Indoor photorealistic 3D mapping using stereo images from SLR cameras

Viktor Kolbe
C3 Technologies
Linköping, Sweden
{*viktor.kolbe*}
@*c3technologies.com*

Folke Isaksson, Thomas Beckman
Saab Bofors Dynamics
Linköping, Sweden
{*folke.isaksson, thomas.beckman*}
@*saabgroup.com*

Thomas B. Schön
Division of Automatic Control
Linköping University
Linköping, Sweden
{*schon*}
@*isy.liu.se*

Abstract

Creating a 3D model from photos require an estimate of the position and orientation (pose) of the camera for each photo that is acquired. This paper presents a method to estimate the camera pose using only image data. The images are acquired at a low frequency using a stereo rig, consisting of two rigidly attached SLR cameras. Features are extracted and an optimization problem is solved for each new stereo image. The results are used to merge multiple stereo images and building a larger model of the scene. The accumulated error after processing 10 images can with the present methods be less than 1.2 mm in translation and 0.1 degrees in rotation.

1. Introduction

The problem of creating 3D-models from images is not new, and many papers have been published on this matter, see for example [1]–[4]. The traditional aim is to produce a real-time system with low resolution video cameras and high image frequency. [5] assumes a world with only straight lines and 90 degree angles. With this information a 3D-model of the world is reconstructed from a single photo. This assumption limits this system to a very small range of environments.

The aim in this paper is to describe a method to compute estimates of the camera pose using stereo image data only. These estimates are then to be used to create a scene model. The department for Sensor Systems at Saab Bofors Dynamics has developed a system for creating three dimensional maps from high resolution aerial photos. The elevation of each point is determined through a stereo calculation between two subsequent photos. In order to determine the distance between the cameras taking the photos, each photo is associated with a position from a GPS and measurements from an inertial navigation system. This data is also used to merge each height map in the software to form a coherent model of the ground.

The same software system has the possibility to process photos of an indoor environment, but one of the problems is to get the exact pose for each photo, this as there is generally no possibility to use GPS indoors. The method described in this paper has been developed during a master thesis project and can be used in this system to determine the pose of each new stereo image.

2. Hardware configuration

Several different hardware setups are possible when trying to solve the problem of creating a 3D model of an indoor environment. Potential sensors besides cameras might be inertial sensors, laser range sensors, ultra wide band and ultrasonic sensors. If it would not be for bad reception in an indoor environment, a GPS would also be a natural sensor to use. Numerous configurations have already been tried, [6] building a model of an urban environment using an inertial navigation system, a GPS, and four video cameras mounted with different viewing angles. Another configuration, used in [3], is three video stereo pairs mounted orthogonally to each other, covering a wider area, to minimize the weak points from a single stereo camera system. Another approach is to use only a single video camera to solve the SLAM problem [7].

This paper presents a solution that uses two calibrated SLR cameras mounted on a stereo rig, see Figure 1. The distance between the cameras has been set to about 20 centimetres to minimize the areas seen by only one camera, but at the same time giving a good-enough measurement of the disparity in an indoor environment. The benefit of using two cameras mounted together instead of a single camera is that it is as easy to calculate the disparity in all images as the relation between the cameras is fixed. The scale in the images is also known since the rig is fixed and the distance between the cameras is known.

An inertial measurement unit (IMU) could give valuable support when positioning the camera, but in this work an IMU is not used. This decision is made to simplify the hardware and minimize the integration problems, and to examine if it is possible to solve this problem with only image data. Due to the problem with reception indoors, a solution with GPS has not been considered.

The lenses used on the cameras are wide-angle fish-eye lenses, which makes the overlapping of images simpler, as

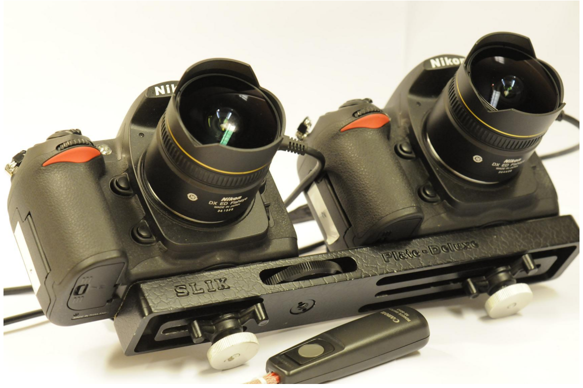


Figure 1: Stereo image registration hardware consisting of two 10 megapixel Nikon D200 DSLR cameras with Nikon 10.5mm f2.8 lenses and a common trigger.

the cameras can move more and still capture overlapping photos with a low frequency.

3. Feature extraction and outlier rejection

The scale-invariant feature transform (SIFT) introduced in [8] is used to extract corresponding features between the four images in two stereo pairs. SIFT is a rotation and affine invariant Harris point operator. Every keypoint is assigned a scale, location and an orientation, which ensures that the keypoints are described in a way that is invariant to location, scale and orientation of the image. The keypoints are converted into descriptor vectors, that can be compared to find similar and matching points.

To remove outliers from the features extracted from the images, the random sample consensus (RANSAC) algorithm has been used. RANSAC can handle a large amount of outliers, and the algorithm is described in [9] p. 118. Some modifications are made to the algorithm to always perform a fixed number of iterations, and a few threshold values are added to decrease the processing time.

4. Algorithms

To estimate the relative pose between each stereo image from the corresponding points, two types of algorithms have been subject to experiments.

4.1. Eight-point algorithm

The eight-point algorithm presented in [10] p. 121 Algorithm 5.1, is a closed form solution that uses SVD to compute the essential matrix. The essential matrix can then be decomposed into a translation and a rotation matrix. This approach together with some improvements, proposed by [11], performed well with fictional data. Unfortunately this method produced results which were completely out of bounds when using real data from the images.

4.2. Newton-Raphson minimization

This method minimizes a cost function, that is presented in Section 5, to iteratively estimate the pose of the stereo images. The cost function is minimized with the Newton-Raphson method, which is a method to successively find better approximations to the roots of a real value function, [12] p. 331. It is an iterative method, where each step is calculated according to

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} \tag{1}$$

The iterations are aborted after a certain number of iterations have been performed, or when the absolute relative error ϵ_a is less than a specified relative error ϵ_s. The absolute relative error is calculated as

$$\epsilon_a = \left| \frac{x_{i+1} - x_i}{x_{i+1}} \right| \tag{2}$$

When starting the iterations, the starting value of x_i, a vector containing all the parameters that is to be estimated, needs to have a value “sufficiently near” the root of the function for the iterations to converge to the global optimum. During this work, “sufficiently near” has meant placing each new stereo image at the position of the last positioned stereo pair.

5. Experiments

Experiments have been made with different cost functions in the minimization described above. The cost function used in the first experiments can be described as

$$\sqrt{\frac{1}{NJ}\sum_{n=1}^{N}\sum_{j=1}^{J}(u_{n,j}^{\text{proj}} - u_{n,j})^2 + (v_{n,j}^{\text{proj}} - v_{n,j})^2} \tag{3}$$

where N is the number of points and J is the number of images in the two stereo pairs, normally 4. The estimation of the position of each point in the 3D space is denoted $[x_n, y_n, z_n]^T$ where $n = 1, \ldots, N$. The position of each point in the image is denoted $[u_{n,j}, v_{n,j}]^T$ where $j = 1, \ldots, J$. The projection of the 3D-point onto the image plane is denoted $\left[u_{n,j}^{\text{proj}}, v_{n,j}^{\text{proj}}\right]^T$.

To measure the result of the experiments, ten stereo images have been taken while holding the camera rig by hand, and rotating 360 degrees. After estimating the pose of all the stereo images in the sequence, the first stereo image is reestimated as being the last in the sequence. The difference in the pose of the first stereo images before and after the estimation is used as a measurement of the accumulated error for all the poses. Some examples of typical errors for different amounts of corresponding points between the images can be seen in Table 1. To estimate the pose with a higher accuracy than about 2 cm in translation and about 0.5 degrees in rotation as an accumulated fault

Table 1: Example of errors for a sequence of 10 stereo images.

Parameter	Number of points		
	22	45	54-90
x (mm)	3.4	8.2	2.2
y (mm)	18.2	15.1	14.2
z (mm)	17.6	11.2	10.9
ψ (deg)	0.30	0.47	0.39
φ (deg)	0.44	0.05	0.05
θ (deg)	0.47	0.27	0.05

after 10 images, a global optimization while identifying a loop closure can be used. In these experiments, loop closures have been identified manually, but this can also be done automatically by comparing the SIFT points. To be able to measure the error in the estimation, the loop is closed after 20 images taken while rotating 720 degrees, but the error of the estimation is measured as before, after 360 degrees of rotation and 10 images. Some examples of accumulated errors in the estimation when using loop closuring can be seen in Table 2. In many of the images used in these experiments, the overlap between stereo images is large enough to allow measurements between each new stereo image and two earlier stereo images in the sequence. This can be described with the cost function

$$\sqrt{\frac{1}{6N_{\text{num}}}\sum_{\forall n\in N}\sum_{j=J_{\text{start}}}^{J_{\text{end}}}(u_{n,j}^{\text{proj}}-u_{n,j})^2+(v_{n,j}^{\text{proj}}-v_{n,j})^2} \quad (4)$$

where $J_{\text{start}} = 2*(I_{\text{nr}}-2)-1$ and $J_{\text{end}} = 2I_{\text{nr}}$. I_{nr} is the number of the stereo pair that is being calculated. N is in this case all the points that connect I_{nr} with $I_{\text{nr}}-1$ and $I_{\text{nr}}-2$, and N_{num} is the number of points in N. As more information is used to position each new stereo pair, the accumulated fault after 10 stereo images decreases, one example of the decrease is shown in Table 3. The result shows notable improvements compared with the result from only matching against one image.

Table 2: Example of errors for a sequence of 10 stereo images, with and without loop closing after 20 stereo images

	Without loop closing	With loop closing
x (mm)	2.0	0.5
y (mm)	5.1	1.1
z (mm)	6.6	1.1
ψ (deg)	0.305	0.065
φ (deg)	0.082	0.096
θ (deg)	0.143	0.090

The estimation in the Newton-Raphson iterations will be improved as more information is given as input. Therefore, the information from every image that has some overlap to the new stereo image, should be used. This can be done by searching for correspondences between SIFT points in the stereo image that is going to be added and the points in all of the already positioned stereo images. The cost function used in this method is

$$\sqrt{\frac{1}{N_{\text{num}}J_{\text{num}}}\sum_{\forall n\in N}\sum_{\forall j\in J}(u_{n,j}^{\text{proj}}-u_{n,j})^2+(v_{n,j}^{\text{proj}}-v_{n,j})^2} \quad (5)$$

where J is all the images that is already positioned and N is all the points connecting the images in J. J_{num} is the number of images already positioned and N_{num} is the number of points in N.

Table 3: Example of errors for a sequence of 10 stereo images, while matching each new stereo image against one or two previous stereo images in the sequence.

	Using 45 points	
Nr of stereo pairs	1	2
x (mm)	8.2	7.1
y (mm)	15.1	6.5
z (mm)	11.2	4.5
ψ (deg)	0.47	0.20
φ (deg)	0.05	0.01
θ (deg)	0.27	0.03

With this approach, there is no longer a need for having the stereo images in a predefined sequence. Instead, each time a new stereo image should be added, the stereo image that has the most correspondences to the already positioned stereo images is selected as the next stereo image to be positioned.

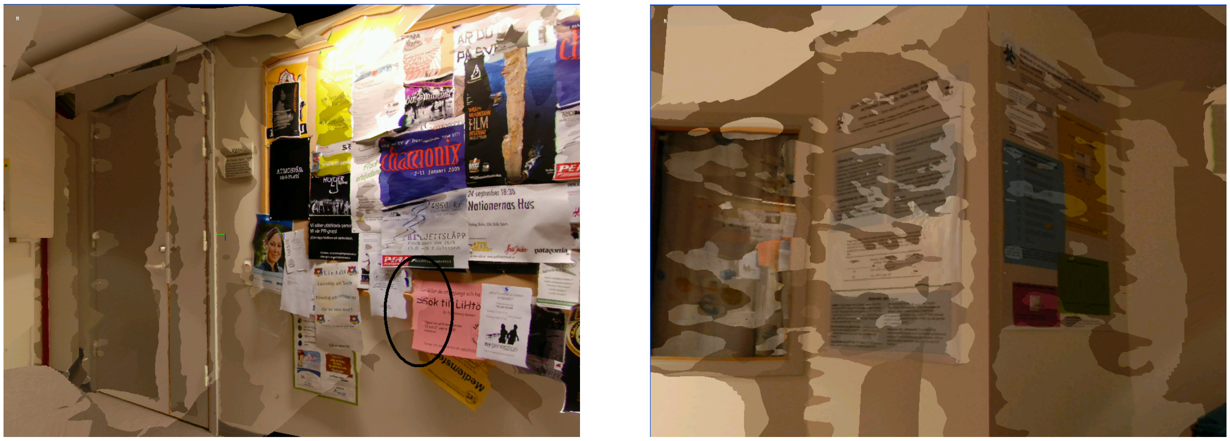
5.1. Results

When the relative poses that are calculated during the experiments are put into the software system to position the models created from the stereo images, the visual error in the pose is small when multiple models are displayed at the same time. There are still visual artifacts that need to be removed, but this does not fall within in the scope of this paper. In Figure 2 two views from a modeled room are depicted. In each view at least three models are displayed at the same time, showing the frontmost surface. The models are speckled because the surfaces compete for visibility and the surfaces are very close to each other, this is an indication of high precision in the poses.

5.2. Future work

During this work there has not been feasible to determine the accuracy of the last method, as all the available information has been used in the calculations. A possible solution would be to position the camera with an industrial robot.

In the two last methods, each new stereo image adds one or several loops in the relations between the cameras. A final optimization of all the positions would increase the accuracy of the result further.



(a) Different stereo images displayed simultaneously. Some errors in the position can be seen in the bottom-right corner of the image.

(b) The surfaces are very close to each other, and the three models will compete for visibility, therefore the exposure changes several times in the same area.

Figure 2: Multiple stereo images are displayed at the same time to form a model.

6. Conclusion

It is shown that the method presented can produce an accuracy of about 1.1 mm in translation and an accuracy of about 0.1 degrees in rotation after sequential and dependent estimation of at least 10 positions using loop closur. It is also shown that the use of relations between more than two images in each step, together with a global optimization using the Newton-Raphson method, can improve the result further. From these poses it is then possible to position separate models created from stereo images to form a larger model.

References

[1] M. Pollefeys, D. Nister, J. M. Frahm, A. Akbarzadeh, P. Mordohai, B. Clipp, C. Engels, D. Gallup, S. J. Kim, P. Merrell, C. Salmi, S. Sinha, B. Talton, L. Wang, Q. Yang, H. Stewenius, R. Yang, G. Welch, and H. Towles, "Detailed real-time urban 3D reconstruction from video," *International Journal of Computer Vision*, vol. 78, no. 2-3, pp. 143–167, 2008.

[2] M. Pollefeys, L. V. Gool, M. Vergauwen, F. Verbiest, K. Cornelis, J. Tops, and R. Koch, "Visual modeling with a hand-held camera," *International Journal of Computer Vision*, vol. 59, no. 3, pp. 207–32, 2004.

[3] C. Netramai and H. Roth, "Real-time photorealistic 3D map building using mobile multiple stereo cameras setup," *3DTV Conference, 2007*, pp. 1–5, May 2007.

[4] P. Biber, H. Andreasson, T. Duckett, and A. Schilling, "3D modeling of indoor environments by a mobile robot with a laser scanner and panoramic camera," in *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, vol. 4, Sendai, Japan, Sep-Oct 2004, pp. 3430–3435.

[5] E. Delage, H. Lee, and A. Ng, "Automatic single-image 3D reconstructions of indoor manhattan world scenes," in *Robotics Research.* Springer Berlin / Heidelberg, 2007, pp. 305–321.

[6] A. Akbarzadeh, J.-M. Frahm, P. Mordohai, B. Clipp, C. Engels, D. Gallup, P. Merrell, M. Phelps, S. Sinha, B. Talton, L. Wang, Q. Yang, H. Stewenius, R. Yang, G. Welch, H. Towles, D. Nister, and M. Pollefeys, "Towards urban 3D reconstruction from video," *Third International Symposium on 3D Data Processing, Visualization, and Transmission*, pp. 1–8, June 2006.

[7] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "Monoslam: Real-time single camera SLAM," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 1052–1067, 2007.

[8] D. Lowe, "Object recognition from local scale-invariant features," *The Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, pp. 1150–1157, Sept. 1999.

[9] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, 2003.

[10] Y. Ma, S. Soatto, J. Kosecká, and S. S. Sastry, *An invitation to 3-D Vision.* Springer, 2004.

[11] R. I. Hartley, "In defence of the 8-point algorithm," in *IEEE International Conference on Computer Vision*, June 1995, pp. 1064–1070.

[12] L. Råde and B. Westergren, *Beta, Mathematics Handbook*. Studentlitteratur, 1988.