2014 IEEE International Conference on Robotics & Automation (ICRA)
Hong Kong Convention and Exhibition Center
May 31 - June 7, 2014. Hong Kong, China

Bayesian Multimodal Integration in a Robot Replicating Human Head and Eye Movements

Marco Antonelli[1], Angel P. del Pobil [1] and Michele Rucci[2]

Abstract— **Autonomous robots need to make sense of their surroundings, recognize objects, detect and, possibly, identify the people around them. Visual perception would be ideally suited for these tasks. Yet, visual perception remains one of the limiting factors of modern robotics: artificial vision systems tend to perform poorly in normal environments, where the scene and the illumination conditions are unpredictable. Evolution has faced similar problems, leading to surprisingly accurate visual capabilities, even in species with very small brains. We argue that the success of biological perception systems relies on three fundamental computational principles: (a) the continual coupling of perception and behavior; (b) the resulting emergence of multimodal cues; (c) and their efficient integration. Building on these computational principles, we describe a humanoid robot that emulates the dynamic strategy by which humans examine a visual scene. The proprioceptive and visual depth cues resulting from this strategy are integrated in statistically optimal manner into a unified representation. We show that this approach yields accurate and robust 3D representations of the observed scene.**

I. Introduction

To efficiently operate in unstructured environments, autonomous robots need to establish reliable and robust representations of the surrounding space. Vision is ideally suited for this task, as light carries rich information about the surfaces it interacts with, and many organisms, including humans, make extensive use of visual information.

However, the process of reconstructing a 3D scene from its projection on a 2D sensor is inherently ambiguous. Depth information is not explicitly available in the images acquired by a camera, but needs to be computed on the basis of cues which represent it implicitly. Extraction of these cues is challenging, as they vary tremendously with the scene and viewing conditions, and they often tend to confound each other when considered individually. As a consequence, successful machine vision applications have typically been restricted to highly controlled environments, where relevant cues can be predicted.

Nature has faced similar problems in the development of organisms, and evolution has led to systems that are extremely efficient in processing visual information. While multiple differences exist between the typical architectures of robotic vision systems and the visual mechanisms of organisms, research in sensory neuroscience has emphasized the importance of three fundamental computational principles. These principles appear to be essential ingredients for successful visual processing in organisms, but are rarely followed in machine perception:

1. Vision and behavior are indissolubly coupled. Unlike many computer vision systems, organisms are not passively exposed to the incoming flow of sensory data, but actively seek useful information by coordinating sensory processing with motor activity. Behavior is always present, even when it is not immediately obvious. In humans and many other species, microscopic head and eye movements occur even in the periods of "visual fixation" the brief intervals (≈300 ms) in between macroscopic relocations of gaze, in which humans acquire visual information. Recent work has shown that these microscopic movements operate a critical reformatting of the spatiotemporal stimulus on the retina [1], are under oculomotor control [2], and contribute to the processing of visual information and the establishment of spatial representations [3], [4], [5].

2. Vision relies on the integration of multiple cues. Although a variety of methods for extracting 3D information have been developed in machine vision, surprisingly little work has been dedicated to the integration of these techniques. As a consequence, machine vision systems often focus on individual cues, an approach bounds to be fragile. In contrast, depth perception in humans and primates is an extremely parallel process which relies on the simultaneous processing of more than 20 visual cues. Importantly, these cues are not restricted to the visual modality. The close coupling between behavioral and visual processes yields depth information directly in the motor and proprioceptive modalities (*e.g.*, vergence and focus adjustments).

3. Cue integration tends to be optimal. A substantial body of work shows that humans often follow a strategy of statistical optimal integration [6], [7], [8], [5]. This approach implies that sensory processes not only extract the relevant cues, but also estimate their reliability on the basis of previously acquired knowledge. It implies a probabilistic model of the environment, which needs to be modified during the course of experience [9]. Similar optimal approaches have been used in robotics and computer vision [10], [11], [12] , but rarely integrates cues from visual, motor, and proprioceptive signals [13], as humans appear to do.

In a series of recent studies, we have focused on the close coupling between perception and action and examined the 3D information that emerges from replicating human head and

This work was supported in part by Ministerio de Ciencia y Innovación (FPI grant BES-2009-027151, DPI2011-27846), by Generalitat Valenciana (PROMETEO/2009/052),by Fundació Caixa-Castello-Bancaixa (P1-1B2011-54), and by NSF grants BCS-1127216 and CCF-0726901.

[1] M. Antonelli and A.P. del Pobil are with Robotic Intelligence Lab, Universitat Jaume I, 12070 Castellón, Spain. antonell,pobil@uji.es

[2] M. Rucci is with Department of Psychological and Brain Sciences and the Graduate Program in Neuroscience, Boston University, Boston, MA 02215, USA. mrucci@bu.edu

eye movements in robots. We have first focused on small camera rotations similar to the fixational eye movements continually performed by humans [14], [3], [15], [16], and, more recently, integrated these movements with small platform rotations similar to fixational head movements [17], [18]. Here, we extend this previous work by coupling the active fixation strategy followed by humans with a processing architecture based on the two other computational principles, the Bayesian integration of the available visual, motor, and proprioceptive cues. The result is a humanoid robot that replicates humans both in the way they acquire and process visual information. We show that the approach leads to robust 3D representations in the peripersonal space.

II. Visual input and motion equations

Let us consider an ideal point light source placed at the position $\vec{P} = [X, Y, Z]^T$, where coordinates are given in a reference frame centered on the camera's optical point and oriented in such a way that the z-axis coincides with the optic axis. Modeling the camera as a pinhole system with focal length f, the point $\vec{P}$ projects on the sensor surface at coordinates $\vec{p} = [x, y]^T = \frac{f}{Z} \cdot [X, Y]^T$.

During active fixation, small compensatory movements of the cameras and neck replicating those performed by humans result in the translational and angular velocities of the optical point $\vec{v} = [v_x, v_y, v_z]^T$ and $\vec{\omega} = [\omega_x, \omega_y, \omega_z]^T$. These movements shifts the image on the camera, so that the point $\vec{P}$ moves with velocity $\dot{\vec{P}} = -\vec{v} - \vec{\omega} \times \vec{P}$. Taking the temporal derivatives of $\vec{p}$ and writing the components of image motion in terms of the instantaneous camera velocity gives the standard optic flow equation [19]:

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} \frac{x \cdot y}{f} & -\frac{x^2+f^2}{f} & y \\ \frac{y^2+f^2}{f} & -\frac{x \cdot y}{f} & -x \end{bmatrix} \cdot \begin{bmatrix} \omega_x \\ \omega_y \\ \omega_z \end{bmatrix} + \frac{1}{Z} \cdot \begin{bmatrix} f & 0 & x \\ 0 & f & y \end{bmatrix} \cdot \begin{bmatrix} v_x \\ v_y \\ v_z \end{bmatrix} = \\ = \begin{bmatrix} r_x \\ r_y \end{bmatrix} + d \cdot \begin{bmatrix} t_x \\ t_y \end{bmatrix} \quad (1)$$

where, we have introduced the disparity d defined as the inverse of the depth Z [20]. In this equation, each component of the optic flow $(\dot{x}, \dot{y})$ is divided into two elements, r and t, which depend on the angular and translational motion of the camera, respectively. It is important to observe that Eq. (1) strictly holds for small velocities only, an assumption that is well justified for the behavior considered in this study.

III. Active maintenance of fixation

During normal fixation, humans continually perform (without being aware of them) microscopic head and eye movements. In this study, we used the APLab humanoid robot to replicate this behavior. Head movements measured in human observers during natural fixation [4] were used as input signals to the motor which controlled the neck movements of the robot. The cameras were actively controlled to counterbalance the effect of head movements and maintain fixation (see fig. 1). This approach resulted in visual input signals similar to those normally experienced by humans during fixation.

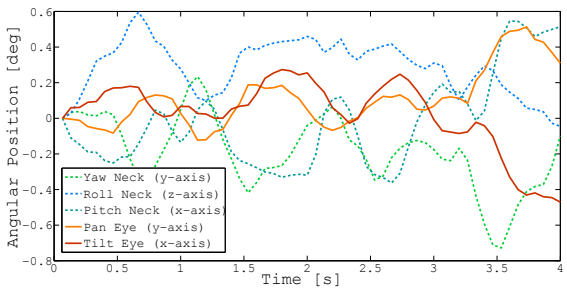


Fig. 1. An example of head and eye movements performed by the robot. The traces represent the signals fed to motors controlling the neck and the left camera.

Use of this approach, requires previous estimation of two components: the optic flow and the velocity of the camera. That is, the establishment of 3D representations requires simultaneous computation of egomotion, two problems that are often treated separately in machine vision. Again, computation of the optic flow is simplified by the small changes in the images resulting from fixational behavior, which allows the use of first order differential methods [21]. The camera velocity can be estimated on the basis of multiple cues: the control signals fed to the robot (the motor signals), the signals measured by the motor encoders (the proprioceptive signals), and/or the way the images acquired by the camera change over time (the visual signals). Since movements are small, an efficient approach requires the integration of more than one of these signals, as explained below.

IV. Bayesian multimodal integration

The model of Bayesian integration used in this study is summarized in Fig. 2. The model is based on two parallel and interacting data streams, each one processing one of the two types of sensory data considered here: the visual input and proprioception. The visual input is processed to extract the optic flow by means of a standard computer vision algorithm [22]. The proprioceptive input is processed by a kinematic model of the robot to provide a first estimate of the camera motion. These two streams are then integrated following a Bayesian approach to solve two problems: the estimation of egomotion and the establishment of a 3D representation of the scene.

A. The proprioceptive flow

This sensory path estimates the motion of the camera by combining proprioceptive cues with the kinematics model of the robot. At each time step, it first computes the homogeneous matrix describing the camera pose (M_t): the 3D position of its optical point and the orientation of the line of sight. Then, it computes the homogeneous matrix M_t^{t-1} given by the difference between the current and previous pose: $M_t^{t-1} = M_t \cdot M_{t-1}^{-1}$. Under the assumption of small

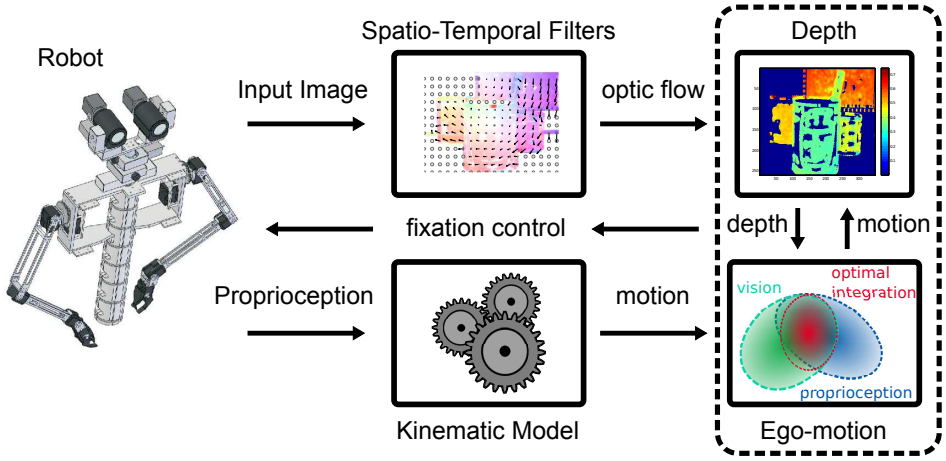


Fig. 2. System architecture. A model of optimal integration combines the multimodal depth cues which emerge during active visual fixation. A humanoid robot performs small coordinated head and eye movements emulating human fixational behavior. The resulting optic flow and proprioceptive signals are optimally integrated to estimate egomotion and establish a 3D representation of the scene.

movements, this difference can be approximated as:

$$M_t^{t-1} = \begin{bmatrix} 1 & -\omega_z & \omega_y & v_x \\ \omega_z & 1 & -\omega_x & v_y \\ -\omega_y & \omega_x & 1 & v_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (2)$$

which directly gives the translational and rotational components of the camera velocity.

Small errors in the proprioceptive data could lead to significant errors in the estimate of egomotion. For this reason, we used a Monte Carlo method to quantify the uncertainty of the estimate. We generated a normal distribution of motor positions with mean equal to the value given at any time by the encoders and variance taken from the motor specifications sheets. The resulting distribution was then filtered through the kinematic model to make a prediction of the position of the optical point and obtain the corresponding distribution. This enabled estimation of the standard deviation of the measurement.

B. The visual flow

The second sensory pathway processed the visual input to extract its optic flow as described above. The optic flow was used both for egomotion estimation as well as for establishing a 3D map of the scene.

Visual estimation of the trajectory of the optical point was obtained by tracking a set of points during fixation. This step is not conceptually necessary, but it was implemented to speed up robotic experiments, as the resulting reduction of dimensionality enabled faster control. The points (features) were selected by means of corner detection algorithm with the requirement of a sufficiently sparse distribution across the image. These corners were then tracked and monitored as proposed by Shi and Tomasi [23].

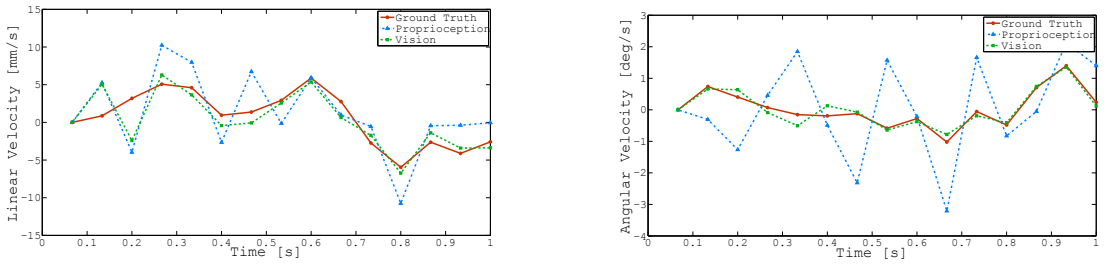
At each time step, the tracking algorithm yields the position (x_i, y_i) of each selected feature and its displacement $(\dot{x}_i, \dot{y}_i)$. If the feature disparity (d_i) is known, the problem of visually estimating ego-motion becomes linear (see Eq. (1)). During active fixation, the distance of each feature from the robot remains approximately constant. Thus, we can use the disparity estimated at the previous time step in Eq. (1) and propagate its uncertainty in the equation.

C. Bayesian integration: egomotion estimation

The Bayesian integration of visual and proprioceptive information in the estimate of egomotion relied on an extended Kalman filter, an approach similar to those of previous studies on structure from motion [24], [25] and visual mono SLAM [11], [26]. The state of the filter was given by the velocity of the camera, the visual location of the features and their disparity (see Eq. (3)). At each time step, the camera velocity was estimated on the basis of proprioception and was plugged in eq. (1) to predict the new position in the image of the tracked features. The predicted position of the features was then compared to the position provided by tracking to obtain the innovation. The Jacobian of equations (3) with respect to the state of the system was then used to update the covariance matrix of the state, which was in turn used together with the covariance of the observation (the position of the features provided by tracking) to compute the Kalman gain.

$$\begin{cases} \vec{\omega}(t) = [-M_{t-1}^t(2,3)\, M_{t-1}^t(1,3) - M_{t-1}^t(1,2)]^T \\ \vec{v}(t) = [M_{t-1}^t(1,4)\, M_{t-1}^t(2,4)\, M_{t-1}^t(3,4)]^T \\ x_1(t) = x_1(t-1) + r_{x_1}(t-1) + d_1(t-1) \cdot t_{x_1}(t-1) \\ y_1(t) = y_1(t-1) + r_{y_1}(t-1) + d_1(t-1) \cdot t_{y_1}(t-1) \\ d_1(t) = d_1(t-1) \\ \quad \dots \\ x_n(t) = x_n(t-1) + r_{x_n}(t-1) + d_n(t-1) \cdot t_{x_n}(t-1) \\ y_n(t) = y_n(t-1) + r_{y_n}(t-1) + d_n(t-1) \cdot t_{y_n}(t-1) \\ d_n(t) = d_n(t-1) \end{cases} \quad (3)$$

Figure 3 shows an example of application of the method to the estimate of both translational (v_x) and rotational (ω_x) velocities. Integration of vision and proprioception significantly improved precision. As expected, the estimation of the translational motion required more iterations to converge



(a) (b)

Fig. 3. Optimal estimation of egomotion. One component of the translational (a) and rotational (b) velocity of the camera (v_x and ω_x; red line) is plotted together with the velocity estimate resulting from proprioception (blue line) and from the optimal integration of proprioception and vision (green line).

than rotational estimates because of its dependence on the unknown depth of the scene.

D. Bayesian integration: 3D representation

The optic flow and the instantaneous velocity of the camera were integrated by an iconic Kalman filter to estimate the disparity of the scene (we used disparity rather than depth to linearize equations [20]). Because of the small amplitude of head movements, relative changes in distance are negligible, and disparity can be considered constant during fixation (see equation 4). We modeled the small error resulting from this as Gaussian noise η_d.

$$d(t+1) = d(t)+\eta_d(t) \tag{4}$$

Disparity can be estimate from the angular and translational velocities of the camera as:

$$\begin{bmatrix}\dot{x}(t)\\ \dot{y}(t)\end{bmatrix} = \begin{bmatrix}t_x(t)\\ t_y(t)\end{bmatrix}\cdot d(t)+\begin{bmatrix}r_x(t)\\ r_y(t)\end{bmatrix}+\big[R(t)\big] \tag{5}$$

where R is the covariance matrix resulting from the computation of the optic flow [27], [18].

V. Experiments and Results

A. Robotic platform and experimental procedures

The approach was tested on the APLab humanoid robot, a platform specifically designed to model the sensory inputs experienced by humans during normal oculomotor behavior [15] (Fig. 4). It consists of torso equipped with a 7 degrees of freedom head, three in the neck (controlling yaw, pitch, and roll rotations) and two in each camera (controlling pan and tilt rotations). Camera movements as well as yaw neck rotations are given by five independent stepper motors. Roll and pitch rotations are controlled by two servo-motors.

Monochromatic images were acquired by the left camera of the robot at a resolution of 1392×1040 pixels. The camera was equipped with a wide-angle lens ($15mm$) to obtain a broad field of view. Images were resized by a factor of two for computational reasons and processed to extract optic flow in a 11×11 pixels neighborhood.

The robot moved so as to replicate head trajectories previously recorded from human observers by means of a customized high-resolution motion tracking system [4]. These signals were resampled at 15 Hz and used to control

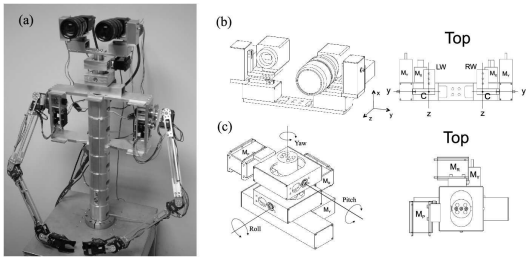


Fig. 4. The APLab humanoid robot. (a) The robot used in this study consists of a torso equipped with a head/eye platform and two arms. (b) Pan-tilt motion for each camera is given by two stepper motors Mx and My. (c) A 3 d.o.f. anthropomorphic neck with two servo motors (MP and MR) and one stepper motor (MY) enabled replication of head movements.

the robot neck motors. Fixation was maintained by the left cameras, which moved to counteract neck rotations and eliminate displacements in the image [28].

The extended Kalman filter was initialized with the values given by proprioception using the Monte Carlo approach described above (see Section IV-A) on a population of 500 points. Disparity was empirically initialized at 1 with initial covariance set to infinite because of the lack of previous depth information. The standard deviation of the optic flow was empirically set to 0.5 pixels.

B. Results

Before testing the method with the robot, we studied its sensitivity and robustness in computer simulations in which the robot and its surrounding scene were modeled by means of OpenRave [29]. We first examined the sensitivity of the method by estimating the minimum distance necessary for discriminating between two different surfaces.

Fig. 5 shows the mean of just noticeable difference for objects at different distances. Two surfaces were initially placed at the same distance D from the robot. They were then separated in depth until the difference between their estimated disparities became discriminable. The just noticeable difference threshold was defined as the depth separation at which the estimated disparity difference exceeded the sum of the two standard deviation of each disparity measurement. The distance D was then systematically varied in the range 0.3-$1.1m$. Results in Fig. 5 show that the approach is highly

sensitive within the space nearby the agent, with threshold approximatively equal to 10% of the actual distance.

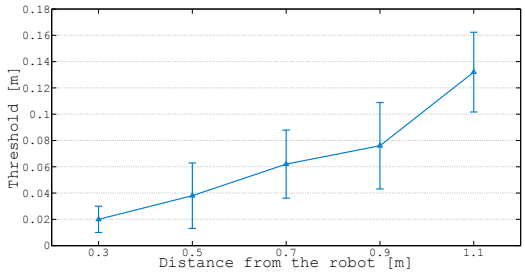


Fig. 5. Sensitivity of the method. Data points represent mean just noticeable differences (the discrimination thresholds between two objects) at different distances from the robot. Error bars represent one standard deviation.

The second simulation tested the precision of the approach. In this case, the robot was presented with a scene consisting of four objects (three mugs and a postcard) at different distances within the range 0.35-0.62m (Fig. 6).

Each data point in Fig. 7 represents an estimate of the distance of one of the objects using a different head movement trace (x axis). These data show that the method enabled consistent estimation of the object distance. Importantly, the depth order was maintained, indicating that the algorithm is capable of establishing faithful representations of the scene.

Fig. 8 shows results obtained in real robotic experiments. In these experiments the robot was presented with different visual scenes, all of them extremely difficult to segment into their constituent objects in the absence of 3D information. Each column in Fig. 8 shows an image of the scene as seen by the robot, and the established 3D map during active fixation. The first column of Fig. 8 shows results with a scene similar to the one studied in simulations. As in the simulations, the robot correctly identifies the presence of four objects on different distance planes. Notice the separation of the pen from the background, a segmentation that would be virtually impossible on the basis of the features present in a single image [30]. The central column shows an example of application of the method to the breaking of camouflage. In this case the object has the same texture of the background, yielding very effective camouflage in each individual image. Use of the parallax present during fixational behavior breaks the camouflage and reveals the object. Finally, the right column in Fig. 8 illustrates the application of the method to a complex scene with strong depth gradients. In this scene, the postcard is aligned at an angle with respect to the optical axis, yielding a depth gradient that is correctly detected by the robot. Again the system correctly identifies the depth order of all objects in the scene.

VI. Conclusions

While performing tasks that require 3D visual information, humans and other primates do not rely on any single individual cue. Rather, they integrate multiple cues into a unified percept of 3D space, and the active mechanisms by

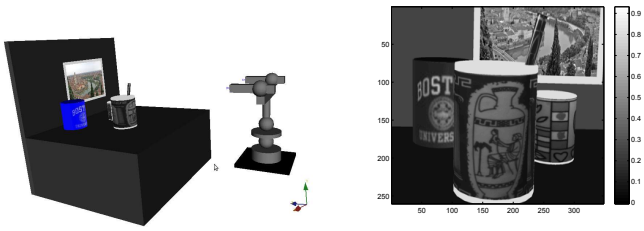


Fig. 6. Simulated environment. (a) Computer simulations modeled both the agent and the scene. (b) An example of a scene as viewed by the agent.

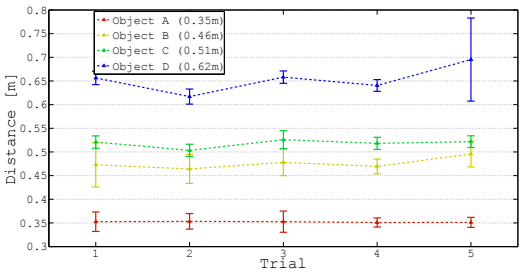


Fig. 7. Accuracy of the method. Estimation of the distance of four objects in the scene. Each trial consisted in the replication of a different trace of head/eye movements recorded from human observers.

which this integration occurs have now started to emerge. Here, we have shown some of the benefits that derive from following similar strategies in a robot that replicates the fixational behavior of humans and optimally integrates the resulting proprioceptive and visual cues. This approach regards the two problems of estimating egomotion and establishing 3D representations as two faces of the same coin and addresses them simultaneously. The resulting tight integration of sensory perception and motor control provides a 3D representation of the scene which is maintained into an image-centered frame of reference. This representation facilitates image segmentation and figure/ground segregation, and can be easily integrated with other visual cues, such as stereopsis, which provides depth information in the same frame of reference.

References

[1] X. Kuang, M. Poletti, J. D. Victor, and M. Rucci, "Temporal encoding of spatial information during active visual fixation," *Current Biology*, vol. 22, no. 6, pp. 510–514, PMCID: PMC3 332 095, 2012.

[2] H. K. Ko, M. Poletti, and M. Rucci, "Microsaccades precisely relocate gaze in a high visual acuity task," *Nature Neuroscience*, vol. 13, pp. 1549–1553, 2010.

[3] F. Santini and M. Rucci, "Active estimation of distance in a robotic system that replicates human eye movement," *Robotics and Autonomous Systems*, vol. 55, no. 2, pp. 107–121, 2007.

[4] M. Aytekin and M. Rucci, "Motion parallax from microscopic head movements during visual fixation," *Vision Research*, vol. 70, pp. 7–17, 2012.

[5] M. Poletti, C. Listorti, and M. Rucci, "Microscopic eye movements compensate for nonhomogeneous vision within the fovea," *Current Biology*, vol. 23, no. 17, pp. 1691–1695, 2013.

[6] D. C. Knill and A. Pouget, "The Bayesian brain: The role of uncertainty in neural coding and computation," *Trends in Neurosciences*, vol. 27, no. 12, pp. 712–719, 2004.

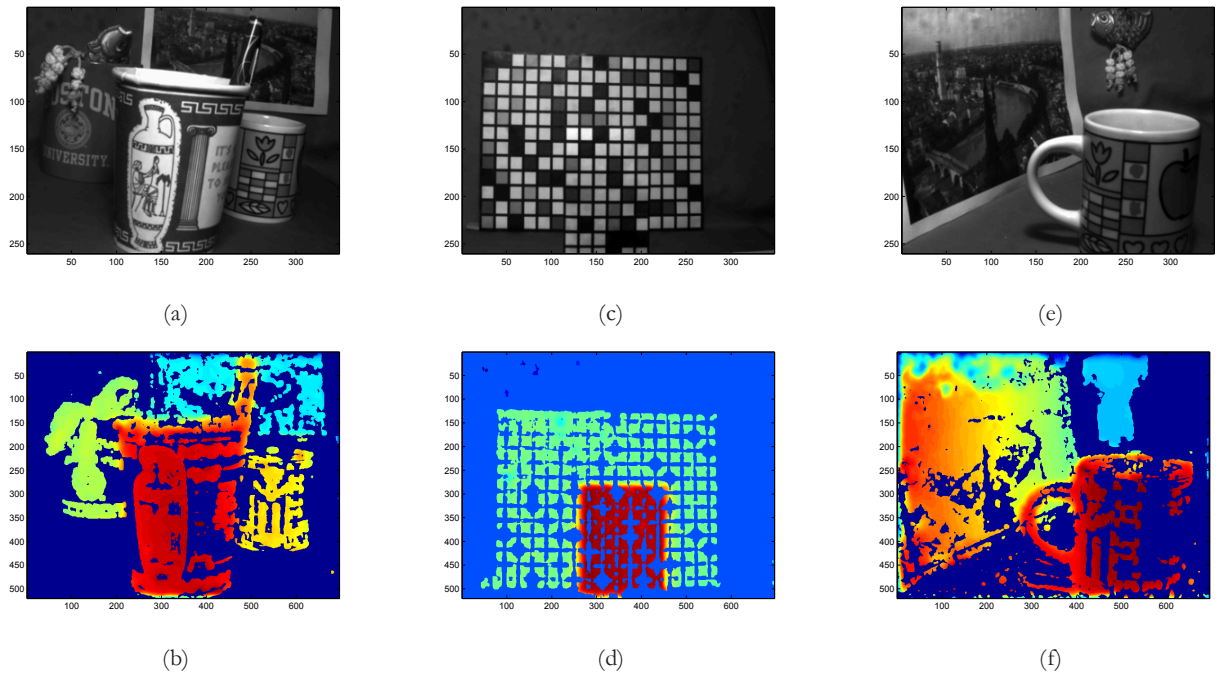


Fig. 8. Results from robotic experiments. Each column compares a scene presented to the robot (top) to the estimated map of distance obtained during active fixation (bottom). (a-b) An example with a scene similar to the one studied in simulations. As in the simulations, the robot correctly identifies the presence of four objects on different distance planes. (c-d) An example of breaking of camouflage and figure/ground segregation. The object is not immediately visible in a single image, but emerges during motor activity. (e-f) An example with a complex scene with depth gradients. The postcard is aligned at an angle with respect to the optical axis, yielding a depth gradient that is correctly detected by the robot.

[7] T. C. A. Freeman, R. A. Champion, and P. A. Warren, "A Bayesian model of perceived head-centered velocity during smooth pursuit eye movement," *Current Biology*, vol. 20, no. 8, pp. 757–762, 2010.

[8] W. S. Geisler, "Contributions of ideal observer theory to vision research," *Vision Research*, vol. 51, no. 7, pp. 771–781, 2011.

[9] J. Fiser, P. Berkes, G. Orbn, and M. Lengyel, "Statistically optimal perception and learning: from behavior to neural representations," *Trends in Cognitive Sciences*, vol. 14, no. 3, pp. 119–130, 2010.

[10] D. M. Wolpert, Z. Ghahramani, and M. I. Jordan, "An internal model for sensorimotor integration," *Science*, vol. 269, no. 5232, pp. 1880–1882, 1995.

[11] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "Monoslam: Real-time single camera slam," *Pattern Anal. Machine Intell., IEEE Trans. on*, vol. 29, no. 6, pp. 1052–1067, 2007.

[12] P. R. Schrater and D. Kersten, "How optimal depth cue integration depends on the task," *International Journal of Computer Vision*, vol. 40, no. 1, pp. 71–89, 2000.

[13] J. F. Ferreira, J. Lobo, P. Bessiere, M. Castelo-Branco, and J. Dias, "A bayesian framework for active artificial perception," *Cybernetics, IEEE Transactions on*, vol. 43, no. 2, pp. 699–711, 2013.

[14] F. Santini and M. Rucci, "Depth perception in an anthropomorphic robot that replicates human eye movements," in *IEEE International Conference on Robotics and Automation*, Orlando, FL, May 2006.

[15] M. Rucci, D. Bullock, and F. Santini, "Integrating robotics and neuroscience: brains for robots, bodies for brains," *Advanced Robotics*, vol. 21, no. 10, pp. 1115–1129, 2007.

[16] F. Santini, R. Nambisan, and M. Rucci, "Active 3d vision in a humanoid robot," *Int. J. Hum. Robot.*, vol. 6, no. 3, pp. 481–503, 2009.

[17] X. Kuang, M. Gibson, B. E. Shi, and M. Rucci, "Active vision during coordinated head/eye movements in a humanoid robot," *Robotics, IEEE Trans. on*, vol. PP, no. 99, pp. 1 –8, 2012.

[18] M. Antontelli, A. del Pobil, and M. Rucci, *Depth Estimation during Fixational Head Movements in a Humanoid Robot*, ser. Lect. Notes Comput. Sc. Springer Berlin Heidelberg, 2013, vol. 7963, pp. 264–273.

[19] L. H. C. Higgins and K. Prazdny, "The Interpretation of a Moving Retinal Image," *Proceedings of the Royal Society of London. Series B, Biological Sciences (1934-1990)*, vol. 208, no. 1173, pp. 385–397, 1980.

[20] L. Matthies, T. Kanade, and R. Szeliski, "Kalman filter-based algorithms for estimating depth from image sequences," *International Journal of Computer Vision*, vol. 3, no. 3, pp. 209—238, 1989.

[21] J. L. Barron, D. J. Fleet, and S. S. Beauchemin, "Performance of optical flow techniques," *International Journal of Computer Vision*, vol. 12, no. 1, pp. 43–77, 1994.

[22] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," 1981, pp. 674–679.

[23] J. Shi and C. Tomasi, "Good features to track," in *PROC CVPR IEEE*. IEEE, 1994, pp. 593–600.

[24] A. Azarbayejani and A. Pentland, "Recursive estimation of motion, structure, and focal length," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 17, no. 6, pp. 562–575, 1995.

[25] A. Chiuso, P. Favaro, H. Jin, and S. Soatto, "Structure from motion causally integrated over time," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 4, pp. 523–535, 2002.

[26] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, "Dtam: Dense tracking and mapping in real-time," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2320–2327.

[27] E. Simoncelli, E. Adelson, and D. Heeger, "Probability distributions of optical flow," in *PROC CVPR IEEE*. IEEE, 1991, pp. 310–315.

[28] F. Chaumette and S. Hutchinson, "Visual servo control, part I: Basic approaches," *IEEE Robot. Autom. Mag.*, vol. 13, pp. 82–90, 2006.

[29] R. Diankov and J. Kuffner, "Openrave: A planning architecture for autonomous robotics," *Robotics Institute, Pittsburgh, PA, Tech. Rep. CMU-RI-TR-08-34*, 2008.

[30] P. Sundberg, T. Brox, M. Maire, P. Arbeláez, and J. Malik, "Occlusion boundary detection and figure/ground assignment from optical flow," in *PROC CVPR IEEE*. IEEE, 2011, pp. 2233–2240.