Monocular Depth from Motion Using a New Closed-Form Solution

Mohamed Hasan and Mohamed Abdellatif

Mechatronics and Robotics Engineering Department
Egypt-Japan University of Science and Technology
{mohamed.hasan,mohamed.abdellatif}@ejust.edu.eg

Abstract. Monocular depth has been found using estimation, closed-form solution and learning techniques. Estimation and closed-form solution compute the depth from motion, while learning techniques calculate the depth using a single image with a depth map as a supervisor. This paper presents a new closed form solution for monocular depth from motion. The proposed method builds on the notation that an interest point in an image of a static scene has a static world location. Camera pose and calibration parameters are used as constraints to provide the depth solution. The proposed method is verified through real experiments on indoor mobile robot platform. The effect of uncertainty in the solution variables is studied and the results are benchmarked to groundtruth.

Keywords: monocular depth, real-time depth from motion.

1 Introduction

Depth calculation is necessary for several applications in computer vision and robotics, such as navigation, obstacle avoidance, visual simultaneous localization and mapping. Traditionally, stereo vision has been used to compute depth through the epipolar geometry. Alternatively, monocular depth is the depth computed from the images captured by a single camera [1]. Computing monocular depth has been introduced through estimation, closed-form solution and learning techniques.

Estimation and closed-form solution techniques compute the depth from the motion of the observing camera. This requires tracking of a number of interest points through the captured monocular sequence of images. Estimation techniques mainly employ Kalman filters to iteratively estimate depth and geometrical structure of the scene [2-7]. However, closed-form solutions impose the geometrical constraints of motion to reduce the number of unknowns and determine the depth in real time [8]. On the other hand, supervised learning has been used to find depth from a single image with a groundtruth depth map acting as a teacher [9-10].

For the applications of mobile robots, depth recovery from visual information is necessary for navigation. The estimation techniques of depth from motion take a number of iterations until settled to a depth value and thus affecting the accuracy of robotic missions like obstacle avoidance, localization and mapping. On contrary,

C.-Y. Su, S. Rakheja, H. Liu (Eds.): ICIRA 2012, Part III, LNAI 7508, pp. 473–483, 2012.

closed-form solutions can compute the depth in real time and thus enable safe navigation of mobile robots. The closed-form solutions depend on several constraints to simplify the depth calculations such as the planar robot motion, the geometrical constraints and the known camera calibration parameters [8, 11]. Murphey *et al* have used the geometrical constraints to find depth from motion using a closed form solution [8]. However, they neglected the camera calibration parameters and the camera rotation information as well.

In this paper, a new closed-form solution for depth from motion is presented. The proposed solution recovers the depth of a static point in real time without the need of iterative estimation. Both the camera pose and the calibration parameters information have been used to compute the depth. The new depth solution builds on the notation that any interest point existing in an image of a static scene, has a static location in world. This constraint has been used to calculate the depth of an image point from just two monocular views of that point. The proposed solution differs from the traditional triangulation methods [12] because it only needs matching for one image point. The paper is organized as follows. The closed-form solution is introduced in the next section. The results of experiments on indoor mobile robot are presented in section 3 along with uncertainty analysis. Discussion of the results and the limitations of the proposed method are given in section 4. Finally, conclusions are drawn in section 5.

2 Closed-Form Solution

The proposed method works under the following assumptions:

- The camera is calibrated.
- The observed features are static.
- The robot pose is known.

The calibrated camera is assumed to move and a sequence of monocular images is captured. Consider a static point feature $p_i^w = (X \quad Y \quad Z)^T$ represented in world coordinates. The camera observes p_i^w twice from two different world positions: r_1^{wc}, r_2^{wc} as shown in Fig. 1. The distances between the camera and p_i^w in the two observations are denoted as h_1^c and h_2^c which are represented in the camera coordinates. Transformation from camera to world coordinates is performed by the rotation matrices: R_1^{wc}, R_2^{wc}which are functions of the camera pitch angles: θ_1, θ_2. The pitch angle is the rotation of the planar robot around the vertical axis Y^c. The projections of p_i^won the image plane are $(u_1, \ v_1)$ in the first image and $(u_2, \ v_2)$ in the other image.

The intrinsic parameters of the camera are given from prior camera calibration. Standard pinhole camera model- without lens distortion- [1] has been used along with these intrinsic parameters. The distance between the camera and an observed point is composed of three Euclidean components; $h^c = (h_x \quad h_y \quad h_z)^T$. Thus, the pixel coordinates of the observed point are given by the camera model:

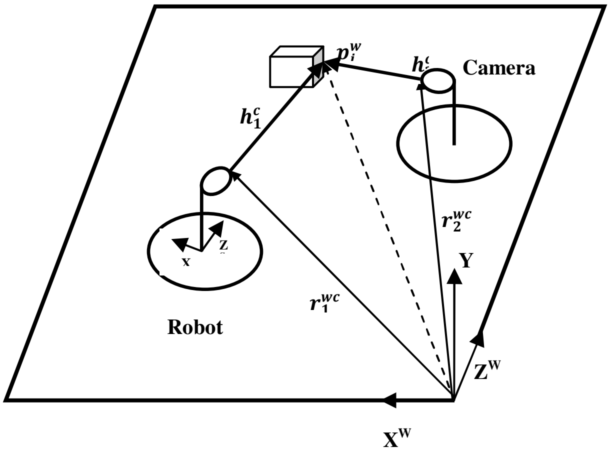


Fig. 1. The mobile robot with the onboard camera observing a feature from two different positions

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} u_0 - \dfrac{f}{d}\dfrac{h_x}{h_z} \\ v_0 - \dfrac{f}{d}\dfrac{h_y}{h_z} \end{pmatrix} \tag{1}$$

where u_0, v_0 are the camera center in pixels, f is the focal length and d is the pixel size.

The world location of the point feature can be determined given the location of the camera and the distance from this location to the point. This may be written for the two different observations as:

$$p_i^w = r_1^{wc} + R_1^{wc} h_1^c \tag{2}$$

$$p_i^w = r_2^{wc} + R_2^{wc} h_2^c \tag{3}$$

where:

$$r_i^{wc} = \begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix},\ R_i^{wc} = \begin{pmatrix} \cos\theta_i & 0 & \sin\theta_i \\ 0 & 1 & 0 \\ -\sin\theta_i & 0 & \cos\theta_i \end{pmatrix} \tag{4}$$

From (1)

$$h_i^c = \begin{pmatrix} h_{xi} \\ h_{yi} \\ h_{zi} \end{pmatrix} = \begin{pmatrix} \frac{d}{f}(u_0 - u_i)h_{zi} \\ \frac{d}{f}(v_0 - v_i)h_{zi} \\ h_{zi} \end{pmatrix} = h_{zi} \begin{pmatrix} a_i \\ b_i \\ 1 \end{pmatrix} \tag{5}$$

where

$$a_i = \frac{d}{f}(u_0 - u_i), \qquad b_i = \frac{d}{f}(v_0 - v_i) \tag{6}$$

and *i*=1, 2 in (4-6).

It can be noted that the observed feature has a fixed world location regardless of camera motion. This constraint has been used to provide solution by equating (2), (3), using (4-6) and rearranging:

$$\begin{pmatrix} a_1 \cos\theta_1 + \sin\theta_1 & -a_2 \cos\theta_2 - \sin\theta_2 \\ b_1 & -b_2 \\ -a_1 \sin\theta_1 + \cos\theta_1 & -a_1 \sin\theta_1 + \cos\theta_1 \end{pmatrix} \begin{pmatrix} h_{z1} \\ h_{z2} \end{pmatrix} = \begin{pmatrix} x_2 - x_1 \\ y_2 - y_1 \\ z_2 - z_1 \end{pmatrix} \tag{7}$$

or equivalently:

$$A\, h_z = B \tag{8}$$

where $h_z = (h_{z1} \quad h_{z2})^T$.

The depth of the feature can now be calculated from (8):

$$h_z = (A^T A)^{-1} A^T B \tag{9}$$

This completes the solution of the depth of the point.

3 Experiments and Results

The proposed depth solution has been verified using an indoor mobile robot platform (Robotino). Camera is fixed on the planar mobile robot as shown in Fig. 1. The robot is translated laterally facing a scene in our lab space while capturing a sequence of monocular images. This robot motion is repeated at three different distances from the scene (*d*$_1$, *d*$_2$, *d*$_3$) resulting in three different scales for each image. In each group of images, distinctive features are detected in the first image and their pixel measurements are recorded. These features are then tracked through the images sequence in each group and the matched measurements are recorded for each feature.

The image coordinates of the feature in the first image and its match in the other image during motion are now known. Also the location and orientation of the camera while capturing both images can be acquired from robot odometry. These data along with camera intrinsic parameters provide a depth solution as given in (9).

The three robot lateral paths start at d_1=3.48m, d_2=2.64m and d_3=1.44m from the nearest point in the scene, respectively. The distance between the nearest and the farthest points in the scene is 1.6m. Twenty five images have been captured during each robot motion for about 92 cm at equidistant points. Thus, three groups of monocular images of a static scene at three different scales are provided. In each group of images, a number of distinctive features is tracked through the sequence and matches for every point are recorded.

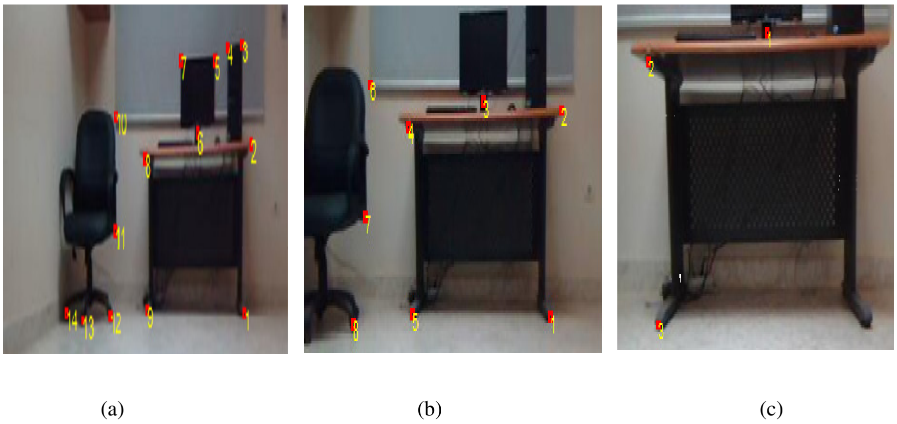
3.1 Manual Matching

In the first experiment, features have been selected and tracked in a manual mode to evaluate the proposed method. For example, the fourth frame from each group of images is shown with the selected points in Fig. 2.

The measurements of the selected features in the first frame are used as (u_1, v_1), while the matched location of the feature in every subsequent image is (u_2, v_2) as in section 2. Depth is calculated using these measurements through (5-7) at every image. The calculated depth of each feature at every frame is used in (2) to calculate the world location of the feature. The *Z*-coordinate of the feature location (depth coordinate) is compared to the measured groundtruth of feature and the error is found. The errors in depth calculations for some of the features in the first group are described in Table 1 Results are shown for the calculations at five different lateral positions.

The effect of the observing distance between camera and feature has been studied for three features in the three groups of measurements. The first feature appears with the feature numbers 6, 3 and 1 in the three groups respectively. The other two features numbers are (8, 4, 2) and (9, 5, 3) respectively. Results of errors in depth calculations are shown in Fig. 3.

The proposed method of depth solution is affected by the uncertainty in points matching (*u, v*), robot pose (*x, z, θ*) and camera calibration (*d/f*). These effects have been studied for the third feature in Fig.3 at the third distance, since this has the minimalist error. The solution parameters associated with this case of minimum error have been selected as reference values. Perturbations have then been made around these reference values and the error in the resulting depth is shown. The error in depth is normalized by the groundtruth depth of the feature. The effects of uncertainty in matching the features through the image sequence have been studied for a range of ±5 pixels. The errors in horizontal matching (*u*-image coordinate) and vertical matching (*v*-image coordinate) are shown in Fig.4. The effect of errors in camera calibration parameters; the ratio (*d/f*) is shown in Fig. 5. Finally, the effect of error in robot pose



(a) (b) (c)

Fig. 2. The fourth frame in the the monocular sequence of the captured scene in the three groups at different distances (a) d_1= 4m. (b) d_2=2.8. (c) d_3=1.6m

Table 1. Depth errors for features in the first group matched manually

P	**Z (m)**	**Error (m) at different frames**					**Mean error (m)**	**% Mean error**
		5	***10***	***15***	***20***	***25***		
1	4.44	-0.004	0.013	-0.017	-0.029	-0.026	-0.012	-0. 270
2	4.33	0.498	-0.074	-0.094	-0.105	0.217	0.088	2.032
3	4.42	-0.026	-0.214	-0.301	-0.112	-0.160	-0.163	-3.687
4	4.42	0.147	-0.056	-0.048	0.023	-0.007	0.011	0.248
5	4.65	0.277	-0.105	-0.150	-0.068	-0.09	-0.028	-0.602
6	4.68	0.462	-0.021	-0.043	-0.065	-0.010	0.064	1.367
7	4.65	0.320	0.093	-0.002	0.025	0.009	0.089	1.913

has been studied through testing error in the robot lateral position, forward position and heading angle (θ) as shown in Fig. 6. The errors in robot pose and calibration parameters are normalized by their original non-noisy value.

3.2 Automatic Matching

The evaluation technique used in section 3.1 has been repeated but with automatic matching of the features through the sequence. The first seven features displayed in Fig. 2 (a) with their depth results in Table 1, are now tracked through the images of the first group. A fixed-size template has been extracted from the first image centered on each feature's image location. This template has been used as the descriptor for the feature and has been tracked using Fast template matching [13].

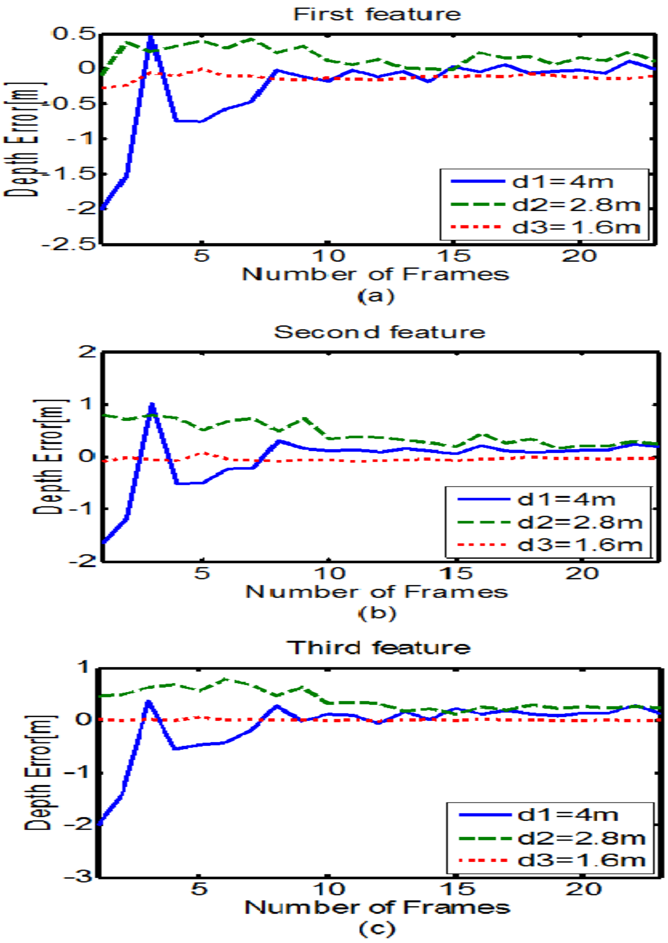


Fig. 3. Depth errors for three features evaluated at the three capturing distances: d_1, d_2, d_3

Fast template matching method implements Normalized Cross Correlation, NCC in frequency domain using pre-computed tables containing the integral of the image and image square over the search window. The normalized form of cross correlation is necessary to overcome the problem that image and templates are not the same size. More important, NCC deals with the variation in the intensity of both the image and templates. Fast Fourier transform is adopted to switch to the frequency domain. Transform computations are efficiently performed using summed-area tables containing the integral of the image (running sum) and image square over the search area.

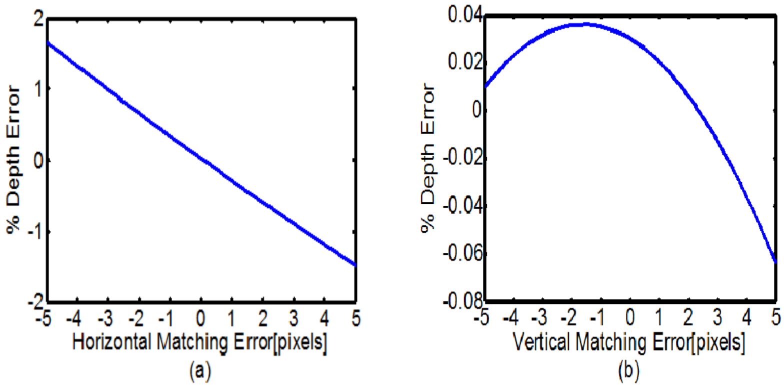


Fig. 4. The effect of error in matching on the depth calculations

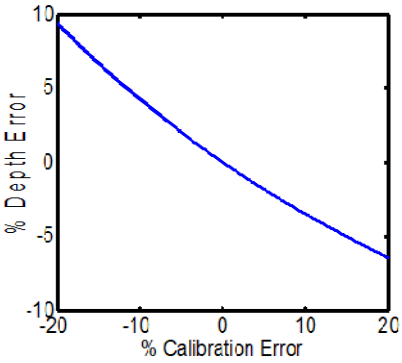


Fig. 5. The effect of error in camera calibration parameters (*d*/*f*) on depth calculations

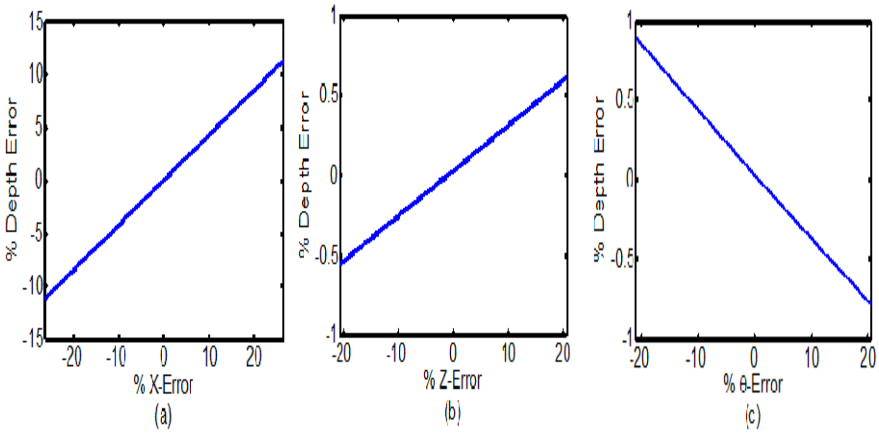


Fig. 6. The effect of error in robot pose on the depth calculations

Every five frames, a matching score is calculated between the templates and the image in order. The image point having the peak of the matching score is compared to an empirically defined threshold (0.85). If the matching score passes the threshold, then this image point is the match of the feature. The results of matching the features are shown in Fig. 7 for three frames in the sequence.

The first three features have not been tracked well through all images. The results of depth calculations of all features are given in Table 2. The effect of matching error is apparent in the large error in the depth of the first three features. The maximum mean error of depth of the other features is 7.60 cm which is 1.63 % of the true depth.

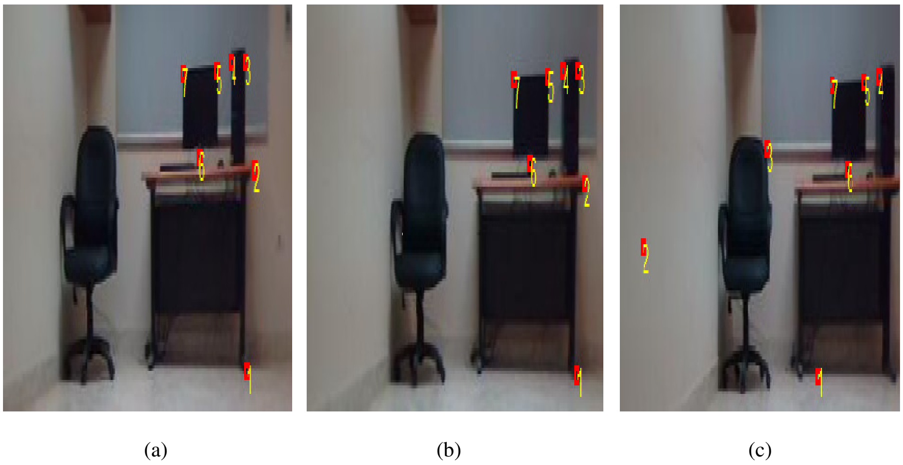

(a) (b) (c)

Fig. 7. Fast template matching results for the first group of images: (a) 5th frame (b) 15th frame (c) 25th fame. The first three features have large error in matching.

Table 2. Results of depth calculations using automatic matching. The first three features have large error in matching and hence in depth.

P	**Z (m)**	**Error (m) at different frames**					**Mean error (m)**	**% Mean error**
		5	***10***	***15***	***20***	***25***		
1	4.44	0.370	0.340	0.354	0.455	-24.58	-4.612	-103
2	4.33	0.682	0.187	0.261	-5.225	-5.510	-1.921	-44.3
3	4.42	-0.098	-0.284	-0.260	-6.846	-7.907	-3.079	-69.6
4	4.42	0.185	-0.138	-0.064	0.100	0.0456	0.025	0.565
5	4.65	0.383	-0.087	-0.103	-0.082	-0.063	0.009	0.193
6	4.68	0.211	-0.167	-0.140	-0.038	-0.088	-0.044	-0.949
7	4.65	0.496	-0.039	-0.051	-0.026	0.004	0.076	1.634

4 Discussion

The error in depth calculations using the proposed method has not exceeded 3.7% in case of manual matching; while with robust automatic matching the depth in error has been below 1.7%. The proposed method depends on observing a feature at two different world positions. The amount of lateral translation between these two positions of observation has great effect on the depth result. It is shown in Table 2 that the amount of translation should not be less than about 20 cm (distance between five images). The mean of the depth error for each feature is shown to be in the range from 1 cm to 16 cm. Fig.3 shows that the depth error is degraded at the tenth frame (40 cm translation). The depth accuracy increases if the camera observes the feature from near distances as already known from stereo vision [1]. These accuracy conditions can be considered suitable for applications like visual navigation of indoor mobile robots.

The effects of uncertainty described in Fig. 4-6 provide the ranges of allowable errors for the proposed method. Due to the lateral motion of the robot observing a static scene, the horizontal image coordinate has a considerable change as clear in Fig.4 (a) compared to the vertical change in (b). These results show that the depth error don't exceed ±2 % (about ±8 cm in this case) for ± 5 pixels in horizontal matching error.

The proposed method depends on the amount of translation between the robot's two observing positions. Consequently, Fig. 5 (a) shows that the error in lateral motion has greater effect than that of the forward motion in (b) or rotational motion in (c). To keep the depth error inside a range of ±2 %, the allowable range for lateral position error is ±5 % (about ±4 cm in this case). To keep the depth error inside the same range of ±2 %, the allowable range for calibration parameters is ±6 % (about 1.8e-004). The depth calculations are thus sensitive to the error in calibration parameters.

5 Conclusions

A new closed-form solution for depth from motion has been introduced. The proposed method makes use of the constraint that the world location of an image point is fixed regardless of motion. The method assumes that the pose and the calibration parameters of the camera are known. Experimental results on an indoor planar mobile robot have been presented to verify the solution. From the discussion given in the previous section, some conclusions can be given about the validity of the proposed method:

- The error in the depth solution with automatic points matching is below 1.7%.
- The error in points matching should be within a range of ± 5 pixels.
- The amount of translation between the camera observing locations should not be less than 20 cm.
- The error of the camera position should not exceed ±5 % of the true position.

Acknowledgement. The first author is supported by a scholarship from the Ministry of Higher Education, Government of Egypt which is gratefully acknowledged.

References

[1] Szeliski, R.: Computer Vision: Algorithms and Applications. Springer, London (2011)

[2] Franke, U., Rabe, C., Badino, H., Gehrig, S.K.: 6D-Vision: Fusion of Stereo and Motion for Robust Environment Perception. In: Kropatsch, W.G., Sablatnig, R., Hanbury, A. (eds.) DAGM 2005. LNCS, vol. 3663, pp. 216–223. Springer, Heidelberg (2005)

[3] Hung, Y., Ho, H.: A Kalman Filter Approach to Direct Depth Estimation Incorporating Surface Structure. IEEE Trans. on Pattern Analysis and Machine Intelligence Pami 21(6), 570–575 (1999)

[4] Szeliski, R.: Bayesian modeling ofuncertainty inlow-level vision. Int. J. of Computer Vision 5(3), 271–301 (1990)

[5] Franke, U., Rabe, C.: Kalman filter based depth from motion with fast convergence. In: IEEE Symp. on Intelligent Vehicles, pp. 181–186 (2005)

[6] Matthies, L., Kanade, T., Szeliski, R.: Kalman filter-based algorithms for estimating depth from image sequences. Int. J. of Computer Vision 3(3), 209–236 (1989)

[7] Civera, J., Davison, A., Montiel, J.: Inverse Depth Parametrization for Monocular SLAM. IEEE Trans. on Robotics 24(5), 932–945 (2008)

[8] Murphey, Y., Chen, J., Crossman, J., Zhang, J., Richardson, P., Sieh, L.: DepthFinder, A Real-time Depth Detection System for Aided Driving. In: IEEE Symp. on Intelligent Vehicles, pp. 122–127 (2000)

[9] Saxena, A., Chung, S., Ng, A.: 3-D Depth Reconstruction from a Single Still Image. Int. J. of Computer Vision 76(1), 53–69 (2008)

[10] Michels, J., Saxena, A., Ng, A.: High Speed Obstacle Avoidance using Monocular Vision and Reinforcement learning. In: Int. Conf. on Machine Learning, pp. 593–600 (2005)

[11] Ortin, D., Montiel, J.: Indoor robot motion based on monocular images. J. Robotica 19(3), 331–342 (2001)

[12] Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision, 2nd edn. Cambridge University Press (2004)

[13] Lewis, J.: Fast Template Matching. In: Vision Interface 1995, Canadian Image Processing and Pattern Recognition Society, pp. 120-123 (1995)