Learning and evaluation of a vergence control system inspired by Hering's law

Flavio Mutti, Cristiano Alessandro, Marco Angioletti, Andrea Bianchi and Giuseppina Gini

Abstract— **We develop a bio-inspired controller for an active stereo vision system based on the Hering's law. We extend a model already proposed in literature in two ways. Firstly we evaluate the performance of the controller, inspecting its capability to foveate a generic feature in the 3D space, and the robustness respect to the initial angular configuration of the stereo system. Secondly we introduce the redundant component of the neck. Using a classical learning method we tune the controller to adapt to the controlled system. We investigate how the redundancy is solved by the learned controller, and show that the performance increases and the controlled stereo system generates human-like trajectories.**

I. Introduction

The problem of controlling an active stereo system is critical. As the complexity of the controlled system increases, the controller equations become more complicated too. The research field has been very active in the last 15 years and methods to effectively employ active vision techniques is surveyed in [3]. Furthermore, many techniques based on biologically plausible models were recently proposed. The typical way to approach a bio-inspired controller is to investigate how the information of the bio-inspired neural network can be processed to achieve the vergence control. This work instead investigates the underling control model that is compatible with neural processing but is placed at an higher level of computation. Moreover, we analyze in a quantitative way the capability of the control system, exploring a wide area of the 3D space. Based on these assumptions we investigate a model derived from the Hering's law of equal innervation [9]. Recent work shows that the Hering model may be plausible at least for disparity-driven vergence and binocular fixation [6]. We propose a learning method to adapt the system to the controlled device, and selectively investigate the quantitative performance in terms of foveation error, and the capability of the system to foveate a 3D point starting from a generic position of the cameras. Then we extend the model introducing the redundant component of the neck that is not present in the classical controlled device.

This paper is organized as follows. In Section II we present some related work, in Section III we present the original model and the performed experiments, in Section IV we present the extended model and the experiments compared with those of the original controller, in Section V we investigate the biological plausibility of the extended model and in Section VI we derive our conclusions.

This research was partially supported by the EU project RobotDoC under 235065 from the 7th Framework Programme (Marie Curie Action ITN)

F. Mutti, G. Gini, M. Angioletti and A. Bianchi are with DEI, Politecnico di Milano, Italy, mutti@elet.polimi.it

C. Alessandro is with Department of Informatics, University of Zurich, Switzerland, alessandro@ifi.uzh.ch

II. Related works

Several approaches and methods to effectively employ active vision techniques are surveyed in [3]. The authors describe problems arising from many applications, e.g. object recognition, tracking, robotic manipulation, localization and mapping. A lot of techniques are proposed to deal with the low-level control strategies to drive the active stereo head. Several approaches propose a bioinspired neural network, based on the disparity energy model, to command vergence and version for foveating tasks.

Wang et al. [13] [14] show the autonomous development of the vergence control, maximizing neural responses through reinforcement learning. Gibaldi et al. [5] show a model that directly extracts the disparity-vergence response without an explicit calculation of the disparity. Moreover, the same authors implement the control strategy for the iCub head to foveate steady or moving object along the depth direction considering only some fixed configurations in the tilt direction [4]. Shimonomura et al. propose an hardware stereo head built with an FPGA and silicon retinas; the vergence system is able to foveate a point processing the disparity computed with the energy model [10]. Tsang et al. [12] show a gaze and vergence control system using the disparity energy model with a vergence-version control with a virtual vergence component. Qu et al. [8] propose a neural model based on the energy model introducing the orientation and scale pooling; they show how the novel features improve the learning curve. Sun et al. [11] demonstrate that the vergence command can be learned starting from a sparse coding paradigm. Other recent approaches addressing the problem of the vergence are based on more classical algorithms, either fuzzy [7] or SIFT [1].

Typically the experimental data are collected only along the depth direction; in our research, instead, we addressed the problem to produce statistics related to a wider space along the three direction in space. Moreover, we introduced the neck redundancy in order to enrich the capability of the control system.

III. Hering-based Model

In this section we introduce the bio-inspired active stereo vision system initially proposed in [9]. The fundamental equations are based on the Hering's law of equal innervation which states that the eyes move by combining the movements of vergence and version [6].

A. Control system

The system is a proportional model which needs to be trained to learn the proportional parameters. The controller is used with a 3 degrees of freedom (DOF) structure with 2 DOF for the pan command for both eyes and 1 DOF for the tilt, as in Figure 1. The fundamental equations are:

$$\dot{\theta}_{version} = K_1(x_L + x_R) \tag{1}$$

$$\dot{\theta}_{vergence} = K_2\delta \tag{2}$$

$$\dot{\theta}_{tilt} = K_3(y_L + y_R) \tag{3}$$

where x_L and x_R are the feature x-position on the left or right image plane and y_L and y_R are the feature y-position on the left or right image plane. The disparity of the projected feature is represented by δ, and $[K_1, K_2, K_3]$ are the parameters that must be estimated.

We can compute the pan and tilt angles as following:

$$\dot{\theta}_r = \dot{\theta}_{version} - \dot{\theta}_{vergence} \tag{4}$$

$$\dot{\theta}_l = \dot{\theta}_{version} + \dot{\theta}_{vergence} \tag{5}$$

$$\dot{\theta}_t = -\dot{\theta}_{tilt} \tag{6}$$

B. Setup

To be as consistent as possible with reality we use the camera model with the same calibration matrix for both eyes:

$$K = \begin{bmatrix} 200 & 0 & 320 \\ 0 & 200 & 240 \\ 0 & 0 & 1 \end{bmatrix}$$

with focal length equal to 200 pixels and with an image plane of 640×480 pixels. This calibration matrix leads to a lens angle of about 100°.

It is worth noting that we use the undistorted non-rectified matrices, taking into account that we deal with an active system and considering the consistency of the camera model.

We define the origin of the neck-frame coincident with the origin of the world frame of reference; the only movement of the neck is given by the tilt activity. The camera positions are defined at 0.2 m of distance to each other along the x-axis, and at 0.2 m along the y-axis of the world frame of reference (see Figure 1). The unity measure of the world frame of reference is meter.

To evaluate the performance of the system we use the following error measure:

$$e_{L/R} = \sqrt{x_{L/R}^2 + y_{L/R}^2} \tag{7}$$

that is the Euclidean distance computed in the image plane between the final feature position in the image plane and the centre of the image plane (in this case we have the origin of the frame of reference of the image plane exactly at the centre of the image plane itself). The subindices L/R refer to the left and right camera respectively. We choose to evaluate the error for the left and the right eye separately to understand if the foveation error vary between the two eyes.

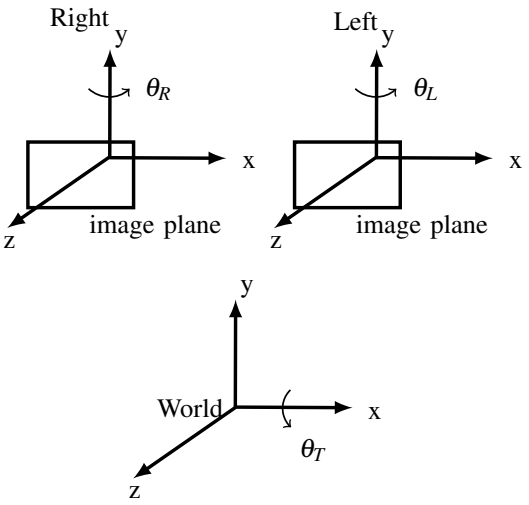

Fig. 1. Frames of reference of the active stereo system with 3 DOF. The tilt movement is executed along the x-axis of the *world* frame, and it rotates the frames of both eyes of $\theta_T[rad]$. Ideally, we define a virtual neck that performs the tilt movement.

C. Learning phase

In this section we propose a method to learn the parameters K_i that guarantee a minimum error $e_{L/R}$ for any desired 3D point to be foveated, independently from the starting position of the stereo camera. The parameters can be learned by performing the following minimisation:

$$c(X,Y,Z) = e_L^2 + e_R^2 + \sum_j |\ \dot{\theta}_l\ |_j + \sum_j |\ \dot{\theta}_r\ |_j + \sum_j |\ \dot{\theta}_t\ |_j$$

$$K = \underset{K_1,K_2,K_3}{\operatorname{argmin}} \sum_x \sum_y \sum_z c(x,y,z)$$

The Euclidean distances in the objective function are needed to evaluate the performance of the system in foveating the desired point; the sum terms are necessary to minimize the lengths of the performed trajectories (and therefore avoiding oscillations around the desired final position).

The objective function is minimised numerically using the gradient descent method; the points used as training set cover most of the view field and can be described as follows:

$$\begin{aligned} x &\in [-100, 100] \quad m \\ y &\in [-100, 100] \quad m \\ z &\in [1, 201] \quad m \end{aligned}$$

with a step of 50 m.

D. Experimental results

The gradient descent minimization of the cost function on the training set lead to the following parameters:

$$K1 = 0.3286 \quad K2 = 0.0859 \quad K3 = 0.1837$$

It is worth noting that the cost function has a lot of local minima but, in our experience, the overall performance of the system is not affected.

To test the performance of the learned control system, we conducted the following experiments:

- *Exploring the 3D space*, that investigates the capability of the active stereo system to foveate points that are not contained in the training set.
- *Testing initial position*, that investigates the capability of the system to foveate a feature in 3D space, regardless the initial joints configuration of the stereo camera. The aim is to investigate the robustness of the system to foveate a feature starting from a generic position.

1) Exploring 3D space: As a first experiment we investigate the capability of the system to foveate a huge set of features (e.g. 3D points) in the 3D space starting from a defined initial position. Based on their 3D positions, the evaluated points (testing sets) can be grouped in three cubes adjacent to the training set:

Along Z direction

$$[-100,100]\times[-100,100]\times[201,401]$$

Along Y direction

$$[-100,100]\times[100,200]\times[1,201]$$

Along X direction

$$[-200,-100]\times[-100,100]\times[1,201]$$

Each of these portions of space is discretised with a step of 10 m in each direction. We do not consider the points that are not projected in either image planes.

Figure 2 shows the errors associated to each point in the 3D space (top row), and the overall error distributions (bottom row).

The mean error associated to the testing set along the Z direction is 1.42 pixel with a variance of 0.33. This result is expected mainly because the projections of the 3D points are closer to the image centre as their distances from the image plane increase. Along the X direction the error increases as the X component increases. Since these points are close to the image planes, their projections are in the border of the images and, consequently, the task of foveating them is more challenging. However, as can be seen from the bottom pane, the errors are distributed in an acceptable error interval; i.e. [2.5;5.5] pixels with an average of 4.33 pixels and a variance of 0.488. Similar considerations can be done for the testing set along the Y direction, where qualitatively the error increases as the Y component of the 3D points increases. The mean error is 3.96 pixels with a variance equal to 0.296.

2) Testing initial position: This experiment presented in this subsection aims at understanding the robustness of the system to foveate a 3D point starting from a generic joint configuration (i.e. θ_l, θ_r and θ_t).

Vergence and version affect the panning command competitively (see Equation 4). To check whether the system is able to perform panning accurately, we evaluated the most problematic region of the 3D space. Indeed, the Z region represents an "easy" case where the points are always projected to the centre of the image, and the Y region does not affect the panning but the tilting. The testing subspace along X direction, used for the experiments, is:

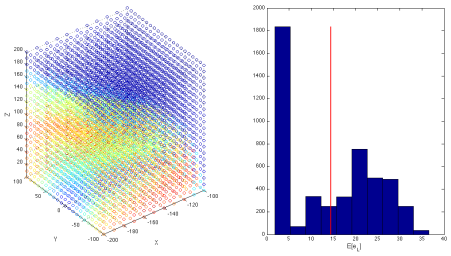

Fig. 3. *Original system*. In the left pane it is shown the mean error associated to each 3D point in the testing set. The mean error is computed considering each plausible initial joints configuration of the head; for each configuration we compute the error to foveate. In the right plane is shown the mean error distribution.

$$[-200,-100]\times[-100,100]\times[1,201]$$

discretised with a step of 10 m on each direction. We let the system foveate each of the testing points starting from each possible joint configuration in the joint space. We defined a range of values for each joint, i.e. $[-60°;60°]$ with a step of $30°$. In total we have 125 different joints configurations. Then, we compute the mean error associated to each 3D point and the results are shown in Figure 3. Qualitatively, the error increases as the Z component of the 3D points decreases (see left pane). Since these points are close to the image planes, their projections are in the border of the images and, consequently, the task of foveating them is more challenging. However, as can be seen from the right pane, the errors are distributed in an acceptable error interval; i.e. [1;35] pixels with an average of 15 pixels.

IV. Extended Model

The model presented so far takes into account only 3 DOF to foveate a generic target in the 3D space. In this section, we extend the model adding a further degree of freedom (i.e. the neck) to improve the performance of the head in the pan activity. Moreover we investigate whether, from a biological point of view, it is possible to infer some similarities between the obtained head trajectories and the stereotypical trajectories performed by primates (eventually humans).

A. Control system

In order to add the additional neck-joint, we investigated different augmented version of control system presented in Sec. III, and for each of them we evaluated the performance.

First of all, we introduced the neck component in accordance with Eq. (1)-(3):

$$\dot{\theta}_{neck} = K_4(x_L + x_R) \tag{8}$$

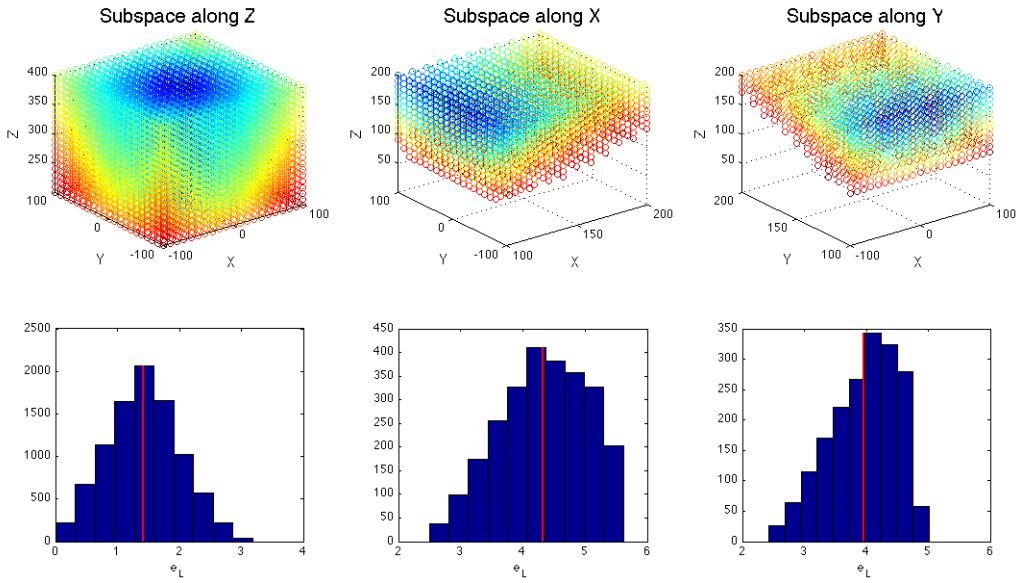

Fig. 2. Error maps computed for the left eye; we experienced very similar error values also for the right eye. Top row: testing sets with the error associated to each foveated 3D point. Bottom row: the error distribution in pixel for the testing set. The red line represents the mean of the error. As we expected the error distribution along the Z direction is lower then along the other directions.

This implies that the neck motions depend on the position of the feature in the image planes. Neck movements only consists of rotations along the Y axis and are independent from the tilting command.

B. Setup

Introducing a new degree of freedom for the neck to make the system redundant, requires to define a chain of roto-translations from the neck to the world frame of reference. The position of a 3D feature (initially defined in the world frame of reference) in the camera frame of reference can be computed as follows:

$$R_W^{L/R} = \quad R_N^{L/R}(\theta_{L/R}) \quad R_H^N(\theta_N) \quad R_W^H(\theta_T) \qquad (9)$$

where $R_W^{L/R}$ is the roto-translation between the world frame of reference and the camera frame of reference (left or right), $R_N^{L/R}(\theta_{L/R})$ is the roto-translation between the neck and the camera frame of reference, $R_H^N(\theta_N)$ is the roto-translation between the head and the neck (defined as the movement along the pan direction) and $R_W^H(\theta_T)$ is the tilting command defined as a rotation of the head frame of reference respect to the world frame. The camera model and the other parameters are defined as in Section III.

C. Neck configurations

In order to compute the angle movements for pan, tilt, and rotation, equations (1)-(3), and (8) have to be combined appropriately. We call *configurations* the different ways to obtain these angle movements.

The configurations are summarised in the table below, and reflects the following ideas:

Configuration 1	*Configuration 2*
$\dot{\theta}_r = \dot{\theta}_{version} - \dot{\theta}_{vergence} + \dot{\theta}_{neck}$ $\dot{\theta}_l = \dot{\theta}_{version} + \dot{\theta}_{vergence} + \dot{\theta}_{neck}$ $\dot{\theta}_t = -\dot{\theta}_{tilt}$ $\dot{\theta}_n = \dot{\theta}_r$	$\dot{\theta}_r = \dot{\theta}_{version} - \dot{\theta}_{vergence} + \dot{\theta}_{neck}$ $\dot{\theta}_l = \dot{\theta}_{version} + \dot{\theta}_{vergence} + \dot{\theta}_{neck}$ $\dot{\theta}_t = -\dot{\theta}_{tilt}$ $\dot{\theta}_n = \dot{\theta}_l$
Configuration 3	*Configuration 4*
$\dot{\theta}_r = \dot{\theta}_{version} - \dot{\theta}_{vergence} + \dot{\theta}_{neck}$ $\dot{\theta}_l = \dot{\theta}_{version} + \dot{\theta}_{vergence} + \dot{\theta}_{neck}$ $\dot{\theta}_t = -\dot{\theta}_{tilt}$ $\dot{\theta}_n = \dot{\theta}_{neck} - \dot{\theta}_{version}$	$\dot{\theta}_r = \dot{\theta}_{version} - \dot{\theta}_{vergence} + \dot{\theta}_{neck}$ $\dot{\theta}_l = \dot{\theta}_{version} + \dot{\theta}_{vergence} + \dot{\theta}_{neck}$ $\dot{\theta}_t = -\dot{\theta}_{tilt}$ $\dot{\theta}_n = \dot{\theta}_{neck}$
Configuration 5	*Configuration 6*
$\dot{\theta}_r = \dot{\theta}_{version} - \dot{\theta}_{vergence}$ $\dot{\theta}_l = \dot{\theta}_{version} + \dot{\theta}_{vergence}$ $\dot{\theta}_t = -\dot{\theta}_{tilt}$ $\dot{\theta}_n = \dot{\theta}_{neck}$	$\dot{\theta}_r = \dot{\theta}_{version} - \dot{\theta}_{vergence}$ $\dot{\theta}_l = \dot{\theta}_{version} + \dot{\theta}_{vergence}$ $\dot{\theta}_t = -\dot{\theta}_{tilt}$ $\dot{\theta}_n = \dot{\theta}_{neck} - \dot{\theta}_{version}$

- The eye movements (pan) could be mediated by the neck component (Configuration 1-4)
- The neck movements (pan direction) could be mediated by vergence and version (Configuration 1,2,3,6)
- The eye and the neck could be independent each other (Configuration 5)

D. Learning phase

We adapted the learning procedure that is used for the 3 DOF system (see Equation III-C) to the new 4 DOF system:

$$c(X,Y,Z) = e_L^2 + e_R^2 + \sum_j |\,\dot{\theta}_l\,|_j + \sum_j |\,\dot{\theta}_r\,|_j + \sum_j |\,\dot{\theta}_t\,|_j + \sum_j |\,\dot{\theta}_n\,|_j$$

$$K = \underset{K_1,K_2,K_3,K_4}{\operatorname{argmin}} \sum_x \sum_y \sum_z c(x,y,z)$$

The minimisation is performed with the same algorithm and on the same training set as in Sec. III.

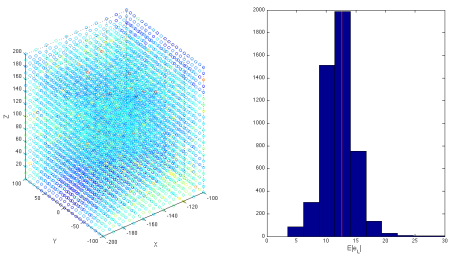

Fig. 5. *Extended system*. In the left pane it is shown the mean error to foveate each 3D point in the testing set. The mean error is computed considering each plausible initial joints configuration of the head. In the right pane it is shown the mean error distribution.

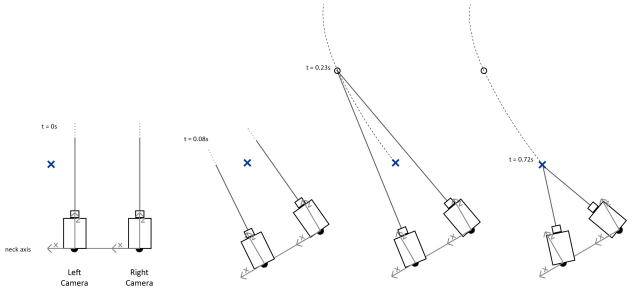

Fig. 6. The trajectories of the cameras performed by the trained extended system. The blue cross represents the 3D feature in space in position [200 0 40]. For graphical reason the image is scaled but it is clearly shown that the system firstly moves the neck and only when the neck is in a steady position the eyes perform the vergence movement.

E. Experimental results

The experiments presented in this section aim to:

- selecting the neck configuration that has the best performance in terms of error in the exploration of the 3D testing space
- comparing the performance with the results collected with the original system

To select the best configuration we compared the results obtained in the experiment "exploring the 3D space". The best configuration was then used to run the experiment "test initial position".

1) Exploring 3D space: We run the experiments for each neck configuration and, comparing mean and variance, we found that the best configuration is the number five with decoupled control between eyes and neck[1]. The testing sets are the same as defined in Sec. III. The obtained parameters K after the training phase of Configuration 5 are:

$$K1 = 0.0167 \quad K2 = 0.5543 \quad K3 = 0.1584 \quad K4 = 0.3542$$

Figure 4 presents the error maps related to Configuration 5. Results seem to be compatible with the performance obtained with the 3 DOF system (see Sec. III and Fig. 2); i.e. the mean errors are 4.33, 3.93 and 1.41 pixel, and the variances are 0.65, 0.32 and 0.34, respectively for the testing sets along the X, Y and Z directions.

2) Test initial position: As shown in Figure 5 the errors are distributed in the interval $[5;20]$ pixels. Compared to the performance of the 3 DOF system (see Sec. III and Fig. 3), the error presents a lower mean and standard deviation. We can therefore conclude that the additional neck-joint provides robustness to the system and, specifically, it reduces the influence of the initial configuration of the head on the performance of the system in foveating a point in space.

V. Discussion

This study proposes a new controller for an active stereo system based on the Hering's law. Quantitative results are presented and a comparison with the original model is provided. In this work we trained the system on a region of the 3D space in front of the cameras, and we evaluated the capability of the system to explore a wider area of the space (in terms of capability to foveate a 3D point); moreover we investigate the robustness of the system with respect to a generic initial joint position of the head. All the test sets are chosen to be in a very far region in space respect to the cameras because what we want to evaluate the capability of the system to react to 3D features that are not projected in the central part of the image plane, and are in general more difficult to foveate.

We investigate different possible control laws for the extended model to take into account the redundancy introduced by the neck; what emerges, comparing the error in foveating 3D points, is that the best performance is obtained when the controllers of the eyes and the one of the neck are decoupled (Configuration 5). Comparing the error illustrated in Figure 2 and 4 it emerges that the mean error and variance associated to the extended system are in general similar to the original ones. Figures 3 and 5 present the experimental results of the initial position of the system. In this case the error of the extended system presents a lower mean and standard deviation. We can therefore conclude that the additional neck-joint provides robustness to the system.

Furthermore, a qualitative analysis of the trajectories of the extended model with decoupled control (see Fig. 6), i.e. Configuration 5, seems to be compatible with some biological results [2].

VI. Conclusion

In this work we presented a vergence-version control system for an active stereo head based on the Hering's law. First, we quantitatively evaluated the performance of the original system previously presented in [9]. We defined a cost function and we trained the system with a classical technique; the obtained results show the robustness and the effectiveness of the controller. Second, we extended the controller adding a neck component that makes the system redundant. We defined different possible configurations of the neck control including coupled/decoupled controls. We extended the cost function and trained the new controller for

[1]The quantitative results can be provided by contacting the authors

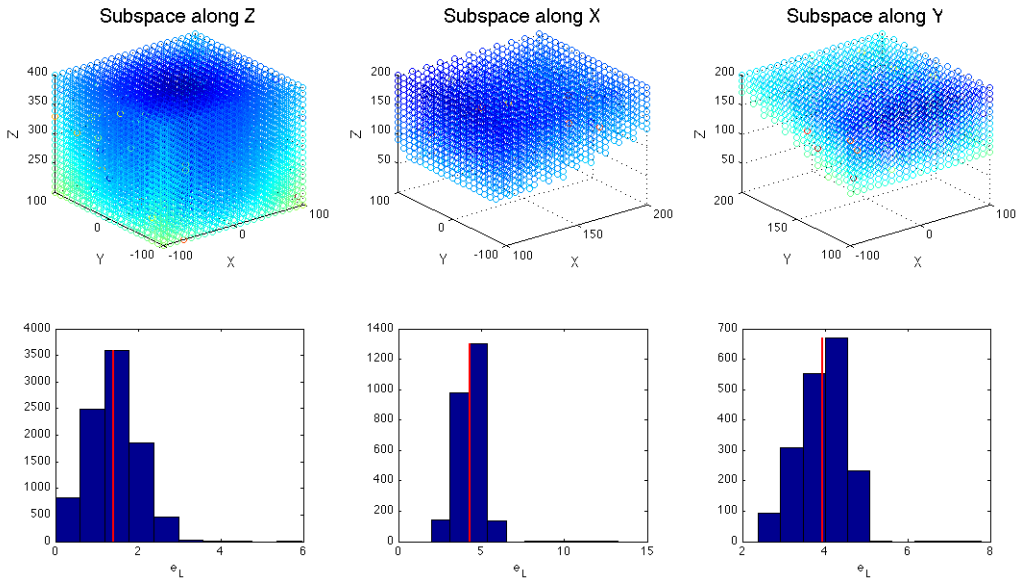

Fig. 4. Error maps computed for the left eye of the extended system with the fifth neck configuration.

each neck configuration. We compared the different neck configurations and chosen the best in terms of obtained performance. We found the best performance with a decoupled control eye-neck. The trajectories generated from this controller are compatible with the human head trajectories in foveating tasks. Moreover, comparing our performance with those of [9], we found that the extended controller solves the redundancy improving the performance and the robustness of the system.

REFERENCES

[1] G. Aragon-Camarasa, H. Fattah, and J. P. Siebert. Towards a unified visual framework in a binocular active robot vision system. *Robotics and Autonomous Systems*, 58(3):276–286, 2010.

[2] L. L. Chen. Head movements evoked by electrical stimulation in the frontal eye field of the monkey: evidence for independent eye and head control. *Journal of Neurophysiology*, 95:3528–3542, 2006.

[3] S. Chen, Y. Li, and N. M. Kwok. Active vision in robotic systems: A survey of recent developments. *The International Journal of Robotics Research*, 2011.

[4] A. Gibaldi, A. Canessa, M. Chessa, S. P. Sabatini, and F. Solari. A neuromorphic control module for real-time vergence eye movements on the icub robot head. In *Proc. 11th IEEE-RAS Int Humanoid Robots (Humanoids) Conf*, pages 543–550, 2011.

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

[6] W. M. King. Binocular coordination of eye movements - hering's law of equal innervation or uniocular control? *European Journal of Neuroscience*, 33:2139–2146, 2011.

[7] N. Kyriakoulis, A. Gasteratos, and S. G. Mouroutsos. Fuzzy vergence control for an active binocular vision system. In *Proc. 7th IEEE Int. Conf. Cybernetic Intelligent Systems CIS 2008*, pages 1–5, 2008.

[8] C. Qu and B. E. Shi. The role of orientation diversity in binocular vergence control. In *Proc. Int Neural Networks (IJCNN) Joint Conf*, pages 2266–2272, 2011.

[9] J. G. Samarawickrama and S. P. Sabatini. Version and vergence control of a stereo camera head by fitting the movement into the Hering's law. In *Proc. Fourth Canadian Conf. Computer and Robot Vision CRV '07*, pages 363–370, 2007.

[10] K. Shimonomura and T. Yagi. Neuromorphic vergence eye movement control of binocular robot vision. In *Proc. IEEE Int Robotics and Biomimetics (ROBIO) Conf*, pages 1774–1779, 2010.

[11] W. Sun and B. E. Shi. Joint development of disparity tuning and vergence control. In *Proc. IEEE Int Development and Learning (ICDL) Conf*, volume 2, pages 1–6, 2011.

[12] E. K. C. Tsang, S. Y. M. Lam, Y. Meng, and B. E. Shi. Neuromorphic implementation of active gaze and vergence control. In *Proc. IEEE Int. Symp. Circuits and Systems ISCAS 2008*, pages 1076–1079, 2008.

[13] Y. Wang and B. E. Shi. Autonomous development of vergence control driven by disparity energy neuron populations. *Neural Computation*, 22:730–751, 2010.

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