close to those that would fall on the human retinas in similar conditions. The iCub head is endowed with two DragonFly cameras with a resolution of 1024×768 pixels, and a frame rate of up to $15fps$ [14]. The mounted lenses have a focal length of $6mm$, that combined with a sensor size of $1/3''$, provide a field of view of $\approx 80°$. Since the stimulus is positioned and moved manually, it is not straightforward to have the ground truth data of its depth. To measure this data, and thus validate the experimental results, we used a Microsoft *Kinect* sensor device, which is endowed with a range camera, developed by PrimeSense, that interprets 3D scene information from a continuously-projected infrared structured light. The device is able to work as a 3D scanner system, and thus to produce a ground truth of the depth of the scene with a precision appropriate for the measurement, and at a frame rate of $30fps$. The whole system runs on a standard PC with an Intel Core i7 CPU 870 @2.93GHz, and 8GB of RAM.

B. Vergence control in real-time

In order to make the system work in real time, the algorithm was implemented in C++, using the Integrated Performance Primitives (Intel IPP), that is a multi-threaded library of functions resorting on low-level optimizations for multicore and multiprocessor computation. From this perspective, they are an optimal tool for image filtering and elaboration. Moreover, while keeping the accuracy of the control, the images where first cropped to 640×480 (that is half of the field of view), and then resized to a resolution of 160×120 pixels. The vergence control signal r_{v_H} yielded by the algorithm is used as a speed control for the eye movements. The effectiveness and stability of the control is ensured by a *PID* controller, whose parameters where manually tuned.

With those policies, the algorithm is able to produce a vergence control updated at a speed of $\approx 40fps$, which corresponds to a vergence control updated at a frame rate even higher than the one supported by the DragonFly cameras, so to ensure a fast updating of the control, and thus a good stability in real-time functioning.

C. Vergence movements along a generic gaze direction

In general, vergence eye movements are tested on humans and robots considering the gaze line directed straight-ahead [2][6][15][16]. This configuration grants a symmetric motor control. With the aim of an active vision system able to explore autonomously the surrounding environment, it is necessary to provide the robot with the ability to achieve the correct vergence movements independently of the gaze direction.

Considering the neck fixed in a random position, the eye system is characterized by three degrees of freedom: a common elevation or tilt angle (V) for the left and the right eyes, and two distinct azimuth or pan angles (H), that allow the eye system to gaze in 3D. From a geometrical point of view, this configuration of nested axes is described by a zero torsion Helmholtz gimbals system, because the elevation axis is fixed, while the pan axes are solidal with each eye, and move with the elevation. This kind of geometry produces many interesting advantages both in term of deriving the correct vergence signals, and in terms of their use for a simple and straightforward control of the fixation. In fact, supposed the mechanical system to be precise, the common tilt axis ensures that the optical axis are always coplanar, thus they intersect in the fixation point, regardless the pan and tilt angles.

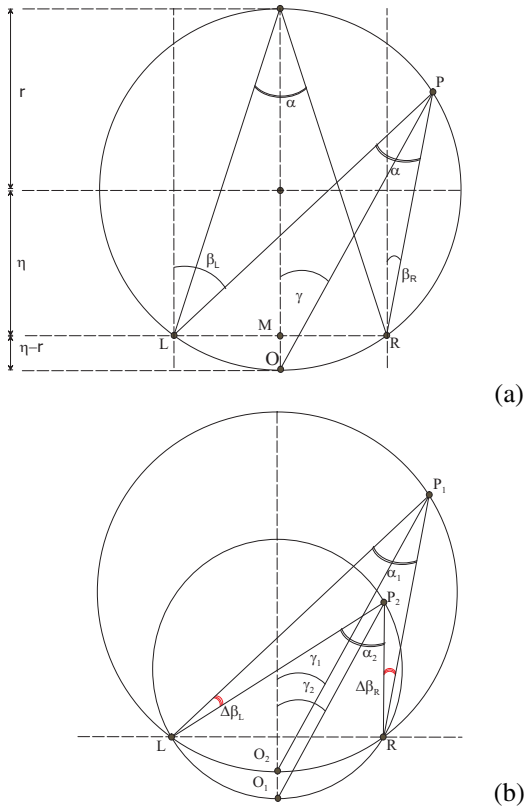


Fig. 1. (a) Cyclopean parametrization of the binocular line of sight in term of version γ and vergence α with respect to the optical axes direction β_L and β_R. (b) Application of a symmetrical vergence control $\Delta\alpha/2$ (in red), from α_1 to α_2, with constant version angle $\gamma_1 = \gamma_2$.

Although we consider the origin of the reference system in the point M, halfway along the interocular axis, the gaze direction γ is more conveniently defined with respect to the point O, that lies at the back of the isovergence circle, that is the circle of radius r passing through the fixation point P and the nodal points L and R, and defines the geometrical locus of the point at zero disparity, (see Fig. 1a). This is not a fixed point, but it changes with the vergence angle α (see Fig. 1b).

This yields a simplified parameterizations of visual direction [17], in which:

$$\begin{cases} \gamma = 1/2(\beta_L + \beta_R) \\ \alpha = \beta_L - \beta_R \end{cases} \quad \text{or} \quad \begin{cases} \beta_L = \gamma - \alpha/2 \\ \beta_L = \gamma + \alpha/2. \end{cases}$$

Thus the vergence control needed to move the fixation point, keeping constant gaze direction γ (see Fig. 1b), is a quantity $\Delta\alpha$ to be applied symmetrically on both the eyes:

$$\Delta\beta_L = \Delta\alpha/2; \;\; \Delta\beta_R = -\Delta\alpha/2. \tag{9}$$

D. Experimental setup

The vergence control implemented on the iCub robot head was first tested in a quantitative way to provide a direct comparison with respect to the psychophysical data [15], and then in a qualitative way, in order to verify the capability of interaction in a real environment. The robot head is kept fixed in a reference position, while the eye position is defined by a specific azimuth and elevation, and it is free to change in terms of vergence angle only. The background stimulus is a surface, perpendicular to the line of sight and characterized by a complex texture (see Fig. 2). The foreground stimulus is an object that can be steady or moving in depth (see Fig. 3). The environment is illuminated with three fluorescent lamps with a total luminous power of $6600lm$. For the quantitative validation, two kinds of visual stimulation were used: (Experiment 1) a stepping-in/stepping-out planar surface (step stimulus), and (Experiment 2) a waving planar surface (sinusoid stimulus). To validate the effectiveness of the control to work in real-word situation, we tested the control with different lighting conditions and complex 3D shape (Experiment 3). The *Kinect* sensor is placed behind the robot head, in order to have a ground truth measure of the position of the stimulus relative the head.

IV. Results and Discussion

The goal of the algorithm is, in case of a steady object, to reach and to maintain a precise and stable fixation along the binocular line of sight, in a short response time. In such a configuration, the binocular image is characterized by zero disparity in the fovea, since the optical axes intersect on the same point on the surface of the object. On the other side, in case of a moving object, the algorithm has to move the fixation point so to track the object in depth. In this way the binocular disparity, even if it is never close to zero, it is kept in a range where its estimate is reliable. The actual position of the fixation point is computed through the position of the motors with respect to a reference position. Analyzing a single track, (see for instance Fig. 4) the fixation point is not always at the correct depth registered by the Kinect sensor device, suggesting a low accuracy in moving the fixation point to the object's surface. Actually we must consider that the DC motors of the eye system have a backlash $\leq 1°$, and this may lead to a biased evaluation of the depth of the fixation point. For instance, if the robot is correctly fixating an object at a depth of $600mm$, the position estimate given by the magnetic encoders is in the range of $520 \div 705mm$, that is a significant error. The error given by the backlash on the motors does not affect the accuracy and the effectiveness of the algorithm because, since it works in a closed visual loop, the control stops when the disparity in the fovea is reduced to a value close to zero, regardless of the real depth of the object, and of the position of the motors. The effectiveness of the vergence movement is than validated through the anaglyph image of the stereo pair (see Fig. 3).

A. Experiment 1

For the step stimulus, the foreground surface was placed at a given distance from the head, along the binocular fixation axis, and removed afterwards (see Fig. 2). The fixation point, starting at the correct depth from the foreground surface, has to move to the background surface and stop there. The distance between the head and the background is fixed at $1400mm$, while the distance between the head and the foreground surface, in order to test steps of different values, varies in a range between $600mm$ and $1200mm$, thus requiring a change of vergence angle from $\approx 0.45°$ ($1200mm$) to $\approx 3.8°$ ($600mm$). The results show that the vergence control is able to discriminate properly the necessity for small movements of the fixation point, in presence of small steps (*i.e.* of small disparities), so as to produce wider movements in case of large steps (*i.e.* of large disparities). In any of the tested steps (see Fig. 4), the control has completed the majority of the vergence movement within $1s$, and the eyes are steady at the new depth within $2s$. The control for a large step (*i.e.* from $600mm$) is less precise than for the trial in which the foreground stimulus is further away (*i.e.* $1200mm$). Indeed, at that distance a small difference in depth means a very small disparity, whereas a

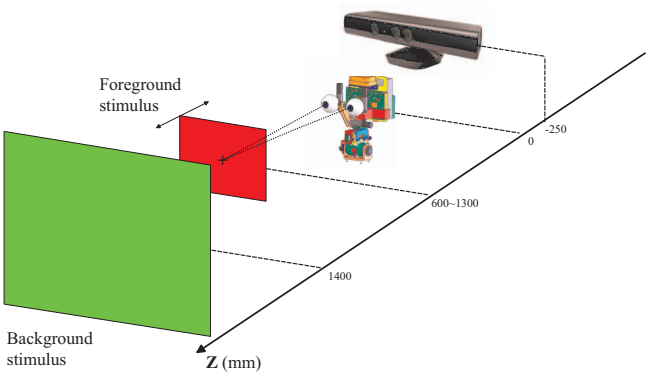
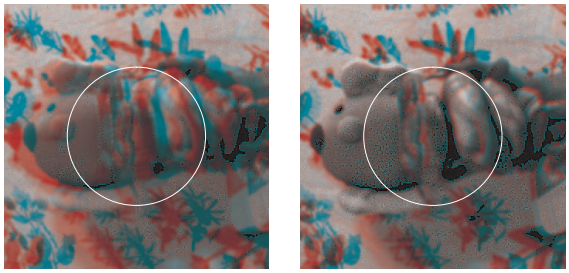


Fig. 2. Experimental setup: the robot head is placed in the reference position, the background stimulus is a surface perpendicular to the line of sight and characterized by a complex texture, the foreground stimulus is an object, steady or moving in depth, while the *Kinect* sensor is placed behind the robot.



Before vergence (a) After vergence (b)

Fig. 3. Anaglyph image of the stereo pair: the left and right gray scale images are superimposed on different color channels. (a) When the eye system is not at the correct vergence angle ($\delta \neq 0$), the anaglyph is "double". (b) When the fixation point lies on the object's surface ($\delta \approx 0$), the left and right images are "fused" in gray scale. The white circle highlight the area where the vergence control is computed.

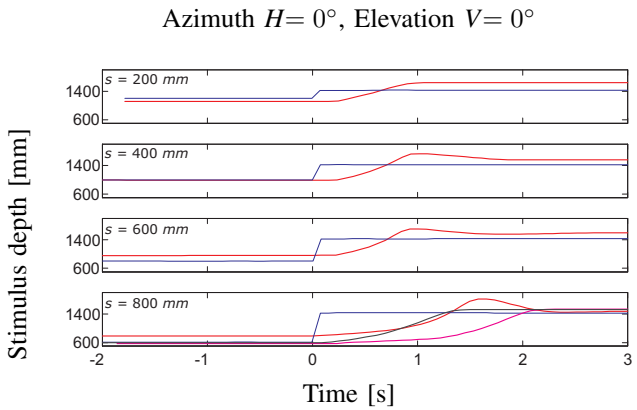


Fig. 4. Stimulus depth (blue line), plotted against the depth of the fixation point of the robot. The background stimulus is at a fixed distance of $1400mm$, while the forground one is positioned at a depth varying in the range of $600-1200mm$ from the robot. The experiment is repeated for a gaze direction of H= 0, and V= 0 (red line), H= 30 and V= 0, and H= 30 and V= 30 (magenta and black lines, shown only for the larger step).

small vergence movement yields a large movement of the fixation point.

B. Experiment 2

In the sinusoid stimulus, in order to test the ability to follow in depth a moving object, the foreground surface oscillates about $800mm$ from the head with an amplitude of $200mm$, thus moving between $600mm$ and $1000mm$, *i.e.* with a change of vergence of $\approx 2.6°$ from the closest to the farthest position. The frequency of the oscillation varies from trial to trial from $30Hz$ to $70Hz$. The resulting control yields an effective

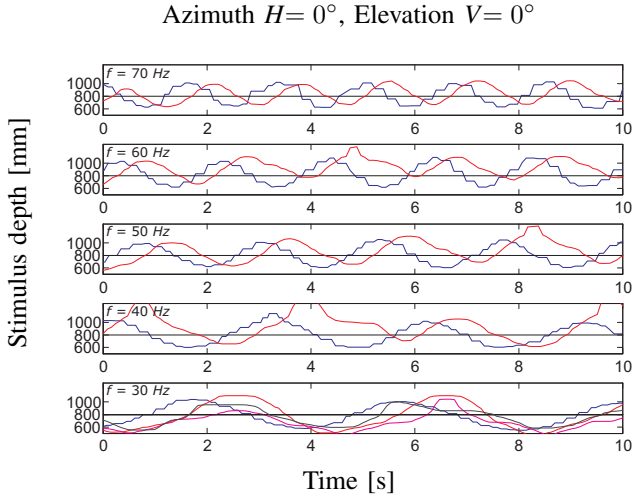


Fig. 5. Stimulus depth (blue line), plotted against the depth of the fixation point of the robot. The stimulus starts at a distance of $800mm$ and oscillates with an amplitude of about $200mm$ at a frequency varying from $30Hz$ to $70Hz$. The experiment is repeated for a gaze direction of H= 0, and V= 0 (red line), H= 30 and V= 0, and H= 30 and V= 30 (magenta and black lines, shown only for the lower frequency).

tracking in depth of the stimulus (see Fig. 5). When the stimulus is moving slowly (bottom rows) the fixation point (red line) follows its depth (blue line) with a small delay. For higher frequencies (top rows) the control seeks to achieve the correct movement, but it is not precise as in the previous cases. The behavioral response to sinusoidal stimuli, closely resembles the psychophysical data [15], showing a fast reaction for higher frequencies (top rows) and accurate, slow and smooth tracking (bottom rows) .

In order to validate the control performance on a real system even when the gaze line is not straight ahead, we tested both the step and sinusoid stimuli for different gaze directions: straight ahead (H= 0°, V= 0°), right (H= 30°, V= 0°), and up-right (H= 30°, V= −30°). The trajectories of the fixation point are qualitatively equivalent, regardless of the gaze direction, for both the stimulations (See Figs. 4 and 5, bottom rows).

Experiments 1 and 2 are shown briefly in a demo video *http://www.youtube.com/watch?v=oNpc_aDZ6uA*. A close-up of the robot cameras shows the size and the precision of the vergence movement, and the anaglyph image of the robot view demonstrates that the fixation point is always close to the stimulus depth. The control is effective even for a slanted plane with respect to the line of sight and in a very near viewing condition.

C. Experiment 3

For what concerns the ability of the robot to work properly in real-word conditions, different situations and stimuli have been considered: 1) variable lighting conditions (moving the light spot, switching off the light); 2) different surface textures (a chessboard, a cartoon, a leaf texture); 3) different surface orientations with respect to the line of sight; 4) complex (*i.e.* not planar and irregular) 3D shapes (a doll, a money box, a ball); 5) deformable objects (a cloth, a hand).

Without the normalization circuits, the vergence control can be effective in stable and defined lighting conditions, only. A single trial was repeated both for the step and the sinusoid stimulus. During the run of the trials, the luminous power was switched from $4400lm$ to $6600lm$ by turning on an additional light. Without the divisive normalization, (see Fig. 6a, top row), even starting from a correct fixation, the change of lighting condition prevent the stability of the control. Furthermore, after the stimulus is moved the fixation point follows it, but since the increasing of the luminous power results in a increasing of the control gain, the control produces wide oscillations around the depth of the stimulus. In a similar way, with the sinusoid stimulus, the fixation point is oscillating about the stimulus depth already when the additional light is off, and the effect increases with the luminous power, leading to high misalignment of the eyes respect to the stimulus (see Fig. 6b, top row). The divisive normalization provides a strong stability of the control that is unaltered by the changing light. Both in the step and in the sinusoid, there are no meaningful changes in the behaviour (see Fig. 6a-b, bottom row).

In [1][2], the behaviour of a verging system, like a monkey or a human being, was tested with random dot stereograms, to obtain the disparity-vergence tuning curves. Indeed, our setup consists of a real 3D environment, whose ground truth disparity data is not available. Since the disparity can be roughly mapped into the depth of the stimulus, to characterize

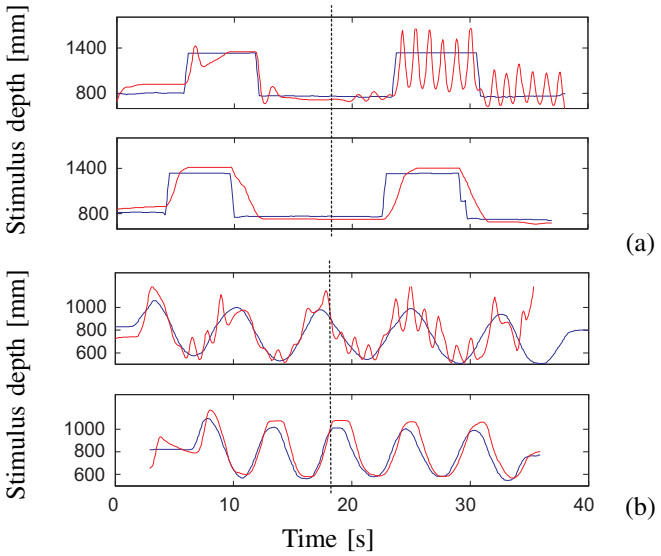


Fig. 6. Single trial for (a) the step and (b) the sinusoid stimulus, without the divisive normalization (top row), or including it (bottom row). During the run of the trials, the luminous power was switched from $4400lm$ to $6600lm$ with an additional light. The black dashed line indicates the switching instant.

the behaviour of our control and to understand the effect of the normalization, we derived a depth-vergence tuning curve.

Starting from a steady fixation point at 600 mm, we moved a frontoparallel surface from a depth of 350 to $1300mm$ several time. The disparity range is approximately $[-2.5\Delta, 2.5\Delta]$. We plotted the vergence control measured $60ms$ after the presentation of the disparity step and its standard deviation, against the magnitude of the step in mm. By repeating the test with a luminous power of 4400 and $6600lm$, the control without the normalization, shows a strong dependence on the variation of the light (see Fig. 7a), in fact with the low light it results in a effective but slow control (blue line), whereas with the more intense light the control is fast but the excess of gain produces high instability (red line). By including the normalization, the control is almost unaffected by the light variation (see Fig. 7b) because it works as a dynamic adaptation of gain. The resulting tuning curves can be qualitatively related to the initial vergence responses to disparity steps in humans and monkeys [1][2] (see inset in Fig. 7a). Similarly we can demonstrate that the normalized control is unaltered with respect to the object's texture (not shown).

When using slanted surfaces (*i.e.* producing a not constant disparity pattern), objects with complex 3D shapes, and deformable objects, the model, gathering disparity information from a perifoveal area (see Fig. 3), is able to provide a vergence control that brings the fixation point at a mean depth with respect to the scene structure, making the system to achieve a fixation characterized by a minimum average disparity.

When the robot has to move the gaze to explore the peripersonal space, starting from the fixation point at a given vergence angle, *i.e.* on a isovergence circle (see Fig. 8, black line), the system is able to yield the correct estimate of the

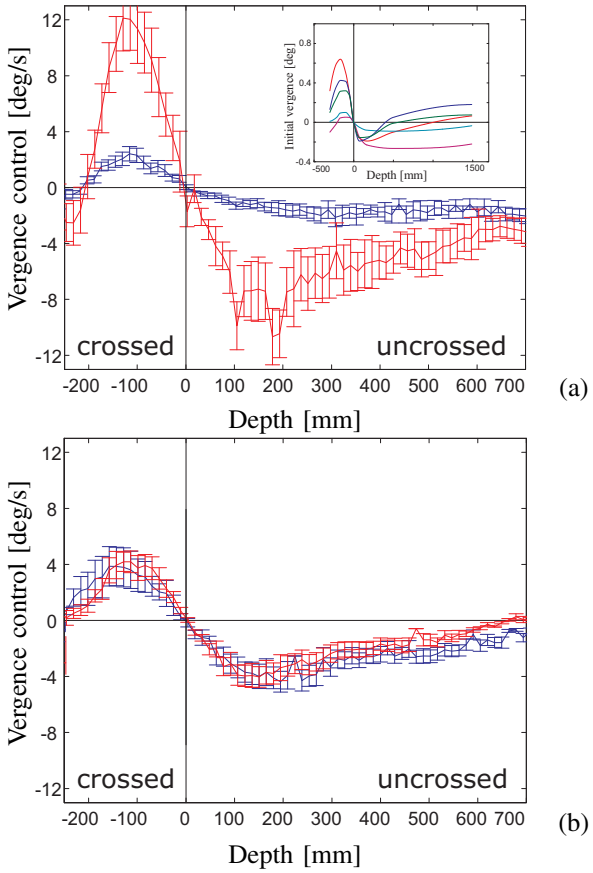


Fig. 7. Depth-vergence tuning curve for the control in case of low luminance (blue line), high luminance (red line), without the divisive normalization (a) or including it (b). The curves are obtained with the initial mean vergence control, measured $60ms$ after the disparity step, plotted against the magnitude of the step in mm. The obtained profiles qualitatively resemble the initial vergence control to disparity step that can be observed in monkeys. The insert in (a) shows this relation between the depth of the stimulus and the initial vergence movement, observed in five different monkeys (adapted from [2]).

vector disparity for an object is inside the blue area, in a way similar to what happens in Panums area. In fact, a correct vergence posture allows a higher overlapping of the left and right images [18], *i.e.* smaller disparities, at least in the region of interest. These disparities, even though an extra computation is required for the epipolar geometry, are easier to be reliably estimated, and can be used for the reconstruction of the 3D structure of the perceived scene [19]. If the object falls outside the blue area, the resulting vergence control is able to produce the correct vergence movements while the retinal disparity produced by the object is within the range $[-2.5\Delta, 2.5\Delta]$ (red area), allowing a proper fixation, and bringing the system to work in a operative condition for the estimation of disparity, from which the depth perception. The effective range of disparities decreased with respect to the one obtained in a simulated environment ($[-3\Delta, 3\Delta]$), because of the vertical disparity offset, due to mechanical inaccuracies.

The last part of the demo video shows the effectiveness of the control in case of random gaze movements in the peripersonal space. When the gaze shift is terminated (blue speed profile), the vergence control (yellow speed profile) yields a movement to nullify the disparity in the fovea.

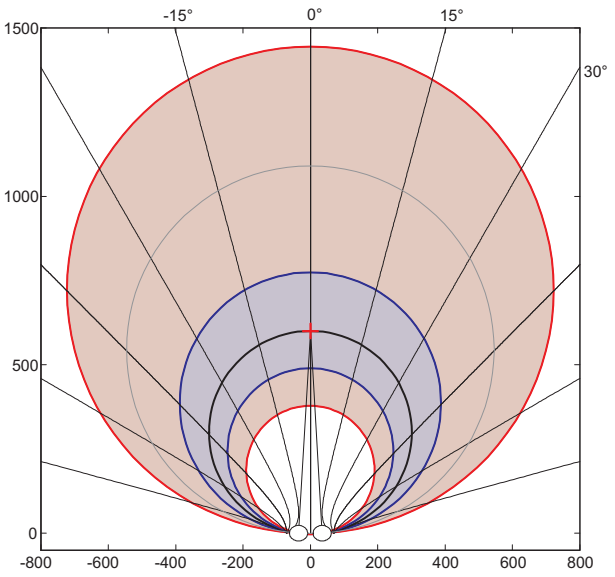


Fig. 8. Range of effectiveness of the algorithm implemented on the iCub stereo head. When the robot is fixating with $6.5°$ of vergence (black circle), *i.e.* at $600mm$ with the gaze straight ahead (red cross), the algorithm is able to provide a reliable estimate of the disparity, while an object is in the blue area, and to provide the correct vergence movement on an object positioned along the gaze line, while it is inside the red area.

V. Conclusion

In this work, we presented the implementation of a cortical like (*i.e.* distributed) real-time control for vergence eye movements, to validate the effectiveness of the model. In real-word situation, the control is able to cope with changing lighting condition and objects with different textures and shapes, and to yield correct vergence movements at any direction of the gaze. These features are essential to provide a robust exploration of the peripersonal space. Furthermore, the obtained behaviour demonstrates the effectiveness of the vergence control in tests similar to psychophysical experiments. Besides the correct vergence signal, the algorithm produces trajectories of the fixation point close to psychophysical data, in term of reaction time, velocity and range of effectiveness.

The flexibility of the distributed phase-based representation of disparity information, and the efficacy of the the divisive normalization circuits, allows a more immediate and efficient/effective use of visual data, even without correction for lens distortion. Triggering eye movements without a complete perception, is a first step toward avoiding a stiff sequentialization of sensorial and motor processes, that is certainly desirable for the development of cognitive abilities of active vision systems. The proposed model, integrating directly early vision modules and motor control, close the perception-action loop in order to give to an artificial system the capability of perceiving the 3D structure of the environment and of coordinating the camera movements to better exploit its potential, with a minimal amount of resources and coping with uncertainties and inaccuracies of real systems.

The further step is to include in the model a space-variant (log-polar) geometry that mimics the topological transformation from the retinal to the cortical domain [20], in order to achieve a higher precision and a larger working range.

Acknowledgment

This work has been partially supported by the EC Project FP7-ICT-217077 "EYESHOTS - Heterogeneous 3D perception across visual fragments" and by the Italian MIUR (PRIN 2008) project "Modelli bio-ispirati per il controllo dei movimenti oculari nella visione attiva e l'esplorazione 3D".

References

[1] G. Masson, C. Busettini, and F. Miles, "Vergence eye movements in response to binocular disparity without depth perception," *Nature*, vol. 389, pp. 283–286, 1997.

[2] A. Takemura, Y. Inoue, C. Quaia, and F. Miles, "Single-unit activity in cortical area mst associated with disparity-vergence eye movements: Evidence for population coding." *J. Physiol.*, vol. 85(5), pp. 2245–2266, 2001.

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

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

[5] A. Franz and J. Triesh, "Emergence of disparity tuning during the development of vergence eye movements," in *International Conference on Development and Learning 2007*, London, 11-13 July 2007, 2007, pp. 31–36.

[6] Y. Wang and B. Shi, "Autonomous development of vergence control driven by disparity energy neuron populations," *Neural Comput.*, vol. 22, pp. 730–751, 2010.

[7] A. Gibaldi, M. Chessa, A. Canessa, S. Sabatini, and F. Solari, "A cortical model for binocular vergence control without explicit calculation of disparity," *Neurocomp.*, vol. 73, pp. 1065–1073, 2010.

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

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

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

[11] M. Chessa, S. 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.

[12] F. Solari, S. Sabatini, and G. Bisio, "Fast technique for phase-based disparity estimation with no explicit calculation of phase," *Elect. Letters*, vol. 37, no. 23, pp. 1382–1383, 2001.

[13] R. Beira, M. Lopes, M. Praga, J. Santos-Victor, A. Bernardino, G. Metta, F. Becchi, and R. Saltaren, "Design of the robot-cub (icub) head," *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.*, pp. 94–100, 2006.

[14] Point-Grey-Research, "Firewire cameras. http://www.ptgrey.com," 2010.

[15] G. Hung, J. Semmlow, and K. Ciuffreda, "A dual-mode dynamic model of the vergence eye movement system," *IEEE Trans. Biomed. Eng.*, vol. 36, no. 11, pp. 1021–1028, 1986.

[16] J. Piater, R. Grupen, and K. Ramamritham, "Learning real-time stereo vergence control," in *Intelligent Control/Intelligent Systems and Semiotics, 1999*, Cambridge, MA , USA, 1999, pp. 272–277.

[17] M. Hansard and R. Horaud, "Cyclopean geometry of binocular vision," *J. Opt. Soc. Am.*, vol. 25, pp. 2357–2369, 2008.

[18] U. Dhond and J. Aggarwal, "Structure from stereo - a review," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 19, no. 6, pp. 1489–1510, 1989.

[19] J. Cardenas-Garcia, H. Yao, and S. Zheng, "3d reconstruction of objects using stereo imaging," *Optics and Lasers in Engineering*, vol. 22, no. 3, pp. 193–213, 1995.

[20] A. Gibaldi, M. Chessa, A. Canessa, S. Sabatini, and F. Solari, "A cortical model for vergence control: advantages of space-variant geometry of the cortical domain." in *COSYNE*, Salt Lake City, Utah, USA, February 2011.