Depth Estimation during Fixational Head Movements in a Humanoid Robot

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

[1] Robotic Intelligence Lab, Universitat Jaume I
12070 Castellón, Spain
{antonell,pobil}@uji.es
[2] Department of Psychology and Graduate Program in Neuroscience
Boston University, Boston, MA 02215, USA
mrucci@bu.edu

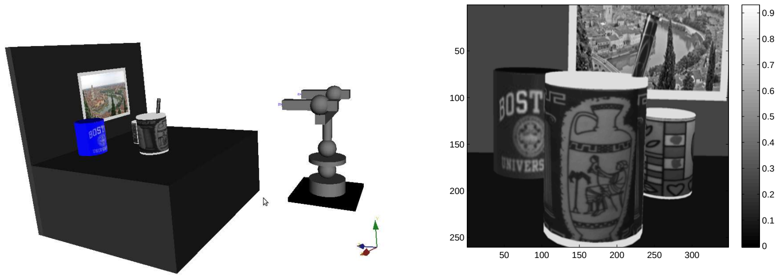
Abstract. Under natural viewing conditions, humans are not aware of continually performing small head and eye movements in the periods in between voluntary relocations of gaze. It has been recently shown that these fixational head movements provide useful depth information in the form of parallax. Here, we replicate this coordinated head and eye movements in a humanoid robot and describe a method for extracting the resulting depth information. Proprioceptive signals are interpreted by means of a kinematic model of the robot to compute the velocity of the camera. The resulting signal is then optimally integrated with the optic flow to estimate depth in the scene. We present the results of simulations which validate the proposed approach.

1 Introduction

Accurate 3D judgments are critical in many computer vision tasks, from visuomotor control of robots to object recognition. Unfortunately, extraction of depth and distance is a complex operation, as this information is lost when the three-dimensional world is projected onto the two dimensional surface of a camera sensor during the process of image acquisition. To circumvent this problem, many techniques have been proposed, but no optimal solution exists, as each method presents both pros and cons. For example, stereopsis—arguably the most common 3D approach in computer vision [2, 7]—requires solving a correspondence problem: the determination of the positions of identical features in the images acquired from cameras at different locations. Decades of research have shown that this is an extremely challenging operation.

A popular 3D approach in computer vision as in biology is depth (or structure) from motion [10], the use of depth/distance information that emerges in a moving agent. The underlying principle is similar to that of stereopsis, but,

* 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) and by Fundació Caixa-Castello-Bancaixa (P1-1B2011-54).



(a) (b)

Fig. 1. Simulated environment. (a) An anthropomorphic head/eye system observes a scene composed by three objects at different distances. (b) The scene as viewed by the robot's camera.

in this case, separate views of the scene are obtained from the same camera at different instants in time, rather than from multiple cameras as in stereo-vision. In this case, the underlying cue to extract is no longer disparity, but motion parallax [12], and if the inter-frame movement of the camera is sufficiently small, the correspondence problem becomes an estimation of the optic flow in the temporal sequence [4]. The use of this cue has the advantages of (a) enabling 3D vision with a single camera; (b) eliminating the need for precise alignment of multiple cameras and complex calibration procedures; and (c) building upon an extensive literature on optic flow computation, which can be directly applied to the estimation of motion parallax.

Research on depth from motion or visual SLAM (simultaneous localization and mapping) has historically focused almost exclusively on relatively large movements of the agent [13, 1, 5, 11]. However, in humans, useful motion parallax also emerges during much smaller movements, such as the minute involuntary head and body movements that humans continually perform during fixation [3]. These small relocations are particularly interesting, as they yield relatively small changes in the images, which greatly facilitate the reliable estimation of motion parallax. Furthermore, this approach opens the possibility for actively closing the sensory-motor control loop to optimize the extraction of 3D information. That is, if the establishment of 3D representations does not occur instantaneously, but is progressively refined based on the integration of new information acquired over the period of fixation, it becomes possible to control the agent on the basis of the available knowledge.

Previous studies have shown that small movements similar to those performed by humans, including small isolated camera rotations [14] and coordinated head/camera rotations [9], provide useful 3D information also in robotic systems. In these previous studies, the authors extracted distance information by means of triangulation, using an approach similar to stereopsis on two images acquired at successive times during fixation. Here, we present a full model for the extraction of the motion parallax that emerges during active fixation in a humanoid robot that replicates the coordinated head/eye movements normally performed by humans. Unlike the previous studies, this model optimally integrates motor/position information with optic flow to continually extract depth from the inflowing temporal sequences of images acquired during fixation and progressively refine 3D representations.

The reminder of the paper is organized as follows. Section 2 describes the optic flow generated by fixational head/eye movements. Section 3 describes the motion of the system. Section 4 summarizes the proposed approach. Finally, in section 5 we report results obtained with simulations of our humanoid robot.

2 Motion equations

This section reviews the equations of the optic flow induced by the camera motion. Let consider an ideal point light source placed at the position $\boldsymbol{P} = [X, Y, Z]^T$ in the frame of reference centered on the camera. This frame of reference is centered in the nodal point of the camera and is oriented in such a way that the z-axis lies on the optic axis. Modeling the camera as a pinhole system with focal length f, the point $\boldsymbol{P}$ is projected on the sensor surface at the coordinate $\boldsymbol{p} = [x, y]^T$, as provided by equation (1).

$$\begin{bmatrix} x \\ y \end{bmatrix} = \frac{f}{Z} \cdot \begin{bmatrix} X \\ Y \end{bmatrix} \tag{1}$$

The movement of both neck and the eye motors induces a motion of the camera that is described by the translational and angular components of the velocity, that we denote with $\boldsymbol{v} = [v_x, v_y, v_z]^T$ and $\boldsymbol{\omega} = [\omega_x, \omega_y, \omega_z]^T$, respectively. This movement causes an apparent motion of the observed scene, so that, the point $\boldsymbol{P}$ is seen as moving with velocity $\dot{\boldsymbol{P}} = -\boldsymbol{v} - \boldsymbol{\omega} \times \boldsymbol{P}$. Taking the temporal derivatives of equation (1) and by writing the components of the apparent motion in term of the instantaneous velocity of the camera, we obtain the classical equation of the optic flow [8], which is reported in equation (2).

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

Each component of the optic flow $(\dot{x}, \dot{y})$ is decomposed into two components, r and t. The former depends on the angular velocity of the camera and the latter on the translational one. It is important to remark that equation (2) is valid for sufficiently small velocities and under the assumption that the scene is static.

3 Coordinated head/eye fixation

This section describes the robot's motor behavior. The robot replicates the strategy by which humans and primates maintain fixation under normal viewing condition. The motors of the neck rotate to generate motion parallax. Indeed, the

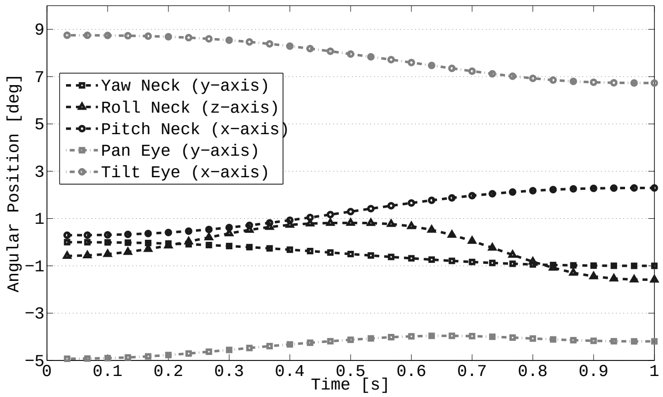


Fig. 2. Angular position of the neck motors (yaw, pitch, roll) and of the left eye motors (pan, tilt) during fixational head movements.

centers of rotation of these motors do not lie on the nodal point of the camera, thus any movements generate rotational and translational velocities. Translational velocities induce image motion that depends on the depth of the scene. Thus, once the optic flow and the camera motion are measured, we compute the depth of the scene directly from equation (2).

However, the measurement process is affected by noise and the magnitude of the signals is small. So that, reliable depth estimation can be achieved by keeping the scene in the field of view and integrating depth cues over time. The scene is maintained in the visual field by means of visual-guided eye movements: the eyes rotate in the opposite direction with respect to the neck to compensate the shift of the gaze.

These compensating movements ensure that the instantaneous velocity of the camera is small, so that, equation (2) keeps its validity. Moreover, the viewed scene change slightly, thus the computational cost required to compute the optic flow decreases.

Figure 1(a) shows a simulated robotic head while is observing a scene composed by four objects (three mugs and a postcard) placed at a different distance. The scene, as viewed by the robot, is shown in Fig. 1(b).

Figure 2 shows an example of fixational eye movements. The neck (yaw,roll and pitch angles) moves accordingly to a minimum jerk trajectory while the left eye (pan and tilt) moves to keep the fixation on the mug that is in the center of the visual field.

4 Model

The proposed model is summarized in Fig. 3. At each time stamp we obtain the head/eye position from proprioceptive cues and the viewed scene from the

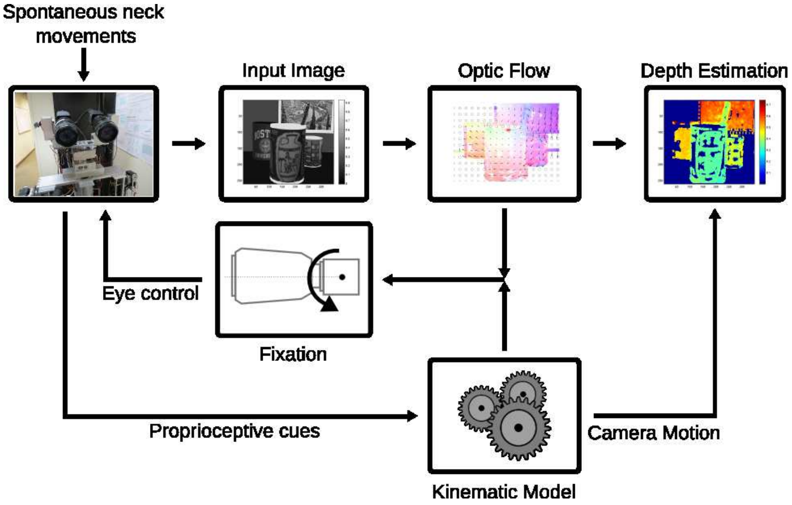


Fig. 3. System architecture. Images are acquired during coordinated head/eye fixation in which the cameras compensate for small random movements of the neck. The optic flow estimated by the images and the camera motion, as estimated by means of proprioceptive signals, are integrated to extract distance information from the scene.

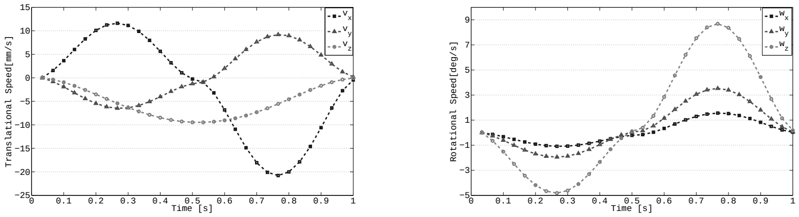
camera. Proprioceptive cues are used by the kinematic model of the robot to provide the motion of the camera. In parallel, the system extracts a dense optic flow from the sequence of images. The iconic filter integrates the camera velocity with the optic flow in order to estimate distance at each point in the scene (see section 4.3). The fixation is maintained by a controller that combines both feed-forward (kinematic model) and feed-back (optic flow) contributions.

4.1 Camera motion

The motion of the camera is provided by proprioceptive cues by using the kinematics model of the robot. Under the assumption of small movements, the camera velocity can be extracted directly from the homogeneous matrix M^t_{t-1}, which describes the displacement of the camera with respect to its previous position. This matrix is a function of the angular positions of the neck and the eye motors and can be approximated as described by equation (3).

$$M^t_{t-1} = \begin{bmatrix} 1 & -\omega_z & \omega_y & v_x \\ \omega_z & 1 & -\omega_x & v_y \\ -\omega_y & \omega_x & 1 & v_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{3}$$

Figure 4 shows the translational and the angular velocities of the nodal point obtained by the Fixational head movements showed in figure 2.



(a) (b)

Fig. 4. (a)Translational velocity ($\boldsymbol{v}$) and (b) angular velocity ($\boldsymbol{\omega}$) of the nodal point as computed using proprioceptive cues.

4.2 Optic Flow

Head fixational movements create an apparent motion of the viewed scene. We extracted the optic flow from two subsequent images using the probabilist version of the Lucas-Kanade algorithm proposed by Simoncelli et al.[15].

The algorithm is based on the assumption that the brightness of the image ($\mathbf{I}$) is constant over time: $\mathbf{I}(x, y, t) = const$. This assumption implies that the first order derivative is zero:

$$\mathbf{I}_x(x, y, t)\cdot\dot{x}+\mathbf{I}_y(x, y, t)\cdot\dot{y}+\mathbf{I}_t(x, y, t) = 0 \tag{4}$$

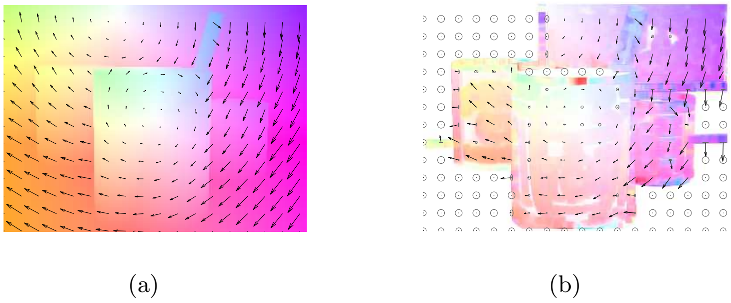
where $\mathbf{I}_x$ and $\mathbf{I}_y$ are the spatial derivatives of the image, $\mathbf{I}_t$ the temporal derivative and $\dot{x}$, $\dot{y}$ the horizontal and vertical components of the optic flow. Eq. (4) provides one constraint for two unknowns, so that, it allows us to compute only the normal component of the optic flow (*aperture problem*). For each pixel, the aperture problem is solved by using the Lucas-Kanade algorithm which assumes the optic flow is constant inside a neighborhood. In this way we set multiple constraints and we find the two unknowns using the least squares method.

The probabilistic version of the algorithm introduces a prior information about the optic flow, and *measurement* and *model* noises. Simoncelli et al.[15] modeled the prior information as a zero mean Gaussian distribution. Conversely, we considered the optic flow as a stochastic constant process. On the other hand, white Gaussian noise (η_t) affects the temporal derivative of the image (*measurement noise*), while the spatial derivatives are assumed to be noise free in order to keep the whole noise Gaussian[15]. The assumption that the velocity is constant inside a patch is usually not verified so that Gaussian noise ($\eta_{\dot{x}}, \eta_{\dot{y}}$) is added to the optic flow. Equation (4) becomes:

$$\mathbf{I}_x\cdot(\dot{x}+\eta_{\dot{x}})+\mathbf{I}_y\cdot(\dot{y}+\eta_{\dot{y}})+\mathbf{I}_t+\eta_t = \mathbf{I}_x\cdot\dot{x}+\mathbf{I}_y\cdot\dot{y}+\mathbf{I}_t+\eta_m = 0 \tag{5}$$

Using this model we can estimate the optic flow using a Kalman filter. The system of equations that describes the state-transition is:

$$\begin{bmatrix}\dot{x}(t+1)\\ \dot{y}(t+1)\end{bmatrix} = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\cdot\begin{bmatrix}\dot{x}(t)\\ \dot{y}(t)\end{bmatrix}+\begin{bmatrix}\xi_{\dot{x}}(t)\\ \xi_{\dot{y}}(t)\end{bmatrix} \tag{6}$$



(a) (b)

Fig. 5. (a) Ideal optic flow. (b) Optic flow estimated by means of Eq.6 and 7.Circles represent one standard deviation.

where $\xi_{\dot{x}}(t)$ and $\xi_{\dot{y}}(t)$ represent the error that we introduce by assuming the optic flow as constant in time.

The system of equations that describes the measurement process is:

$$\begin{bmatrix} \mathbf{I}_{t1}(t) \\ \mathbf{I}_{t2}t) \\ \dots \\ \mathbf{I}_{tn}(t) \end{bmatrix} = \begin{bmatrix} \mathbf{I}_{x1}(t) & \mathbf{I}_{y1}(t) \\ \mathbf{I}_{x2}(t) & \mathbf{I}_{y2}(t) \\ \dots & \\ \mathbf{I}_{xn}(t) & \mathbf{I}_{yn}(t) \end{bmatrix} \cdot \begin{bmatrix} \dot{x}(t) \\ \dot{y}(t) \end{bmatrix} + \begin{bmatrix} \eta_{m1}(t) \\ \eta_{m2}(t) \\ \dots \\ \eta_{mn}(t) \end{bmatrix} \quad (7)$$

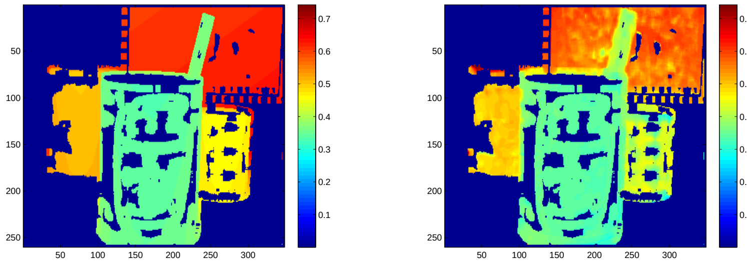
where the subscripts $1, 2, \dots, n$ denote the pixels inside the neighborhood.

Figure 5 compares the theoretical optic flow (Fig. 5(a)) and the estimated one (Fig. 5(b)). Different colors represent different orientations while the intensity is proportional to the magnitude of the optic flow. The circumferences in figure 5(b) represent the standard deviation of the optic flow. We can observer small radii in high textured regions and big radii in homogeneous regions. Moreover, the standard deviation assumes an elliptic shape in correspondence of the edges.

4.3 Iconic depth map

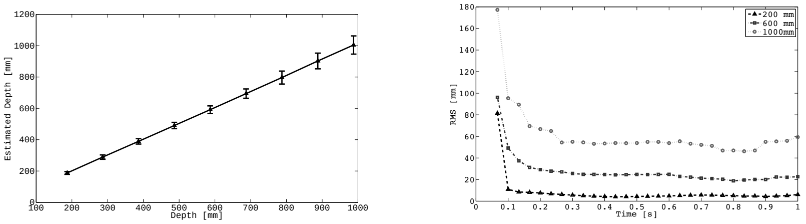
The optic flow and the instantaneous velocity of the camera are used by an iconic Kalman-filter to estimate the inverse of depth, also called disparity (d). Disparity is employed instead of the depth to work with a linear system [10]. Thanks to the small amplitude of the head movements, changes of the distance are negligible with respect to the distance of the objects. Also, the fixation movements keep the visual features practically in the same visual position, so that each pixel-centered filter observes features that are almost at the same depth. Indeed, in the acquired sequence of images the optic flow is always smaller than 2 pixels. For this reason we consider the disparity of the scene as constant in time (see equation 8). The small error due to this assumption is modeled by Gaussian noise η_d.

$$d(t+1) = d(t) + \eta_d(t) \quad (8)$$



(a) (b)

Fig. 6. Resulting map of distance in the scene. The ideal map (a) is compared to the map obtained from Eq. 8 and 9 (b). Distance estimation is possible only within textured regions.



(a) (b)

Fig. 7. Accuracy of the method. (a) Estimation of a single object (a cup) placed in front of the left camera at a variable distance in the range 0.2-1 m. The robot's neck moved as shown in Fig. 2 while the camera compensated to maintain fixation on the object. Data points represent means ± std. (b) Dynamics of error variance with targets at three different distances.

The measurement of the disparity is time-variant and depends on the angular and translational velocities of the camera:

$$\begin{bmatrix}\dot{x}(t)\\ \dot{y}(t)\end{bmatrix} = \begin{bmatrix}t_x(t)\\ t_y(t)\end{bmatrix} \cdot d(t) + \begin{bmatrix}r_x(t)\\ r_y(t)\end{bmatrix} + \begin{bmatrix}R(t)\end{bmatrix} \tag{9}$$

The covariance matrix of the error R is the covariance matrix provided by the computation of the optic flow described in the previous section.

Figure 6 shows the true depth of the scene (Fig. 6(a)) and the depth computed by the proposed algorithm (Fig. 6(b)). The depth is shown only where the result of the algorithm is reliable, that is, in the textured regions.

5 Experiments and Results

The simulation was performed using OpenRave [6] and we implemented the code in MATLAB. We simulated a robotic head (Fig. 1(a)) that moved with 7 degrees of freedom (d.o.f.s), three in the neck (yaw, roll and pitch) and two in each eye (pan and tilt). We simulated a vision system that acquires images at 30 frame/s from a high-resolution (1392 × 1040 pixels) monochrome camera. The pixel size of the camera was set almost as small as the cones in the retina ($4.65\mu m$). The focal length was set to $15mm$ and we did not take into account lens distortion.

During the experiments, the images acquired from the simulated environment were resized to 348 × 260 pixels to reduce computational time. For each pixel the optic flow was measured into a neighborhood of 7 × 7 pixels. The state transition error is represented by a white Gaussian noise with a standard deviation ($\xi_{\dot{x}}$ and $\xi_{\dot{y}}$) of 0.5 pixels. On the other hand, the standard deviation of the white Gaussian noises that affect the measurements, that is $\eta_{\dot{x}} = \eta_{\dot{y}}$ and η_t, were set to 0.002 and 0.1 pixels respectively. At the first frame the optic flow was initialized to zero and with a big standard deviation (10 pixels). A similar initialization was used also for the depth estimator, so that at the beginning we treated the whole scene as background (zero disparity). The algorithm was tested by executing fixational head movements described in section 3. While the neck moved following the trajectory shown in Fig. 2, the eye moved to keep the fixation.

Figures 5 and 6, described in the previous sections, provide a qualitative demonstration of the functioning of the algorithm. However, in figure 6(b) we can note that the error of the estimation grows with the depth of the object. In this section we provide some quantitatively results obtained in simulation. The experimental setup was composed of only a mug which was placed at variable distance within a range of a meter. For each scene we executed the fixational behavior described above.

Figure 7(a) shows the estimated depth (mean and standard deviation) for the mug placed at nine different distances. As expected, the error increases with the depth and, at a distance of 1 m, the standard deviation of the measure is approximately 100mm (10%). The proposed system improves the performance of previous work, in which the error at the same distance was around 150 mm [14, 9]. Moreover, our algorithm works with smaller interframe displacements and with 4 times smaller images. Figure 7(b) shows how the estimation error evolves with the time. The plot is shown for the object placed at three different distances (0.2 m, 0.6 m and 1 m). The result shows that the convergence time increases with the depth of the object. However, in the worst case (1 m), it converges in less than ten frames, that is, 0.3 s with a frame rate of 30 fps.

6 Conclusions

We have presented a novel framework that combines visual, motor, and proprioceptive signals to extract distance/depth information in a humanoid robot. The robot replicates the coordinated head and eye movements that human normally perform while maintaining fixation. Because of these movements, depth information emerges in the form of parallax. A probabilistic filter in our model extracts the optic flow from the sequence of the incoming images and combines it with proprioceptive cues to extract 3D information. Results obtained by means of simulations have shown that the methods is efficient and robust. This method can be used alone or combined with other cues, such as stereopsis, to obtain more reliable 3D vision systems.

References

1. Aloimonos, Y., Duric, Z.: Estimating the heading direction using normal flow. International Journal of Computer Vision 13(1), 33–56 (1994)
2. Ayache, N.: Artificial vision for mobile robots - stereo vision and multisensory perception. MIT Press (1991)
3. Aytekin, M., Rucci, M.: Motion parallax from microscopic head movements during visual fixation. Vision Research (Aug 2012)
4. Barron, J.L., Fleet, D.J., Beauchemin, S.S.: Performance of optical flow techniques. International Journal of Computer Vision 12(1), 43–77 (1994)
5. Davison, A.J., Reid, I.D., Molton, N.D., Stasse, O.: Monoslam: Real-time single camera slam. Pattern Analysis and Machine Intelligence, IEEE Transactions on 29(6), 1052–1067 (2007)
6. Diankov, R., Kuffner, J.: Openrave: A planning architecture for autonomous robotics. Robotics Institute, Pittsburgh, PA, Tech. Rep. CMU-RI-TR-08-34 (2008)
7. Faugeras, O.D., Luong, Q.T., Papadopoulo, T.: The geometry of multiple images - the laws that govern the formation of multiple images of a scene and some of their applications. MIT Press (2001)
8. Higgins, L.H.C., Prazdny, K.: The Interpretation of a Moving Retinal Image. Proceedings of the Royal Society of London. Series B, Biological Sciences (1934-1990) 208(1173), 385–397 (1980)
9. Kuang, X., Gibson, M., Shi, B.E., Rucci, M.: Active vision during coordinated head/eye movements in a humanoid robot. Robotics, IEEE Transactions on PP(99), 1 –8 (2012)
10. Matthies, L., Kanade, T., Szeliski, R.: Kalman filter-based algorithms for estimating depth from image sequences. International Journal of Computer Vision 3(3), 209—238 (1989)
11. Ramachandran, M., Veeraraghavan, A., Chellappa, R.: A fast bilinear structure from motion algorithm using a video sequence and inertial sensors. IEEE Trans. Pattern Anal. Mach. Intell. 33(1), 186–193 (2011)
12. Rogers, B., Graham, M.: Motion parallax as an independent cue for depth perception. Perception 8(2), 125–134 (1979)
13. Sandini, G., Tistarelli, M.: Active tracking strategy for monocular depth inference over multiple frames. Pattern Analysis and Machine Intelligence, IEEE Transactions on DOI - 10.1109/34.41380 12(1), 13–27 (1990)
14. Santini, F., Rucci, M.: Active estimation of distance in a robotic system that replicates human eye movement. Robotics and Autonomous Systems 55(2), 107–121 (2007)
15. Simoncelli, E., Adelson, E., Heeger, D.: Probability distributions of optical flow. In: Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on. pp. 310–315. IEEE (1991)