An Information Theoretic Approach for Next Best View Planning in 3-D Reconstruction

Stefan Wenhardt, Benjamin Deutsch,*
Joachim Hornegger, Heinrich Niemann
Chair for Pattern Recognition,
Friedrich-Alexander University of Erlangen
Martensstr. 3, 91058 Erlangen, Germany

Joachim Denzler
Chair for Digital Image Processing
Friedrich Schiller University of Jena
Ernst-Abbe-Platz 2, 07743 Jena

published in: The 18th International Conference on Pattern Recognition, Proceedings, pp. 103-106

Abstract

We present an algorithm for optimal view point selection for 3-D reconstruction of an object using 2-D image points. Since the image points are noisy, a Kalman filter is used to obtain the best estimate of the object's geometry. This Kalman filter allows us to efficiently predict the effect of any given camera position on the uncertainty, and therefore quality, of the estimate. By choosing a suitable optimization criterion, we are able to determine the camera positions which minimize our reconstruction error.

We verify our results using two experiments with real images: one experiment uses a calibration pattern for comparison to a ground-truth state, the other reconstructs a real world object.

1. Introduction

We study the problem of finding the next best view in 3-D reconstruction. Unlike other works, which use geometrical approaches [8, 9] and range scanners [2, 12, 14], we extract 2-D points from camera images using feature point tracking and combine them to a probabilistic representation of 3-D object points, using the Kalman filter.

Approaches using the Kalman filter for 3-D reconstruction have already been published [7, 10, 16], as have been works aiming to obtain the best views by minimizing the entropy in state estimation problems [1, 4, 5]. In this paper we will show how these approaches can be combined to improve the 3-D reconstruction by planning the next best view.

We evaluate our proposed method with two scenarios with mobile cameras. The first uses a camera mounted on a robot arm with multiple degrees of freedom (SCORBOT, fig. 1, left), the second uses a camera on a turntable with a tilting arm (fig. 1, right).

Figure 1. SCORBOT (left) and turn table (right)

This paper is organized as follows: Section 2 gives an overview of 3-D reconstruction with the Kalman filter. Section 3 applies the methods of uncertainty reduction to this reconstruction, and lists several constraints. This application is evaluated in section 4 in two separate experiments with real cameras. Finally, section 5 concludes this work and describes opportunities for future research.

2. 3-D Reconstruction with the Kalman Filter

This section describes the task of 3-D reconstruction as a state estimation problem. The 3-D reconstruction is represented by a list of 3-D points, concatenated to the *state vector* $\boldsymbol{x} \in \mathbb{R}^n$. The camera makes an observation $\boldsymbol{o}_t \in \mathbb{R}^m$ consisting of 2-D projections of the state points at each time

*This work was partly funded by the German Research Foundation (DFG) under grant SFB 603, TP B2. Only the authors are responsible for the content.

step t. The dimensions n and m are constant in time, since we consider only feature points which can be tracked all the time. Since camera images are noisy, this observation is assumed to contain normally distributed noise. The observation equation is therefore

$$\boldsymbol{o}_t = \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{c}_t) + \boldsymbol{w}, \tag{1}$$

where $\boldsymbol{g}(\cdot,\cdot)$ is the observation function, $\boldsymbol{c}_t$ is the camera's view point at time t, and $\boldsymbol{w}$ is a zero-mean additive Gaussian noise process with covariance matrix $\boldsymbol{W}$.

The Kalman filter uses these observations to maintain an estimate of the state vector as a Gaussian distribution $\mathcal{N}(\widehat{\boldsymbol{x}}_t, \boldsymbol{P}_t)$. A starting distribution $\mathcal{N}(\widehat{\boldsymbol{x}}_0, \boldsymbol{P}_0)$ is typically obtained by an initial triangulation. The information of each new image is used to update the state estimate:

$$\widehat{\boldsymbol{x}}_t = \widehat{\boldsymbol{x}}_{t-1} + \boldsymbol{K}_t\left(\boldsymbol{o}_t - \boldsymbol{g}(\widehat{\boldsymbol{x}}_{t-1}, \boldsymbol{c}_t)\right) \tag{2}$$

$$\boldsymbol{P}_t = \left(\boldsymbol{I} - \boldsymbol{K}_t\boldsymbol{G}_t(\boldsymbol{c}_t)\right)\boldsymbol{P}_{t-1}, \tag{3}$$

where $\boldsymbol{G}_t(\boldsymbol{c}_t)$ is the Jacobian of $\boldsymbol{g}(\widehat{\boldsymbol{x}}_{t-1}, \boldsymbol{c}_t)$, $\boldsymbol{I}$ is the identity matrix, $\boldsymbol{K}_t$ is the Kalman gain matrix, defined as

$$\boldsymbol{K}_t = \boldsymbol{P}_{t-1}\boldsymbol{G}_t^{\mathrm{T}}(\boldsymbol{c}_t)\big(\boldsymbol{G}_t(\boldsymbol{c}_t)\boldsymbol{P}_{t-1}\boldsymbol{G}_t^{\mathrm{T}}(\boldsymbol{c}_t)+\boldsymbol{W}\big)^{-1}. \tag{4}$$

The Jacobian is necessary because $\boldsymbol{g}(\cdot,\cdot)$ is not linear, as the classical Kalman filter requires; a Kalman filter with this linearization is called an *extended Kalman filter* [6] [1].

These steps allow both the current best estimate $\widehat{\boldsymbol{x}}_t$ and the uncertainty of this estimate, in the form of the covariance matrix $\boldsymbol{P}_t$, to be maintained.

3. Next Best View Point Selection

3.1. View Point Evaluation

The goal of next best view selection is to find the optimal next view point $\boldsymbol{c}_t^*$ to improve the reconstruction accuracy. One optimality criterion is to reduce the uncertainty in the state estimation, which is measured in information theory by its entropy $H_{\boldsymbol{c}_t}(\boldsymbol{x})$. This entropy, however, has to be calculated *a priori* to optimize the view before obtaining a new image. Therefore, we need to determine the expected entropy $H_{\boldsymbol{c}_t}(\boldsymbol{x}|\boldsymbol{o}_t)$, also called *conditional entropy* [11]. Let $\boldsymbol{p}(\boldsymbol{x}|\boldsymbol{O}_t, \boldsymbol{C}_t)$ be the probability density function of the state $\boldsymbol{x}$ after acquiring images from the view points $\boldsymbol{C}_t = \boldsymbol{c}_0, \boldsymbol{c}_1, \ldots \boldsymbol{c}_t$ and obtaining observations $\boldsymbol{O}_t = \boldsymbol{o}_0, \boldsymbol{o}_1, \ldots \boldsymbol{o}_t$. The expected entropy is the mean of the entropy of $\boldsymbol{x}$ over all observations:

$$H_{\boldsymbol{c}_t}(\boldsymbol{x}|\boldsymbol{o}_t) = \int \boldsymbol{p}(\boldsymbol{o}_t|\boldsymbol{c}_t) H_{\boldsymbol{c}_t}(\boldsymbol{x}) d\boldsymbol{o}_t \tag{5}$$

$$H_{\boldsymbol{c}_t}(\boldsymbol{x}) = -\int \boldsymbol{p}(\boldsymbol{x}|\boldsymbol{O}_t, \boldsymbol{C}_t) \log \boldsymbol{p}(\boldsymbol{x}|\boldsymbol{O}_t, \boldsymbol{C}_t) d\boldsymbol{x} \tag{6}$$

[1] Another difference to the classical Kalman filter is that $\boldsymbol{x}$ is time-invariant, and so no prediction step is performed.

The optimality criterion is to determine the view point $\boldsymbol{c}_t^*$ which minimizes the conditional entropy:

$$\boldsymbol{c}_t^* = \operatorname{argmin}_{\boldsymbol{c}_t} H_{\boldsymbol{c}_t}(\boldsymbol{x}_t|\boldsymbol{o}_t) \tag{7}$$

Since the state estimate is in the form of a normal distribution, $\boldsymbol{x} \sim \mathcal{N}(\widehat{\boldsymbol{x}}_t, \boldsymbol{P}_t)$, its entropy has the closed form

$$H_{\boldsymbol{c}_t}(\boldsymbol{x}) = \frac{n}{2} + \frac{1}{2}\log(2\pi^n|\boldsymbol{P}_t|), \tag{8}$$

where $|\cdot|$ denotes the determinant of a matrix. Since the covariance matrix $\boldsymbol{P}_t$ as calculated in eq. (3) depends on $\boldsymbol{c}_t$ but *not* on $\boldsymbol{o}_t$, we can simplify eq. (5) by pulling $H_{\boldsymbol{c}_t}(\boldsymbol{x})$ out of the integral. The remaining integral now integrates a probability density function and is therefore 1. If we further consider the monotony of the logarithm, and disregard constant terms and factors, the optimality criterion (7) becomes

$$\boldsymbol{c}_t^* = \operatorname{argmin}_{\boldsymbol{c}_t} |\boldsymbol{P}_t|. \tag{9}$$

3.2. Constraints For The Next Best View

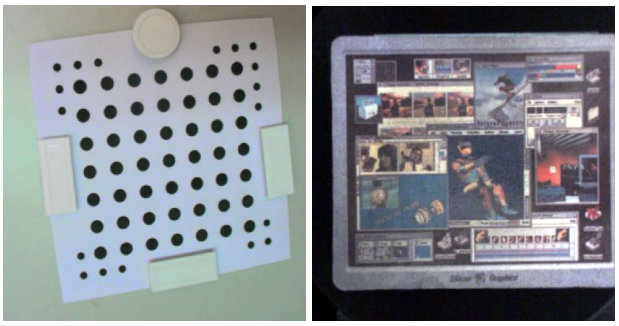
Some constraints on $\boldsymbol{c}_t$ must be considered. Not every optimal view point, in the sense of (9), results in a usable image. Some examples of effects that can make a view point completely or partly unusable are:

- Visibility: The observed 3-D points must be visible in the image. This may fail because points lie outside the field of view of the camera, or because they are occluded by parts of the object or by the robot arm. We consider only the constraint resulting from the limited field of view. Self occlusions which appear in non-planar objects will be considered in future work (cf. section 5).
- Reachability: The view point must be reachable by the robot. To ensure this, we use the 4 by 4 *Denavit-Hartenberg* matrix [3], which depends on the angles of the rotation axes and the distances between the joints, to calculate the transformation to a fixed world coordinate system. Since the lengths are fixed, only the angles are relevant.

We now search for the optimal view point $\boldsymbol{c}_t^*$ with an exhausitve search over the discretely sampled action space.For each sample, we calculate both $|\boldsymbol{P}_t|$ and the expected observation. If the expected observation contains image points outside the field of view, we discard this sample. The best-rated undiscarded sample is the next best view.

4. Experimental Results

We verify our approach for next best view planning with real world experiments. We use a Sony DFW-VL500



(a) calibration pattern (b) mouse pad

Figure 2. Images taken during the experiments (2(a) from first and 2(b) from second)

firewire camera, whose intrinsic parameters were calibrated by Tsai's algorithm [15]. The camera is moved by the SCORBOT in the first experiment and by the turn table with tilting arm in the second one (cf. Fig. 1). The extrinsic parameters are calculated by the Denavit-Hartenberg matrix and the hand-eye transformation, which is acquired by the algorithm of Schmidt [13]. The first experiment reconstructs a calibration pattern, the second a mouse pad.

In both experiments, we start with an initial estimation, obtained by triangulating from an image pair from two view points. This gives us an initial estimate of $\boldsymbol{x}$. The initial covariance matrix $\boldsymbol{P}_0$ is set to a diagonal matrix, as we assume that the uncertainty is equal in each direction.

To evaluate the expected uncertainty, we calculate the determinant of $\boldsymbol{P}_t$ (eq. (3)). The Jacobian $\boldsymbol{G}_t(\boldsymbol{c}_t)$ of the observation function depends on the axis values of the robot and must be calculated for each candidate view point. The computation time for the next best view for the SCORBOT (5 degrees of freedom, due to its 5 axes, 384000 view points analyzed, 49 3-D points) is about 9 minutes on a system with an Pentium IV processor with 3 GHz, and 1 GB RAM, and about 45 seconds for the turntable (2 degrees of freedom, 2000 view points analyzed, 50 3-D points). The computation time is linear in the number of points.

4.1. Reconstructing a Calibration Pattern

A calibration pattern (cf. Fig. 2) is viewed from the top of the SCORBOT. The pattern simplifies the acquisition of 2-D points, and allows us to compare our results with ground truth data. After the initialization, we start the optimization process to take new images from the optimal view point.

Table 4.1 shows the results for the first 5 iterations in the optimized case and a non-optimized one. The images

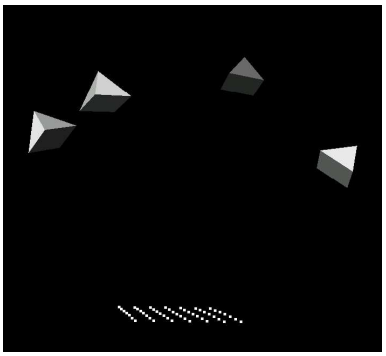


Figure 3. View points for reconstruction of the calibration pattern

for the non-optimized view points were taken by alternating between the two initial positions.

By construction, the determinant of $\boldsymbol{P}_t$ is reduced faster in the optimized case than in the non-optimized case. Additionally, the mean of the errors of all points decreases after each time step, except for some outliers. This rise in error is not a contradiction to the decrease in uncertainty, since the Kalman filter cannot judge the quality of an observation.

The view points are shown in Fig. 3. After the initialization steps (middle top) the optimized view points lie as expected: the cameras are opposite each other and the angle between each line of sight is approx. 90 degrees.

Table 1. First experiment: μ_t is the mean of the difference between reconstructed points and the ground truth data in mm, σ_t is the standard deviation of this error, $|\boldsymbol{P}_t|$ is the determinant of the covariance matrix. We display the values for the optimized and a non-optimized view point sequence.

	optimized			non-optimized		
t	μ_t	σ_t	$\lvert\boldsymbol{P}_t\rvert$	μ_t	σ_t	$\lvert\boldsymbol{P}_t\rvert$
1	0.132	0.080	7.281	0.132	0.080	7.281
2	0.128	0.079	1.762	0.125	0.072	3.338
3	0.115	0.062	0.705	0.128	0.073	1.468
4	0.108	0.062	0.385	0.129	0.074	0.905
5	0.107	0.061	0.244	0.127	0.074	0.531

4.2. Reconstructing a Mouse Pad

In this experiment we use a mouse pad (cf. Fig. 2), requiring us to track feature points during movement, using the algorithm of Zinsser [17]. However, only the tracked points from the optimal positions are used to update the

state estimation. Integration of the points tracked *en route* to the optimal positions is possible, but this would prevent a comparison of two view point sequences due to a diverging number of integrated observations.

Table 4.2 shows the root mean square error between the reconstructed 3-D points and their regression plane, as well as the trend of the covariance matrix $\boldsymbol{P}_t$, for the first 5 iterations. We compare the values from the optimized view points to an experiment with view points uniformly distributed on a circle perpendicular to the rotation axis of the turn table, and to one completely random view point sequence on the half sphere. The error decreases fastest in the optimized case, signifying a measurable benefit from view point optimization.

Table 2. Second experiment: μ_t is the mean of the root mean square error of the points to their regression plane in mm, $|\boldsymbol{P}_t|$ the determinant of the covariance matrix after each iteration. The optimized, one uniform and one random view point sequence are shown.

	optimized		circle		random	
t	μ_t	$\lvert\boldsymbol{P}_t\rvert$	μ_t	$\lvert\boldsymbol{P}_t\rvert$	μ_t	$\lvert\boldsymbol{P}_t\rvert$
1	0.073	8.62	0.073	8.62	0.073	8.65
2	0.050	1.75	0.041	1.98	0.054	2.76
3	0.033	0.636	0.038	0.845	0.043	1.20
4	0.030	0.315	0.038	0.428	0.041	0.479
5	0.026	0.175	0.041	0.235	0.041	0.329

5. Conclusion And Future Work

We have presented an approach for planning the next best view for 3-D reconstruction. To evaluate view points, we reshape the problem of 3-D reconstruction to the problem of estimating a partially observable system with an extended Kalman filter.

The Kalman filter allows us to predict the expected uncertainty of the state estimate, measured by the entropy, for any view point. By selecting the next view with the least expected entropy, we can minimize the uncertainty, and thereby the state estimation error, of our reconstruction. The approach was tested in experiments with real images, which show the error decreasing faster than with non-optimized view points.

In the future, we will expand this approach to the problem of self-occlusions. Without handling this constraint, our approach cannot reconstruct complex objects in an optimal way. This requires an object surface to be estimated, in order to determine where tracked points may be occluded by other surface parts.

References

[1] T. Arbel and F. Ferrie. Entropy-based gaze planning. *Image and Vision Computing*, 19(11):779–786, September 2001.

[2] J. E. Banta, L. R. Wong, C. Dumont, and M. A. Abidi. A next-best-view system for autonomous 3-d object reconstruction. *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, 30(5):589–598, 2000.

[3] J. J. Craig. *Introduction to Robotics: Mechanics and Control*. Prentice Hall, Upper Saddle River, USA, 3-rd edition, 2004.

[4] J. Denzler and C. Brown. An information theoretic approach to optimal sensor data selection for state estimation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(2):145–157, 2002.

[5] J. Denzler, M. Zobel, and H. Niemann. Information Theoretic Focal Length Selection for Real-Time Active 3-D Object Tracking . In *International Conference on Computer Vision* , pages 400–407, Nice, France, 2003. IEEE Computer Society Press.

[6] A. Gelb. *Applied Optimal Estimation*. MIT Press, Cambridge, 1974.

[7] Y. Hung and H. Ho. A Kalman filter approach to direct depth estimation incorporating surface structure. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21(6):570–575, June 1999.

[8] K. N. Kutulakos and C. R. Dyer. Recovering shape by purposive viewpoint adjustment. *International Journal of Computer Vision*, 12(2-3):113–136, 1994.

[9] E. Marchand and F. Chaumette. Active vision for complete scene reconstruction and exploration. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21(1):65–72, 1999.

[10] L. Matthies, T. Kanade, and R. Szeliski. Kalman filter-based algorithms for estimating depth from image sequences. *International Journal of Computer Vision*, 3(3):209–238, 1989.

[11] A. Papoulis. *Probability Random Variables, and Stochastic Processes*. McGraw-Hill, Inc, Singapore, 4-th edition, 2002.

[12] R. Pito. A solution to the next best view problem for automated surface acquisition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21(10):1016–1030, 1999.

[13] J. Schmidt, F. Vogt, and H. Niemann. Vector Quantization Based Data Selection for Hand-Eye Calibration . In *Vision, Modeling, and Visualization 2004* , pages 21–28, Stanford, USA, 2004. Aka / IOS Press, Berlin, Amsterdam.

[14] W. R. Scott, G. Roth, and J.-F. Rivest. View planning for automated three-dimensional object reconstruction and inspection. *ACM Computing Surveys*, 35(1):64–96, 2003.

[15] R. Tsai. A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses. *IEEE Journal on Robotics and Automation*, 3(4):323–344, Aug. 1987.

[16] Y.-K. Yu, K.-H. Wong, and M. M.-Y. Chang. Recursive 3d model reconstruction based on Kalman filtering. *IEEE Transactions on Systems, Man and Cybernetics - Part B*, 35(3):587–592, July 2003.

[17] T. Zinßer, C. Gräßl, and H. Niemann. Efficient Feature Tracking for Long Video Sequences . In *Pattern Recognition, 26th DAGM Symposium* , Lecture Notes in Computer Science , pages 326–333, Tübingen, 2004. Springer, Berlin, Heidelberg, New York.