Int J Comput Vis (2017) 121:281–302
DOI 10.1007/s11263-016-0936-z

A Portable Bio-Inspired Architecture for Efficient Robotic Vergence Control

Agostino Gibaldi[1] · **Mauricio Vanegas**[1] · **Andrea Canessa**[1] · **Silvio P. Sabatini**[1]

Received: 5 December 2014 / Accepted: 18 July 2016 / Published online: 8 August 2016

Abstract In stereoscopic vision, the ability of perceiving the three-dimensional structure of the surrounding environment is subordinated to a precise and effective motor control for the binocular coordination of the eyes/cameras. If, on the one side, the binocular coordination of camera movements is a complicating factor, on the other side, a proper vergence control, acting on the binocular disparity, facilitates the binocular fusion and the subsequent stereoscopic perception process. In real-world situations, an effective vergence control requires further features other than real time capabilities: real robot systems are indeed characterized by mechanical and geometrical imprecision that affect the binocular vision, and the illumination conditions are changeable and unpredictable. Moreover, in order to allow an effective visual exploration of the peripersonal space, it is necessary to cope with different gaze directions and provide a large working space. The proposed control strategy resorts to a neuromimetic approach that provides a distributed representation of disparity information. The vergence posture is obtained by an open-loop and a closed-loop control, which directly interacts with saccadic control. Before saccade, the open-loop component is computed in correspondence of the saccade target region, to obtain a vergence correction to be applied simultaneously with the saccade. At fixation, the closed-loop component drives the binocular disparity to zero in a foveal region. The obtained vergence servos are able to actively drive both the horizontal and the vertical alignment of the optical axes on the object of interest, thus ensuring a correct vergence posture. Experimental tests were purposely designed to measure the performance of the control in the peripersonal space, and were performed on three different robot platforms. The results demonstrated that the proposed approach yields real-time and effective vergence camera movements on a visual stimulus in a wide working range, regardless of the illumination in the environment and the geometry of the system.

Keywords Vergence control · Active vision · Stereo vision · Binocular energy models · Neuromorphic architectures

Communicated by S. Soatto.

✉ Agostino Gibaldi
agostino.gibaldi@unige.it
http://www.pspc.unige.it

Mauricio Vanegas
http://www.pspc.unige.it

Andrea Canessa
http://www.pspc.unige.it

Silvio P. Sabatini
http://www.pspc.unige.it

[1] Physical Structure of Perception and Computation Group, Department of Informatics, Bioengineering, Robotics and System Engineering, University of Genoa, Genoa, Italy

1 Introduction

From the very beginning to nowadays, the perceptual process of an active vision system is considered to be an "exploratory and searching" activity (Aloimonos et al. 1988; Bajcsy et al. 2016) in which the system modifies the cameras' viewpoints in order to exploit its full potential.

Considering a binocular active visual system with a vergent stereo geometry, a common approach to guide binocular fovation is to select the target on the two retinas, and to compute the camera control as a combination of a *conjugate* version movement, to shift the binocular line of sight towards the object of attention, and a *disconjugate* vergence, to properly align the optical axes in depth (Enright 1998;

Samarawickrama and Sabatini 2007; Zhang and Phuan 2009; Muhammad and Spratling 2015). From this perspective, dynamic vergence comes to be an instrumental resource not just for an expansion of the visual field, but mainly because at close distances the eyes' coordination ensures binocular fusion, providing a powerful and reliable source of information for visually guided behaviours.

The numerous advantages of a proper vergence control basically arise from a reduction of the search space for the stereo correspondences and from the consequent simplification of the related algorithms. In a static scene (*i.e.* stationary objects), vergence overcomes the problem of large disparities and allows the computation of the fixation point (Culverhouse et al. 2009; Belhaoua et al. 2010) and of the absolute depth map by a triangulation algorithm (Hansen and Sommer 1996). If the disparities close to the fixation point are kept small, segmentation (Mishra et al. 2009), stereo matching and object recognition (Das and Ahuja 1995; Belhaoua et al. 2010) problems are facilitated. A correct vergence posture is also instrumental to visual attention and scene understanding (Rea et al. 2014). In a dynamic scene (*e.g.* moving objects or egomotion) the vergence angle can be actively exploited in smooth pursuit tasks, to improve the segmentation capabilities of a moving object (Coombs and Brown 1993; Bernardino and Santos-Victor 1998; Choi et al. 2003), like a robot hand (Dankers et al. 2007), to continuously estimate the epipolar geometry of the vision system (Bjorkman and Eklundh 2002; Pauwels and Van Hulle 2012) and to improve its estimation (Monaco et al. 2009), as well as for navigation purpose (Konolige 1998; Knight and Reid 2006).

Aiming to drive the vergence control of a real system in a complex and dynamic three-dimension (3D) environment, the salient features needed for a proper binocular coordination can be summarized as follows: (1) real-time capabilities with short reaction time, (2) significant accuracy and precision, (3) the ability to work at different gaze directions with a wide-angle working space, (4) robustness to both the geometrical imprecision of an actual camera system, and to changes in the environmental illumination conditions, and (5) the ability to provide the correct vergence correction for a version movement in 3D. It is a common practice evaluating the performance of the vergence algorithms with respect to the first two features, only (Theimer and Mallot 1994; Ching et al. 1995; Daniilidis et al. 1996; Bernardino and Santos-Victor 1998; Marfil et al. 2003; Bana and Lee 2007; Tsang et al. 2008; Shimonomura and Yagi 2010; Kyriakoulis et al. 2010). Moreover, the problem of vergence elicited by retinal disparity is commonly cast as a horizontal one-dimensional (1D) problem, and the effectiveness of vergence control is shown with the eyes fixating straight-ahead. This assumption is an oversimplification of the problem that hardly holds for an active system. Indeed, in vergent geometry, binocular disparity becomes a two-dimensional (2D) feature with horizontal and vertical components (Howard and Rogers 2002; Hansard and Horaud 2010). Moreover, with a real robot head the optical axes cannot be considered vertically aligned, so that this misalignment introduces a further vertical disparity pedestal that depends on the gaze and on the vergence angle.

In this paper, we present a biologically-inspired architecture for efficient vergence control that relies on a population of cortical-like oriented binocular disparity detectors (Gibaldi et al. 2010, 2011). The control is able to cope with the vertical disparity originated both by the vergent geometry and by the misalignment of the optical axes. The implemented read-out mechanism of the disparity population code is able to produce adequate vergence control servos that satisfy the requirements posed by real-world applications. The adopted single-scale approach yields a good accuracy of the control with a minimal amount of resources, and thus provides the real-time behaviour needed for robot control. To overcome instability effects that result from unpredictable changeable lighting condition, the architecture includes a divisive normalization circuit (Gibaldi et al. 2010, 2011).

With respect to our previous works (Gibaldi et al. 2010, 2011, 2012), the architecture includes two novel extensions, which, to the best of our knowledge, have never been implemented and validated on real stereo heads. First, the distributed cortical representation has been exploited to provide also control servos for the vertical vergence, in order to simultaneously drive the horizontal and the vertical alignment of the optical axes. Second, during 3D world exploration, a vergence correction is computed on the object to be fixated before the saccade. This correction is applied as an open-loop vergence control, simultaneous with the saccade, to provide, already after the first saccade, a vergence posture closer to the desired one.

By exploiting the recent technologies for range data acquisition (RGB-D cameras), we designed a set of experiments that allowed us to quantitatively assess the effectiveness of the algorithm in terms of stability, accuracy and precision, as well as the working range of the control and of fixation point's trajectories in the 3D space. The flexibility and robustness of the algorithm have been proved by testing it on three robot stereo heads with major differences in their geometrical and optical characteristics. The results, comparable across the different robot platforms, evidenced that the proposed architectural solution for vergence control allows for an effective binocular coordination of the cameras in the real environment.

The paper is organized as follows: in Sect. 2 we frame the mathematical formulation of binocular coordination of camera movements for Tilt-Pan and Pan-Tilt systems; in Sect. 3 we review the relevant state-of-the-art algorithms for vergence control; in Sect. 4 we present our network architecture for the closed-loop horizontal and vertical vergence control and for open-loop version control; in Sect. 5 we first explain

the implementation of the algorithm in real time, and the characteristics of the different used stereo heads, and subsequently we detail the design of the experimental setup and discuss the results; finally, in Sect. 6, we draw the conclusions.

2 Binocular Camera Fixation in 3D

2.1 Monocular Fixation

Let us consider a camera that has to perform a fixation task. The required control to move the camera for fixating in a certain direction is directly computed by the current camera orientation and by the coordinates of the target point on the image plane ($\mathbf{x}$). The classical geometric configurations of the motors to move a camera are mainly of two kinds: the Tilt-Pan and the Pan-Tilt systems.

2.1.1 Tilt-Pan

From a geometrical point of view, a Tilt-Pan stereo head is described by the Helmholtz gimbal system rotation sequence (Van den Berg 1995). Defined a head-fixed reference frames $\langle h \rangle = \{\mathbf{h}_X, \mathbf{h}_Y, \mathbf{h}_Z\}$, we perform first a rotation by an angle H^{HZ} around the $\mathbf{h}_Y$ axis, followed by a rotation by an angle V^{HZ} around the $\mathbf{h}_X$ axis. The superscript HZ stands for Helmholtz. This leads to a final camera orientation that, for the sake of simplicity, can be described using the rotation vector notation:

$$\mathbf{q}^{HZ} = \mathbf{q}(H^{HZ}, V^{HZ}) = \left(\tan\frac{V^{HZ}}{2}, \tan\frac{H^{HZ}}{2}, \tan\frac{V^{HZ}}{2}\tan\frac{H^{HZ}}{2}\right) \quad (1)$$

that represent the vertical, horizontal and torsional components of the rotation axis from the primary (reference) camera position to the current one, respectively. In particular, the torsional component is not directly controllable in tilt-pan geometry, and it is defined by the H and V angles. In this case, suppose an actual camera orientation $\mathbf{q_1}(H_1, V_1)$ and a target fixation point $\mathbf{x}$ in the image plane, expressed in Helmholtz system by its angular coordinates $\underline{\angle \mathbf{x}} = (H_\mathbf{x}, V_\mathbf{x})$, where $\underline{\angle \mathbf{x}}$ stands for the horizontal and vertical angles subtended by the point $\mathbf{x}$ (see Fig. 1a, left). On this basis, a proper transformation is required to map the angular position of the target $(H_\mathbf{x}, V_\mathbf{x})$ on the image plane into the motor commands to move the camera to the new position (H_2, V_2). The details are reported in the Appendix.

2.1.2 Pan-Tilt

Pan-Tilt systems are described by the Fick gimbal sequence (Van den Berg 1995): first a rotation by an angle V^{FK} around the $\mathbf{h}_X$ axis and then a rotation by an angle H^{FK} around the $\mathbf{h}_Y$ axis. The superscript FK stands for Fick. Also in this case, it is possible to express the final orientation as a rotation vector:

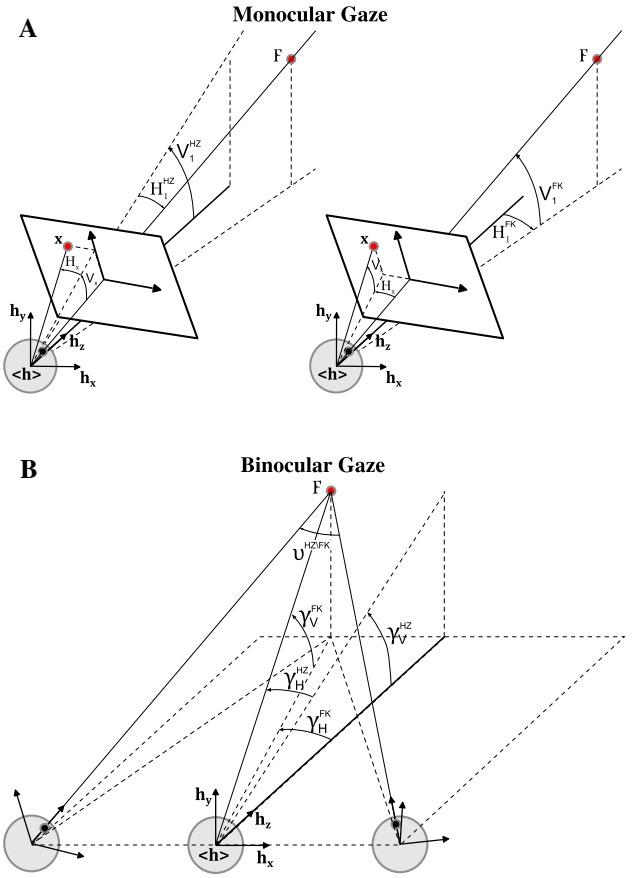


Fig. 1 Representation of the monocular (**a**) and binocular (**b**) gaze direction for the Tilt-Pan (HZ) and Pan-Tilt (FK) reference frames

$$\mathbf{q}^{FK} = \mathbf{q}(H^{FK}, V^{FK}) = \left(\tan\frac{V^{FK}}{2}, \tan\frac{H^{FK}}{2}, -\tan\frac{V^{FK}}{2}\tan\frac{H^{FK}}{2}\right) \quad (2)$$

which, as for Eq. 1, represents the vertical, horizontal and torsional components of the rotation axis from the primary (reference) camera position to the current one, respectively. Again, the torsional component is not directly controllable in pan-tilt geometry, and it is defined by the H and V angles. In this case, suppose an actual camera orientation $\mathbf{q_1}(H_1, V_1)$ and a target fixation point $\mathbf{x}$ in the image plane, expressed in Fick system by its angular coordinates $\underline{\angle \mathbf{x}} = (H_\mathbf{x}, V_\mathbf{x})$ (see Fig. 1a, right). Similarly to the Tilt-Pan system, a proper transformation is required to move the camera to the new fixation position (H_2, V_2). The details are reported in the Appendix.

2.2 Binocular Coordination

The monocular strategies presented above could be straightforwardly extended to the binocular case, where two cameras have to fixate the same 3D point. Though, it is convenient not to consider the two cameras as moved by two monocularly computed controls (see Eqs. 18 and 19). For a real efficient coordinated binocular movement, it is a better solution considering camera positions as not independently controlled but driven by a combination of two binocular controls, in the spirit of the Hering's law of equal innervation (Hering 1868). Those controls act in common on both the cameras, one in a conjugate and the other in a disconjugate fashion. These are commonly referred as a *conjugate version* and a *disconjugate vergence* control.

Commonly in the literature, studies on vergence control are grounded on the assumption of a zero version control (a straight ahead binocular *gaze direction*), both in living systems (Hung et al. 1986; Takemura et al. 2001) and robots (Theimer and Mallot 1994; Daniilidis et al. 1996; Piater et al. 1999; Shimonomura and Yagi 2010; Wang and Shi 2010). In such a configuration, the geometrical characterization of the system is rather straightforward: the optical axes of the two eyes intersect in a point along the gaze direction, and vergence is defined by the angle defined by these axes. Moreover, a change of vergence is a simple rotation of each eye about its vertical axis, in opposite directions, and is provided by an antisymmetric motor control.

More generally, aiming to an active vision system capable of autonomously exploring the surrounding environment, it is necessary to provide the robot with the ability of achieving the correct vergence movements independently of the gaze direction. From this perspective, it is possible to transforms the right ($\mathbf{q}_R$) and left camera ($\mathbf{q}_L$) rotation vectors into a pair of binocular rotation vectors: the conjugate rotation vector ($\mathbf{q}_\gamma$) and the disjunctive rotation vector ($\mathbf{q}_\nu$) (Rijn and Berg 1993; Minken et al. 1994). Each binocular camera position is thus described as the result of conjugate and disconjugate rotations:

$$\begin{aligned} \mathbf{q}_L &= \mathbf{q}_\nu \otimes \mathbf{q}_\gamma \\ \mathbf{q}_R &= -\mathbf{q}_\nu \otimes \mathbf{q}_\gamma \end{aligned} \tag{3}$$

where $\otimes$ indicate the rotation vector product (Rijn and Berg 1993), and $\mathbf{q}_\gamma$ and $\mathbf{q}_\nu$ are defined in the Appendix (Eqs. 18 and 19):

$$\begin{aligned} \mathbf{q}_\gamma &= \mathbf{q}_\gamma(\gamma_{\mathrm{H}}, \gamma_{\mathrm{V}}) \\ \mathbf{q}_\nu &= \mathbf{q}_\nu(\nu_{\mathrm{H}}, \nu_{\mathrm{V}}) \end{aligned} \tag{4}$$

for Tilt-Pan system and Pan-Tilt system, respectively, according to Fig. 1b. We call γ_{H} and ν_{H} the horizontal version and vergence angles; analogously, γ_{V} and ν_{V} are the vertical version and vergence angles. Assuming to know the projections of a 3D point in space on the left $\mathbf{x_L}$ and the right $\mathbf{x_R}$ camera image planes, commonly the version and vergence control are computed respectively by the average and by the difference of their coordinates (Samarawickrama and Sabatini 2007; Zhang and Phuan 2009; Muhammad and Spratling 2015):

$$(\gamma_{\mathrm{H}}, \gamma_{\mathrm{V}}) = (\underline{\angle \mathbf{x_L}} + \underline{\angle \mathbf{x_R}})/2 \tag{5}$$

$$(\nu_{\mathrm{H}}, \nu_{\mathrm{V}}) = \underline{\angle \mathbf{x_L}} - \underline{\angle \mathbf{x_R}} \tag{6}$$

In order to perform binocularly coordinated eye movements, the variation of those angles will be driven by the control signals that we will define in Sect. 4.

Despite the theoretical correctness of this approach, it is worth considering that, in the real world, the target does not project as a single "point" on the image plane, but it results in an image region with a shape and an extension. In monocular version control, some errors can be tolerated, provided that the fixation point lands not too far from the centroid of the target, and the segmentation approach provides sufficiently robust information. Conversely, a correct vergence posture requires a fine alignment of the optical axes with pixel accuracy. Though, in binocular viewing, object segmentation may not be able to provide the required accuracy, due to different issues: (1) differences in the optics and sensitivity of the left and right cameras that might prevent an equivalent segmentation on the two images, (2) the different perspective resulting by the horizontal separation of the two cameras that might result in considerably different segmented shapes in the two images, particularly in close viewing, and when the object to be segmented has a complex 3D shape. Accordingly, to improve vergence accuracy and robustness, different methodologies have been proposed, which exploit other visual features other than the segmented object. In the next section, we will review the literature, with a particular emphasis on the approaches based on the binocular disparity.

3 Related Works

Disparity-based vergence requires solving the stereo correspondence problem and planning/executing disconjugate (i.e., in opposite direction) rotations of the eyes in order to nullify the final disparity of the fixated target. A number of different algorithms have been proposed to solve the problem of vergence eye movement guidance, based on the binocular disparity. The first distinction that is worth evidencing is that the image can be processed in Cartesian coordinates or in log-polar coordinates after a space-variant mapping that resembles the mapping from the retinal to the cortical plane in

the primary visual cortex (Schwartz 1977). This transformation allows us to compress the peripheral information, thus reducing the computational load, while preserving the resolution in the foveal region. Independently of the mapping adopted, the algorithms can be grouped mainly in three categories: (1) those based on the extraction of corresponding features in the stereo image, (2) those based on a measure of the local correlation or (3) those based on the local phase difference between the left and right images.

Referring to the feature extraction, many algorithms (*e.g.* Peng et al. 2000; Samarawickrama and Sabatini 2007; Zhang and Phuan 2009) are first based on a color segmentation of the object of interest in the stereo image. Even if such an approach is not designed to cope with a real and unstructured environment, some works (*e.g.* Peng et al. 2000; Samarawickrama and Sabatini 2007) present an interesting and effective integration of vergence with saccadic movements and vestibulo-ocular reflexes. In Bana and Lee (2007) a saliency map is built on stereoscopic cues. The difference in the image position of the segmented features, interpreted as disparity, is used to correct the vergence position. Also, the appropriate gaze control in 3D can be computed exploiting object recognition (Antonelli et al. 2014) or saliency (Rea et al. 2014) on the stereoscopic pair.

The second category exploits a correlation index, as the sum of the squared distances or the normalized cross-correlation between left and right image portions, to compute an estimate of the disparity map (Abbott and Ahuja 1988; Ching et al. 1995; Yamato 1999; Marfil et al. 2003; Choi et al. 2003; Kyriakoulis et al. 2010; Rea et al. 2014). The disparity can be thus used to drive the eye position towards the correct vergence. As pointed out by Bernardino and Santos-Victor (1996), the sum of squared distances is very sensitive to luminance changes, making the use of this approach problematic in real world conditions. Thus, the normalized cross correlation is usually used. This method is commonly adopted and applied in the log-polar domain, by exploiting the cortical magnification to provide a better vergence (Bernardino and Santos-Victor 1996; Capurro et al. 1997; Bernardino and Santos-Victor 1998; Manzotti et al. 2001; Zhang and Tay 2011). Avoiding a direct computation of the disparity map, the approach is based on a global correlation index that is maximum when the two perspectives coincide under the log-polar transformation. Since this index is proportional to the absolute value of the disparity, it gives no cue on whether it is crossed or uncrossed, *i.e.* if a divergence or a convergence movement is required. The correct direction is obtained by performing an initial exploratory movement and inverting or maintaining the movement, according to the correlation index. The third category comprises the approaches that grounds on stereo image differences in the Fourier domain. Whereas few methods are based on cepstral filtering (Olson and Coombs 1991; Taylor et al. 1994), or fast Fourier transform (Marefat et al. 1997), a more appealing and widely used approach exploits a local approximation of the Fourier Shift Theorem. According to such approximation, the disparity $\delta(\mathbf{x})$ is estimated as the 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). The resulting disparity around the target as well as its temporal flux (Theimer and Mallot 1994; Daniilidis et al. 1996; Kim et al. 2000) is eventually used to adjust the vergence angle. Despite the stability of the phase difference signal, the drawback of this approach derives from the fact that the disparity information is reliable within a limited range only (Fleet and Jepson 1993). Moreover, the vergence issue is commonly thought as a 1D problem related to the horizontal disparity only, and only few authors have taken into account its 2D nature (Theimer and Mallot 1994; Daniilidis et al. 1996; Chumerin et al. 2010; Gibaldi et al. 2011; Qu and Shi 2011; Gibaldi et al. 2012), and the effect of the vertical disparity on the vergence control (Rambold and Miles 2008; Gibaldi et al. 2010).

As the available computational power increases, it has became feasible to rely of a neuromimetic approach to build on populations of disparity detectors whose overall responses can be directly used to extract effective vergence servos, without requiring any intermediate step for reconstructing the three-dimensional layout of the scene (Tsang et al. 2008; Shimonomura and Yagi 2010). A further development, which is becoming popular in the last years, is to exploit the population approach as a substrate for learning algorithms that are able either to gather effective vergence servos from a predetermined population of neurons (Franz and Triesch 2007; Wang and Shi 2010; Chumerin et al. 2010; Wang and Shi 2011; Gibaldi et al. 2013, 2015, ?), or to concurrently learn an efficient coding of disparity representation and how to exploit it for vergence movements of active stereo heads (Sun and Shi 2011; Zhao et al. 2012; Lonini et al. 2013).

In this work, we adopted the phase-difference approach, based on a population of *oriented* disparity detectors, because of its robustness and its capability to cope with vertical disparity components, which are desirable features for a vergence control system working in real-world conditions.

4 Bio-Inspired Vergence Control

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 both *horizontal* and *vertical* disparities are primary cues to drive vergence movements (Cumming and Parker 1997; Masson et al. 1997; Takemura et al. 2001). Therefore, as anticipated above, 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 $\boldsymbol{\delta}(\mathbf{x}) = [\delta_{\mathrm{H}}(\mathbf{x}), \delta_{\mathrm{V}}(\mathbf{x})]^T$ between the left and right images can be related to a phase shift $\mathbf{k}_0^T \boldsymbol{\delta}(\mathbf{x})$ in the local spectrum of the bandpassed image pairs, where $\mathbf{k}_0$ is the peak spatial frequency of the band-pass spatial filter. 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 local image structure is intrinsically 1D, with a dominant orientation θ_s (let us think to an oriented edge or to an oriented grating with frequency vector $\mathbf{k}_s = (k_s \cos\theta_s, k_s \sin\theta_s)^T$, as extreme cases), the aperture problem (Morgan and Castet 1997) restricts detectable disparity to the direction orthogonal to the edge (*i.e.*, to the direction of the dominant frequency vector $\mathbf{k}_s$). 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 frequency interval and it is gathered by the bandwidth of a single activated orientation channel. This is a realistic assumption for a relatively large number of orientation channels.

When the image structure is intrinsically 2D (let us think to 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 band-pass 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 along the direction of the filter's frequency.

By considering a complete set of oriented filters, we can derive the projected disparities in the directions of all the frequency components of a multi-channel band-pass representation, and obtain the full disparity vector by intersection of constraints (Theimer and Mallot 1994), thus solving the aperture problem. Without measurement errors, the full disparity vector $\boldsymbol{\delta}(\mathbf{x})$ can be recovered from at least two projections $\delta_\theta(\mathbf{x})$, which are not linearly dependent. Taking into account phase difference measurement errors, the redundancy of more than two projections can be used to minimize the mean square error for $\boldsymbol{\delta}(\mathbf{x})$:

$$\boldsymbol{\delta}(\mathbf{x}) = \underset{\boldsymbol{\delta}(\mathbf{x})}{\operatorname{argmin}} \left\{ \sum_\theta c_\theta(\mathbf{x}) \left(\delta_\theta(\mathbf{x}) - \frac{\mathbf{k}_0^T}{k_0} \boldsymbol{\delta}(\mathbf{x}) \right)^2 \right\} \tag{7}$$

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.

However, relying vergence control upon the explicit solution of the disparity aperture problem restricts the efficacy of the control within the model's detectability of the magnitude of the 2D disparity vector, which ultimately depends on the size of the filter used. The operative drawback of this approach is to limit the capability of the vergence control within the range in which the system is already able of producing a correct perception of the scene, and thus when, paradoxically, the vergence movements are not crucial. Differently, we have already evidenced (Gibaldi et al. 2010) that one can achieve a more flexible and efficient control of vergence without explicitly computing the disparity map, but relying upon a distributed representation of disparity, so to have a system able to cope with larger disparities and to achieve stable fixations. In the following, for the sake of completeness, we will summarize the grounding model to focus then on the advantage of the approach for the joint control of horizontal and vertical vergence.

4.1 Population Coding of Disparity

At any given scale, and for each spatial orientation channel θ, the phase-based disparity estimation approach presented in the previous paragraphs implies *explicit* measurements of the local phase difference in the image pairs, from which we obtain the *direct* measure of the binocular disparity component δ_θ. Differently, we can consider a distributed approach in which the binocular disparity 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, simple and complex cells in the primary visual cortex (V1), implement early mechanisms for stereopsis (DeAngelis et al. 1993). In the model, each simple cell neural response, for a given pixel position $\mathbf{x}_0$, is obtained as the squared output of the scalar product between the binocular receptive field (h^L, h^R) and the stereo image pair (I^L, I^R):

$$\begin{aligned} &r_s(\mathbf{x}_0; \theta, \Delta\psi) \\ &\quad = \left(< h^L(\mathbf{x} - \mathbf{x}_0), I^L(\mathbf{x}) > + < h^R(\mathbf{x} - \mathbf{x}_0), I^R(\mathbf{x}) > \right)^2 \end{aligned} \tag{8}$$

with h^L and h^R the real-valued Gabor receptive fields oriented along θ, and defined by:

$$\begin{aligned} h^L(\mathbf{x}) &\triangleq h^L(\mathbf{x}; \theta, \psi^L) = A \exp\left(-\frac{\mathbf{x}^T\mathbf{x}}{2\sigma^2} \right) \cos(\mathbf{k}_0^T\mathbf{x} + \psi^L) \\ h^R(\mathbf{x}) &\triangleq h^R(\mathbf{x}; \theta, \psi^R) = A \exp\left(-\frac{\mathbf{x}^T\mathbf{x}}{2\sigma^2} \right) \cos(\mathbf{k}_0^T\mathbf{x} + \psi^R) \end{aligned} \tag{9}$$

where $\mathbf{k}_0 = (k_0 \cos\theta, k_0 \sin\theta)^T$ is the oriented spatial frequency vector, k_0 is the radial peak frequency, A is a proper normalization constant, σ relates to the spatial extension of the receptive fields, ψ^L and ψ^R are the phases that characterize the binocular RF profile; $\Delta\psi = \psi^L - \psi^R$. The peak radial frequency k_0 and the width σ of the Gaussian envelope are linked by the constant relative bandwidth factor β (in octave) as:

$$\sigma = \frac{1}{k_0}\left(\frac{2^\beta + 1}{2^\beta - 1}\right). \tag{10}$$

Typically, $\beta \in [0.8, 1.2]$ and in our model we fixed $\beta = 1$.

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 are defined as the sum of the responses of a pair of simple cells (r_s and $r_{s,q}$) in quadrature relationship:

$$r_c(\mathbf{x}_0) = r_s(\mathbf{x}_0) + r_{s,q}(\mathbf{x}_0) = r_c(\mathbf{x}_0; \delta_\mathrm{H}, \delta_\mathrm{V}). \tag{11}$$

The resulting complex cells show a tuning to a 2D disparity vector $\boldsymbol{\delta} = (\delta_\mathrm{H}, \delta_\mathrm{V})^T$ oriented along the direction orthogonal to θ; the interocular phase difference $\Delta\psi$ defines the preferred disparity along that direction, to which the complex cell is selective, that is $\delta^\theta_{pref} = \lfloor \Delta\psi \rfloor_{2\pi}/k_0$, where $\lfloor\cdot\rfloor_{2\pi}$ denotes the principal value of its argument. Due to the wrap-around problem of the phase outside the range $(-\pi, \pi]$, 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 receptive fields: $\pm\Delta = \delta^\theta_{pref}|_{\Delta\psi=\pm\pi} = \pm\pi/k_0$.

On this basis, a distributed representation of disparity can be obtained by a population of oriented binocular energy detectors (Qian 1994; Fleet et al. 1996) centered in each image pixel and characterized by a multichannel spatial frequency vector $\mathbf{k}_{0,i} = (k_0 \cos\theta_i, k_0 \sin\theta_i)^T$ with $i = 1, \ldots, N$. To have a full coverage of the 2D disparity, as required by the vergent geometry, we equally spaced both θ and $\Delta\psi$ in $[0, \pi) : \theta_i = i\pi/N$, $i = 0 \ldots 7$, $N = 8$, and $\Delta\psi_j = 2j\pi/M - \pi$, $j = 0, \ldots 7$. $M = 8$.

4.2 Horizontal Vergence Control

On the basis of such a representation, an effective control of horizontal vergence movement can be achieved by a local weighting of the complex cell responses. Since the meaningful information for vergence comes from the perifoveal part of the image only, we used the response from a spatial neighborhood around the image center ($\mathbf{x} = 0$). In case of an anticipative vergence correction before a saccade, we will consider $\mathbf{x}$ as the center of the image of the next-to-be-fixated object. The spatial neighborhood Ω is defined by a Gaussian profile centered in $\mathbf{x}$ and with a standard deviation of 1.5°. Accordingly, the horizontal vergence control v_H can be expressed as:

$$v_\mathrm{H} = \sum_{\mathbf{x}\in\Omega}\sum_{i=1}^{N}\sum_{j=1}^{M} w_\mathrm{H}^{ij} r_c^{ij}(\mathbf{x}). \tag{12}$$

For the sake of compactness, in the subsequent formulas we will replace the spatial average $\sum_{\mathbf{x}\in\Omega} r_c^{ij}(\mathbf{x})$ with $\bar{r}_c^{ij}$.

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_H^{ij} are obtained by the minimization of a functional that considers both the features:

$$w_\mathrm{H}^{ij} = \underset{w_\mathrm{H}^{ij}}{\mathrm{argmin}} \left\{ \left\| \sum_{i=1}^{N}\sum_{j=1}^{M} w_\mathrm{H}^{ij} \bar{r}_c^{ij}(\delta_\mathrm{H}, 0) - v_\mathrm{H} \right\|^2 + \lambda \left\| \sum_{i=1}^{N}\sum_{j=1}^{M} (w_\mathrm{H}^{ij} - 1)\bar{r}_c^{ij}(0, \delta_\mathrm{V}) \right\|^2 \right\} \tag{13}$$

where $r_c^{ij}(\delta_\mathrm{H}, 0)$ and $r_c^{ij}(0, \delta_\mathrm{V})$ are the disparity-tuning curves" of the population of complex cells, v_H is the desired disparity-vergence signal to δ_H, and $\lambda > 0$ balances the relevance of the first term (sensitivity to δ_H) over the second (insensitivity to δ_V).

The insensitivity to δ_V is an important feature that allows the control to work in the case of a disparity pattern with significant vertical components, *e.g.* as the ones occurring when fixating a slanted surface or when the gaze direction is oblique. Working with a real robot head, we have also to cope with mechanical errors. Misalignment of the optical sensors or of the optical axes, produce unpredictable values of δ_V in the stereo images that compromise the effectiveness of the control. Thus, in designing the weight vector w_H^{ij}, the factor λ can be used to modulate vergence control with a stronger or weaker insensitivity to vertical disparity. The resulting system yields a vergence control that is effective in a range of horizontal disparities about three times the theoretical one, *i.e.* $[-3\Delta, 3\Delta]$, and is unaffected by the vertical component of disparity, in a range of $\pm\Delta$ (see Fig. 2, left).

4.3 Combined Horizontal and Vertical Vergence Control

The horizontal vergence control is effective for robot head designed with a Tilt-Pan geometry, *i.e.* when the vertical alignment of the optical axes cannot be controlled, but it is granted by the geometry of the system (see Sec. 2). Considering a Pan-Tilt head that explores the peripersonal space with saccadic movements, the optical axes are free to move with four degrees of freedom, so a combined horizontal and

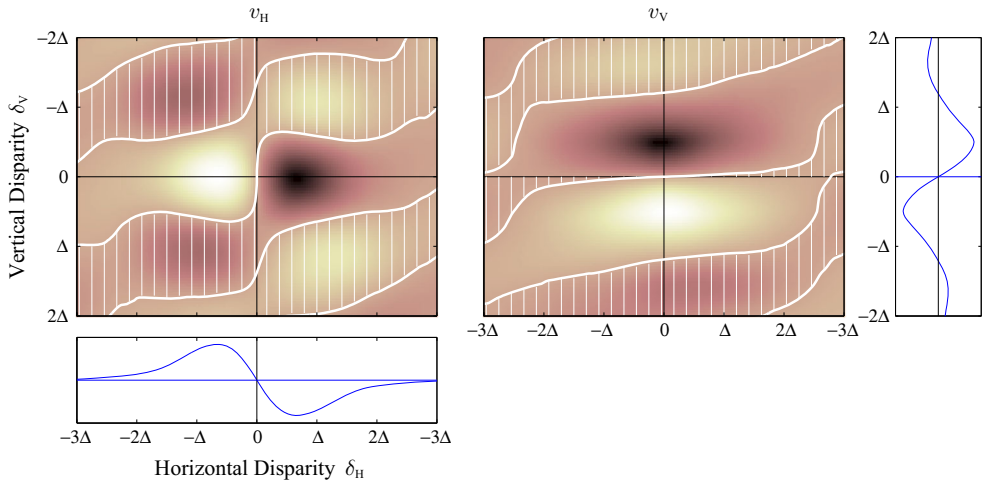


Fig. 2 Response surface to vector disparity in a range of $\delta_H \in [-3\Delta, 3\Delta]$, $\delta_V \in [-2\Delta, 2\Delta]$ of the *horizontal* (v_H) and *vertical* (v_V) vergence control. The *white contour lines* indicate where the control changes sign, and the *hatched areas* over the surface plot indicates where the control is not effective. The control is better characterized by the horizontal and vertical cross sections

vergence control is necessary to provide a proper vergence posture with the fixation point lying on the object of interest.

Since the population of disparity detectors used to derive the control of vergence is tuned to oriented disparity (Chessa et al. 2009; Gibaldi et al. 2010), we can obtain the proper control for vertical vergence from the same resources used for the horizontal one. From a computational point of view, the weight vector w_V^{ij} can be straightforwardly obtained by inverting the terms $r_c^{ij}(\delta_H)$ and $r_c^{ij}(\delta_V)$ in Eq. 13, and by designing a desired servo to vertical disparity v_V. Relying on the disparity statistics reported in Liu et al. (2008), the desired profile v_V is designed alike the profile for horizontal vergence, but with a smaller working range. Accordingly, whereas the horizontal vergence requires a control that is effective for a wide range of δ_H and robust to (*i.e.* as much as possible independent of) a relatively small range of δ_V, the control for vertical vergence requires complementary characteristics. The resulting weight w_V^{ij} are able to yield a control that is insensitive to horizontal disparity and is modulated by the vertical one, providing the desired behaviour (see Fig. 2, right).

In this way, it is possible to compute two separated control signals v_H and v_H for horizontal and vertical vergence, that are able to nullify the δ_H and δ_H in a foveal region, respectively. From these two scalar quantities it is possible to define a 3D vergence control which is applied in closed-loop to drive the vergence angle:

$$\begin{bmatrix} \dot{v}_H \\ \dot{v}_V \end{bmatrix} = \begin{bmatrix} \alpha_H & 0 \\ 0 & \alpha_V \end{bmatrix} \begin{bmatrix} v_H(\delta_H, \delta_V) \\ v_V(\delta_H, \delta_V) \end{bmatrix} \tag{14}$$

where α_H and α_V are the horizontal and vertical constant gains, respectively. On this basis, horizontal and vertical vergence angle changes are proportional to v_H and v_V signals, but it is just these signals that model the complex coupling between the horizontal and vertical disparity.

4.4 Stability Against Illumination Changes and Interocular Differences: Normalization Circuits

The vergence control so obtained is able to provide the correct behaviour in a range that, for both horizontal and vertical disparities, is wider than the one supported by the neural architecture for the computation of a reliable disparity map. Yet, further problems have to be taken into account in real-word situations. A desired feature for a real-word system for vergence control concerns the capability of coping with changeable and unpredictable illumination conditions. The illumination may not be diffuse but coming from a single and bright source, thus providing dark shades and bright areas in the environment, the light source may move or change of intensity, the object itself may move and tilt with respect to the light source, thus drastically modifying the illumination. Moreover, significant differences might be present between the left and right images, due to imprecision of the two optics or different sensitivity of the sensors, which eventually affect binocular energy approaches (Ogale et al. 2005). The robustness of the control against these issues has been obtained by implementing a binocular and a monocular normalization stage.

Binocular Normalization Regarding the changeable illumination of the environment, it is mandatory to consider that the energy model provides a quadratic dependence of the complex cell responses on the energy of the binocular image. In fact, from the Fourier transform of the monocular images $\tilde{I}^L(\omega)$ and $\tilde{I}^R(\omega)$, assuming that, for the sake of simplicity,

locally $\tilde{I}^L \approx \tilde{I}^R \approx \tilde{I}$, the response of a complex cell (see Eq. 11) becomes:

$$r_c(\mathbf{x}) \approx \frac{16\pi^4 \left|\tilde{I}\right|^2}{\sigma^4} \left(1 + e^{-\frac{|\delta|^2}{\sigma^2}} + 2e^{-\frac{|\delta|^2}{2\sigma^2}} \cos(\mathbf{k}_0^T \boldsymbol{\delta} - \Delta\psi)\right). \tag{15}$$

According to Eq. 15, the module of the Fourier transform $|\tilde{I}|^2$, acts as a multiplicative gain on the vergence control. Consequently, if the images have a low power spectrum within the receptive field bandwidths, it results in a slowdown of the control, whereas a high value of $|\tilde{I}|^2$ produces overshoots and oscillations of the fixation point around the correct position in depth. In order to remove the dependence of the vergence control signal on the energy of the image, we included a divisive normalization stage. Such an extension of the binocular energy model was introduced to explain the response saturation to interocular contrast of the complex cell response (Fleet et al. 1996), but yields interesting effects on the amplitude of the population response to natural binocular images. Accordingly, the response of each complex cell is divided by a normalization factor E_{bin}, obtained by pooling the activity of the complex cells over all the phases and orientations:

$$\begin{aligned} E_{bin}(\mathbf{x}) &= \int_0^{\pi} \int_{-\pi}^{\pi} r_c(\mathbf{x}; \theta, \Delta\psi)\, d\Delta\psi\, d\theta \\ &= \frac{32\pi^5}{\sigma^4} \left(1 + e^{-\frac{|\delta|^2}{\sigma^2}}\right) \left|\tilde{I}\right|^2. \end{aligned} \tag{16}$$

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

$$\begin{aligned} r_{c,n}(\mathbf{x}) &= r_c(\mathbf{x}) / E_{bin}(\mathbf{x}) \\ &= 1/2\pi \left(1 + \cos(\Delta\psi - \mathbf{k_0^T}\boldsymbol{\delta}) \operatorname{sech}\left(\frac{|\boldsymbol{\delta}|^2}{2\sigma^2}\right)\right). \end{aligned} \tag{17}$$

From an operative point of view, since locally $\tilde{I}^L \approx \tilde{I}^R$, the normalization factor is computed for each retinal location $\mathbf{x}$ as a summation of $E_{bin}(\mathbf{x})$ over a neighborhood, and weighted by a Gaussian function centered in $(\mathbf{x})$, and defined by the same variance σ^2 of the receptive fields.

Monocular Normalization The classical binocular energy model has been extended to account for the invariance of the V1 complex cells responses to the interocular contrast of the input images (Fleet et al. 1996). This extension has an interesting functional effect for vergence control. Similarly to binocular normalization, a monocular normalization factor $E_{mon}(\mathbf{x})$ can be obtained by pooling the activity of the monocular simple cells over all the phases and orientations, defined by Eq. 8, and used as the normalization term. Such an approach allows us to normalize the response of the monocular simple cells in each camera, thus removing the dependence of the complex cell response on the interocular contrast differences.

Thereby, the resulting vergence control, being derived from a linear summation of the responses of the disparity detectors, results to be not affected by the image luminance and interocular contrast differences. In case of real images, the divisive normalization stages render the vergence control robust under different and extreme working conditions.

4.5 Integrated Vergence and Version Control

In biological active vision, during the visual exploration of the 3D environment, an accurate eye posture on the object of interest is usually obtained by saccades followed by a vergence refinement (Enright 1998; Rea et al. 2014; Bajcsy et al. 2016). From this perspective, we implemented a control strategy composed of a sequence of two movements. First, an open-loop version movement in 3D is performed if the target position is far from the fovea, in order to bring the binocular gaze direction towards the object of interest. Second, a small correction is performed in cascade, so to refine the vergence posture and to reduce the disparity in the image area corresponding to the object of interest (*e.g.* see Fig. 9). Whereas different methodologies can be used to define the target of the binocular fixation on the left and right image planes (Samarawickrama and Sabatini 2007; Ruesch et al. 2008; Mishra et al. 2009; Rea et al. 2014; Antonelli et al. 2014; Beuth and Hamker 2015), for the sake of simplicity we used a color segmentation based on the Water Shed algorithm (OpenCV implementation Bradski and Kaehler 2008). The algorithm computes a binary mask corresponding to a selected color, *i.e.* the target object.

In order to compare the effectiveness of the proposed methodology with other approaches, we implemented three different control strategies.

INTEGRATED SACCADE- VERGENCE (ISV) The object of interest is first segmented on the binocular image. The first version movement is computed by integrating the version control computed by the mean target centroids on the left ($\mathbf{x_L}$) and right ($\mathbf{x_R}$) image plane (Eq. 5), with the disparity-vergence control (Eq. 14) computed in correspondence of the target area before the version movement. After the version movement, the object of interest, which is now in a foveal area, is again segmented on the binocular image. The version movement is thus followed by a vergence refinement, performed by the closed-loop disparity-vergence control computed on the target area by Eq. 14.

POST- SACCADIC VERGENCE (PSV) - After segmenting the object of interest on the binocular image, the first ver-

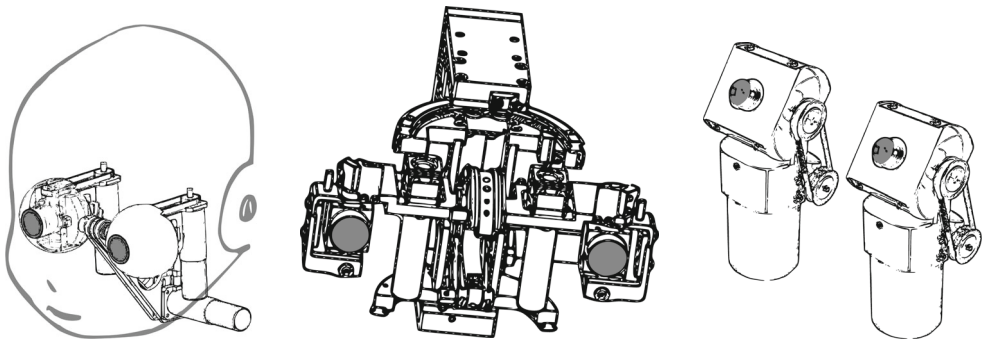


Fig. 3 Representation of the robotic platforms used.

sion movement is computed by the target centroids ($\mathbf{x_L}$,$\mathbf{x_R}$), according to Eqs. 5, 6. Once the version movement is completed, the image area corresponding to the target object is computed, and the vergence refinement is obtained by a closed-loop disparity-vergence movement (Eq. 14), as for the ISV strategy.

DOUBLE SACCADE (DS) A simple but effective approach is to perform two consecutive version movements towards the binocular target. The centroids of the segmented object of interest ($\mathbf{x_L}$,$\mathbf{x_R}$) are used according to Eqs. 5, 6, as for the PSV strategy. Considering that binocular gaze direction is correctly directed on the object of interest already after the first saccade, a second saccade is performed, almost exclusively providing a refinement of the vergence posture.

The actual Pan and Tilt motor controls are applied on the robot head on the basis of the equations reported in the Appendix, depending on the kinematics of the platform used, that is Eq. 18 for the Tilt-Pan system (iCub and Koala) and Eq. 19 for the Pan-Tilt system.

A simple implementation (C/C++ with OpenCV) of the integrated saccade-vergence control, released for research use only, can be found at: http://www.pspc.unige.it/Code/index.html.

5 Results

5.1 Implementation on Robot Platforms

In order to make the system work in real-time, the algorithm was implemented in `C/C++`, using the Integrated Performance Primitives (Intel IPP), that is a multi-threaded library of functions that rely on low-level optimizations for multi-core and multi-processor computation. From this perspective, they are an optimal computational engine for image filtering and elaboration.

The algorithm has been implemented considering filters of 43×43 pixels, with eight orientations distributed in the range $(0, \pi]$ and eight phase shifts within $(-\pi, \pi]$. To reduce the computational load and achieve real time capabilities, while preserving the accuracy of the control, the stereo images were re-sized to a resolution of 160×120 pixels. The whole system runs on a standard PC with an Intel Core i7 CPU 870 @2.93 GHz, and 8GB of RAM.

With those policies, the algorithm is able to compute the vergence control at ≈40 fps, so to ensure a fast updating and thus a good stability in real-time functioning.

The vergence control signal $\mathbf{v}$ provided by the algorithm is used as a speed control for the camera movements (cf. Eq. 14) . For motor control, the effectiveness and stability of the control is ensured by a *PID* controller, whose parameters were manually tuned. The geometry of a robotic platform, might or might not allow for a direct control of the vertical alignment of the optical axes, depending on its design characteristics. Accordingly, if the vertical alignment is controllable, the v_V is used directly to modify the motor position. Conversely, if the vertical alignment of the optical axes is not actively controllable, the vertical vergence control is used to reduce the vertical disparity pedestal is reduced performing an online image rectification.

In order to demonstrate the effectiveness and the portability of the proposed control on different systems and in real word situations, we tested the algorithm with three robot stereo heads (see Fig. 3), with different optical and kinematics characteristics (see Table 1): a iCub stereo head (Beira et al. 2006), the SEARISE stereo head (Vanegas et al. 2012), and a K-Team Koala stereo head (K-Team-Corp 2010).

iCub Stereo Head The iCub robot system has been engineered to serve as a research tool for embodied cognition, visuomotor coordination, and development (Beira et al. 2006). Being designed to resemble the human head, it has the interesting feature of a baseline of 70 mm, *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

Table 1 Mechanical and camera characteristics for the robotic platforms used

	Geom.	Baseline (mm)	Cameras	Weight (Kg)	Res.	FC (mm)	FOV	FPS
iCub	Tilt-Pan	70	DragonFly	≈0.07	1024 × 768	6	80° × 60°	15
SEARISE	Tilt-Pan	320	Baumer TXG13, Fijinon TV-Z	≈0.5	1280 × 960	7.3	44° × 34°	15
Koala	Pan-Tilt	110	Logitech HD C-310	≈0.2	640 × 480	6.4	43° × 32°	30

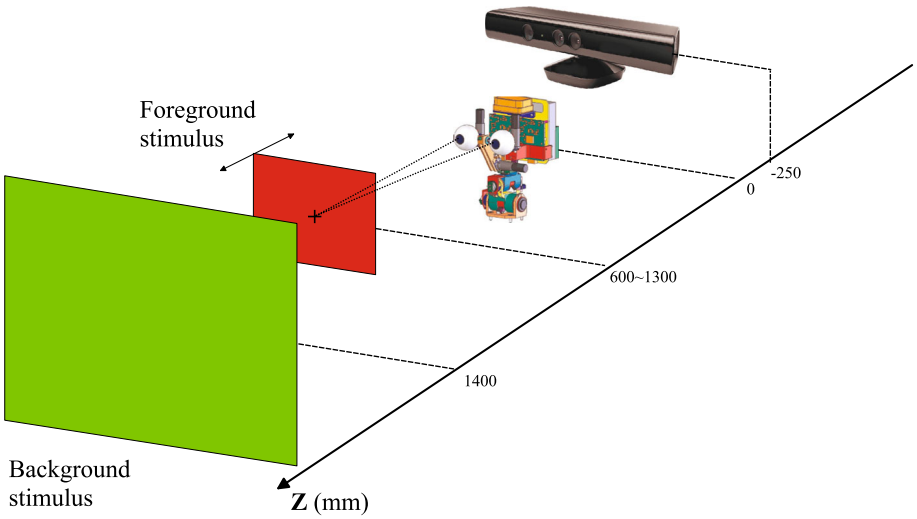


Fig. 4 Experimental setup: the background stimulus (a surface with a complex texture, *perpendicular* to the line of sight) is placed at a distance of 1400 mm with respect to the robot head, while the foreground stimulus, steady or moving in depth, layies in a range between 600 and 1300 mm. A *Kinect* sensor is placed 250 mm behind the robot in order to allow a proper measure of the distance of the stimulus with respect tot he head

two DragonFly cameras with a resolution of 1024 × 768 pixels, and a frame rate of up to 15 fps (Point-Grey-Research 2010). The mounted lenses have a focal length of 6mm, that combined with a sensor size of $1/3''$, provide a field of view of ≈ 80°.

SEARISE Stereo Head The SEARISE platform is a trinocular robotic head designed for video-surveillance purposes within the homonym EU project (Vanegas et al. 2012) with 5 degrees of freedom (a common tilt movement, and independent zoom and pan movements for left and right cameras). The system is endowed with three cameras, a central one that is kept fixed and has a wide viewing angle, and the other two that are active in a tilt-pan configuration, with variable focal length. The active cameras (Baumer TXG13 mounting Fijinon TV-Z optics) are used at a resolution of 1280 × 960, a frame rate of 15 fps, and with a fixed focal length of 7.3 mm, which with a sensor size of $1/3''$, provide a field of view of ≈44°. Since both the cameras and the optics have a large dimension, in order to prevent the optics to collide, the head has been designed with a large baseline (320 mm). Being out of the purpose of this paper, the fixed wide angle camera is not used.

Koala Stereo Head Koala is a mid-size robot designed for real-world applications by the K-Team Corporation (K-Team-Corp 2010). Differently from the majority of stereo heads, the system adopts a Pan-Tilt geometry with separate tilt for the left and right cameras, *i.e.* it has 4 degrees of freedom. The original cameras have been replaced by two low-cost commercial cameras (Logitech HD Webcam C310), used with a resolution of 640 × 480, and a frame rate of 30 fps. The focal length is 6.4 mm, which with a sensor size of of $1/4''$, provides a field of view of ≈43°.

5.2 Experimental Setup

The implemented vergence control module was qualitatively and quantitatively tested, in order to verify the interaction capabilities in a real environment. In the experimental setup, the robot head is kept fixed in a reference position, the camera position is defined by a specific azimuth and elevation, and it is free to change in terms of vergence angle only (see Fig. 4). The background stimulus is a surface, perpendicular to the line of sight and characterized by a complex texture, whereas the foreground stimulus is an object that can be steady or moving in depth. The environment is illuminated by three fluorescent lamps with a total luminous power of 6600 lm.

At each instant, the ground truth position of the stimulus with respect to the robot head was measured by means of a *Kinect* sensor, placed behind the robot head itself. The precision of the measure is ensured by a specific calibration procedure Canessa et al. (2014).

For the qualitative validation (Experiment 1), two kinds of visual stimulations were used: a stepping-in/stepping-out planar surface (step stimulus) and a waving planar surface

(sinusoid stimulus). Experiment 2 is designed to evaluate the accuracy and precision of the achieved vergence posture, whereas Experiment 3 measures the actual working range of the algorithm in the 3D space. In Experiment 4 the control was tested with different lighting conditions. Experiment 5 evaluates the effectiveness of binocular coordination in a visual exploration task.

5.3 Experiment 1: Accuracy and Precision of the Control

The experiment has the purpose of testing,for the different robot platforms, the accuracy of the algorithm in providing a correct vergence posture, regardless of the starting position. Hence, to compare the performance at different gaze directions, we repeated the experiment for the primary ($\gamma_H = 0°, \gamma_V = 0°$), a secondary ($\gamma_H = 30°, \gamma_V = 0°$) and a tertiary position ($\gamma_H = 30°, \gamma_V = 20°$). The robot is in front of a surface that is perpendicular to the line of sight, placed at a vergence distance of $\approx 8°$, that in the primary position is 500 mm for the iCub head, 2290 mm for the SEARISE head and 810 mm for the Koala. The robot starts from 250 different vergence postures, randomly chosen between $\approx 4°$ and $\approx 12°$, and the control has to move the fixation point towards the correct vergence posture.

The accuracy has been measured as the residual disparity at the end of the vergence movement (see Table 2). The iCub and SEARISE heads, relying on the same Tilt-Pan geometry, show similar behaviour. The residual δ_H has a very small mean value ($<0.05°$) and standard deviation ($\approx 0.1°$) for both the primary and secondary positions. Since the vertical vergence is not controllable, due to the common tilt of the cameras, the residual δ_V can be used as a measure of vertical alignment of the optical axes. Indeed, the results show how for the iCub head the vertical alignment of the optical axes remains almost unchanged in the secondary position, whereas for the SEARISE head there is a misalignment that introduces a vertical disparity offset. Nevertheless, the control is able to provide a correct vergence fixation, reducing the horizontal vergence error to a very small value. Since the camera tilt is common, the tertiary position has not been tested, because it would produce a vergence position identical to the secondary one. It is worth pointing out how the residual horizontal disparity is correlated with the starting vergence angle. In fact, Fig. 5 left and center, shows how a divergence movement results in a negative error, whereas a convergence movement results in a positive error. This occurs when the fixation point approaches to the desired position and the control slows down till the needed movement becomes smaller than the resolution of the motors.

Different considerations can be done for the Pan-Tilt Koala head (see Table 2, third row), since the additional degree of freedom results in a more complicated situation both in term of the disparity pattern (Hansard and Horaud 2007) and of the vergence control. In primary and secondary positions, the performance is slightly degraded with respect to the Tilt-Pan heads, for both the residual δ_H and δ_V, due to the additional degree of freedom. In tertiary position, for a Pan-Tilt head the disparity pattern has a significant vertical component, and the vertical vergence control is necessary to properly align the optical axes on the target. Nevertheless,

Table 2 Accuracy and precision of the control along a gaze direction

		$\gamma_H = 0°, \gamma_V = 0°$	$\gamma_H = 30°, \gamma_V = 0°$	$\gamma_H = 30°, \gamma_V = 20°$
iCub (with v_V)	δ_H	-0.010 ± 0.106	-0.043 ± 0.115	–
	δ_V	-0.008 ± 0.006	0.005 ± 0.006	–
iCub (without v_V)	δ_H	-0.010 ± 0.106	-0.043 ± 0.115	–
	δ_V	-0.026 ± 0.005	0.004 ± 0.005	–
SEARISE (with v_V)	δ_H	0.022 ± 0.093	0.040 ± 0.112	–
	δ_V	0.009 ± 0.007	-0.011 ± 0.010	-
SEARISE (without v_V)	δ_H	0.022 ± 0.093	0.040 ± 0.112	–
	δ_V	0.024 ± 0.013	-0.306 ± 0.025	–
Koala (with v_V)	δ_H	-0.065 ± 0.146	0.097 ± 0.215	0.057 ± 0.232
	δ_V	0.113 ± 0.130	0.113 ± 0.218	-0.103 ± 0.328
Koala (without v_V)	δ_H	0.128 ± 0.162	-0.132 ± 0.233	-0.304 ± 0.225
	δ_V	-0.148 ± 0.145	0.136 ± 0.121	0.410 ± 0.279

The table reports the mean and standard deviation of the residual horizontal (δ_H) and vertical (δ_V) disparity, measured in [deg], over 250 trials for each robot platform. The experiment is achieved for the primary ($\gamma_H = 0°, \gamma_V = 0°$), a secondary ($\gamma_H = 30°, \gamma_V = 0°$) and a tertiary ($\gamma_H = 30°, \gamma_V = 20°$) position of gaze. The experiment is repeated using or not using the vertical vergence control. The vertical vergence signal for the Koala head is used directly as motor control, whereas for the iCub and SEARISE heads it is exploited for an online image rectification of vertical disparity

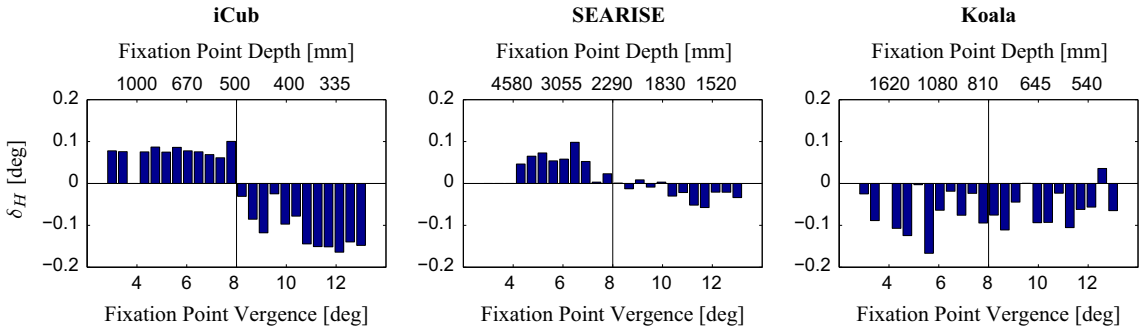


Fig. 5 Accuracy of the vergence control on the different robot platforms (from *left* to *right*: iCub, SEARISE and Koala), measured while verging on a stimulus placed at a vergence distance of $\approx 8°$. The mean residual vergence error (Y axis) is plotted against the starting one (*bottom* X axis) or depth (*top* X axis)

the control provides similar performances to the primary and secondary positions, except for the standard deviation of the δ_V that is increased to 0.33°.

In order to evidence the relevance of the vertical vergence control on the vergence posture, we repeated the tests on the Koala head with the disabled control (see Table 2, bottom row). On the one hand, in primary and secondary positions, *i.e.* where the tilt angle for the cameras are equal and the vertical control is not needed, the performance of the system is comparable to the case for which the control is used. On the other hand, in tertiary position, even if the control moves the system towards the correct posture, the residual disparity is biased by a notable negative value for the horizontal component and a positive value for the vertical one, *i.e.* leading to a wrong vergence posture.

5.4 Experiment 2: Range of Effectiveness

While exploring the environment, the robot agent needs to move the fixation point from close to far targets and *vice-versa* in the peripersonal space and farther. This experiment has the purpose to assess the range of effectiveness of the proposed control at different azimuth and elevation angles. Starting from a reference fixation distance of 8° of vergence, we moved a surface perpendicular to the gaze direction from far to close distances several times keeping the fixation point steady. The range of effectiveness is defined by the maximum (farthest) and minimum (closest) distance from the fixation point where the control is able to produce the correct vergence movement. To completely characterize the control, we tested different azimuths (γ_H from $-30°$ to 30° at steps of $-15°$) and elevations (γ_V equal to $-20°$, 0° and 20°). The experiment was conducted on the iCub and Koala heads only, because on SEARISE, due to the large baseline, we would have required to move the stimulus to more than 10m away.

As predicted by the theoretical working range of the model (Gibaldi et al. 2012), the algorithm is able to provide the correct control in the peripersonal space (see Fig. 6, left). Due to the δ_V offset, the actual range (see Fig. 6, right, reddish area) is slightly smaller than the theoretical one (blueish area). Whereas in the parameter space the operating ranges of the two heads are comparable, the different geometry of the Koala head (large baseline and smaller field of view) results in a smaller working range compared to the iCub one.

5.5 Experiment 3: Invariance to Luminance, Texture, and Interocular Differences

Experiment 4 has been designed to evidence the role of the proposed monocular and binocular normalization stages in coping with variable illumination conditions and interocular image differences.

To quantitatively assess the performance of our approach under variations of external conditions, we can derive an approximation of the disparity-vergence curves associated to the control signal. In experimental neurophysiology and psychophysics (*e.g.* see Masson et al. 1997; Takemura et al. 2001), such curves are obtained in controlled situations by measuring the triggered vergence in response to random dot stereograms. In real 3D environments (as in our set-up), the ground-truth disparity is not available and we can only exploit the rough relationship between the disparity and the depth of the stimulus to obtain *depth-vergence* curves so to characterize the behaviour of the control and to understand the effect of the normalization.

Accordingly, the control was measured by presenting a stimulus moving in depth on different trials, while the eyes are fixed at a reference depth (see Fig. 7). Two different approaches were employed to evidence the effect of the binocular and monocular normalization stages. For the binocular one, the test was repeated with different luminous powers (4400 and 6600 lm). The control without the normalization exhibits a strong dependence on the light variation (see Fig. 7a). In fact, under low illumination it results in an effective but slow control (blue line), whereas under a more intense illumination the control is fast but the excess of gain yields a high instability (red line). In Fig. 7b, the control is almost unaffected by the illumination changes because the normal-

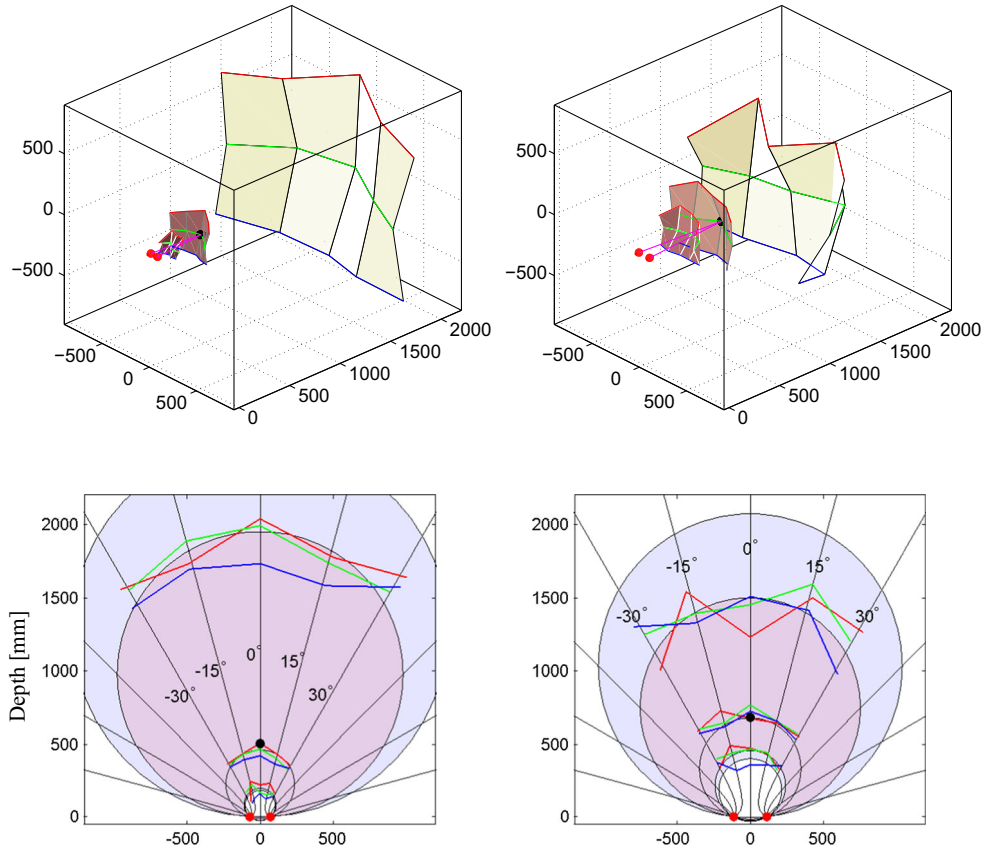


Fig. 6 The near and far vergence working bounds in the 3D space (*top*) for the iCub (*left*) and Koala (*right*) stereo heads. The robots are fixating at a vergence distance of 8° with the azimuth γ_H spanning between −30 and 30° at steps of 15°, and the elevation γ_V set at −20° (*red*), 0° (*green*) and 20° (*blue*). The top view (*bottom*) highlights the actual range (*purple small circle*) versus the range predicted by the model (*azure large circle*) (Color figure online)

ization works as a dynamic gain adaptation mechanism. An example of the effect of the binocular normalization stage in dynamic environment can be found in Gibaldi et al. (2011).

For the monocular normalization, the test was repeated by artificially varying the interocular contrast, from camera initialization parameters, in order to obtain a contrast ratio of ≈1, 1.5 and 2, between the left and right images. Without monocular normalization, the interocular difference reduces the stereo pair correlation, thus resulting in an effective but slow control (see Fig. 7c). Similarly to the binocular normalization, the monocular one restores the balance between the left and right images, thus removing the dependence of the control on interocular difference (see Fig. 7d).

The resulting depth-vergence curves can be qualitatively related to the initial vergence responses to disparity steps in humans and monkeys (Masson et al. 1997; Takemura et al. 2001) (see inset in Fig. 7a).

5.6 Experiment 4: Vergence in Dynamic Environment

This experiment demonstrates the ability of the control system to reach and to maintain a precise and stable fixation on a steady object, so as to move the fixation point in order to track an object that moves in depth. The results are shown for the iCub head (see Fig. 8). 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 along different gaze directions, as in Experiment 1.

The first case was tested with a step stimulus, in which a foreground surface is placed at a given distance from the head, along the binocular gaze direction, and removed afterward (see Fig. 4). The fixation point, starting at the depth of the foreground surface, has to move to the background surface and to stop there. The distance of the background is fixed at 1400 mm from the head, whereas the distance between the head and the foreground surface varies in a range between 600 mm and 1200 mm, thus requiring a change of the vergence angle from ≈0.45° (1200 mm) to ≈3.8° (600 mm).

The results show that the vergence control is able to properly discriminate the necessity for small movements of the fixation point, in presence of small steps (*i.e.* small disparities), so as to produce wider movements in case of large steps (*i.e.* large disparities). In all the tested steps the onset of the movement has a single frame of delay, the control completes

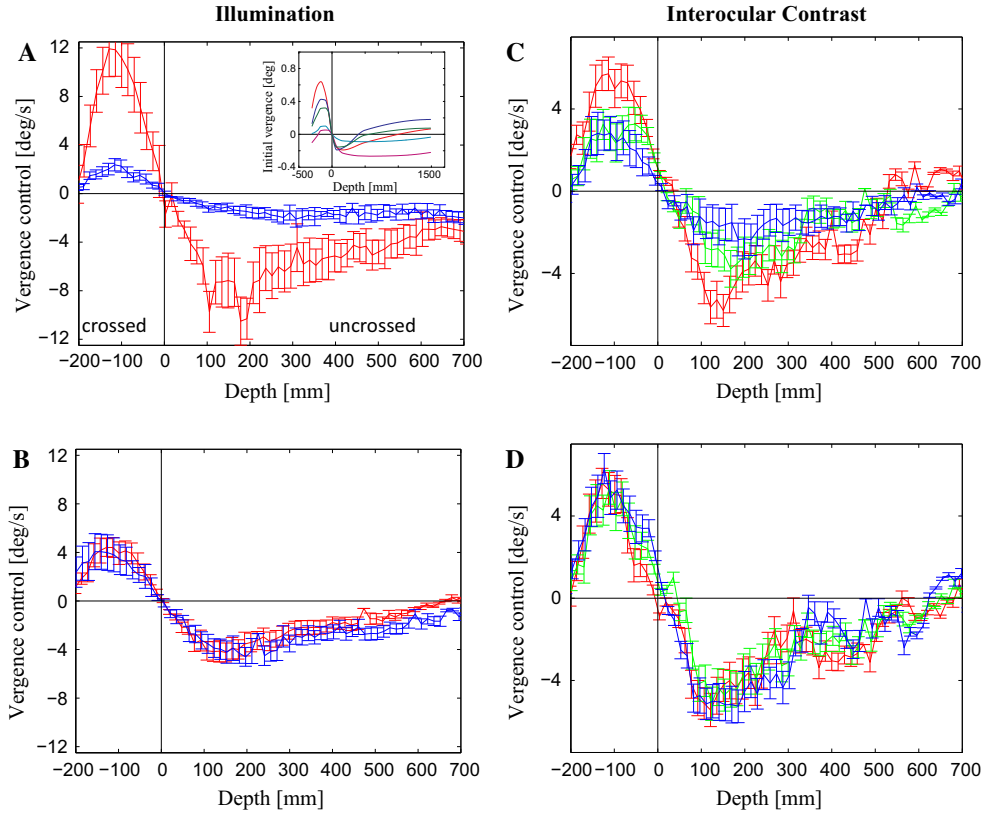


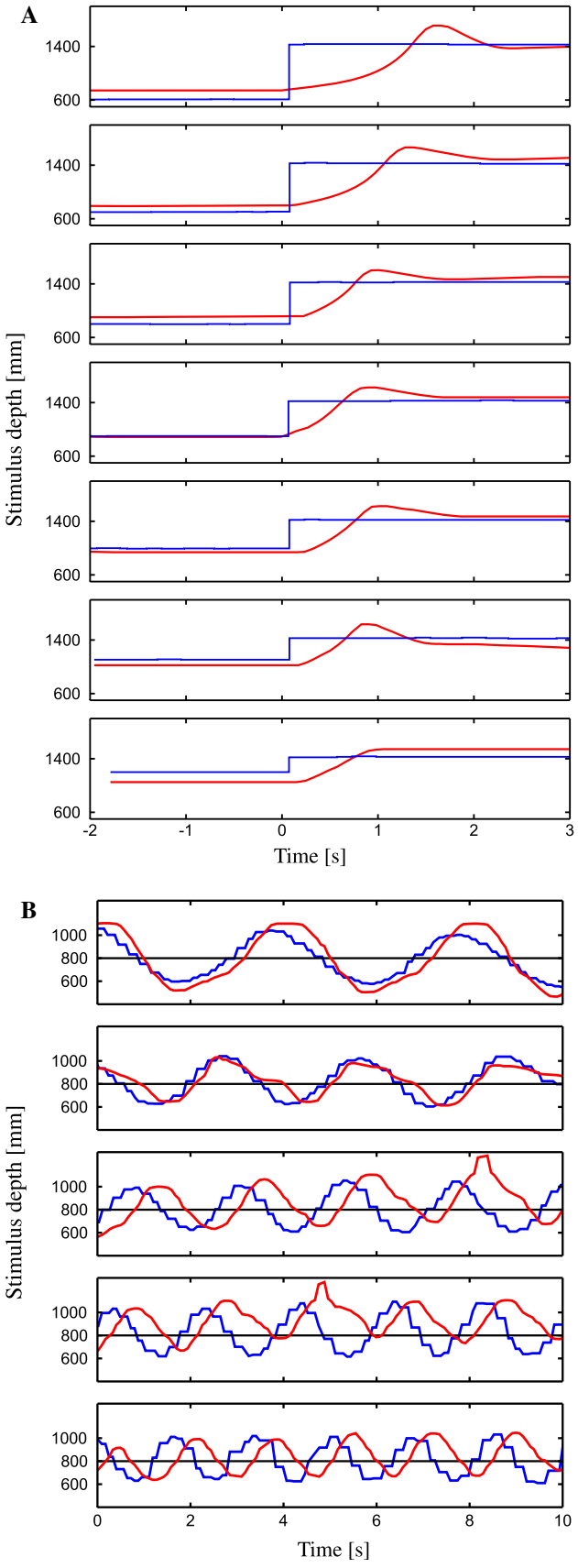
Fig. 7 Effect of the binocular and monocular energy normalization on depth-vergence curves. The curves are obtained with the initial mean vergence control, measured 60 ms after the disparity step, and are plotted against the magnitude of the step in mm. **a**, **b** Depth-vergence curves (mean and standard deviation over different trials) obtained at varying illumination condition, in case of low illumination (*blue line*) and high illumination (*red line*), without the binocular divisive normalization (**a**) or including it (**b**). **c**, **d** Depth-vergence curves (mean and standard deviation over different trials) obtained at varying interocular contrast for a contrast ratio between the *left* and *right* images of 1 (*red curve*), 1.5 (*green curve*), and 2 (*blue curve*), without the monocular divisive normalization (**c**) or including it (**d**).The obtained profiles qualitatively resemble the relation between the depth of the stimulus and the initial vergence movement, observed in monkeys, as shown by the *inset* in subfigure *a* (adapted from Takemura et al. (2001)) (Color figure online)

the majority of the vergence movements within 1 s, and the eyes are steady at the new depth within 2 s. Moreover, the effectiveness of the control is not altered by considering different gaze directions, providing the correct fixation even when the fixation is far away from the straight-ahead (see Fig. 8a).

In order to test the ability to follow in depth a moving object, a sinusoidal stimulus was used, where the foreground surface oscillates in depth about a distance of 800 mm from the head with an amplitude of 200 mm, thus moving between 600 mm and 1000 mm, *i.e.* with a change of vergence of ≈2.6° from the closest to the farthest position. The frequency of the oscillation varies from trial to trial from 30 to 70 Hz. The control yields an effective tracking in depth of the stimulus (see Fig. 8b). When the stimulus is moving slowly (top rows) the fixation point (red line) follows its depth (blue line) with a small delay, whereas a higher motion frequency (bottom rows) results in a slightly larger delay. The behavioral response to sinusoidal stimuli, closely resembles the psychophysical data (Hung et al. 1986), exhibiting both a fast reaction and a slow and smooth tracking. Moreover, as for the step stimulation, the vergence control maintains its effectiveness in following a moving stimulus, regardless of the gaze direction.

5.7 Experiment 5: Visual Exploration of the 3D Scene

The validity of the proposed methodology for enabling an active exploration of the 3D environment was assessed in a simple environment, in order to ensure a reliable object segmentation (see Fig. 9). The environment consists of five objects, labeled with numbers from 1 to 5, and placed on a white background. Each object consists of a convex plate of a specific color, with a size approximately 80 × 40 mm, and are placed at a distance ranging between 500 and 800 mm from the robot head, thus covering and horizontal field of view of ≈5.7–9.1° and a and vertical one of ≈2.8–4.6°. The objects were placed within the visual scene at different gaze



◀ **Fig. 8** Vergence trajectories achieved on the iCub stereo head with azimuth $\gamma_H = 30°$ and elevation $\gamma_V = 20°$. The stimulus depth (*blue line*) is plotted against the depth of the fixation point of the robot for a step (**a**) and sinusoid (**b**) experiment. In the step experiment, the background stimulus is at a fixed distance of 1400 mm, while the foreground one is positioned at a depth varying in the range of 600–1200 mm from the robot. In the sinusoid experiment, the stimulus starts at a distance of 800 mm and oscillates with an amplitude of about 200 mm at a frequency varying from 30 Hz (*top*) to 70 Hz (*bottom*) (Color figure online)

directions and depths with respect to the robot head, so that the robot, while fixating each one of them, is able to see the other four objects.

The experiment was performed with the three control strategies presented in Sect. 4.5, in order to obtain a performance comparison of the proposed integrated saccade-vergence control ISV with other strategies commonly used in literature, like PSV and DS. The anaglyph images are obtained superimposing the images from the left (red) and right (green and blue) cameras on different color channels, in order to provide a simple evaluation of the binocular alignment. In fact, if the system reaches the correct vergence posture, the object of interest turns from a red/green double image into a fused gray level image.

Figure 10 shows the trajectory of the binocular fixation, performed on the iCub head, point during a fixation movement between objects #1 and #3 (mean trajectory and 95 % confidence limit over 500 trial), obtained by the three control strategies. The DS and PSV trajectories (see Fig. 10b, c) show how the first saccade is accurate in moving the binocular gaze direction towards the selected object, whereas the obtained vergence is not precise, and the error can be mitigated by a second saccade (DS) or by a closed loop vergence control (PSV). Interestingly, the ISV strategy is able to provide a more correct vergence posture already with the first saccade (see Fig. 10a), and only small vergence refinements are required.

In order to provide more general conclusions, the experiment was repeated with a random succession among the 5 targets, for 500 trials. The experiment was conducted on the iCub and Koala stereo heads, but not on the SEARISE, because it would have required a too large workspace. Figure 11 shows the absolute residual horizontal disparity resulting after the first saccade and after the vergence correction, for the three control strategies, obtained on the iCub platform. Table 3 reports the mean and standard deviation of the absolute residual horizontal (δ_H) and vertical (δ_V) disparity and the eccentricity of the target centroid (ecc) with respect to the center of the image, measured in [deg].

On the iCub head, by using the DS strategy, the first saccade is able to provide the correct gaze direction and roughly the correct vergence posture. The second saccade partially reduces the residual horizontal disparity (see Fig. 11c), whereas the closed-loop disparity-vergence control used by

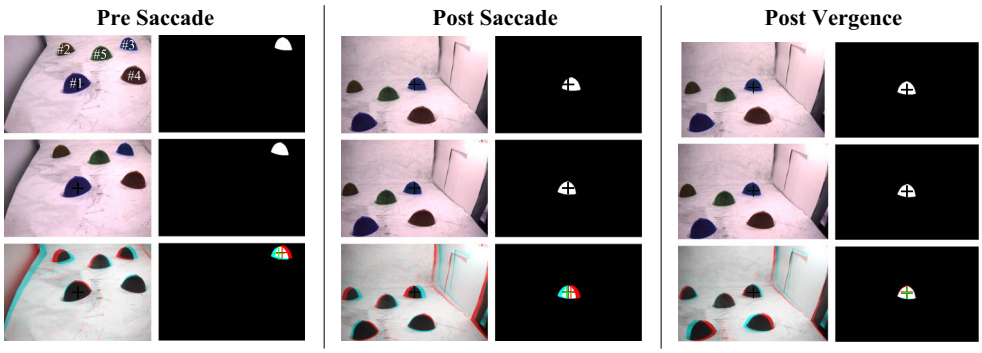


Fig. 9 Example of a complete fixational movement in 3D, from object #1 to object #3. The *figure* shows three instants of the fixational movement, corresponding to the position before the first saccade while fixating object #1 (Pre Saccade), after the first saccade towards object #3 (Post Saccade), and after the vergence correction on object #3 (Post Vergence). *Figures* in the *panels* show from *top* to *bottom* the *left* and *right color images* acquired by the robot cameras and the resulting anaglyph image, together with the masks computed while searching object #3, and the anaglyph mask image. The position of the centroids of the segmented object is marked on the anaglyph image with *red* and *green crosses*, whereas the center of the image is marked by a *black cross* (Color figure online)

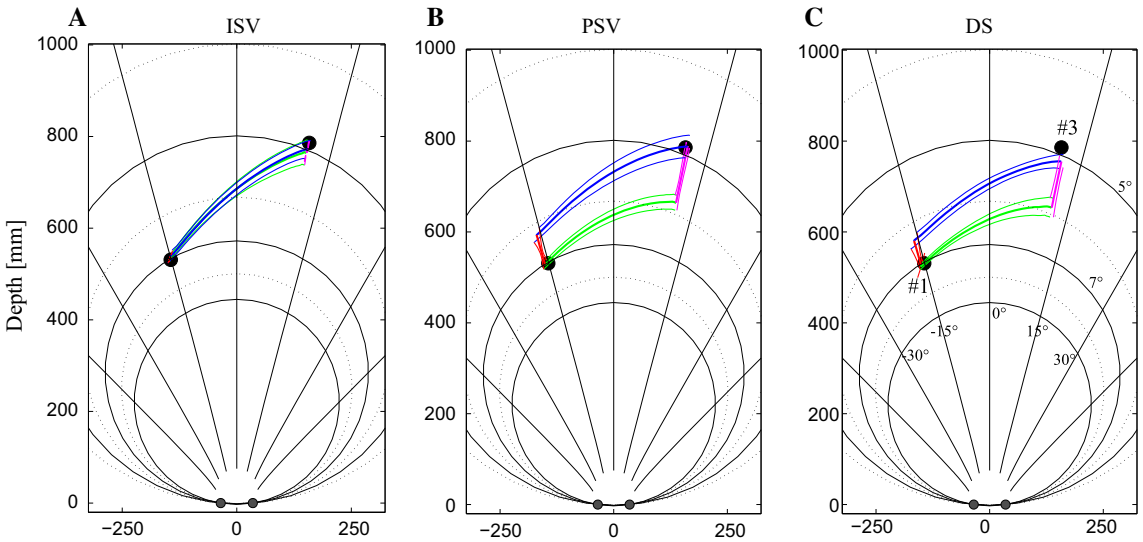


Fig. 10 Trajectory of the fixation point during a 3D version-vergence movement between two objects, performed on the iCub head. The figures show the mean trajectory (*solid thick line*) and the 95 % confidence limit (*solid thin lines*), computed from the encoder positions, over 500 saccades forth and back between the objects #1 and #3 (see Fig. 9). Object#1 is placed (according to a Helmholtz reference frame) approximately at $\gamma_H = -15°$, $\gamma_V = 10°$ and $v_H = 7°$, object #2 is at $\gamma_H = 12.5°$, $\gamma_V = -7.5°$ and $v_H = 5°$. Green color is for the *forth* saccade trajectory from object #1 to #3, whereas purple is the vergence adjustment on object #3 after saccade. *Blue color* is for the *back* saccade trajectory from object #3 to #1, while red is the vergence adjustment on object #1 after saccade. The final position is obtained using the three different methodologies for the binocular coordination presented in Sect. 5.7, *i.e.* the ISV (**a**), the PSV (**b**) and the DS (**c**) (Color figure online)

the ISV and PSV strategies provides a better performance (see Fig. 11a, b). The proposed ISV methodology is able to provide a better vergence posture already with the first saccade, preventing large vergence errors (see Fig. 11a). Moreover, since vertical vergence is not achieved by the DS strategy, using the first two strategies yields lower residual vertical disparity.

On the Koala head we observed equivalent performances, which are summarized in Table 3. The second saccade of the DS strategy reduces the residual horizontal disparity only partially, and better performances are obtained in closed-loop by the PSV and the ISV strategies. Regarding the residual vertical disparity, a slightly worse performance is obtained with respect to the iCub head. From this perspective, it is worth considering that the additional degree of freedom of the Tilt-Pan geometry an complicates the required control, especially for binocular coordination. Nevertheless, the closed-loop control used by the PSV and ISV strategies provides a better performance than DS also for the vertical disparities. Differently from what occurs on the iCub platform, with the first saccade the ISV strategy yields almost equivalent results to those obtained with the PSV and DS strategies.

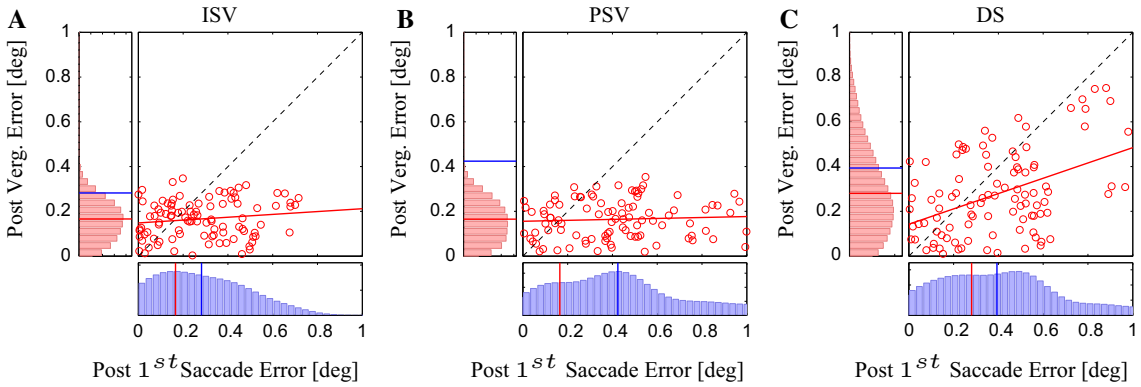


Fig. 11 Accuracy and precision of the control during 3D fixation. Scatter plots comparing the absolute residual horizontal disparity, measured in [deg], after the first saccadic movement and the one after the vergence correction, over 500 trials, for the iCub robot platforms. Each graph represents the scatter plots together with the linear regression line, for the three different methodologies of binocular camera coordination presented in Sect. 5.7, *i.e.* the ISV (**a**), the PSV (**b**) and the DS (**c**). The insets below each figure show the distribution (*blue histogram*) and mean (*blue vertical line*) of the residual disparity after the first saccadic movement, whereas the insets on the left report the distributions (*red histogram*) and mean (*red vertical line*) after the vergence correction (Color figure online)

Table 3 Accuracy and precision of the control during 3D fixation

		INT.VERG			POSTSACC.VERG.			DOUB.SACC.		
		Post 1st Sacc.	Post. Verg.	Post 1st Sacc.	Post. Verg.	Post 1st Sacc.	Post 2nd Sacc.			
iCub	δ_H	0.428 ± 0.283	0.258 ± 0.189	0.169 ± 0.084	0.435 ± 0.274	0.397 ± 0.292	0.165 ± 0.090	0.411 ± 0.312	0.375 ± 0.233	0.258 ± 0.174
	δ_V	0.071 ± 0.058	0.023 ± 0.009	0.009 ± 0.006	0.068 ± 0.056	0.022 ± 0.008	0.008 ± 0.006	0.075 ± 0.048	0.024 ± 0.008	0.022 ± 0.007
	ecc	14.52 ± 6.64	0.48 ± 0.24	0.49 ± 0.21	14.97 ± 6.86	0.47 ± 0.29	0.51 ± 0.29	14.01 ± 6.33	0.46 ± 0.24	0.36 ± 0.25
Koala	δ_H	0.987 ± 0.480	0.254 ± 0.226	0.148 ± 0.138	0.955 ± 0.521	0.243 ± 0.213	0.146 ± 0.141	0.925 ± 0.473	0.232 ± 0.198	0.221 ± 0.143
	δ_V	0.480 ± 0.226	0.061 ± 0.040	0.025 ± 0.027	0.510 ± 0.241	0.058 ± 0.044	0.024 ± 0.028	0.497 ± 0.215	0.052 ± 0.054	0.048 ± 0.026
	ecc	12.82 ± 5.47	0.35 ± 0.048	0.38 ± 0.042	12.92 ± 5.63	0.41 ± 0.41	0.43 ± 0.39	12.13 ± 5.95	0.36 ± 0.41	0.28 ± 0.35

The table reports the mean and standard deviation, computed over 500 trials, of the absolute residual horizontal (δ_H) and vertical (δ_V) disparity and the eccentricity of the target centroid (ecc) with respect to the center of the image, measured in [deg], for the iCub and Koala robot platforms. The residual disparity was measured before the first saccadic movement, once saccade has been completed, and after the correction of the fixational position. Three different methodologies were used for the binocular coordination of the cameras, the ISV , the PSV and the DS, as described in Sec. 5.7. For the Koala head, the vertical vergence signal is used directly as the motor control, whereas, for the iCub and Searise heads, it is exploited for an on-line image rectification of vertical disparity

5.8 Video

The results of the implemented experiments have been resumed in a demo video performed on the iCub stereo head (see http://www.youtube.com/watch?v=viO-SMzpHxo). A close-up of the robot cameras shows the size and the precision of the vergence movement, whereas the anaglyph image, built from the left and right images acquired by the robot cameras, shows how the fixation point is always close to the stimulus depth. The video shows in the following cases: (1) verging on a steady or moving surface perpendicular to the line of sight, (2) verging on a slanted surface with changing illumination, (3) verging on complex and deformable objects, (4) combined saccade-vergence in peripersonal space.

6 Conclusions

In this work, we presented the implementation of a bio-inspired control for the binocular coordination of camera movements, which is able to provide real-word functional operativity on different robot platforms, thus allowing for an active binocular exploration of the environment in the peripersonal space.

The open-loop control integrates information about the position of the binocular target on the image plane with disparity information, thus providing an effective vergence correction to be integrated within the binocular version movement in the 3D environment. The closed-loop control provides an accurate vergence refinement on the foveal target. The robustness and stability of the distributed phase-based representation of the binocular disparity information is at the base of the effectiveness of the vergence movements in a complex and dynamic environment. The phase information is directly used to control the vergence angle, without requiring any intermediate step for reconstructing the three-dimensional layout of the environment, or requiring camera calibration. The resulting vergence posture allows the binocular visual system to actively reduce the search space for the

vector disparity, easing the stereo correspondence process, and thus a reliable computation of depth. In static situations the fixation point is able to switch to and to remain steady on the surface of the object of interest, whereas in dynamic conditions it is able to follow in depth objects that move along different gaze directions.

In real-word situation, the control is proven to be robust to optical and geometrical imprecision, as well as unpredictable environmental changes. Divisive normalization stages allow the control to cope (1) with the changing lighting condition, (2) objects with different textures and (3) deformable shapes, as well as with (4) possible optical differences between the two cameras. The oriented disparity channels play a key role in coping with the vertical disparities that arise from the imprecision of a real Tilt-Pan robot head and in obtaining a vertical alignment of the eyes in Pan-Tilt heads. It is also worth noting that the implemented experimental setups are effective in providing a quantitative characterization of the vergence performance, and, in principles, it can be adapted for psychophysical experiment on humans and primates.

The proposed model, integrating directly early vision modules and motor control, closes the perception-action loop allowing a more immediate and efficient/effective use of the visual data. The binocular camera coordination enhances the perception of depth, allowing an artificial system to better exploit its potential, with a minimal amount of resources and coping with uncertainties of a real environment and with the inaccuracies of real systems. As a conclusion, the resulting architecture can be easily implemented on stereo heads with major differences in their kinematic and dynamic characteristics, providing an effective binocular coordination without the necessity of an accurate knowledge of the system kinematics.

Acknowledgements This work has been partially supported by the EC Project FP7-ICT-217077 EYESHOTS - Heterogeneous 3D perception across visual fragments (see http://www.eyeshots.it), and by the EU project FP7-ICT-215866 SEARISE (see http://www.searise.eu)

Appendix

The Pan movement of a camera is obtained by a horizontal rotation $\mathbf{q_{H_1}}$ around the $\mathbf{h}_Y$ axis:

$$\mathbf{q_{H_1}} = \left(0, \tan\frac{H_I}{2}, 0\right)$$

while the Tilt movement is obtained by a vertical rotation $\mathbf{q_{V_1}}$ around the $\mathbf{h}_Y$ axis:

$$\mathbf{q_{V_1}} = \left(\tan\frac{V_I}{2}, 0, 0\right).$$

The expressions for a combined Pan and Tilt movement directly depend on the geometry of the considered system.

Tilt-Pan System A Tilt-Pan stereo head is described by the Helmholtz gimbal system rotation sequence (Van den Berg 1995):

$$\begin{aligned}\mathbf{q_1}(H_I, V_I) &= \mathbf{q_{V_1}} \otimes \mathbf{q_{H_1}} \\ &= \left(\tan\frac{V_I}{2}, \tan\frac{H_I}{2}, \tan\frac{V_I}{2}\tan\frac{H_I}{2}\right)\end{aligned}$$

Given a fixation point $\mathbf{x}$ in the rotated camera reference frame, described by its versor in Helmholtz coordinates:

$$\mathbf{x} = (-\sin\mathbf{H_x}, \cos\mathbf{H_x}\sin\mathbf{V_x}, \cos\mathbf{H_x}\cos\mathbf{V_x})$$

it is possible to express $\mathbf{x}$ in the head reference frame as:

$$\mathbf{x_2} = \mathbf{q_1} \otimes \mathbf{x} \otimes -\mathbf{q_1}.$$

Given the new versor $\mathbf{x_2}$ it is possible to calculate the associate Helmholtz angle as:

$$\begin{aligned}\tan H_2 &= \frac{x_{2x}}{\sqrt[2]{x_{2y}^2 + x_{2z}^2}} \\ &= \frac{\sin H_I \cos H_\mathbf{x} \cos V_\mathbf{x} + \cos H_I \sin H_\mathbf{x}}{\sqrt[2]{(\cos H_I \cos H_\mathbf{x} \cos V_\mathbf{x} - \sin H_I \sin H_\mathbf{x})^2 + (\cos H_\mathbf{x} \sin V_\mathbf{x})^2}} \\ \tan V_2 &= \frac{x_{2y}}{x_{2z}} \\ &= \frac{\cos H_I \cos H_\mathbf{x} \cos V_\mathbf{x} - \sin H_I \sin H_\mathbf{x} + \cos H_\mathbf{x} \sin V_\mathbf{x} \cot V_I}{\cos H_I \cos H_\mathbf{x} \cos V_\mathbf{x} \cot V_I - \sin H_I \sin H_\mathbf{x} \cot V_I + \cos H_\mathbf{x} \sin V_\mathbf{x}}\end{aligned} \tag{18}$$

Pan-Tilt System A Pan-Tilt stereo head is described by the Fick gimbal system rotation sequence (Van den Berg 1995):

$$\begin{aligned}\mathbf{q_1}(H_I, V_I) &= \mathbf{q_{V_1}} \otimes \mathbf{q_{H_1}} \\ &= \left(\tan\frac{V_I}{2}, \tan\frac{H_I}{2}, -\tan\frac{V_I}{2}\tan\frac{H_I}{2}\right).\end{aligned}$$

Given a fixation point $\mathbf{x}$ in the rotated camera reference frame, described by its versor in Fick coordinate:

$$\mathbf{x} = (\cos V_\mathbf{x} \sin H_\mathbf{x}, \sin V_\mathbf{x}, \cos H_\mathbf{x} \cos V_\mathbf{x})$$

it is possible to express $\mathbf{x}$ in the head reference frame as:

$$\mathbf{x_2} = \mathbf{q_1} \otimes \mathbf{x} \otimes -\mathbf{q_1}$$

Given the new versor $\mathbf{x_2}$ it is possible to calculate the associate Fick angle as:

$$\begin{aligned}\tan H_2 &= \frac{x_{2x}}{x_{2z}} \\ &= -\tan V_1 - \frac{\sec^2 H_1}{\cos V_1 \cot H_{\mathbf{x}} + \tan H_1 + \csc H_{\mathbf{x}} \sin V_1 \tan V_{\mathbf{x}}} \\ \tan V_2 &= \frac{x_{2y}}{\sqrt[2]{x_{2x}^2 + x_{2z}^2}} \\ &= \frac{\cos V_1 \sin V_{\mathbf{x}} - \sin V_1 \cos H_{\mathbf{x}} \cos V_{\mathbf{x}}}{\sqrt[2]{(\cos V_1 \cos H_{\mathbf{x}} \cos V_{\mathbf{x}} + \sin V_1 \sin V_{\mathbf{x}})^2 + (\sin H_{\mathbf{x}} \cos V_{\mathbf{x}})^2}}.\end{aligned} \tag{19}$$

References

Abbott, A. L., Ahuja, N. (1988). Surface reconstruction by dynamic integration of focus, camera vergence, and stereo. In *Second international conference on computer vision* (pp. 532–543). IEEE.

Aloimonos, J., Weiss, I., & Bandyopadhyay, A. (1988). Active vision. *International Journal of Computer Vision*, *1*(4), 333–356.

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

Bajcsy, R., Aloimonos, Y., Tsotsos, J. K. (2016). Revisiting active perception. arXiv preprint arXiv:1603.02729.

Bana, S., & Lee, M. (2007). *Biologically motivated vergence control system based on stereo saliency map model* (pp. 513–530). Pose Estimation and Tracking: Scene Reconstruction.

Beira, R., Lopes, M., Praga, M., Santos-Victor, J., Bernardino, A., Metta, G., Becchi, F., Saltaren, R. (2006). Design of the robot-cub (iCub) head. *Proceedings 2006 IEEE international conference on robotics and automation, 2006* (pp. 94–100). ICRA 2006.

Belhaoua, A., Kohler, S., & Hirsch, E. (2010). Error evaluation in a stereovision-based 3D reconstruction system. *Journal on Image and Video Processing*, *2010*, 2.

Bernardino, A., Santos-Victor, J. (1996). Vergence control for robotic heads using log-polar images. In *Intelligent robots and systems*.

Bernardino, A., & Santos-Victor, J. (1998). Visual behaviours for binocular tracking. *Robotics and Autonomous Systems*, *25*(3), 137–146.

Beuth, F., & Hamker, F. H. (2015). A mechanistic cortical microcircuit of attention for amplification, normalization and suppression. *Vision Research*, *116*, 241–257.

Bjorkman, M., Eklundh, J.O. (2002). A real-time system for epipolar geometry and ego-motion estimation. In: *Proceedings of IEEE conference on computer vision and pattern recognition, 2000*. (Vol. 2, pp. 506–513).

Bradski, G., & Kaehler, A. (2008). *Learning OpenCV: Computer vision with the OpenCV library*. Sebastopol: O'Reilly Media, Inc.

Canessa, A., Chessa, M., Gibaldi, A., Sabatini, S. P., & Solari, F. (2014). Calibrated depth and color cameras for accurate 3d interaction in a stereoscopic augmented reality environment. *Journal of Visual Communication and Image Representation*, *25*(1), 227–237.

Capurro, C., Panerai, F., & Sandini, G. (1997). Dynamic vergence using log-polar images. *International Journal of Computer Vision*, *24*(1), 79–94.

Chessa, M., Sabatini, S. P., Solari, F. (2009). A fast joint bioinspired algorithm for optic flow and two-dimensional disparity estimation. In *Proceedings of the ICVS '09* (pp. 184–193). Berlin: Springer.

Ching, W. S., Toh, P. S., & Er, M. H. (1995). Robust vergence with concurrent identification of occlusion and specular highlights. *Computer Vision and Image Understanding*, *62*(3), 298–308.

Choi, I., Yoon, J., Lee, Y., Chien, S. (2003). Stereo system for tracking moving object using log-polar transformation and zero disparity filtering. In *Computer analysis of images and patterns* (pp. 182–189). Berlin: Springer.

Chumerin, N., Gibaldi, A., Sabatini, S. P., & Van Hulle, M. (2010). Learning eye vergence control from a distributed disparity representation. *International Journal of Neural Systems*, *20*, 267–278.

Coombs, D., & Brown, C. (1993). Real-time binocular smooth pursuit. *International Journal of Computer Vision*, *11*(2), 147–164.

Culverhouse, P., Martin, S., Talloneau, R., Rodier, T., Hughes, N., Gibbons, P., Bugmann, G. (2009). Vision processing on the bunny robot humanoid robot. In *Proceedings of the 4th workshop on humanoid soccer robots a workshop of the 2009 IEEE-RAS international conference on humanoid robots (Humanoids 2009)* (pp. 60–65). Paris.

Cumming, B., & Parker, A. (1997). Responses of primary visual cortical neurons to binocular disparity without depth perception. *Nature*, *389*, 280–283.

Daniilidis, K., Hansen, M., & Sommer, G. (1996). Real time pursuit and vergence control with an active binocular head. In G. Schmidt & F. Freyberger (Eds.), *Autonome mobile systeme 1996, Informatik Aktuell* (pp. 78–87). Berlin: Springer.

Dankers, A., Barnes, N., & Zelinsky, A. (2007). MAP ZDF segmentation and tracking using active stereo vision: Hand tracking case study. *Computer Vision and Image Understanding*, *108*(1), 74–86.

Das, S., & Ahuja, N. (1995). Performance analysis of stereo, vergence, and focus as depth cues for active vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *17*(12), 1213–1219.

DeAngelis, G., Ohzawa, I., & Freeman, R. (1993). Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. II: Linearity of temporal and spatial summation. *Journal of Neurophysiology*, *69*, 1118–1135.

Enright, J. (1998). Monocularly programmed human saccades during vergence changes? *The Journal of Physiology*, *512*(1), 235–250.

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

Fleet, D., Wagner, H., & Heeger, D. (1996). *Modelling binocular neurons in the primary visual cortex*. Cambridge: Cambridge University Press.

Franz, A., Triesch, J. (2007). *Emergence of disparity tuning during the development of vergence eye movements* (pp. 31–36).

Gibaldi, A., Canessa, A., Chessa, M., Sabatini, S.P., Solari, F. (2011). A neuromorphic control module for real-time vergence eye movements on the icub robot head. In *11th IEEE-RAS international conference on humanoid robots, 2011* (pp. 1065–1073).

Gibaldi, A., Canessa, A., Chessa, M., Solari, F., Sabatini, S. P. (2012). A neural model for coordinated control of horizontal and vertical alignment of the eyes in three-dimensional space. In *4th IEEE RAS & EMBS international conference on biomedical robotics and biomechatronics (BioRob), 2012* (pp. 955–960). IEEE.

Gibaldi, A., Canessa, A., Chessa, M., Solari, F., Sabatini, S.P. (2013). Population coding for a reward-modulated hebbian learning of vergence control. In *The 2013 international joint conference on neural networks (IJCNN)* (pp. 1–8). IEEE.

Gibaldi, A., Canessa, A., Sabatini, S. (2015). Vergence control learning through real v1 disparity tuning curves. In *2015 7th International IEEE/EMBS conference on neural engineering (NER)* (pp. 332–335). IEEE.

Gibaldi, A., Canessa, A., Solari, F., & Sabatini, S. (2015). Autonomous learning of disparity-vergence behavior through distributed coding and population reward: Basic mechanisms and real-world condi-

tioning on a robot stereo head. *Robotics and Autonomous Systems*, *71*, 23–34.
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.
Hansard, M., & Horaud, R. (2007). Patterns of binocular disparity for a fixating observer. In F. Mele & G. Ramella (Eds.), *Advances in brain, vision, and artificial intelligence*. Berlin: Springer.
Hansard, M., & Horaud, R. (2010). Cyclo-rotation models for eyes and cameras. *IEEE Transactions on Systems, Man, and Cybernetics Cybernetics*, *40*, 151–161.
Hansen, M., Sommer, G. (1996). Active depth estimation with gaze and vergence control using Gabor filters. In *Proceedings of the 13th international conference on pattern recognition* (Vol. 1, pp. 287–291). doi:10.1109/ICPR.1996.546035.
Hering, E. (1868). *The theory of binocular vision*. New York: Plenum Press.
Howard, I., & Rogers, B. (2002). *Seeing in depth*. Toronto: I. Porteous.
Hung, G., Semmlow, J., & Ciuffreda, K. (1986). A dual-mode dynamic model of the vergence eye movement system. *IEEE Transactions on Biomedical Engineering*, *36*(11), 1021–1028.
K-Team-Corp. (2010). K-team mobile robotics. http://www.k-team.com
Kim, H. J., Yoo, M. H., Lee, S. W. (2000). A control model for vergence movement on a stereo robotic head using disparity flux. In *Proceedings of 15th international conference on pattern recognition, 2000* (Vol. 4, pp. 491–494). IEEE
Knight, J., & Reid, I. (2006). Automated alignment of robotic pan-tilt camera units using vision. *International Journal of Computer Vision*, *68*(3), 219–237.
Konolige, K. (1998). Small vision systems: Hardware and implementation. In K. Konolige (Ed.), *Robotics research* (pp. 203–212). London: Springer.
Kyriakoulis, N., Gasteratos, A., & Mouroutsos, S. (2010). An adaptive fuzzy system for the control of the vergence angle on a robotic head. *Journal of Intelligent and Fuzzy Systems*, *21*(6), 385–394.
Liu, Y., Bovik, A., & Cormack, L. (2008). Disparity statistics in natural scenes. *Journal of Vision*, *8*(11), 1–14.
Lonini, L., Zhao, Y., Chandrashekhariah, P., Shi, B.E., Triesch, J. (2013). Autonomous learning of active multi-scale binocular vision. In *2013 IEEE third joint international conference on development and learning and epigenetic robotics (ICDL)* (pp. 1–6). IEEE.
Manzotti, R., Gasteratos, A., Metta, G., & Sandini, G. (2001). Disparity estimation on log-polar images and vergence control. *Computer Vision and Image Understanding*, *83*(2), 97–117.
Marefat, M., Wu, L., & Yang, C. (1997). Gaze stabilization in active vision-I. vergence error extraction. *Pattern recognition*, *30*(11), 1829–1842.
Marfil, R., Urdiales, C., Rodriguez, J., & Sandoval, F. (2003). Automatic vergence control based on hierarchical segmentation of stereo pairs. *IJIST*, *13*(4), 224–233.
Masson, G., Busettini, C., & Miles, F. (1997). Vergence eye movements in response to binocular disparity without depth perception. *Nature*, *389*, 283–286.
Minken, A., Gielen, C., & Gisbergen, J. V. (1994). An alternative three-dimensional interpretation of hering's equal-innervation law for version and vergence eye movements. *Vision Research*, *35*(1), 93–102.
Mishra, A., Aloimonos, Y., Fah, C. L. (2009). Active segmentation with fixation. In *2009 IEEE 12th international conference on computer vision* (pp. 468–475).
Monaco, J., Bovik, A., & Cormack, L. (2009). Active, foveated, uncalibrated stereovision. *International Journal of Computer Vision*, *85*(2), 192–207.
Morgan, M. J., & Castet, E. (1997). The aperture problem in stereopsis. *Vision Research*, *37*, 2737–2744.
Muhammad, W., & Spratling, M. (2015). A neural model of binocular saccade planning and vergence control. *Adaptive Behavior*, *23*(5), 265–282.
Ogale, A., Aloimonos, Y. (2005). Robust contrast invariant stereo correspondence. In *Proceedings of the 2005 IEEE international conference on robotics and automation, 2005. ICRA 2005*. (pp. 819–824). IEEE
Olson, T., & Coombs, D. (1991). Real-time vergence control for binocular robots. *International Journal of Computer Vision*, *7*(1), 67–89.
Pauwels, K., & Van Hulle, M. (2012). Head-centric disparity and epipolar geometry estimation from a population of binocular energy neurons. *International Journal of Neural Systems*, *22*(03), 1250007.
Peng, J., Srikaew, A., Wilkes, M., Kawamura, K., Peters, A. (2000). An active vision system for mobile robots. In *2000 IEEE international conference on systems, man, and cybernetics* (Vol. 2, pp. 1472–1477). IEEE
Piater, J., Grupen, R., Ramamritham, K. (1999). Learning real-time stereo vergence control. In *Intelligent control/intelligent systems and semiotics, 1999* (pp. 272–277). Cambridge.
Point-Grey-Research. (2010). Firewire cameras. http://www.ptgrey.com.
Qian., N. (1994). Computing stereo disparity and motion with known binocular cell properties. *Neural Computation*, *6*(3), 390–404.
Qu, C., Shi, B. (2011). The role of orientation diversity in binocular vergence control. In Proceedings *The 2011 international joint conference on neural networks, 2011* (pp. 2266–2272).
Rambold, H., & Miles, F. (2008). A human vergence eye movements to oblique disparity stimuli: evidence for an anisotropy favoring horizontal disparities. *Vision Research*, *48*, 2006–2019.
Rea, F., Sandini, G., Metta, G. (2014). Motor biases in visual attention for a humanoid robot. In *2014 14th IEEE-RAS international conference on humanoid robots (Humanoids)* (pp. 779–786). IEEE
Ruesch, J., Lopes, M., Bernardino, A., Hornstein, J., Santos-Victor, J., Pfeifer, R. (2008). Multimodal saliency-based bottom-up attention a framework for the humanoid robot icub. In *IEEE international conference on robotics and automation, 2008. ICRA 2008* (pp. 3962–967). IEEE.
Samarawickrama, J., Sabatini, S. (2007). Version and vergence control of a stereo camera head by fitting the movement into the Hering's law. In *Fourth Canadian conference on computer and robot vision, 2007. CRV'07* (pp. 363–370). IEEE.
Sanger, T. (1988). Stereo disparity computation using Gabor filters. *Biological Cybernetics*, *59*, 405–418.
Schwartz, E. (1977). Spatial mapping in the primate sensory projection: Analytic structure and relevance to perception. *Biological Cybernetics*, *25*, 181–194.
Shimonomura, K., Yagi, T. (2010). Neuromorphic vergence eye movement control of binocular robot vision. In *2010 IEEE international conference on robotics and biomimetics (ROBIO)* (pp. 1774–1779). IEEE.
Sun, W., Shi, B. (2011). Joint development of disparity tuning and vergence control. In *2011 IEEE international conference on development and learning (ICDL)* (pp. 1–6).
Takemura, A., Inoue, Y., Quaia, C., & Miles, F. (2001). Single-unit activity in cortical area mst associated with disparity-vergence eye movements: Evidence for population coding. *Journal of Neurophysiology*, *85*(5), 2245–2266.
Taylor, J., Olson, T., Martin, W. (1994). Accurate vergence control in complex scenes. In *1994 IEEE computer society conference on computer vision and pattern recognition, 1994. Proceedings CVPR'94* (pp. 540–545). IEEE

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., Lam, S., Meng, Y., Shi, B. (2008). Neuromorphic implementation of active gaze and vergence control. In *IEEE international symposium on circuits and systems, 2008. ISCAS 2008* (pp. 1076–1079). IEEE

Van den Berg, A. (1995). Kinematics of eye movement control. *Proceedings of the Royal Society of London B: Biological Sciences*, *260*, 191–197.

Van Rijn, L., & Van den Berg, A. (1993). Binocular eye orientation during fixations: Listing's law extended to include eye vergence. *Vision Research*, *33*(5/6), 691–708.

Vanegas, M., Chessa, M., Solari, F., & Sabatini, S. P. (2012). *Surveillance applications, machine vision—Applications and systems, chap*. InTech: Bio-inspired active vision paradigms in surveillance applications.

Wang, Y., & Shi, B. (2010). Autonomous development of vergence control driven by disparity energy neuron populations. *Neural Computation*, *22*, 730–751.

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

Yamato, J. (1999). A layered control system for stereo vision head with vergence. In *1999 IEEE international conference on systems, man, and cybernetics, 1999. IEEE SMC'99 conference proceedings* (Vol. 2, pp. 836–841). IEEE.

Zhang, X., & Phuan, A. L. T. (2009). A physical system for binocular vision through saccade generation and vergence control. *Cybernetics and Systems: An International Journal*, *40*(6), 549–568.

Zhang, X., & Tay, L. (2011). A spatial variant approach for vergence control in complex scenes. *Image and Vision Computing*, *29*(1), 64–77.

Zhao, Y., Rothkopf, C., Triesch, J., Shi, B. (2012). A unified model of the joint development of disparity selectivity and vergence control. In *IEEE 8th International Conference on Development and Learning*.