Depth Perception in an Anthropomorphic Robot that Replicates Human Eye Movements

Fabrizio Santini
Dept. of Cognitive and Neural Systems
Boston University
Boston, Massachusetts 02215
Email: santini@cns.bu.edu

Michele Rucci
Dept. of Cognitive and Neural Systems
Boston University
Boston, Massachusetts 02215
Email: rucci@cns.bu.edu

***Abstract*— In the eyes of many species, the optical nodal points of the cornea and lens do not lie on the axes of rotation of the eye. During eye movements, this lack of alignment produces depth information in the form of an oculomotor parallax. That is, a redirection of gaze shifts the projection of an object on the retina by an amount that depends not only on the amplitude of the rotation of the eye, but also on the distance of the object with respect to the observer. Species as diverse as the chameleon and the sandlance critically rely on this depth cue to estimate distance. An oculomotor parallax is present also in the human eye and, during natural eye movements, it produces retinal shifts that are well within the range of sensitivity of the human visual system. We have developed an anthropomorphic robot equipped with a pan/tilt head specifically designed to reproduce the oculomotor parallax present in the human eye. We show that replication of sequences of human eye movements with this robot produces accurate estimation of distance. In robotic vision it is often debated whether the dynamic analysis of a visual scene by means of a mobile camera presents advantages with respect to the static analysis provided by a stationary camera with a wide field of view. This study shows that by generating depth information, replication of the dynamic strategy by which humans scan a visual scene greatly facilitates the processes of figure/ground segregation and image segmentation, two of the hardest tasks of machine vision.**

I. Introduction

Depth perception is one of the most investigated aspects of biological and machine vision. Accurate estimation of the distance of objects and surfaces with respect to the agent is crucial for most operations involving interactions with the environment, as in the motor tasks of navigation and manipulation. Depth information also provides a fundamental contribution to image segmentation, one of the most serious challenges faced by machine vision systems.

As an agent navigates through the environment, cues of distance in the visual modality emerge in the form of motion parallax, *i.e.* the apparent motion of stationary objects [1], [2]. While motion parallax is most evident for large movements of the agent, in species with mobile eyes, a small parallax is also generated by eye movements. In the eyes of most species, the optical nodal points are not coincident with the center of rotation. Therefore, during a relocation of gaze, the projection of an object on the retina moves by an amount that depends both on the amplitude of the rotation and on the distance of the object from the observer (see Fig. 1). Species as diverse as the chameleon and the sandlance, for which the optics of the cornea and lens maximize the distance between nodal points and the center of rotation, rely on this cue to judge distance [3], [4]. A similar parallax is also present in the eyes of primates, and during the normal scanning of a visual scene it produces retinal shifts that are well within the range of human visual acuity [5], [6], [7].

In robotic vision, since the process of reconstructing a 3-D scene from its projections on a 2-D sensor is inherently ambiguous, a variety of techniques have been investigated. Proposed methods range from the comparison of images taken from different points of view (stereoscopic vision) [8], [9] or in different instants of time (depth from motion) [10], [11] to consideration of a priori knowledge of the scene and its structure (depth from shading, size, occlusions, etc.) [12], [13], [14], [15]. In particular, a number of studies have examined the parallax that emerges during large movements of a camera as those produced by a vision system mounted on a mobile platform [16], [17], [18], [19], [20], [10]. These conditions amplify the motion parallax. No previous study, however, has specifically focused on the parallax produced by rotating the camera of a stationary head/eye system. Camera rotations produce a parallax similar to that of the eye if the nodal points of the optical system do not lie on the axes of rotations. Since such a misalignment occurs unless it is intentionally eliminated by careful specification of the optical system's and mechanical characteristics, an oculomotor parallax is present in virtually every pan/tilt unit used in robotics.

This study investigates the use of the oculomotor parallax for the visual estimation of distance in robotics. We use an anthropomorphic robotic system to reproduce the parallax present in the human eye. We show that the oculomotor parallax that emerges during the small relocations of gaze that characterize human oculomotor activity provides reliable depth information within a range of nearby distances.

II. Camera rotation and distance estimation

As illustrated in Fig. 2, to reproduce the oculomotor parallax present in the human eye, we developed a head/eye system in which the distances between center of rotation, nodal points, and sensor surfaces closely replicated the arrangement of the eye.

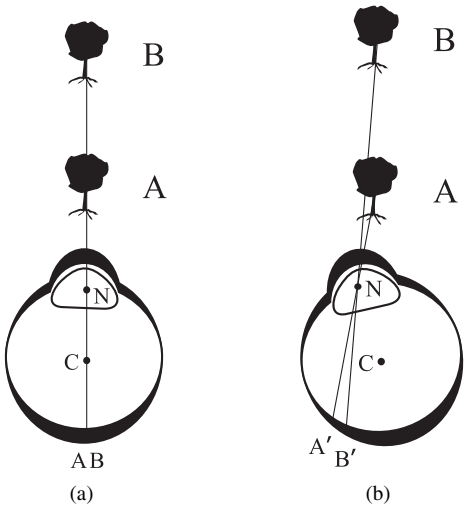


Fig. 1. Depth information produced by the oculomotor parallax. In the eye of many species, the center of rotation C is not coincident with the position of the nodal point N. A rotation of the eye shifts the projection of an object on the retina by an amount that depends both on the rotation magnitude and on the distance of the object. The two panels (a) and (b) illustrate the retinal projections of the points A and B before and after a rotation.

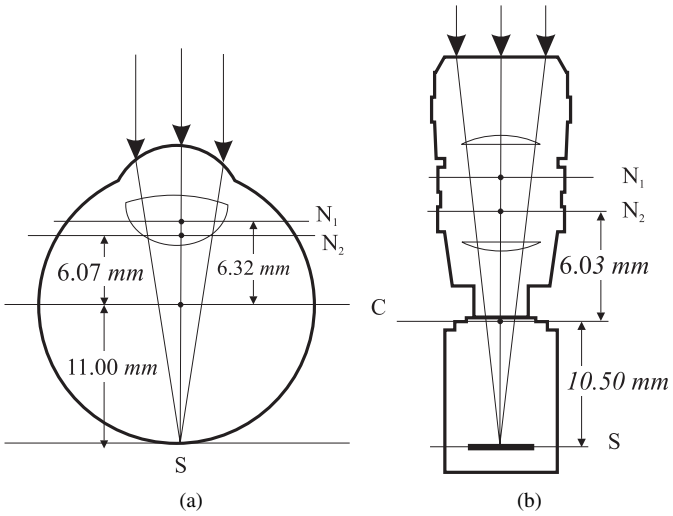


Fig. 2. Optical systems of the eye and the anthropomorphic robot used in this study. N_1 and N_2 represent the lens nodal points. C is the center of rotation. S represents the sensor plane.

In this section we analyze the oculomotor parallax of an infinitesimally small light source that produces a projection of a single point on the sensor. As shown in Fig. 3, we assume this Point Light Source (PLS) to be located in front of the sensor at distance d_A from the center of rotation C of the camera at position $A = (-d_A \sin\alpha,\ d_A \cos\alpha)$.

In the case of an optical system with a single nodal point (i.e. only the nodal point N_1), the projection of the PLS, $\tilde{x}_A$, is given by the intersection between the sensor surface and the line AN_1. This line verifies the equation:

$$x(d_A \cos\alpha - CN_1) + z\ d_A \sin\alpha - d_A\ CN_1 \sin\alpha = 0$$

In a two-nodal-point system, as in the case of the approximation of a camera with a thick lens, a ray of light going through the first nodal point, N_1, with an angle $\tilde{\alpha}$ exits the lens by the second nodal point, N_2, with the same angle. Thus, the projection of the PLS on the receptor array, x_A, is given by the intersection of the sensor surface and the line parallel to AN_1 going through N_2. This line has equation:

$$x(d_A \cos\alpha - CN_1) + z\ d_A \sin\alpha - d_A\ CN_2 \sin\alpha = 0 \quad (1)$$

From (1), considering that the sensor plane is located at $z = -d_C$, it is possible to explicitly obtain the projection x_A of the PLS on the sensor as a function of its distance d_A and eccentricity α:

$$x_A = f(d_A, \alpha) = \frac{d_A f \sin\alpha}{d_A \cos\alpha - CN_1} \quad (2)$$

where $f = d_C + CN_2$ represents the focal length of the lens.

Eq. (2) shows how the projection x_A depends on the PLS distance d_A from the center of rotation C. The origin of the oculomotor parallax lies in this dependence. Indeed, this equation is equivalent to the oculomotor parallax generated by a rotation of the camera by an angle α that brings the PLS from position A' on the optical axis to its current position A (see Fig. 3).

$$d_A = \frac{x_A\ CN_1}{x_A \cos\alpha - f \sin\alpha} \quad (3)$$

However, given a specific value x_A, there are an infinite number of possible PLS locations that verify (3). In fact, every point on the line AN_1 yields the same projection, although different points differ in their distances and eccentricities. Whereas the location of the retinal projection of the PLS, x_A, can be directly measured, d_A and α are not known.

To disambiguate the position of an object in space, this study proposes an active approach that relies on the change in x_A following a rotation of the sensor. Let x_A and x'_A indicate the projections of a PLS on the sensor before and after a rotation $\Delta\alpha$. For each of these two measurements, (2) establishes a relationship between possible values of the PLS eccentricity α and distance d_A. Since a rotation around C does not change the PLS distance d_A, the two estimates of distance obtained from (3) before and after a rotation of known amplitude $\Delta\alpha$ can be equated:

$$\frac{x_A\ CN_1}{x_A \cos\alpha - f \sin\alpha} = \frac{x'_A CN_1}{x'_A \cos(\alpha + \Delta\alpha) - f \sin(\alpha + \Delta\alpha)} \quad (4)$$

Eq. (4) yields an analytical expression of α as a function of the rotation amplitude $\Delta\alpha$ and the PLS projections:

$$\alpha = \arctan\left(\frac{x'_A\,[1 - \cos\Delta\alpha] + f \sin\Delta\alpha}{f\left[\frac{x'_A}{x_A} - \cos\Delta\alpha\right] - x'_A \sin\Delta\alpha}\right) \quad (5)$$

Substitution in (3) of the value for α obtained from (5) gives the distance of A.

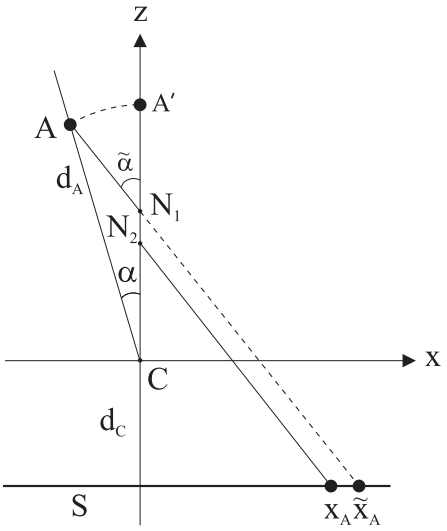


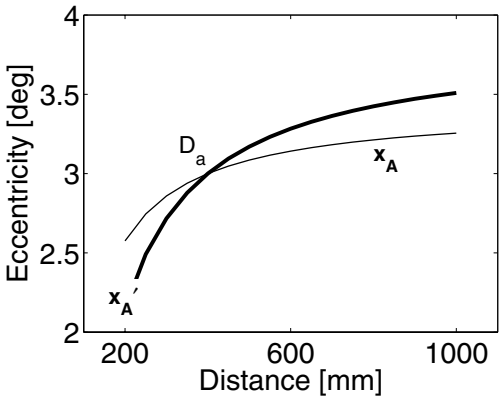
Fig. 3. Oculomotor parallax in the robot's camera. x_A identifies the projection on the sensor of a PLS at distance d_A and eccentricity α.

Fig. 4a provides an example of the two curves verifying (2) before and after a rotation of $3°$. These graphs were obtained by simulating the projection of a PLS located at an eccentricity of $3°$ and a distance of $40\ cm$. Each curve represents all possible combinations of values distance-eccentricity that produced either x_A or x'_A. Since a rotation around C does not change the distance between the camera and the PLS, the two curves intersect at a single point D_A. This point identifies the spatial coordinates of the PLS.

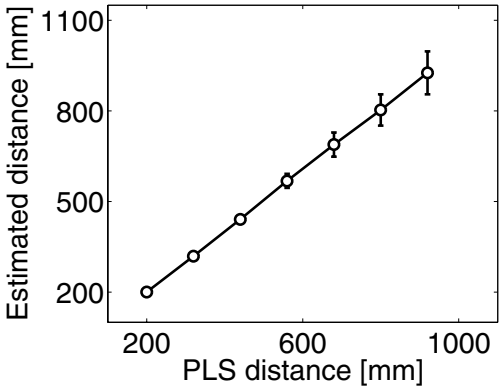
Fig. 4b shows the accuracy of the method in simulations that replicated the robotic system used in our experiments. The projections of a PLS before and after a rotation of the camera were simulated by means of (2). The eccentricity and distance of the PLS were recovered by means of (5) and (3). To examine the impact of measurement errors, random Gaussian noise with zero mean and $5\ \mu m$ standard deviation was superimposed on the parallax. Each data point represents an average over 100 repetitions of the experiment, each performed a rotation of $3°$. The estimated distance of the target is plotted as a function of its simulated distance.

As shown by these data, the oculomotor parallax provides extremely accurate estimation of distance for targets located within a near range of distances. In these simulations, the estimates varied from $20.0 \pm 0.4\ cm$ for a target located at $20\ cm$ to $102.4 \pm 18.9\ cm$ for a target at $100\ cm$. That is, the mean percentage error in evaluating the absolute distance of a target at $100\ cm$ was only 2%. As illustrated by Fig. 4b, the range of applicability of the method depends on the accuracy with which the parallax can be measured. As in the case of the camera, the impact of measurement noise becomes more pronounced for distant targets that produce a small parallax.

It is important to observe that in real-world applications large rotations do not necessary result in more accurate estimates of distance. Large rotations introduce a significant degree of distortion in the way objects appear in images



(a)



(b)

Fig. 4. Distance estimation on the basis of the oculomotor parallax. Data are results from simulations in which the distance of a PLS target was estimated by rotating the camera by $3°$. (*a*) The intersection between the loci of all possible points producing the measured projections on the sensor before and after a rotation disambiguates the distance of the target. (*b*) Application of the proposed method to estimating the position of targets at various distances from the robot. Error bars represent one standard deviation. Error bars are not visible for nearby targets due to the small variability of the measurements.

acquired before and after the movements. This distortion complicates the identification of corresponding features and results in a high level of measurement noise. In practice, the range of rotation amplitudes that provide useful distance information is bounded by the resolution of the sensor on one side and the solution of the correspondence problem on the other. On one extreme, very small rotations produce parallaxes that are too tiny to be reliably detected. On the other, large rotations produce parallaxes that cannot be reliably measured.

III. Robotic setup

The experiments of this study were conducted using the humanoid robot shown in Fig. 5. The head/eye system of this robot, referred to in this work as the "robotic oculomotor system", was developed to replicate the visual input signals to the human retina during eye movements. The oculomotor system consisted of two mobile cameras, each with two degrees of freedom. Motion was provided by two pan-tilt units (Directed Perception, Burlingame, CA) supported by a custom

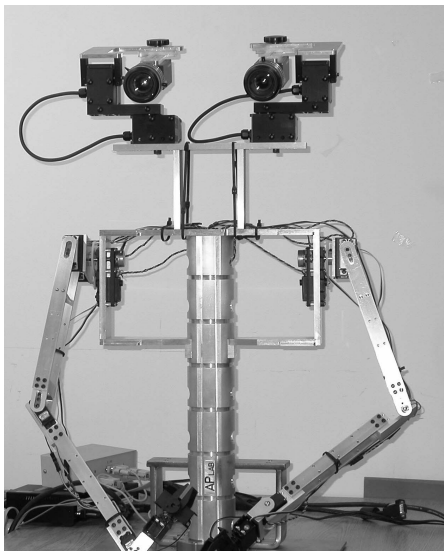


Fig. 5. The robot used in the experiments. The vision system of a humanoid robot replicated the oculomotor parallax present in the human eye. In this system, the relative positions of nodal points, center of rotation, and sensor surface followed the arrangement of the eye as shown in Fig. 2.

designed aluminum frame. Units were digitally controlled by proprietary microprocessors that ensured movements with a precision higher than $1'$. This degree of accuracy is comparable to the spatial resolution of the eyetracker used to acquire the sequences of eye movements that controlled the robot in the experiments (see Section IV). By means of a nodal adapter, the two units were mounted so that the pan and tilt axes of rotation intersected at a point C. Specifically designed aluminum wings enabled positioning of the center of rotation between the sensor plane S and the nodal points of the camera, as occurs in the human eye. In this study, d_C (the distance between the center of rotation C and the sensor S) was set to $10.5\ mm$, which is close to the value of $11\ mm$ measured in the human eye.

The system was equipped with two digital cameras (Pulnix, Sunnyvale, CA) each with a $11.5-69\ mm$ zoom lens. Cameras possessed 640×484 CCD sensors (photoreceptor size $9\ \mu m$) and allowed an acquisition rate of $120\ frames/s$. Images were acquired by means of a fast frame-grabber (Datacube, Danvers, MA). Only one of the two mobile cameras was used in the experiments described in this work.

In a zoom lens, the positions of the nodal points depend on the focal length. It should be observed that in both the eye and the camera, the projection of the stimulus on the receptor surface is strongly sensitive to the position of N_2 and relatively unaffected by the location of N_1. Indeed, in (2), CN_1 appears at the denominator together with the much larger term d_A. For this reason, by means of a preliminary calibration procedure, we estimated the locations of N_1 and N_2. We chose the focal length that positioned N_2 as close as possible to the location specified in Gullstrand's eye model of $CN_2 = 6.03\ mm$. Selection of a focal length $f = 16\ mm$ yielded $CN_2 = 5.72\ mm$, thus producing an oculomotor parallax very similar to that present in the human eye.

IV. Experimental results

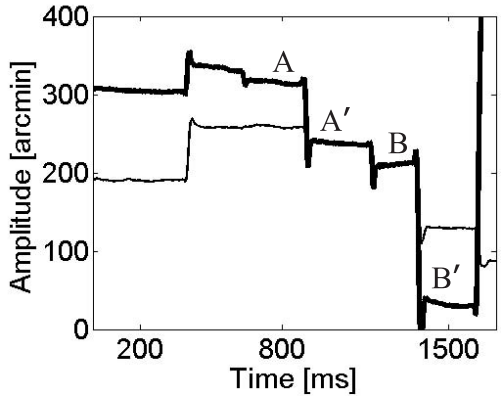
In a series of experiments we examined the depth information emerging when the robot replicated sequences of human eye movements. In this study we focused on rotations around the pan axis, as there is ample evidence that humans rely mainly on horizontal disparities in their evaluation of distance. However, since the cardinal points of the model in Fig. 3 are located on the optical axis z, the method can be applied to both the horizontal and vertical displacements produced by rotations along the pan and tilt axes.

Eye movements were recorded by means of a Dual Purkinje Image (DPI) eyetracker (Fourward Technologies, Buena Vista, VA). This high-resolution device (see Fig. 6a) estimates rotations of the eye by measuring differences in the first and fourth corneal reflections (the Purkinje images) of an infrared beam. It achieves spatial and temporal resolutions of $1'$ and $1\ ms$, respectively. subjects viewed the scene with the right eye, while the left eye was covered by an opaque eyepatch. Eye movement data were first low-pass filtered and then sampled at $1\ kHz$. The eye movements of a subject were recorded for a period of $10\ s$ while viewing the same scene presented to the robot. Recorded eye movements (shown in Fig. 6b) were used as motor commands for the robot. In transforming the oculomotor signals into motor commands, the velocity of movement was reduced by properly expanding the time axis. In this way, it was possible to operate the robot within a range of velocities that could be achieved reliably, without jeopardizing the spatial accuracy with which eye movements were reproduced. A preliminary calibration in which the subject fixated on a number of predetermined points in the scene was used to find a linear correspondence between voltages generated by the eyetracker and motor signals fed to the robot. This calibration ensured that the images acquired by the camera were centered on the points fixated by the subject.

The oculomotor parallax at various locations in the scene was evaluated on an array of 20×20 rectangular patches, each composed of 32×24 pixels. Normalized cross-correlation of each patch with the images sampled before and after rotations of the camera (points A and A', B and B' in Fig. 6) provided an estimate of the oculomotor parallax at the patch location. Every sample of the oculomotor parallax was subsequently converted into an estimate of distance by means of the robot model (Eqs. 2-5). In this way, an estimate of distance was obtained for a number of equispaced locations in the image.

In this study, correlations were estimated before and after camera rotations larger than 1°. However, the high spatial sensitivity of the DPI eyetracker allowed also discrimination of the small rotations that are not visible with most other eye-tracking systems [21]. Because of their small amplitude, replicating these fixational saccades by means of a robotic system is a challenging task. In the absence of sophisticated image processing, the smallest parallax that can be reliably discriminated in the images acquired by the robot is determined

(a)



(b)

Fig. 6. Measuring the eye movements of human subjects. (a) A high-resolution Dual-Purkinje-Image eyetracker recorded oculomotor activity while subjects examined the same scene presented to the robot. The scene was viewed monocularly with the right eye, while the left eye was covered by an opaque eye-patch. Head movements were prevented by means of a chin-rest. (b) Example of recorded eye movements (thick line: horizontal displacement; thin line: vertical displacement). The oculomotor parallax was evaluated on the basis of pairs of images acquired in correspondence of saccades larger than 1°. The letters mark the instants in time at which images were acquired.

by the size of photoreceptors in the camera. On the contrary, the resolving power of the visual system goes far beyond the limit set by the size of retinal receptors. Humans can reliably detect misalignments as small as $2 - 5''$ even though the theoretical resolution given by receptor spacing in the eye is approximately $50''$. Enlarging fixational eye movements by a factor that compensates for this difference between the camera and the eye clearly defeats the purpose of replicating fixational saccades in the robot, as it transforms these movements into large rotations. For this reason, fixational saccades were not considered in this study.

Fig. 7 shows an example of scene used in the experiments. This scene was composed of 5 objects located at various distances. Visual occlusions, together with similarities in the colors and textures of the objects made segmentation of the scene difficult. Fig. 8 shows estimates of distance obtained on the basis of the oculomotor parallax. The mean amplitude of these saccades in the recorded sequence was $4.76 \pm 2.60^\circ$ ($N = 9$). Data points in Fig. 8b represent the mean distance

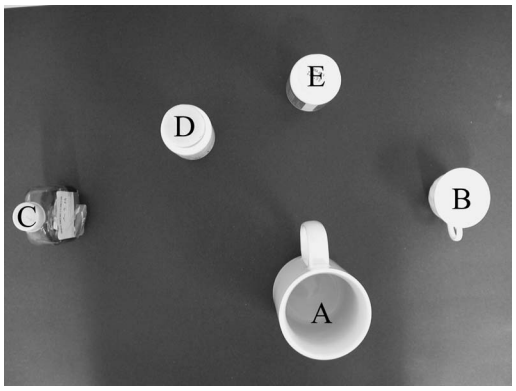


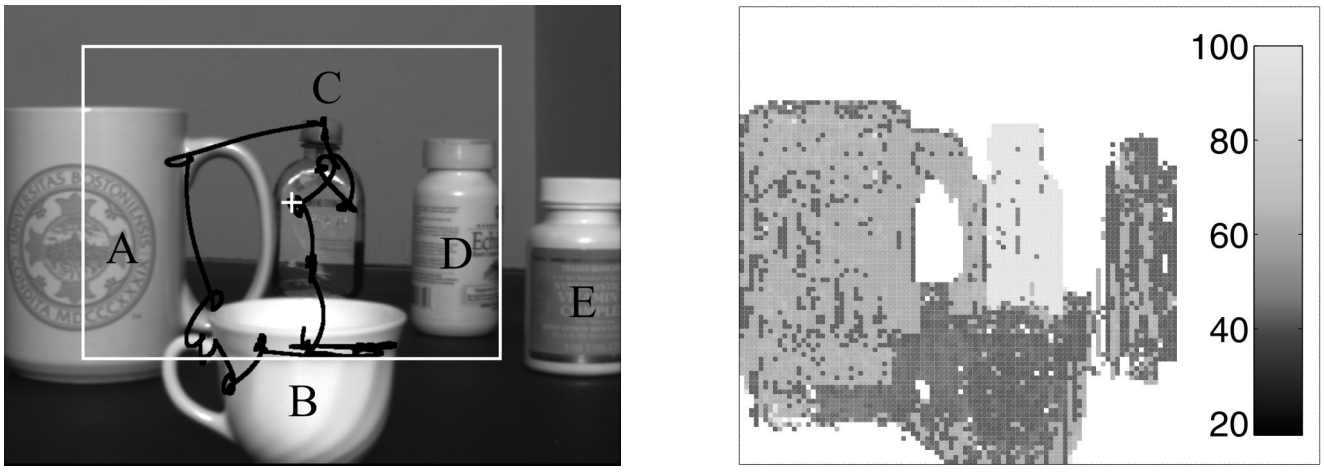
Fig. 7. Estimating distance with sequences of human eye movements. Several objects were placed at various distances in front of the robot to compose a complex scene. The distances of the various objects were: $A = 59\ cm$; $B = 43\ cm$; $C = 88\ cm$; $D = 74\ cm$; $E = 64\ cm$.

estimated for each object. Since the oculomotor parallax could be measured only for saccades that maintained an object within the field of view of the camera, averages were evaluated over a different number of measurements for each object. Means were calculated over time (the 9 saccades composing the sequence of eye movements shown in Fig. 8a). Given the limited field of view of the camera, the recorded sequence of eye movements did not allow estimation of the oculomotor parallax for object E, which was located at the margins of the scene. The data in Fig. 8b show that the method produced good estimates of distances when a sufficient number of measurements was available. The white areas in Fig. 8b correspond to the uniform surfaces of the table and the background, which did not produce measurable parallax. These data show that the relocations of gaze that characterize the normal oculomotor activity of human subjects produce depth information that facilitates the visual segregation of individual objects.

V. Conclusions

A critical function of a visual system is to estimate the distance of objects and surfaces with respect to the agent. In this study we have shown that the oculomotor parallax produced by rotations of the camera is a powerful cue for depth estimation. In a system that actively scans a visual scene, the local estimates of distance provided by the oculomotor parallax can be directly integrated with processes based on other visual cues to produce robust algorithms of image segmentation.

Two important features differentiate this study from previous research on depth from motion. A first novel aspect regards the specific cue used to estimate distance. While previous publications have analyzed the motion parallax that emerges from large relocations of a vision system, this is the first study to explicitly focus on the parallax produced by camera rotations. This is an important difference, given that such a parallax is present in virtually every head/eye system used in robotics. The results of this paper show that this parallax provides accurate depth information within the



(a) (b)

Fig. 8. Estimating distance with sequences of human eye movements. (*a*) A trace of recorded eye movements (dark lines) is shown superimposed to the scene. The white rectangle represents the field of view of the camera. (*b*) Segmentation of the visual scene based on information of distance. The distance of the objects from the robot is expressed in centimeters.

space in proximity to the agent. This cue can be reliably used to control local motor interactions with the scene, as in the operations of object reaching and grasping.

A second critical difference with respect to previous studies is the emphasis of this research on emulating human oculomotor behavior. While a motion parallax occurs in most head/eye systems, this study analyzed the case of a robot specifically designed to reproduce the oculomotor parallax present in the human eye. Control of this system with recorded traces of eye movements enabled direct replication of human oculomotor activity. Under natural viewing conditions humans tend to relocate their gaze by small shifts. Most saccades have amplitudes within a few degrees. This study shows that such a scanning strategy produces oculomotor parallaxes that can be reliably measured and used to predict the distances of objects and surfaces.

It is often debated whether a dynamic relocation of the direction of gaze presents advantages over the static analysis provided by a stationary camera with a wide field of view. As shown by this study, an active system can exploit the 3D information produced by the oculomotor parallax to simplify challenging visual tasks such as figure-ground segregation and image segmentation.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation grant CCF-0432104.

REFERENCES

[1] H. von Helmholtz, *Treatise on Physiological Optics*. Dover, New York, U.S.A.: J. P. C. Southall (Ed.), 1909/1962.

[2] E. Gibson, *et al.*, "Motion parallax as a determinant of perceived depth," *J. Exp. Psychol.*, vol. 58, pp. 40–51, 1959.

[3] J. Pettigrew, S. Collin, and M. Ott, "Convergence of specialised behaviour, eye movements and visual optics in the sandlance (teleostei) and the chameleon (reptilia)," *Curr. Biol.*, vol. 9, no. 8, pp. 421–424, 1999.

[4] M. Land, "Fast-focus telephoto eye," *Nature*, vol. 373, pp. 658–659, 1995.

[5] I. Hadami, G. Ishai, and M. Gur, "Visual stability and space perception in monocular vision: Mathematical model," *J. Opt. Soc. Am. A*, vol. 70, pp. 60–65, 1980.

[6] A. Mapp and H. Ono, "The rhino-optical phenomenon: Ocular parallax and the visible field beyond the nose," *Vision Res.*, vol. 26, pp. 1163–1165, 1986.

[7] G. Bingham, "Optical flow from eye movement with head immobilized: Ocular occlusion beyond the nose," *Vision Res.*, vol. 33, no. 5/6, pp. 777–789, 1993.

[8] N. Ayache, *Artificial Vision for Mobile Robots: Stereo Vision and Multisensory Perception*. Cambridge, Massachusetts, U.S.A.: MIT Press, 1991.

[9] O. Faugeras, *The Geometry of Multiple Images: The Laws that Govern the Formation of Multiple Images of a Scene and Some of Their Applications*. Cambridge, Massachusetts, U.S.A.: MIT Press, 2001.

[10] J. Weng, T. Huang, and N. Ahuja, *Motion and Structure From Image Sequences*. New York, U.S.A.: Springer-Verlag, 1993.

[11] J. Oliensis, "A critique of structure-from-motion algorithms," *Comput. Vis. Image Und.*, vol. 80, no. 2, pp. 172–214, 2000.

[12] A. Torralba and A. Oliva, "Depth estimation from image structure," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 9, pp. 1226–1238, 2002.

[13] B. Super and A. Bovik, "Shape from texture using local spectral moments," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 17, no. 4, pp. 333–343, 1995.

[14] R. Zhang, *et al.*, "Shape from shading: A survey," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 21, no. 8, pp. 690–706, 1999.

[15] A. Pentland, "A new sense for depth of field," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 9, pp. 523–531, 1987.

[16] G. Sandini and M. Tistarelli, "Active tracking strategy for monocular depth inference over multiple frames," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 12, pp. 13–27, 1990.

[17] A. Dalmia and M. Trivedi, "High-speed extraction of 3D structure of selectable quality using a translating camera," *Comput. Vis. Image Und.*, vol. 64, no. 1, pp. 97–110, 1996.

[18] Y. Hung and H. Ho, "A Kalman filter approach to direct depth estimation incorporating surface structure," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 21, no. 6, pp. 570–575, 1999.

[19] Y. Aloimonos and Z. Duric, "Estimating the heading direction using normal flow," *I. J. Comp. Vision*, vol. 13, pp. 33–56, 1994.

[20] G. Adiv, "Determining three-dimensional motion and structure from optical flow generated by several moving objects," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 7, no. 4, pp. 384–401, 1985.

[21] R. Steinman, *et al.*, "Miniature eye movement," *Science*, vol. 181, pp. 810–819, 1973.