Determining an initial image pair for fixing the scale of a 3d reconstruction from an image sequence

Christian Beder and Richard Steffen

Institute for Photogrammetry
Bonn University, Germany
beder@ipb.uni-bonn.de
rsteffen@uni-bonn.de

Abstract. Algorithms for metric 3d reconstruction of scenes from calibrated image sequences always require an initialization phase for fixing the scale of the reconstruction. Usually this is done by selecting two frames from the sequence and fixing the length of their base-line. In this paper a quality measure, that is based on the uncertainty of the reconstructed scene points, for the selection of such a stable image pair is proposed. Based on this quality measure a fully automatic initialization phase for simultaneous localization and mapping algorithms is derived. The proposed algorithm runs in real-time and some results for synthetic as well as real image sequences are shown.

1 Introduction

In recent years the fully automatic 3d reconstruction of scenes and camera trajectories from monocular image sequences has received a lot of attention. In the early work of [7] and [6], the extraction of feature points together with their uncertainty represented by covariance matrices was developed. More recently, feature extraction and tracking of features across image sequences was improved by [24], [14] and [13]. This reliable feature extraction methods enabled the 3d reconstruction from image sequences (e.g. [1], [29],[18]) using robust estimation of the epipolar geometry. The use of self-calibration techniques (cf. [20],[19]) or prior knowledge of the internal camera calibration leads to a metric 3d reconstruction, that is defined up to a similarity transformation. In the calibrated case efficient real-time algorithms, that are also able to cope with planar scenes were developed by [17] and [25]. Starting from this prerequisites the field of real-time simultaneous localization and mapping has recently emerged and was given much attention by many researchers (e.g. [4],[3],[5],[27],[15],[23]).

As the calibrated 3d reconstruction is only defined up to a similarity transformation, somehow fixing the scale is an important task. Although scale is a gauge parameter and therefore does not affect the overall accuracy of the reconstruction it does affect the stability of the reconstruction algorithms and must therefore be chosen carefully. Usually this is done by initially selecting two reference images

and fixing the length of their base-line. Those key-frames are selected based on image sharpness and disparity (cf. [16]), based on the distribution of matched points in the images (cf. [12]), based on selecting the most appropriate motion model (cf. [28],[21],[22]) or based on evaluating the bundle-adjustment of the whole sequence (cf. [26]).

The contribution of this work is to present a statistically motivated measure for the quality of the pair of reference images. Based on this quality measure an efficient algorithm is proposed, that automates the manual setting of initial number of frames or the initialization phase required for example by the approach of [4]. It turns out, that, in case of known internal camera calibration, this is a very efficient alternative to the model selection approach of [28],[21] and [22], who proposed to decide when the base-line is large enough by checking, if the image pair is related only by a homography or the full epipolar constraint. A drawback of this approach is, that it cannot handle planar objects, which our approach can.

To achieve this goal, another subject of recent computer vision research is employed. It has been studied by [2] [11], [10],[8] and [9], how uncertainties can be efficiently represented and propagated for geometric reasoning tasks involving projective geometric entities. Especially the work of [9], who showed how covariance matrices easily transform for various projective geometric constructions, plays a key role in this work.

The paper is organized as follows: In section 2 first the shape of confidence ellipsoids of scene points resulting from a given point correspondence and camera pose is exploited. Based on this shape, more explicitly its roundness, a measure for the quality of the image pair for the task of fixing the scale is derived. In section 3 an algorithm is outlined, that is used to determine the optimal image pair for fixing the scale of a 3d reconstruction. Finally some results on simulated and real image sequences are shown in section 4.

2 Confidence Ellipsoids of Scene Points

Now the propagation of uncertainty from two measured corresponding image points on the reconstructed scene point is analyzed. If a scene point $\mathbf{X}$ is observed by two projective cameras P' and P'', the image coordinates are

$$\mathbf{x}' \cong \mathsf{P}'\mathbf{X} \tag{1}$$

and

$$\mathbf{x}'' \cong \mathsf{P}''\mathbf{X} \tag{2}$$

Denoting with $\mathsf{S}(\cdot)$ the first two rows of the skew-symmetric matrix inducing the cross-product

$$\mathsf{S}(\mathbf{x}) = \begin{pmatrix} 0 & -x_3 & x_2 \\ x_3 & 0 & -x_1 \end{pmatrix} \tag{3}$$

the two conditions can be written as

$$\mathsf{S}(\mathbf{x}')\mathsf{P}'\mathbf{X} = -\mathsf{S}(\mathsf{P}'\mathbf{X})\mathbf{x}' = \mathbf{0} \tag{4}$$

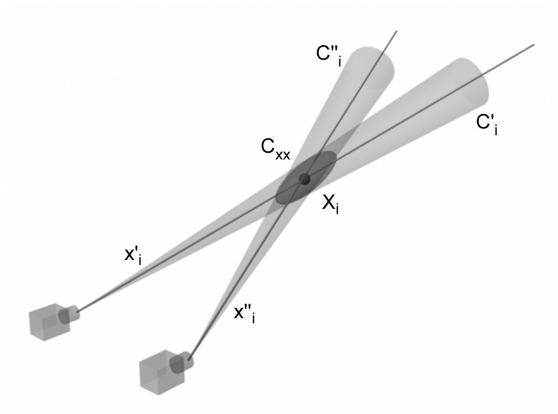

Fig. 1. Scene geometry: Projecting rays of two corresponding image points $\mathbf{x}'_i$ and $\mathbf{x}''_i$ together with their uncertainties C'_i and C''_i are observed by two cameras. The corresponding scene point $\mathbf{X}_i$ has the uncertainty ellipsoid C_{XX}. The roundness of this object, i.e. the ratio of its smallest and longest axis, is a measure of the quality of the scene geometry.

and

$$\mathsf{S}(\mathbf{x}'')\mathsf{P}''\mathbf{X} = -\mathsf{S}(\mathsf{P}''\mathbf{X})\mathbf{x}'' = \mathbf{0} \tag{5}$$

if the image points are not at infinity. Both expressions are linear in the scene point as well as in the image points, i.e.

$$\underbrace{\begin{pmatrix} \mathsf{S}(\mathbf{x}')\mathsf{P}' \\ \mathsf{S}(\mathbf{x}'')\mathsf{P}'' \end{pmatrix}}_{\substack{\mathsf{A} \\ 4\times 4}} \mathbf{X} = \mathbf{0} \tag{6}$$

and

$$\underbrace{\begin{pmatrix} -\mathsf{S}(\mathsf{P}'\mathbf{X}) & 0 \\ 0 & -\mathsf{S}(\mathsf{P}''\mathbf{X}) \end{pmatrix}}_{\substack{\mathsf{B} \\ 4\times 6}} \begin{pmatrix} \mathbf{x}' \\ \mathbf{x}'' \end{pmatrix} = \mathbf{0} \tag{7}$$

Now the scene point coordinates and the two image point coordinates are assumed to be random variables. Note that, as all three quantities are homogeneous, the covariance matrices of their distributions are singular. Let the covariance matrices of the image points $\mathbf{x}'$ and $\mathbf{x}''$ be given by C' and C'' respectively, then it has been shown by [9], that the covariance matrix C_{XX} of the distribution of the scene point coordinates $\mathbf{X}$ is proportional to the upper left 4×4-submatrix

$$\mathsf{C}_{XX} = \left(N^{-1}\right)_{1:4,1:4} \tag{8}$$

of the inverse of

$$\underset{5\times5}{\mathsf{N}} = \begin{pmatrix} \mathsf{A}^T \left(\mathsf{B} \begin{pmatrix} \mathsf{C}' & \mathsf{0} \\ \mathsf{0} & \mathsf{C}'' \end{pmatrix} \mathsf{B}^T \right)^{-1} \mathsf{A} & \mathbf{X} \\ \mathbf{X}^T & 0 \end{pmatrix} \tag{9}$$

Note, that no specific distribution must be assumed, as all arguments regard only the second moments. We have neglected the effect of the uncertainty of the projection matrices P' and P'' here, as the relative orientation of the two cameras is determined by many points, so that it is of superior precision compared to a single point.

Now the effect of normalizing the homogeneous vector $\mathbf{X} = [\boldsymbol{X}_0^T, X_h]^T$ to Euclidean coordinates on its covariance matrix C_{XX} is analyzed. For this to be meaningful it is assumed, that the cameras are calibrated, so that the reconstruction is Euclidean, i.e. defined up to a similarity transformation. The Jacobian of a division of $\boldsymbol{X}_0$ by X_h is

$$\mathsf{J}_e = \frac{\partial \boldsymbol{X}_0 / X_h}{\partial \boldsymbol{X}_0} = \frac{1}{X_h} \left(\mathsf{I}_{3\times3} - \frac{\boldsymbol{X}_0}{X_h} \right) \tag{10}$$

and hence by linear error propagation the covariance matrix of the distribution of the corresponding Euclidean coordinates is

$$\mathsf{C}^{(e)} = \mathsf{J}_e \mathsf{C}_{XX} \mathsf{J}_e^T \tag{11}$$

The roundness of the confidence ellipsoid is directly related to the condition number of the 3d reconstruction of the point. Therefore it is a measure, how well the two camera poses are suited for the task of 3d reconstruction. Hence, using the singular value decomposition of this covariance matrix

$$\mathsf{C}^{(e)} = \mathsf{U} \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix} \mathsf{V}^T \tag{12}$$

its roundness is defined as the square root of the quotient of the smallest and the largest singular value

$$R = \sqrt{\frac{\lambda_3}{\lambda_1}} \tag{13}$$

This measure lies between zero and one, is invariant to scale changes and only depends on the relative geometry of the two camera poses, the image points and the object. If the two camera centers are the same, it is equal to zero. If the object is equally far away from the two cameras and the projecting rays of the image points are orthogonal and their covariance matrices are identical and isotropic, it is equal to $\sqrt{\frac{1}{2}}$. This is the maximum under the assumption of isotropic covariance matrices. The maximum of one is reached for the same configuration as before except for the covariance matrices of the projecting rays. The principal axis of this covariance matrices must therefore be aligned with the epipolar plane with the extension perpendicular to it being $\sqrt{2}$ times the extension perpendicular to the viewing direction.

3 Determining the optimal image pair

Now the roundness measure of the previous section will be put into the context of finding the optimal image pair for fixing the scale of a 3d reconstruction. Of course the global optimal solution can only be found by checking all image pairs. As our intended application is the real-time initialization of a simultaneous localization and mapping system, the proposed algorithm fixes the first frame of the sequence and terminates, when the first acceptable second frame is reached. The acceptability will be determined via the roundness of the confidence ellipsoids of the reconstructed scene points. The details are as follows:

1. Fix the first image of the sequence and let its projection matrix be
$$\mathsf{P}' = [\mathsf{I}|\mathbf{0}]$$
2. Extract the interest points $\mathbf{q}'_i$ together with their covariance matrices $\mathsf{C}_{q'_i q'_i}$ from this image and apply the known camera calibration matrix K to the image coordinates and their covariance matrices, yielding the directions
$$\mathbf{x}'_i = \mathsf{K}^{-1}\mathbf{q}'_i$$
and their covariance matrices
$$\mathsf{C}'_i = \mathsf{K}^{-1}\mathsf{C}_{q'_i q'_i}\mathsf{K}^{-T}$$
3. For each new image of the sequence do the following
   (a) Extract the interest points $\mathbf{q}''_i$ together with their covariance matrices $\mathsf{C}_{q''_i q''_i}$ from this new image and apply the known camera calibration matrix yielding again the directions
   $$\mathbf{x}''_i = \mathsf{K}^{-1}\mathbf{q}''_i$$
   and their covariance matrices
   $$\mathsf{C}''_i = \mathsf{K}^{-1}\mathsf{C}_{q''_i q''_i}\mathsf{K}^{-T}$$
   (b) Determine the point correspondences $\mathbf{x}'_i \leftrightarrow \mathbf{x}''_i$ and relative orientation R, $\mathbf{t}$ to the first image of the sequence according to the algorithm proposed in [17]. The camera matrix for the current image is then
   $$\mathsf{P}'' = [R|-\mathbf{t}]$$
   (c) Determine the scene point positions $\mathbf{X}_i$ for each found correspondence by forward intersection. This can for example be done by solving the homogeneous equation system (6) using the singular value decomposition of the matrix A.
   (d) Determine the roundness R_i (cf. equation (13)) of each scene point $\mathbf{X}_i$'s confidence ellipsoid as outlined in the previous section.
   (e) If the mean roundness is above a given threshold T, use this image pair to fix the scale of the reconstruction and continue with the main application.

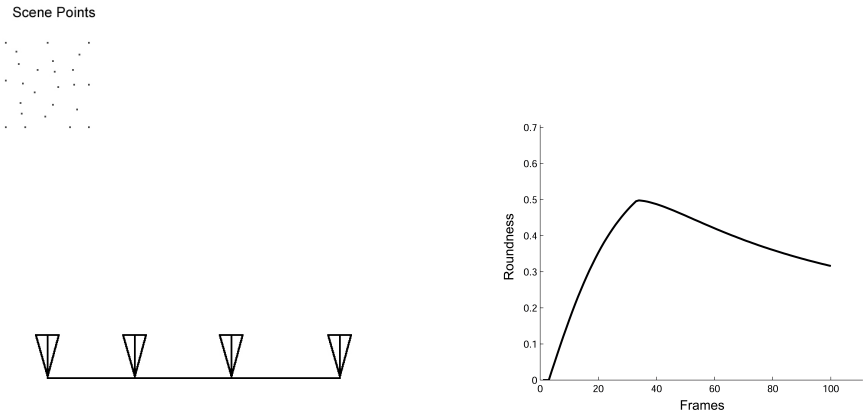

Fig. 2. Left: Synthetic image sequence trajectory of the translation experiment, where the camera faces the object and is moved to the right. Right: Roundness measure against video frame for the synthetic translation experiment. A maximum is reached at the frame, where the angle of the rays is approximately 35°. As the distance to the object increases, the roundness decreases again.

4 Results

To evaluate the usefulness of the proposed roundness measure, experiments on synthetic as well as real image sequences were carried out. The setup for the synthetic experiments is shown on the left hand sides of figure 2 and figure 3. The cameras where assumed to be normalized and the image points where assumed to have isotropic and equal covariance matrices. Note, that by definition the overall scale is irrelevant, since the proposed roundness measure only depends on the relative scales and is therefore scale-invariant.

In the first experiment, depicted in figure 2, the camera was facing the object and then moved to the right. The resulting roundness measure is shown on the right hand side in figure 2. It can be observed, that a maximum roundness is reached, where the angle of the projecting rays is approximately 35°. As the distance of the second camera to the object increases, the roundness decreases again. The optimal image pair, i.e. the pair yielding highest stability, is therefore not only dependent on the intersection angle of the projecting rays, but also on the relative distances of the cameras to the object.

The second synthetic experiment was to rotate the camera around the object at equal distance as depicted on the left hand side of figure 3. The resulting roundness measure is shown on the right hand side in figure 3. It can be seen, that it directly corresponds to the rotation angle and the maximum of $\sqrt{\frac{1}{2}}$ is reached at an angle of 90°, where the intersection of the rays is optimal for the accuracy of the 3d reconstruction.

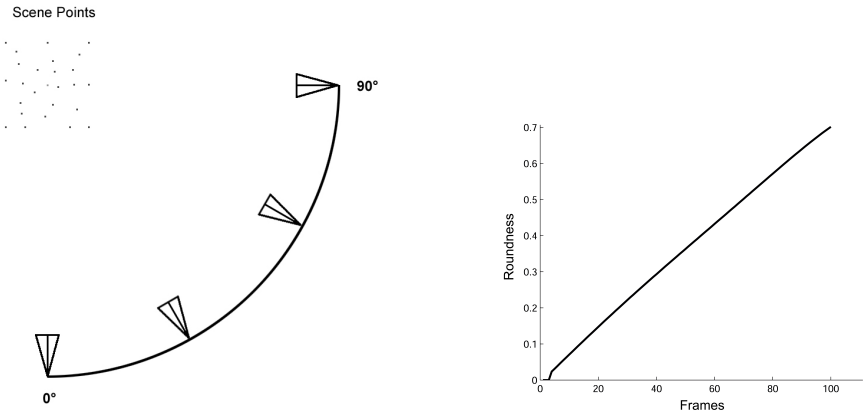

Fig. 3. Left: Synthetic image sequence trajectory of the rotation experiment, where the camera is moved at equal distance around the object. Right: Roundness measure against video frame for the synthetic rotation experiment. The maximum of $\sqrt{\frac{1}{2}}$ is reached at the angle of 90°.

Finally a real image sequence was taken using a cheap hand-held consumer web-cam. Two exemplary frames are depicted in figure 4. Features were extracted and tracked and the roundness measure was computed for each frame with respect to the first image, which is depicted on the left hand side in figure 4. The resulting roundness measure is shown in figure 5. In the first 25 frames the camera was not moved, so that the roundness measure stays near to zero. When the camera starts moving, the expected accuracy of the depth of the 3d reconstruction, and hence the proposed roundness measure, is increasing. After about 110 frames the roundness measure reached the threshold $T = \sqrt{\frac{1}{10}}$. This threshold is not a critical parameter but a minimal requirement stemming from the goal of achieving a condition number of approximately 10 for the 3d reconstruction. The corresponding last frame is depicted on the right hand side in figure 4. It can be seen, that still enough corresponding points can be identified, so that the determination of the relative orientation between the frames is not an issue. Note also, that all processing was performed in real-time.

5 Conclusion

A fully automatic real-time algorithm for initially fixing the scale of the 3d reconstruction in simultaneous localization and mapping applications with calibrated cameras was proposed. As the metric reconstruction is fixed up to a similarity transformation in the calibrated case, the shape of the confidence ellipsoids of reconstructed scene points is a meaningful quantity. The roundness of this confidence ellipsoids can be used to decide, when the accuracy of the

Fig. 4. Left: First image of the real image sequence. Right: Last image of the real image sequence, where the roundness of the scene point covariance matrices was sufficiently high.

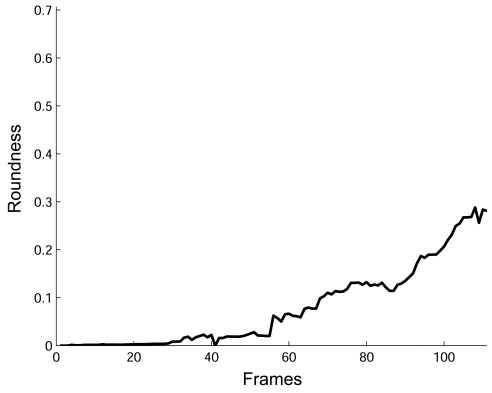

Fig. 5. Roundness measure against video frame for the real image sequence. The camera was not moved for the first 25 frames. The threshold value of $T = \sqrt{\frac{1}{10}}$ was reached after the movement was sufficiently large on the frame depicted on the right in figure 4.

reconstruction is most stable, as it is directly related to the condition number of the 3d reconstruction of the point. Hence, choosing the image pair for which this roundness is maximal is the most stable choice for initially fixing the scale of a 3d reconstruction.

The proposed algorithm was demonstrated to work on synthetic as well as real monocular image sequences. Since the most complex operations are only the inversion of a 5×5-matrix and two singular value decompositions of a 4×4- and a 3×3-matrix, the dominant part of the computation time is taken by the feature extraction and tracking, enabling a real-time initialization phase.

References

1. Paul A. Beardsley, Philip H. S. Torr, and Andrew Zisserman. 3d model acquisition from extended image sequences. In *Proc. of ECCV (2)*, pages 683–695, 1996.
2. A. Criminisi, I. Reid, and A. Zisserman. A plane measuring device. *Image and Vision Computing*, 17(8):625–634, 1999.
3. A. J. Davison and D. W. Murray. Simultaneous localisation and map-building using active vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):865–880, July 2002.
4. A.J. Davison. Real-time simultaneous localisation and mapping with a single camera. In *Proc. International Conference on Computer Vision, Nice*, pages 1403–1410, October 2003.
5. J. Diebel, K. Reuterswrd, S. Thrun, J. Davis, and R. Gupta. Simultaneous localization and mapping with active stereo vision. In *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3436–3443, 2004.
6. W. Förstner. A framework for low level feature extraction. In *Proc. of European Conference on Computer Vision*, pages 383–394, 1994.
7. C.G. Harris and M.J. Stephens. A combined corner and edge detector. In *Fourth Alvey Vision Conference*, pages 147–151, 1988.
8. S. Heuel and W. Förstner. Matching, reconstructing and grouping 3d lines from multiple views using uncertain projective geometry. In *CVPR '01*. IEEE, 2001.
9. Stephan Heuel. *Uncertain Projective Geometry - Statistical Reasoning for Polyhedral Object Reconstruction*, volume 3008 of *LNCS*. Springer, 2004.
10. Kenichi Kanatani. Uncertainty modeling and model selection for geometric inference. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 26(10):1307–1319, 2004.
11. Kenichi Kanatani and Daniel D. Morris. Gauges and gauge transformations for uncertainty description of geometric structure with indeterminacy. *IEEE Transactions on Information Theory*, 47(5):2017–2028, July 2001.
12. Reinhard Koch, Marc Pollefeys, and Luc Van Gool. Robust calibration and 3d geometric modeling from large collections of uncalibrated images. In W. Förstner, J. Buhmann, A. Faber, and P. Faber, editors, *Proceedings of the DAGM*, Informatik Aktuell, pages 412–420. Springer, 1999.
13. David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004.
14. J. Matas, O.Chum, M.Urban, and T.Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *BMVC*, pages 384–393, 2002.
15. Jason Meltzer, Rakesh Gupta, Ming-Hsuan Yang, and Stefano Soatto. Simultaneous localization and mapping using multiple view feature descriptors. In *Proc. of IROS*, 2004.
16. David Nistér. Frame decimation for structure and motion. In *SMILE*, pages 17–34, 2000.
17. David Nistér. An efficient solution to the five-point relative pose problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(6):756–777, 2004.
18. M. Pollefeys, R. Koch, M. Vergauwen, and L. Van Gool. Automated reconstruction of 3d scenes from sequences of images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 55(4):251–267, 2000.
19. Marc Pollefeys and Luc Van Gool. A stratified approach to metric self-calibration. In *Proc. CVPR*, pages 407–412, 1997.

20. Marc Pollefeys and Luc Van Gool. Stratified self-calibration with the modulus constraint. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21(8):707–724, 1999.
21. Marc Pollefeys, Luc Van Gool, Maarten Vergauwen, Kurt Cornelis, Frank Verbiest, and Jan Tops. Video-to-3d. In *Proceedings of Photogrammetric Computer Vision*, 2002.
22. Jason Repko and Marc Pollefeys. 3d models from extended uncalibrated video sequences: Addressing key-frame selection and projective drift. In *Proc. of 3DIM*, 2005.
23. S. Se, D. G. Lowe, and J. J. Little. Vision-Based Global Localization and Mapping for Mobile Robots. *IEEE Transactions on Robotics*, 21(3):364–375, 2005.
24. Jianbo Shi and Carlo Tomasi. Good features to track. In *1994 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'94)*, pages 593 – 600, 1994.
25. Henrik Stewenius, Christopher Engels, and David Nister. Recent developments on direct relative orientation. *ISPRS Journal*. to appear.
26. Thorsten Thormählen, Hellward Broszio, and Axel Weissenfeld. Keyframe selection for camera motion and structure estimation from multiple views. In *Proceedings of European Conference on Computer Vision*, pages 523–535, 2004.
27. S. Thrun, M. Montemerlo, D. Koller, B. Wegbreit, J. Nieto, and E. Nebot. Fastslam: An efficient solution to the simultaneous localization and mapping problem with unknown data association. *Journal of Machine Learning Research*. To appear.
28. Philip H. S. Torr. Bayesian model estimation and selection for epipolar geometry and generic manifold fitting. *International Journal of Computer Vision*, 50(1):35–61, 2002.
29. Philip H. S. Torr and Andrew Zisserman. Feature based methods for structure and motion estimation. In *Workshop on Vision Algorithms*, pages 278–294, 1999.