On Minimizing Errors in 3D Reconstruction for Stereo Camera Systems[1]

S. Wenhardt[1], J. Denzler[2], and H. Niemann[1]

[1] *Chair for Pattern Recognition, University of Erlangen-Nürnberg, Martensstr. 3, 91058 Erlangen, Germany*
e-mail: wenhardt@informatik.uni-erlangen.de
[2] *Computer Vision Group, Faculty of Mathematics and Informatics, University of Passau, 94030 Passau, Germany*
e-mail: denzler@fmi.uni-passau.de

Abstract—Active reconstruction of three-dimensional (3D) surfaces deals with the control of camera viewpoints to minimize error and uncertainty in the reconstructed shape of an object. In this paper, we develop a mathematical relationship between the setup and focal lengths of a stereo camera system and the corresponding error in 3D reconstruction of a given surface. We explicitly model the noise in the image plane, which can be interpreted as pixel noise or as uncertainty in the localization of corresponding point features. The results can be used to plan sensor positioning, e.g., using information theoretic concepts for optimal sensor data selection.

INTRODUCTION

In the past, more and more areas in computer vision have been benefiting from active processing strategies, which means that the sensor data is acquired in an active, purposive way. Viewpoint selection for object recognition [1], actively controlling the focal length during object tracking [2], and sequential sensor data selection for state estimation in general [3] are examples.

Besides these mentioned areas, to date only a few approaches are known that suggest active sensor data selection for three-dimensional (3D) reconstruction of surfaces and objects, for example, for range image data [4]. Obviously, for reconstructing the surface of an unknown object, the viewpoints of the recorded images strongly influence the resulting accuracy and robustness of the reconstruction. This observation is true independently of the chosen approach for 3D reconstruction (stereo, factorization method, trifocal tensor). The quality mainly depends on the surface normal, the ex- and intrinsic parameters of the camera, and noise. Therefore, the question arises: is it possible to come up with a relationship between the selected views and the error and uncertainty of the reconstructed surface of an object? The long-term benefit of such an approach consists of the possibility to apply information theoretic methods for sequential sensor data selection [3] to 3D reconstruction as well. With this goal in mind, in this paper we investigate the influence of the parameters of a stereo camera system on error in reconstruction of a surface, taking explicitly into account noise in the image acquisition and feature extraction process. To the best of our knowledge, such an investigation has not been done before.

The paper is structured as follows: first, we describe the setup for 3D reconstruction using a stereo camera system. Then, we present a mathematical development of the error in reconstruction, taking noise in the image plane explicitly into account. We map the problem of optimal stereo positioning to an optimization problem. This will be analyzed to obtain the optimal focal length and the optimal baseline in a normalized stereo system. Later, we look at stereo systems with one rotation parameter and optimize this rotation. This paper ends with a conclusion and an outlook on future work.

PROBLEM OF 3D RECONSTRUCTION ON A NORMALIZED STEREO SYSTEM

First, we explain what we understand by a normalized stereo system: it consists of two cameras, which have the same orientation, and translation is possible only in the *x*-direction (cf. Fig. 1). The points O_l and O_r are the optical centers. Each camera has its own coordinate system, with the *x*- and *z*-axes indexed by "l" for

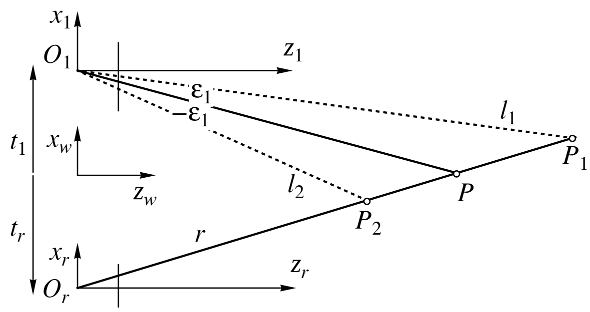


Fig. 1. Norm. stereo system with errors by triangulation.

[1] This article was submitted by the authors in English.

Received October 25, 2004

Pattern Recognition and Image Analysis, Vol. 15, No. 2, 2005, pp. 468–471.

left and "r" for right camera. $\mathbf{t}_l$ and $\mathbf{t}_r$ are the translations of the cameras from the world coordinate system.

$$\|\mathbf{t}\| := \|\mathbf{t}_l - \mathbf{t}_r\| = |t_l - t_r| \tag{1}$$

is called the baseline. For the triangulation, we have to know all parameters, i.e., the translation, focal length, and image coordinates. But for real world data, disturbances occur, which results in a triangulation error. We analyze whether there is a configuration of modifiable parameters for which the error is minimal.

MODELING OF THE ERROR

There are many choices of how to model the disturbances and measure the error. Here, we will assume that there is an error in one image plane (Fig. 1), e.g., caused by an inaccurate solution of the correspondence problem. That is, we select points in one image—these points are exact—and then we search for the corresponding points in the second one. So, errors can occur only in the second image. We do not specify a statistical distribution of the error, but we model the maximal error, i.e., the worst case. Minimization of the error function means minimization of triangulation error if the maximal error occurs. Further on, the other parameters are assumed to be exact, and, for the sake of better understanding, all y-coordinates are set to zero, because the lines cannot be skew in the plane. We define the maximal error in the x-direction to be $\pm\varepsilon_1$, cf. Fig. 1. We define error e as

$$e = \|\mathbf{P}_1 - \mathbf{P}_2\|^2. \tag{2}$$

An optimal 3D reconstruction means that we have to minimize error function e with respect to the free parameters of our stereo camera system. For that, we have to derive the error function. Therefore, we have to calculate the coordinates of point $\mathbf{P}_1$, which is the intersection of lines of sight r from the right camera system and the disturbed l_1 from the left (cf. Fig. 1). The linear equation for r in the world coordinate system is

$$x_w = -\frac{t_r - x_P}{z_P} z_w + t_r, \tag{3}$$

where $(x_P\ z_P)$ are the coordinates of $\mathbf{P}$. With respect to the equations on perspective projection with focal length f_1, we can see that the linear equation for l_1 is

$$x_w = \left(-\frac{t_1 - x_P}{z_P} + \frac{\varepsilon_1}{f_1}\right) z_w + t_1. \tag{4}$$

From Eqs. (3) and (4), we calculate $\mathbf{P}_1$:

$$\mathbf{P}_1 = \begin{pmatrix} \dfrac{(t_1 - t_r) z_P f_1}{(t_1 - t_r) f_1 - \varepsilon_1 z_P} \\ -\dfrac{(t_1 - t_r)(t_r - x_P) f_1}{(t_1 - t_r) f_1 - \varepsilon_1 z_P} + t_r \end{pmatrix}. \tag{5}$$

The coordinates for point $\mathbf{P}_2$ can be calculated in the same way. Thus, for e we get

$$e = \frac{4 f_1^2 \varepsilon_1^2 z_P^2 (t_r - t_1)^2 ((t_r - x_P)^2 + z_P^2)}{((t_r - t_1)^2 f_1^2 - \varepsilon_1^2 z_P^2)}. \tag{6}$$

OPTIMIZATION OF FOCAL LENGTH

In our active vision stereo system, we can modify the focal length, translations in the x-direction, and rotations around the y-axis to improve 3D reconstruction, i.e., to minimize the error function. If we ignore the visibility, i.e., assuming infinite image planes, we can analyze all parameters separately. First, we analyze the influence of the focal length. Therefore, we differentiate e with respect to the focal length:

$$\frac{\partial e}{\partial f_l} = \frac{z_P^2 \varepsilon_1^2 (t_r - t_l)^2 ((t_r - x_P)^2 + z_P^2)((t_r - t_l)^2 f_l^2 + z_P^2 \varepsilon_1^2)}{-0.125 f_l^{-1} ((t_r - t_l)^2 g_l^2 - z_P^2 \varepsilon_1^2)^3}. \tag{7}$$

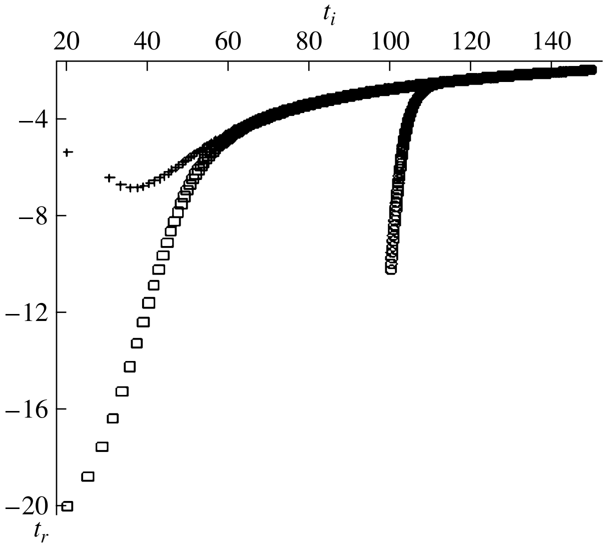
We can show that, for $f_l \in\ ]0, z_P\varepsilon_1/(t_l - t_r)[$, the point $\mathbf{P}_1$ lies behind the cameras. Thus, the relevant interval for the focal length is $f_l \in\ ]z_P\varepsilon_1/(t_l - t_r), \infty[$. For $f_l > z_P\varepsilon_1/(t_l - t_r)$, the first derivate is negative, i.e., error func- 1
tion e is strictly monotonically decreasing and there is no minimum. We conclude that, for a real camera system, the focal length should be chosen to be as large as possible, so that the object is all in the image, to improve the 3D reconstruction. This is also true for more than one point, because the error function is then the sum of all errors (6) and the sum of monotonically decreasing functions is monotonically decreasing.

OPTIMIZATION OF TRANSLATIONS

To minimize error e, the gradient of e with respect to translations t_l and t_r, which are given with respect to a fixed world coordinate system, must be zero. We obtain a nonlinear system of equations with polynomials of degree 5. This is generally not solvable by radicals [5], so we try to find a minimum by numeric analysis.

We search for a minimum with the gradient descent method. In Fig. 2, we plotted $(t_l\ t_r)$, shown by different symbols for different initializations, and 2
iterated 1000 times.

We observe that translation t_r converges to a value near zero and t_l becomes larger in each step. The trajectories converge to an asymptote. This seems to be the same asymptote for all tested initializations, for differ- 2
ent values of z_P, f_l or ε_1. Only if $x_P \neq 0$ is the asymptote shifted by x_P.



2 **Fig. 2.** Trajectories for translations: The initializations for (t_l, t_r) for the cross symbol are (20, –20), for the box are (20, –5), and for circle are (100, –10) under the assumptions f_l = 1, P = (0 15), and ε_1 = 1/2.

It is an already well-known result that a larger baseline is better than a smaller one. In general, for $t_l \longrightarrow \infty$, e becomes zero:

$$\lim_{t_l \to \infty} e = 0. \tag{8}$$

But not only the length of the baseline is relevant for reconstruction: e.g., for $t_l = -t_r = 100$, $e = 28.8$, and for $t_l = 110$, $t_r = -10$, $e = 2.6$, although in the first case the baseline is twice as large. Furthermore, an infinite baseline does not imply that e is zero:

$$\lim_{t_r \to \infty} e = 4\varepsilon_1^2 z_P^2 / f_1^2. \tag{9}$$

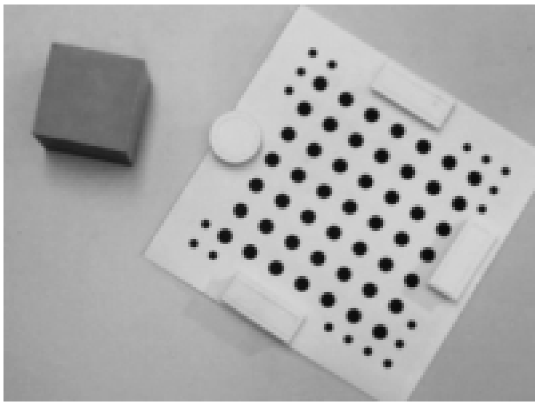


Fig. 3. Typical experiment image.

Thus, we conclude that, in addition to the baseline, the position between cameras and points is an important factor for 3D reconstruction, as well.

If we want to reconstruct more than one point, the error will be the sum of e for the coordinates of different $\mathbf{P_i}$. The problem is more complex, because each error for one point depends on its coordinates (x_{Pi} z_{Pi}), and we can see in Eq. (6) that z_{Pi} has a strong influence on the value of e. Thus, points with large z components result in a large error and, therefore, they have more influence on the minimization procedure.

OPTIMIZATION OF ROTATION

For example, if we use pan-tilt cameras, there are two rotations. Therefore, we introduce a rotation around the y-axis that is perpendicular to the x–z plane in Fig. 1. If the error is only in one camera, the rotation of the other is irrelevant. So, we consider only rotation of the left camera by angle α. Then, the error function is

$$e = \frac{\varepsilon_1^2 (t_r - t_1)^2 ((t_1 - x_P)\sin\alpha - z_P \cos\alpha)^2 ((t_r - x_P)^2 + z_P^2)}{0.25 f_l^{-2} ((t_r - t_1)^2 f_1^2 - \varepsilon_1^2 (z_P \cos\alpha + (t_1 - x_P)\sin\alpha))^2}. \tag{10}$$

Symbolic differentiation of Eq. (10) with respect to α and computing the zero crossings are possible. Due to lack of space, we must omit the complicated term for the derivative. We investigate the solution for $f_l = 1$, $\mathbf{P}$ = (0 15), $\varepsilon_1 = 1/2$, $t_l = 5$, and $t_r = -5$. For $\alpha = 0$, this is equivalent to the configuration of Fig. 1. There are two minima in $\alpha_1 = 1.89$ and $\alpha_2 = -1.25$ (values in radian). For α_1, $\mathbf{P}_1$ is behind the camera. Thus, the left camera must be rotated counterclockwise by about 71°. The camera should not rotated toward, but turned away, while the object is in the image.

Minimization by camera rotation for more than one point is similar to the translation case: large values of z_P result in large e. Therefore, points at a larger distance have more influence on the minimization procedure and will bias the optimal solution for the rotation angle.

EXPERIMENTAL RESULTS

In this section, we present the first experimental results to show the influence of focal length and translation on quality. We took images of a calibration pattern and a cube (cf. Fig. 3). We calibrated the cameras with the calibration pattern and reconstructed 49 points on it (experiment 1). In this case, we can verify the triangulation results with ground truth data. Later, we reconstructed all seven visible corners of the cube and calculated the edge lengths, which we compared with the true value (experiment 2). In Table 1, the first value

Experimental results (focal length is in pixels, the other values in mm)

	$\|\|t\|\| = 51$	$\|\|t\|\| = 63$	$\|\|t\|\| = 201$	$\|\|t\|\| = 326$
$f_1 = 763$	6.8/28	4.5/25	1.5/9.9	1.0/5.6
$f_1 = 1155$	1.1/13	1.0/8.3	0.4/2.2	0.3/2.3
$f_1 = 1487$	0.8/0.8	0.6/0.11	0.3/0.73	0.2/0.4

in each cell is the mean difference between the real and reconstructed points in experiment 1. The second value is the mean difference of the measured edge lengths and the correct one (60 mm) in experiment 2.

In the theory sections, we showed that, if translation or focal length increases, the error decreases. Therefore, the largest errors are at the top left in Table 1 and the smallest ones should be at bottom right, but in experiment 2 there are two outliers (for $||t|| = 63$, $f_l = 1487$ and $||t|| = 201$, $f_l = 1155$). A possible reason for these outliers is that detection of points that are not on the top side of the cube is quite inaccurate. However, if we ignore the outliers, we can see that the error decreases if focal length increases (cf. columns of Table 1) or translation increases (cf. rows of Table 1). We therefore imply that the prediction of the theory is true and important in real world experiments.

CONCLUSIONS

It is obvious that, for 3D reconstruction, not every recorded view is equally useful. We used a stereo system for our analysis and specified which parameters 3D reconstruction depends on. There are unchangeable parameters and parameters modifiable by an active vision system. The main question was what configuration of parameters results in a good triangulation.

First, we analyzed the influence on focal length. We could analytically prove that the error strictly monotonically decreases if the focal length increases.

Second, we looked at the influence of translations. We observed that a large baseline decreases the error, but that the error also depends on the position between points and cameras.

We also analyzed the effects of rotations. The result was that the camera should not turn toward, but away from, the object.

In our future work, we will extend our results to setups of cameras that are not restricted, i.e., arbitrary positions of the cameras will be allowed. Later, we will include the problem of visibility and the correspondence problem, which are important constraints in real applications, in our theory. With these results, we will be able to apply an already-approved framework for optimal sensor data acquisition to the problem of active 3D reconstruction.

REFERENCES

1. F. Deinzer, J. Denzler, and H. Niemann, “Viewpoint Selection—Planning Optimal Sequences of Views for Object Recognition,” in *Computer Analysis of Images and Patterns*, Ed. by N. Petkov and M. A. Westenberg, CAIP’03 (Springer, Heidelberg, 2003), pp. 65–73.
2. 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*, pp. 400–407.
3. J. Denzler and C. M. Brown, “Information Theoretic Sensor Data Selection for Active Object Recognition and State Estimation,” IEEE Transactions on Pattern Analysis and Machine Intelligence **24** (2), 145–157 (2002).
4. M. K. Reed and P. K. Allen, “Constrained-Based Sensor Planning for Scene Modeling,” IEEE Transactions on Pattern Analysis and Machine Intelligence **22** (12), 1460–1466 (2000).
5. N. Jacobsen, *Lectures in Abstract Algebra: Vol. III—Theory of Fields and Galois Theory* (D. van Nostrand Company, 1964).

SPELL: 1. derivate, 2. initializations