Proceedings of the 2006 IEEE International Conference on Robotics and Automation
Orlando, Florida - May 2006

Foveated vision systems with two cameras per eye

Aleš Ude[1,3] Chris Gaskett[4] Gordon Cheng[1,2]

[1]ATR Comput. Neurosc. Lab.
Dept. of Humanoid Robotics
and Computational Neuroscience
2-2-2 Hikaridai, Seika-cho
Soraku-gun, Kyoto, Japan

[2]Japan Science and Technology
Agency, ICORP Computational
Brain Project
2-2-2 Hikaridai, Seika-cho
Soraku-gun, Kyoto, Japan

[3]Jožef Stefan Institute
Dept. of Automatics,
Biocybernetics and Robotics
Jamova 39, 1000 Ljubljana
Slovenia

[4]James Cook University
School of Information
Technology
Cairns, Queensland
Australia

***Abstract*— In this paper we discuss active humanoid vision systems that realize foveation using two rigidly connected cameras in each eye. We present an exhaustive analysis of the relationship between the positions of the observed point in the foveal and peripheral view with respect to the intrinsic and extrinsic parameters of both cameras and 3-D point position. Based on these results we propose a control scheme that can be used to maintain the view of the observed object in the foveal image using information from the peripheral view. Experimental results showing the effectiveness of the proposed foveation control are also provided.**

I. Introduction

Designers of a number of humanoid robots attempted to mimic the foveated structure of the human eye. Foveation is useful because, firstly, it enables the robot to monitor and explore its surroundings in images with wider field of view and sparsely distributed pixels, thereby increasing the efficiency of the search process. Secondly, it makes it possible to simultaneously extract additional information – once the object of interest appears in the fovea – from the foveal area, which has a denser pixel distribution and contains more detail.

Approaches proposed to mimic the foveated structure of biological vision systems include the use of two cameras per eye [1]–[4], i. e. a narrow-angle foveal camera and a wide-angle camera for peripheral vision; lenses with space-variant resolution [5], i. e. a very high definition area in the fovea and a coarse resolution in the periphery; and space-variant log-polar sensors with retina-like distribution of photo-receptors [6]. It is also possible to implement log-polar sensors by transforming standard images into log-polar images in software [7], but this approach requires the use of high definition cameras to get the benefit of varying resolution. Systems with zoom lenses have some of the advantages of foveated vision, but cannot simultaneously acquire wide angle and high resolution images.

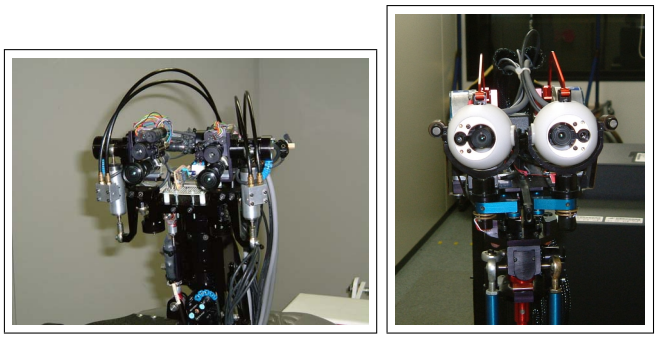

Fig. 1. Example heads with two cameras per eye setup for foveation. The left head has foveal cameras above the peripheral cameras, while the foveal cameras on the right head are located on the outer side of peripheral cameras. The peripheral and foveal cameras are rigidly connected with parallel optical axes. The motor system of each eye consists of two independent degrees of freedom.

While log-polar-sensors and lenses with space-variant resolution are conceptually appealing, they are difficult to construct and prevent us from using high-quality standard cameras and lenses. High-definition cameras are problematic for real-time humanoid vision because of the form factor and bandwidth. We therefore follow the first approach and use two cameras per eye to realize foveation. This allows us to use miniature cameras like the ones in Fig. 1, which improves the dynamic properties of the oculomotor system.

There are several visual tasks that can benefit from foveated vision. One of the most prominent among them is object recognition. Object recognition requires the robot to detect objects in dynamic environments and to control the eye gaze to get the objects into the fovea and to keep them there. Once these tasks are accomplished, the robot can determine the identity of the object by processing foveal views. In this paper we analyze foveation in terms of intrinsic and extrinsic parameters of a two cameras per eye foveation setup and derive the formulas needed to center the foveal view on the object of interest based on information coming from the peripheral view. We also present a motor control system that can be used to accomplish this task.

II. Foveation Model

Two issues need to be considered when analyzing the foveation setup with two cameras:

1) Given a 3-D point that projects onto the center of the foveal image, where will the point be projected onto the peripheral image? This will be the ideal position in the periphery for foveation.
2) If a 3-D point projects onto the peripheral image away from the ideal position described above, how far is the projection of the point from the center of the foveal image?

A. Camera Model

For our theoretical analysis, we model both cameras by a standard pinhole camera model. We denote a 3-D point by

$\boldsymbol{M} = \begin{bmatrix} X & Y & Z \end{bmatrix}^T$ and a 2-D point by $\boldsymbol{m} = \begin{bmatrix} x & y \end{bmatrix}^T$. Let $\tilde{\boldsymbol{M}} = \begin{bmatrix} X & Y & Z & 1 \end{bmatrix}^T$ and $\tilde{\boldsymbol{m}} = \begin{bmatrix} x & y & 1 \end{bmatrix}^T$ be the homogeneous coordinates of $\boldsymbol{M}$ and $\boldsymbol{m}$, respectively. The relationship between a 3-D point $\boldsymbol{M}$ and its projection $\boldsymbol{m}$ is then given by [8]

$$s\tilde{\boldsymbol{m}} = \boldsymbol{A} \begin{bmatrix} \boldsymbol{R} & \boldsymbol{t} \end{bmatrix} \tilde{\boldsymbol{M}}, \tag{1}$$

where s is an arbitrary scale factor, $\boldsymbol{R}$ and $\boldsymbol{t}$ are the extrinsic parameters denoting the rotation and translation that relate the world coordinate system to the camera coordinate system and $\boldsymbol{A}$ is the intrinsic matrix

$$\boldsymbol{A} = \begin{bmatrix} \alpha & \gamma & x_0 \\ 0 & \beta & y_0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{2}$$

α and β are the scale factors, γ is the parameter describing the skewness of the two image axes, and (x_0, y_0) is the principal point.

In the following we assume without loss of generality that the origin of the image coordinate system coincides with the principal point (x_0, y_0), thus $x_0 = y_0 = 0$. Note that on a real camera the principal point does not coincide with the image center in pixel coordinates exactly. However, since the distortion effects are smallest around the principal point and since making this assumption significantly simplifies the equations, it makes sense to attempt to bring the point of interest to the position that projects onto the principal point of the foveal camera and not to the precise image center. Taking a standard video camera producing 640 x 480 images, the distance of the principal point from the image center in pixel coordinates is usually less than 10 pixels.

The pinhole camera model (1) does not consider the effects of lens distortion. Such an assumption is justified for foveal cameras, which are equipped with lenses with relatively long focal lengths that normally do not exhibit noticeable distortion effects. This is especially true because the distortion function is usually dominated by radial components [8], [9]. Hence the distortion effects are larger at the edges than in the center of an image and therefore have only limited effects on foveation. Conversely, to achieve wide field of view, peripheral cameras need to have lenses with shorter focal lengths. Cameras with such lenses often produce significantly distorted images. However, the distortion can be corrected in a preprocessing step using a suitable distortion correction procedure, e. g. the one described in [8]. Equation (1) is valid for the distortion-corrected pixels and we conclude that we do not need to consider the distortion effects in our analysis.

B. Principal point in the fovea and peripheral images

We denote by $\boldsymbol{A}_f$, $\boldsymbol{R}_f$, $\boldsymbol{t}_f$ and $\boldsymbol{A}_p$, $\boldsymbol{R}_p$, $\boldsymbol{t}_p$ the intrinsic and extrinsic parameters of the foveal and peripheral camera, respectively. Lets now assume that the world coordinate system is aligned with the coordinate system of the foveal camera. In this case we have $\boldsymbol{R}_f = \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix, and $\boldsymbol{t}_f = 0$. Let $\hat{\boldsymbol{t}}$ be the position of the origin of the peripheral coordinate system expressed in the foveal coordinate system and let $\hat{\boldsymbol{R}}$ be the rotation matrix that rotates the basis vectors of the peripheral coordinate system into the basis vectors of the foveal coordinate system. We then have

$$\boldsymbol{R}_p\boldsymbol{M} + \boldsymbol{t}_p = \hat{\boldsymbol{R}}(\boldsymbol{M} - \hat{\boldsymbol{t}}). \tag{3}$$

The projections of a 3-D point $\boldsymbol{M}$ onto the foveal and peripheral image are then given by

$$x_f = \frac{\alpha_f X + \gamma_f Y}{Z}, \tag{4}$$

$$y_f = \frac{\beta_f Y}{Z}, \tag{5}$$

and

$$x_p = \frac{\alpha_p \boldsymbol{r}_1 \cdot (\boldsymbol{M} - \hat{\boldsymbol{t}}) + \gamma_p \boldsymbol{r}_2 \cdot (\boldsymbol{M} - \hat{\boldsymbol{t}})}{\boldsymbol{r}_3 \cdot (\boldsymbol{M} - \hat{\boldsymbol{t}})}, \tag{6}$$

$$y_p = \frac{\beta_p \boldsymbol{r}_2 \cdot (\boldsymbol{M} - \hat{\boldsymbol{t}})}{\boldsymbol{r}_3 \cdot (\boldsymbol{M} - \hat{\boldsymbol{t}})}, \tag{7}$$

where $\boldsymbol{r}_1$, $\boldsymbol{r}_2$, and $\boldsymbol{r}_3$ are the rows of the rotation matrix $\hat{\boldsymbol{R}} = \begin{bmatrix} \boldsymbol{r}_1^T & \boldsymbol{r}_2^T & \boldsymbol{r}_3^T \end{bmatrix}^T$. $\boldsymbol{M}$ projects onto the principal point in the fovea if $x_f = y_f = 0$. Assuming that $\boldsymbol{M}$ is in front of the camera, hence $Z > 0$, we obtain from Eq. (4) and (5) that $X = Y = 0$, which means that the point must lie on the optical axis of the foveal camera. Inserting this into Eq. (6) and (7), we obtain the following expression for the ideal position $(\hat{x}_p, \hat{y}_p)$ in the peripheral image that results in the projection onto the principal point in the foveal image

$$\hat{x}_p = \frac{\alpha_p \boldsymbol{r}_1 \cdot \boldsymbol{t} + \gamma_p \boldsymbol{r}_2 \cdot \boldsymbol{t} - (\alpha_p r_{13} + \gamma_p r_{23})Z}{\boldsymbol{r}_3 \cdot \boldsymbol{t} - r_{33}Z}, \tag{8}$$

$$\hat{y}_p = \frac{\beta_p \boldsymbol{r}_2 \cdot \boldsymbol{t} - \beta_p r_{23} Z}{\boldsymbol{r}_3 \boldsymbol{t} - r_{33} Z}, \tag{9}$$

where $\begin{bmatrix} r_{13} & r_{23} & r_{33} \end{bmatrix}^T$ is the third column of $\hat{\boldsymbol{R}}$. Note that the ideal position in the periphery is independent from the intrinsic parameters of the foveal camera. It depends, however, on the distance of the point of interest from the foveation setup.

C. Displacement from the ideal position

Lets assume now that the 3-D point of interest $\boldsymbol{M}$ projects onto a pixel away from the principal point in foveal image by displacement (D_x, D_y). From (4) and (5) we have

$$D_x = \frac{\alpha_f X + \gamma_f Y}{Z}, \tag{10}$$

$$D_y = \frac{\beta_f Y}{Z}, \tag{11}$$

thus

$$X = \frac{D_x - \gamma_f D_y / \beta_f}{\alpha_f} Z, \tag{12}$$

$$Y = \frac{D_y}{\beta_f} Z. \tag{13}$$

Point $\begin{bmatrix} 0 & 0 & Z \end{bmatrix}^T$ is the point on the optical axis which is closest to $\boldsymbol{M}$. It projects onto $(\hat{x}_p, \hat{y}_p)$ in the peripheral view. We define (d_x, d_y) to be the displacement of the projection of $\boldsymbol{M}$ from this point in the peripheral view and we would like to

express (D_x, D_y) in terms of (d_x, d_y). We have the following relationship

$$s \begin{bmatrix} \hat{x}_p + d_x \\ \hat{y}_p + d_y \\ 1 \end{bmatrix} = \boldsymbol{A}_p \begin{bmatrix} \left(r_{11}\left(D_x - \gamma_f D_y/\beta_f\right)/\alpha_f + \right. \\ \left. r_{12} D_y/\beta_f + r_{13}\right) Z - \boldsymbol{r}_1 \cdot \boldsymbol{t} \\ \left(r_{21}\left(D_x - \gamma_f D_y/\beta_f\right)/\alpha_f + \right. \\ \left. r_{22} D_y/\beta_f + r_{23}\right) Z - \boldsymbol{r}_2 \cdot \boldsymbol{t} \\ \left(r_{31}\left(D_x - \gamma_f D_y/\beta_f\right)/\alpha_f + \right. \\ \left. r_{32} D_y/\beta_f + r_{33}\right) Z - \boldsymbol{r}_3 \cdot \boldsymbol{t} \end{bmatrix}. \tag{14}$$

By subtracting (8) and (9) from (14), we can obtain a rather complex expression for the error in the periphery (d_x, d_y) in terms of the error in the fovea (D_x, D_y). This expression also depends on the distance Z of the point of interest from the camera setup. Fortunately, the result can be simplified by making some reasonable assumptions about the foveation setup. It is common to construct a foveated camera system in such a way that the optical axes of the peripheral and foveal camera are parallel. No calibration is needed to achieve this, only the cradles of both cameras must be built with sufficient precision. Standard industrial cameras are constructed precisely enough to support such an arrangement, which is used in most fixed foveation systems (like in Fig. 1). In this case we have $r_{31} = r_{32} = r_{13} = r_{23} = 0$, $r_{33} = 1$ and Eq. (14) becomes

$$s \begin{bmatrix} \hat{x}_p + d_x \\ \hat{y}_p + d_y \\ 1 \end{bmatrix} = \boldsymbol{A}_p \begin{bmatrix} \left(r_{11}\left(D_x - \gamma_f D_y/\beta_f\right)/\alpha_f + \right. \\ \left. r_{12} D_y/\beta_f\right) Z - \boldsymbol{r}_1 \cdot \boldsymbol{t} \\ \left(r_{21}\left(D_x - \gamma_f D_y/\beta_f\right)/\alpha_f + \right. \\ \left. r_{22} D_y/\beta_f\right) Z - \boldsymbol{r}_2 \cdot \boldsymbol{t} \\ Z - t_z \end{bmatrix}, \tag{15}$$

where $\boldsymbol{t} = \begin{bmatrix} t_x & t_y & t_z \end{bmatrix}^T$. The denominator in (8) and (9) coincides with the third component in Eq. (15), hence subtracting (8) and (9) from (15) results in

$$d_x = \frac{Z}{Z - t_z}\left(\frac{\alpha_p r_{11} + \gamma_p r_{21}}{\alpha_f} D_x + \frac{-r_{11}\alpha_p\gamma_f + r_{12}\alpha_p^2 - r_{21}\gamma_p\gamma_f + r_{22}\gamma_p^2}{\alpha_f\beta_f} D_y\right), \tag{16}$$

$$d_y = \frac{Z}{Z - t_z}\left(\frac{r_{21}\beta_p}{\alpha_f} D_x + \frac{\beta_p\left(-r_{21}\gamma_f/\alpha_f + r_{22}\right)}{\beta_f} D_y\right). \tag{17}$$

It is trivial to invert this equation system to obtain the expression for the error in the fovea in terms of the error in the periphery and distance Z. We omit the details here and only present results with further simplifying assumptions. Lets assume that both cameras are completely aligned, i. e. $r_{21} = r_{12} = 0$ and $r_{11} = r_{22} = 1$ (no rotation around the optical axis). In this case we can calculate a simpler relationship between the error displacements in the foveal and peripheral images

$$D_x = \frac{Z - t_z}{Z} \cdot \frac{\alpha_f}{\alpha_p}\left(d_x + \frac{\alpha_p\gamma_f - \gamma_p\alpha_f}{\alpha_f\beta_p} d_y\right), \tag{18}$$

$$D_y = \frac{Z - t_z}{Z} \cdot \frac{\beta_f}{\beta_p} d_y. \tag{19}$$

The skew parameters γ_p and γ_f are normally much smaller than α_p, α_f, β_p, and β_f. Similarly, the displacement of the camera t_z is usually much smaller than the distance Z of the point of interest from the camera. Note that it is difficult to achieve $t_z = 0$ on a practical camera system because it is not easy to determine the exact positions of the projection centers and to place the cameras accordingly. Thus, for totally aligned cameras we have the following approximation

$$D_x \approx \frac{\alpha_f}{\alpha_p} d_x, \tag{20}$$

$$D_y \approx \frac{\beta_f}{\beta_p} d_y. \tag{21}$$

This means that the error from the ideal displacement in the peripheral image is scaled in the fovea by the ratio of focal lengths. This approximation is exact for perfect pinhole cameras ($\gamma_p = \gamma_f = 0$) with precisely aligned coordinate systems, i. e. $\hat{\boldsymbol{R}} = \boldsymbol{I}$ and $t_z = 0$. Since the focal length of the foveal camera is always greater than the focal length of the peripheral camera, i. e. $\alpha_p, \beta_p < \alpha_f, \beta_f$, the deviation from the principal point in the fovea is greater than the deviation from the ideal position in the peripheral image. This is, of course, an expected result.

D. Analysis of foveated vision systems

Making the same assumptions as when calculating (15), we obtain from Eq. (8) and (9) the following expression for the ideal position in the peripheral image

$$\hat{x}_p = \frac{\alpha_p \boldsymbol{r}_1 \cdot \boldsymbol{t} + \gamma_p \boldsymbol{r}_2 \cdot \boldsymbol{t}}{t_z - Z} \approx -\frac{\alpha_p \boldsymbol{r}_1 \cdot \boldsymbol{t}}{Z}, \tag{22}$$

$$\hat{y}_p = \frac{\beta_p \boldsymbol{r}_2 \cdot \boldsymbol{t}}{t_z - Z} \approx -\frac{\beta_p \boldsymbol{r}_2 \cdot \boldsymbol{t}}{Z}. \tag{23}$$

We can again neglect the influence of γ_p, which is always significantly smaller than α_p and β_p. Note, however, that it is important that the cradles are built precisely and that the optical axes are aligned accurately because in Eq. (8) and (9) the zero terms r_{13} and r_{23} are multiplied by Z, which is normally large. Hence if the system is not built precisely, the above approximations are not valid. This is intuitive because if optical axes diverge, the foveal image will not overlap with the peripheral image as the distance increases and it is impossible to get a point into the fovea based on information from the peripheral image. The above equation system can be further simplified by assuming totally aligned pinhole cameras ($r_{21} = r_{12} = 0$ and $r_{11} = r_{22} = 1$, $t_z = 0$, $\gamma_p = 0$), which results in

$$\hat{x}_p = -\frac{\alpha_p t_x}{Z}, \tag{24}$$

$$\hat{y}_p = -\frac{\beta_p t_y}{Z}. \tag{25}$$

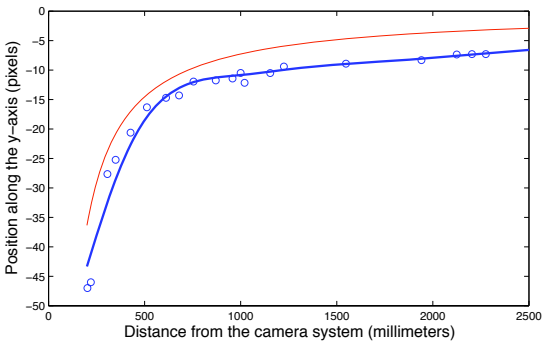

Fig. 2. Red curve: $\hat{y}_p$ with respect to the distance of the object from the camera system as calculated by Eq. (25) (totally aligned, ideal pinhole cameras, $\hat{\boldsymbol{R}} = \boldsymbol{I}$, $t_x = t_z = 0$, $t_y = 25$, $\alpha_p = \beta_p = 290.9$). Blue curve: $\hat{y}_p$ determined experimentally by placing the object manually at various distances so that it projects on the center of the foveal image. At each such configuration we measured the object's position in peripheral image and its distance from the eye (using stereo vision). The blue circles show these measurements.

In our foveation setup, the peripheral cameras are equipped with 3mm lenses and with CCD chips of size 6.6×4.4 millimeters, while the foveal cameras are equipped with 12mm lenses and with CCD chips of size 3.3×2.2 millimeters. The distance between them is about 25mm along the y-axis ($t_x \approx 0$, $t_y \approx 25$). Theoretically, the scaling factors of such cameras are $\alpha_p = \beta_p \approx 290.9$ and $\alpha_f = \beta_f \approx 1306.8$ when the cameras are calibrated for images of size 640×480.

Fig. 2 shows the variation of $\hat{y}_p$ with respect to Z under such assumptions and proves that our system indeed exhibits such characteristics. For $Z = 1$ meter, the ideal position in the periperal image is given by $\hat{x}_p = 0$ and $\hat{y}_p = -290.9 \times 25/1000 = -7.3$ pixels. For objects further away, $\hat{y}_p$ tends to zero. From Eq. (25) it follows that the necessary displacement doubles to -14.6 when $Z \approx 498$mm. Hence, if we fix Z to 1 meter and observe objects more than 0.5m away from the camera, the systematic error in the peripheral images will be less than 7.3 pixels. Eq. (21) tells us that the displacement from the central position in the foveal view will be at most $1306.8/290.9 \times 7.3 \approx 32.8$ pixels, hence we are still relatively close to the principal point in the foveal image. Note that fixing the distance Z is equivalent to replacing the perspective projection with the orthographic projection in our model.

III. Foveation Control

There are various ways to get the object of interest into the robot's fovea. One is to calibrate the system and update $(\hat{x}_p, \hat{y}_p)$ with respect to the current 3-D position using either full Eq. (8) and (9) or the simplified version (22) and (23). This is not practical on highly dynamic humanoid robots because it makes an unrealistic assumption that we can maintain the calibration of the eyes during fast eye movements. The alternative provided by the discussion of Section II-D is to fix the distance Z to a constant value Z^*, which results in a constant displacement $(\hat{x}_p, \hat{y}_p) = (x_p^*, y_p^*)$. We can then maintain the view of the object of interest in foveal images by fixating on the object at (x_p^*, y_p^*) in the peripheral views. We have shown that for a foveation system like ours, such an approximation causes the object located more than half a meter away from the eye to be at most 32.8 pixels away from the principal point. This is sufficient to maintain the object's view in the fovea. Although our practical foveation setup differs from the theoretical model due to inaccuracies in its construction, we obtained good results by experimentally setting $\hat{x}_p = 0$, $\hat{y}_p = -9.7$ pixels while moving the objects more than 1 meter away from the robot. We could improve the accuracy by switching the control input data from the peripheral cameras to the foveal cameras once the object appears in the fovea. It is, however, not trivial to ensure smooth operation in such a system because a fast-moving object can quickly disappear from the fovea, thus requiring frequent switching between the two operation modes.

We therefore developed a control system whose primary goal is to maintain foveation based on 2-D information from peripheral views. The developed system attempts to maintain the view of the object in foveal views of both eyes simultaneously. The secondary goal of the system is to enhance the appearance of the humanoid through mimicking aspects of human movement: human eyes follow object movement, but without head and body movement have a limited range; thus, the robot's control system supports its eye movements through head and body movements. Altogether we used 10 degrees of freedom (4 on the eyes, 3 on the head, and 3 on the torso) to maintain the view of the object. It was important for the control system not to rely on exact knowledge of the robot's kinematics since the action of the robot's joints varies over time due to joint wear-and-tear and maintenance activities.

The robot's primary mechanism for maintaining the view of the object of interest is eye movement: the control system continuously alters the pan and tilt of each eye to keep the object near the center of the corresponding view (i.e. visual servo control [10]). Independent eye motion is acceptable when the object is being tracked properly in both peripheral views, but looks rather unnatural when one eye loses its view of the object while the other eye continues to roam. Our solution is to introduce a gentle cross-coupling between a camera's view and the control of the other eye. Thus, when a camera's view of the target is lost, its corresponding eye continues to move, fairly slowly, under the influence of the other camera's view. As well as appearing natural, such eye movements improve the likelihood of re-locating the object.

We consider that the task of the robot's head is to assist the eyes by increasing the viewable area and avoiding unnatural eye poses. Similarly, we consider that the task of the robot's body is to assist the head, further increasing the viewable area and avoiding unnatural head poses.

To aid in coordinating the joints, we assign a relaxation position to each joint and 2-D object position. The relaxation position for the object is at (x_p^*, y_p^*) and the eyes' task is to bring the object to that position. The relaxation position for the 4 eye joints is to face forward, and the head's task is to bring the eyes to that position. Further, the 3 neck joints have a relaxation position, and the torso's task is to bring the head to that position. For example, if the object of interest is up

and to the left, the eyes would tilt up and pan left, causing the head would tilt up and turn left, and the torso to lean back and turn.

The complete control system is implemented as a network of PD controllers expressing the assistive relationships. The PD controllers are based on simplified mappings between visual coordinates and joint angles, which are described below, rather than on a full kinematic model. Such mappings are sufficient because the system is closed-loop and can make corrective movements to converge towards the desired configuration.

We define the *desired change* for self-relaxation, D, for each joint,

$$D_{joint} = \left(\theta^*_{joint} - \theta_{joint}\right) - K_d \dot{\theta}_{joint}, \quad (26)$$

where K_d is the derivative gain for joints; θ is the current joint angle; $\dot{\theta}$ is the current joint angular velocity, and the asterisk indicates the relaxation position. The derivative components help to compensate for the speed of the object and assisted joints.

The desired change for the object position is:

$$D_{\mathsf{X}object} = \left(x^*_p - x_{object}\right) - K_{dv} \dot{x}_{object}, \quad (27)$$

where K_{dv} is the derivative gain for 2-D object position; X represents the x pixels axis; and x_{object} is 2-D object position in pixels.

The purpose of the *left eye pan* ($\mathsf{L}EP$) joint is to move the target into the center of the left camera's field of view:

$$\begin{aligned}\hat{\dot{\theta}}_{\mathsf{L}EP} = K_p \times \Big[& K_{\text{relaxation}} D_{\mathsf{L}EP} \\ & - K_{\text{target}\rightarrow EP} K_v C_{\mathsf{L}object} D_{\mathsf{LX}object} \\ & + K_{\text{cross-target}\rightarrow EP} K_v C_{\mathsf{R}object} D_{\mathsf{RX}object} \Big], \end{aligned} \quad (28)$$

where $\hat{\dot{\theta}}_{\mathsf{L}EP}$ is the new target velocity for the joint; L and R represent left and right; K_p is the proportional gain; K_v is the proportional gain for 2-D object position; C_{object} is the tracking confidence for the object; and the gain $K_{\text{cross-target}\rightarrow EP} < K_{\text{target}\rightarrow EP}$.

The purpose of the *left eye tilt* ($\mathsf{L}ET$) joint is to move the target into the center of the left camera's field of view:

$$\begin{aligned}\hat{\dot{\theta}}_{\mathsf{L}ET} = K_p \times \Big[& K_{\text{relaxation}} D_{\mathsf{L}ET} \\ & - K_{\text{target}\rightarrow ET} K_v C_{\mathsf{L}object} D_{\mathsf{LY}object} \\ & - K_{\text{cross-target}\rightarrow ET} K_v C_{\mathsf{R}object} D_{\mathsf{RY}object} \Big]. \end{aligned} \quad (29)$$

The equations for the right eye pan and tilt joints are the same as for the left, except that L becomes R and vice versa.

Head nod joint (HN) assists the eye tilt joints:

$$\begin{aligned}\hat{\dot{\theta}}_{HN} = K_p \times \Big[& K_{\text{relaxation}} D_{HN} \\ & - K_{ET\rightarrow HN} \left(D_{\mathsf{L}ET} + D_{\mathsf{R}ET}\right) \Big]. \end{aligned} \quad (30)$$

The *head tilt joint* (HT), which tilts the head from side to side, moves to assist the pan (EP) and equalize the tilt (ET) of the eyes:

$$\begin{aligned}\hat{\dot{\theta}}_{HT} = K_p \times \Big[& K_{\text{relaxation}} D_{HT} \\ & - K_{EP\rightarrow HT} \left(D_{\mathsf{L}EP} - D_{\mathsf{R}EP}\right) \\ & - K_{ET\rightarrow HT} \left(D_{\mathsf{L}ET} - D_{\mathsf{R}ET}\right) \Big]. \end{aligned} \quad (31)$$

The *torso flexion-extension joint* (TFE) assists the head nod joint:

$$\hat{\dot{\theta}}_{TFE} = K_p \times \Big[K_{\text{relaxation}} D_{TFE} - K_{HN\rightarrow TFE} D_{HN} \Big]. \quad (32)$$

We omit the control rules for *head rotate joint* (HR), *torso rotate joint* (TR) and *torso abduction-adduction joint* (TAA). These controllers are defined equivalently to the controllers for the head nod joint and torso flexion-extension joint, the only difference being that they support different preceding degrees of freedom.

A. Control experiments

The graphs in Fig. 3 show the assistive relationships between different degrees of freedom. Even though the eye joints often hit the joint limits, the robot could still maintain the view of the object by making use of head movements. This is also useful if one of the joints fails; the robot can still function, although with degraded performance.

To demonstrate the reliability of the pursuit strategy, we measured the velocity of object motion while the robot attempted to maintain its view in the fovea. The object was moved in front of the robot by an experimenter. For this purpose we attached three active markers to the object and measured their motion with the optical tracking system Visualeyze. Since the rotational component of motion was

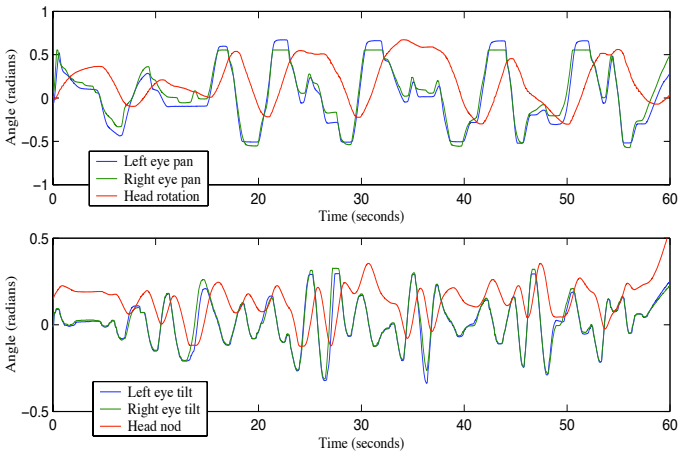

Fig. 3. Eye (blue and green trajectories) and head (red trajectories) joints during foveation. On our robot both eye pans and head rotation represent rotations around axes parallel to the vertical body axis (left - right change of viewing direction), while both eye tilts and head nod represent rotations around axes parallel to the shoulder axis (up - down change of view). These are the most important movements for foveation. As expected left and right eye pans as well as left and right eye tilts follow similar trajectories to maintain the direction of view. Due to the assistive relationship between eye pans and head rotation, the head rotation joint follows the pan angles of both eyes. There is a similar relationship between eye tilts and head nod motion.

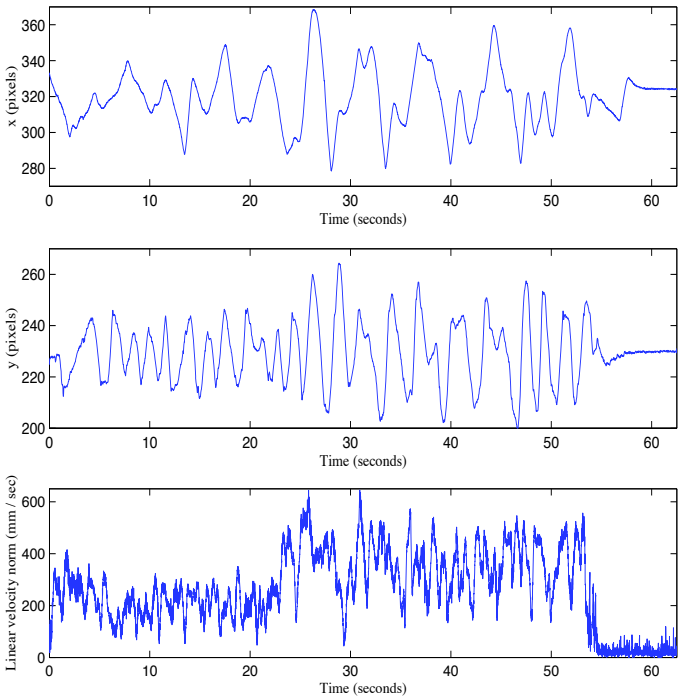

Fig. 4. Effect of object velocity on accuracy of foveation. The upper two graphs show object position in one of the peripheral images. The desired image position was set at $(320, 230.3)$. It was slightly displaced from the image center, which was at $(320, 240)$, to account for the displacement of foveal cameras from the peripheral ones. The lower graph depicts the norm of linear velocity of the tracked object.

negligible, we estimated only linear velocity of object motion. Not surprisingly, the robot was more successful at maintaining the view of the object when the object motion was slower (see Fig. 4). In the graphs the object velocity increases after approximately 23 seconds. This causes larger deviations from the ideal position in the peripheral view. Nevertheless, even with fairly fast movements (more than 0.5 m/sec), the robot was still able to maintain the view of the object using information from peripheral cameras. Although the accuracy in this case is often not sufficient to maintain the view of the object in the fovea, the robot can at least restart the image analysis once the object returns to the fovea. As shown in the graphs, the robot quickly directs its eyes towards the object and maintains its foveal view after the object comes to a standstill.

Our experiments demonstrate that the proposed control strategy is successful at smoothly pursuing objects of interest in peripheral images and that it can maintain the view of the object in the fovea. An example for foveated control is shown in Fig. 5. We successfully applied the proposed foveation control system to solve complex visual tasks including object recognition in dynamic environments [11].

IV. Conclusion

This paper presents an exhaustive analysis of foveated vision systems using two rigidly connected cameras in each eye. We derived fully general formulas that express the relationship between 3-D point projections in foveal and peripheral views and also provided simplified versions under reasonable assumptions. Based on our theoretical analysis we showed – for standard foveation setups consisting of cameras with parallel optical axes – that we can maintain the view of an object in the fovea using solely information from peripheral views and without making use of 3-D information, which is often difficult to compute on a highly dynamic humanoid system. A suitable control system, which can exploit the redundancies of the humanoid, was also presented. Such an analysis has not been done before. It is important because it supports the design and control of the two cameras per eye setup, which is currently the most commonly used approach to realize foveation on humanoid robots.

Fig. 5. Maintaing the object view in foveal image based on information from the peripheral view

We intend to investigate in the future how to make use of information from the foveal view for more accurate servoing towards the object of interest. The main problem that needs to be addressed is how to ensure smooth motion behavior when switching between controllers using information from different views that provide information at different scales.

References

[1] B. Scassellati, "A binocular, foveated active vision system," MIT, Artificial Intelligence Laboratory, Tech. Rep. A.I. Memo No. 1628, 1999.

[2] C. G. Atkeson, J. G. Hale, F. Pollick, M. Riley, S. Kotosaka, S. Schaal, T. Shibata, G. Tevatia, A. Ude, S. Vijayakumar, and M. Kawato, "Using humanoid robots to study human behavior," *IEEE Intelligent Systems*, vol. 15, no. 4, pp. 46–56, 2000.

[3] T. Shibata, S. Vijayakumar, J. Jörg Conradt, and S. Schaal, "Biomimetic oculomotor control," *Adaptive Behavior*, vol. 9, pp. 189–208, 2001.

[4] H. Kozima and H. Yano, "A robot that learns to communicate with human caregivers," in *Proc. Int. Workshop on Epigenetic Robotics*, Lund, Sweden, 2001.

[5] S. Rougeaux and Y. Kuniyoshi, "Robust tracking by a humanoid vision system," in *Proc. IAPR First Int. Workshop on Humanoid and Friendly Robotics*, Tsukuba, Japan, 1998.

[6] G. Sandini and G. Metta, *Sensors and Sensing in Biology and Engineering*. Wien-New York: Springer-Verlag, 2003, ch. Retina-like sensors: motivations, technology and applications.

[7] G. Engel, D. N. Greve, J. M. Lubin, and E. L. Schwartz, "Space-variant active vision and visually guided robotics: Design and construction of a high-performance miniature vehicle," in *Proc. 12th IAPR Int. Conf. Pattern Recognition. Vol. 2 - Conf. B: Computer Vision & Image Processing*, Jerusalem, Israel, 1994, pp. 487 – 490.

[8] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Analysis Machine Intell.*, vol. 22, no. 11, pp. 1330–1334, 2000.

[9] R. Y. Tsai, "A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses," *IEEE J. Robotics and Automation*, vol. 3, no. 4, pp. 323–344, 1987.

[10] S. Hutchinson, G. D. Hager, and P. I. Corke, "A tutorial on visual servo control," *IEEE Trans. Robotics and Automation*, vol. 12, no. 5, pp. 651–670, 1996.

[11] A. Ude, C. G. Atkeson, and G. Cheng, "Combining peripheral and foveal humanoid vision to detect, pursue, recognize and act," in *Proc. 2003 IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, Las Vegas, Nevada, 2003, pp. 2173–2178.