Image Uncertainty and Pose Estimation in 3D Euclidian Space

Brian Wettegren, Lars Bjerre Christensen, Bodo Rosenhahn,
Oliver Granert and Norbert Krüger

February 28, 2005

Abstract

We describe a problem of a sucessful 3D–2D pose estimation algorithm when it is applied in scenarios with large depth variation. In this case image uncertainty is inhomomogenuously reflected in the Euclidian space where the constraint equations are formulated. We introduce a scaling of the constraint equations that equalizes this inhomogenity. We can show that we can reduce the error significantly in outdoor scenarios with large depth discontinuities.

1 Introduction

The estimation of the motion of rigid bodies (rigid body motion, RBM) is an important sub–problem in computer vision for tasks such as object recognition [3], mutiple view reconstuction [7] and disambiguation of visual representations [11]. It is also important in the context of robot navigation since the ego–motion of a person or vehicle in a static scene can be described by an RBM. The mathematical formalization of this kind of motion has been studied for a long while (see, e.g., [2, 9]). An RBM can be described as a six–dimensional manifold consisting of a translation (parametrised by the three coefficients $\mathbf{t} = (t_1, t_2, t_3)$) and a rotation (parametrised by $\mathbf{r} = (r_1, r_2, r_3)$). It describes the transformation of a 3D entity[1] $\mathbf{e}$ in the first frame to a 3D entity $\mathbf{e}'$ in the second frame

$$RBM^{(\mathbf{t},\mathbf{r})}(\mathbf{e}) = \mathbf{e}^{'}. \tag{1}$$

A camera projects a scene to a 2D chip. Therefore it is often convenient to work with entities that are extracted from a 2D image. However, there occur many applications in which prior object knowledge does exist. For example in industrial robot applications CAD descriptions of objects may be available (see, e.g., [4]). 3D information can also be extracted from image sequences beforehand through stereo as done in this paper. This requires then an RBM estimation algorithm that can work on entities of different dimensions: The

[1] In the following 3D entities are printed in boldface while 2D entities are printed normal.

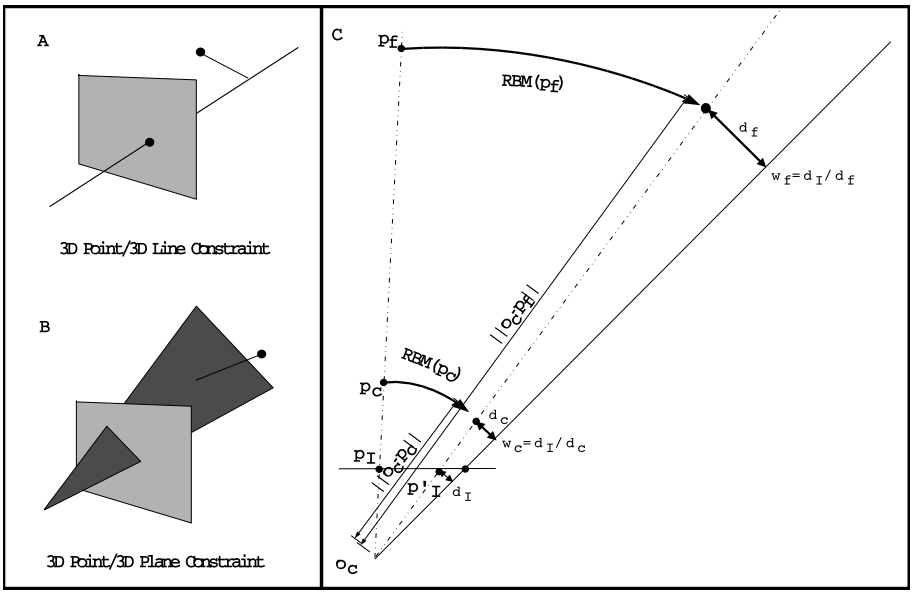

Figure 1: a) Knowing the camera gemometry a 3D line can be generated from an image point and the optical center of the camera. The 3D point/3D line constraint realizes the shortest Euclidian distance between the 3D Point and the 3D line. b) From an image line a 3D plane can be generated. The 3D point/3D plane constraint realizes the shortest Euclidian distance between the 3D Point and the 3D plane. c) Inhomogenity of Uncertainty in the Euclidian Space. The difference of a projection of a 3D point and the extracted feature corresponding to this point in the image leads to a deviation d_I. However, in case this projection comes from a far point $\mathbf{p_f}$ it results in a much larger distance d_f between the generated line and $\mathbf{p_f}$ than in case of a close point $\mathbf{p_c}$ (with associated distance d_c). An appropriate scaling factor that equalizes these differences for these two points is $w_c = d_I/d_c$ and $w_f = d_I/d_f$ respectively.

3D object knowledge needs to be aligned with 2D entities in an image of this object. The problem of computing the RBM from correspondences between 3D object and 2D image entities is commonly referred to as 3D–2D pose estimation problem [6, 13]. In mathematical terms we have the following kind of constraint equations:

$$P(RBM(\mathbf{e})) = e', \tag{2}$$

where P represents the perspective projection.

There exist approaches (in the following called 'projective' approaches) that formalize constraints directly on equation (2) (see, e.g., [1]). An alternative is, instead of formalising the pose estimation problem in the image plane, to associate a 3D entity to each 2D entity: For example, a 2D image point together

with the optical center of the camera spans a 3D line (see figure 1a) and an image line together with the optical center generates a 3D plane (see figure 1b). In case of a 2D point p we denote the 3D line that is generated in this way by $\mathbf{L}(p)$. Now the RBM estimmation problem can be formulated for 3D entities

$$RBM^{(\mathbf{t},\mathbf{r})}(\mathbf{p}) \in \mathbf{L}(p).$$

where $\mathbf{p}$ is the 3D Point. Such an Euclidian formulation has been applied by, e.g., [14, 15, 5, 13]. They have coded the RBM estimation problem in a twist representation. The RBM can then be computed iteratively on a linearized approximation of the RBM.

This approach is elegant, since it deals with the full perspective projection. It works in the space where the RBM takes place (i.e., the Euclidian space) and also allows for nicely interpretable constraint equations which basically represent the Euclidian distance between the 3D entities (see figure 1,a,b). It can also deal with any kind of camera model (orthographic, perspective, paraperspective, ...): For switching between these camera models only the reconstruction of the entities change but not the actual constraint equations.

We have been sucessfully working with this algorithm which is turned out to be numerically stable and fast [10]. It is also straightforward to implement and the meaning of constraints and entities is well defined (which will become important for our improvement of the algorithm). However, one problem of such a formulation is that when dealing with natural scenes uncertainties are associated to the image features used as correspondences. These uncertainties can be for example caused by unprecise positioning or the calibration of cameras. These image uncertainties lead to an inhomogeneity in the constraint equations: The estimation of feature attributes of entities with large depth cause a higher uncertainty in the constraint equations than that of entities at a close distance. This is caused by the fact that the constraint equations are formulated on entities in the 3D–Euclidian space which however originate from 2D entities which uncertainties reproject back to the Euclidian space in a non-homogenuous way. Thus, correspondences of entities with large distance would have higher influence in the constraint equations (see figure 1c).

In this paper, we demonstrate the effect of this inhomogenity on the example of RBM estimation from stereo sequences: We can show that for scenes with large depth variation, although we get a good reduction of the error measured in the 3D constraints this can lead to quite significant errors in the 2D projections. We then introduce a scaling of the constraint equations that elliminates the inhomogenity and we can show that we achieve better results for scenes with large depth variation but not scenes with small depth variation.

The paper is structured as following: In section 2, we briefly describe the 3D–2D pose estimation algorithm. In section 4 we describe our modification of the algorithm. In section 3, we introduce the scenario in which our algorithm is applied and in section 5 we show the effect of our scaling.

2 Constraint Equations

Following [14, 15, 5, 13] an RBM can be represented as

$$RBM = e^{\tilde{\xi}\alpha} = \sum_{n=0}^{\infty} \frac{1}{n!}(\tilde{\xi}\alpha)^n \tag{3}$$

with $\tilde{\xi}$ being the 4×4 matrix

$$\tilde{\xi} = \begin{pmatrix} \tilde{w} & -\tilde{w}\mathbf{q} + \lambda\mathbf{w} \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -w_3 & w_2 & w_3q_2 - w_2q_3 + \lambda w_1 \\ w_3 & 0 & -w_1 & w_1q_3 - w_3q_1 + \lambda w_2 \\ -w_2 & w_1 & 0 & w_2q_1 - w_1q_2 + \lambda w_2 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

with $\tilde{w}$ being the direction of the line around which the rotation is performed, $\mathbf{q}$ being a point on this line λ being the translation along the line. A straight forward linearisation is given by $e^{\xi\alpha} \approx (I_{4\times4} + \alpha\tilde{\xi})$. We can represent a 3D point $\mathbf{p} = (p_1, p_2, p_3)$ by the null space of a set of equations

$$\mathbf{F^p}(\mathbf{x}) = \begin{pmatrix} 1 & 0 & 0 & -p_1 \\ 0 & 1 & 0 & -p_2 \\ 0 & 0 & 1 & -p_3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \tag{4}$$

Note that the value $||\mathbf{F^p}(\mathbf{x})||$ represents the Euclidian distance between $\mathbf{x}$ and $\mathbf{p}$. This will is important to derive interpretable constraint equations.

A 3D line $\mathbf{L}$ can be expressed as two 3D vectors $\mathbf{r}, \mathbf{m}$. The vector $\mathbf{r}$ describes the direction and $\mathbf{m}$ describes the moment which is the cross product of a point $\mathbf{p}$ on the line and the direction $\mathbf{m} = \mathbf{p} \times \mathbf{r}$. $\mathbf{r}$ and $\mathbf{m}$ are called Plücker coordinates. The null space of the equation $\mathbf{x} \times \mathbf{r} - \mathbf{m} = \mathbf{0}$ is the set of all points on the line. In matrix form this reads

$$\mathbf{F^L}(\mathbf{x}) = \begin{pmatrix} 0 & r_x & -r_y & -m_x \\ -r_z & 0 & r_x & -m_y \\ r_y & -r_x & 0 & -m_z \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{pmatrix} = 0 \tag{5}$$

Note that the value $||\mathbf{F}^L(\mathbf{x})||$ can be interpreted as the Euclidian distance between the point (x_1, x_2, x_3) and the closest point on the line to (x_1, x_2, x_3) [8, 13].

We now want to formulate constraints between 2D image entities and 3D object entities. Given a 3D point $\mathbf{p}$ and a 2D point p we first generate the 3D line $\mathbf{L}(\mathbf{r}, \mathbf{m})$ that is generated by the optical center and the image point (see figure 1b).[2] Now the constraint reads:

$$\mathbf{F^{L}}^{(p)} \left((I_{4\times4} + \alpha\tilde{\xi})\mathbf{p} \right) = 0. \tag{6}$$

[2]Note that the line $\mathbf{L}$ depends on the camera parameters.

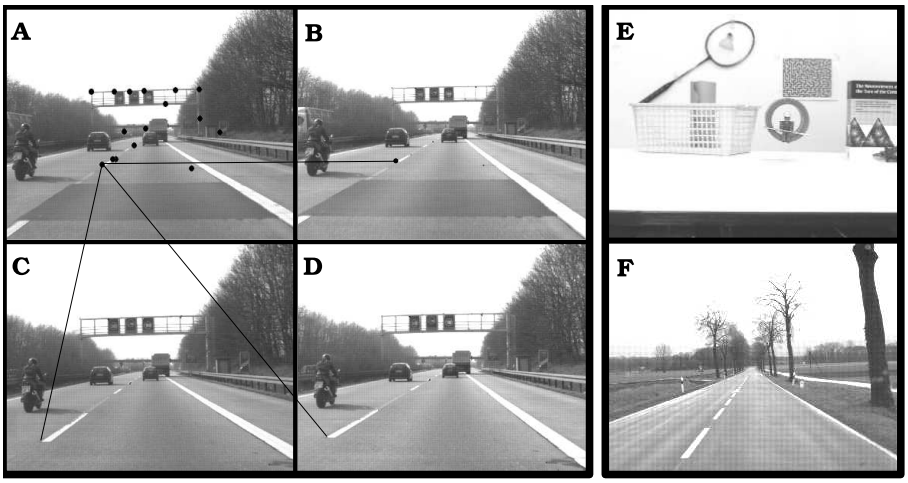

Figure 2: A,B,C,D: The scenario for RBM estimation. A,B: Left and right image of the first frame. Some of the used correspondences are displayed in A. C,D: Left and right image of the second frame. E,F: Two other scenes used for testing the pose estimation algorithm. E: Lab Scene without large differences in depth. F: Another outdoor scene.

Note that although we have 3 equations for one correspondence the matrix is of rank 2 resulting in 2 constraints. For different correspondences we get more equations. This results in a system of linear equations which solution becomes optimized iteratively (for details see [5, 12]).

3 Ego-motion estimation from Stereo Sequences

We apply the pose estimation algorithm in the context of egomotion estimation from stereo sequences (see figure 2A–D). Here we do not have any model knowledge about the scene. Therefore the 3D entities need to be computed from stereo correspondences. We provide manually derived correspondences in two consecutive stereo frames for a number of 3D points. For each 3D points we therefore get four projections, two in the first and also two in the second frame (see figure 2A,B,C.D). From the correspondences in the first frame we compute a 3D point and the correspondences in the second frame result in two 3D lines for which two constraint equations (6) can be derived.

We measured the image distances between manually determined points and points projected after the computed RBM has been performed. We noticed that for the points close to the camera there occur in average large differences. We expect that this inhomogenity results from the inhomogenity in the constraint equations.

4 Scaling of Constraint Equations according to Image Uncertainty

In the context of ego-motion estimation from stereo sequences we are faced with uncertainties in the 3D model as well as in the feature extraction. Both uncertainties are caused by the unprecision in the positioning of the corresponding 2D points. First, it results in an unprecision of stereo reconstruction.[3] Second, it leads to an unprecision in the reconstruction of the 3D line from the 2D point. Since we deal with relativeley small motions compared to depth variation in the scene we can assume that both uncertainties lead to similar distributions and can be handled by the same mechanism.

We replace equation (6) by

$$\frac{1}{w_p}\mathbf{F}^{\mathbf{L}(p)}\left((I_{3\times 3}+\tilde{\xi}\alpha)\mathbf{p}\right)=0. \tag{7}$$

where w_p is computed by

$$w_p=\frac{1}{||\mathbf{o_c}-\mathbf{RBM}(\mathbf{p})||} \tag{8}$$

where $\mathbf{o_c}$ is the optical center of the camera. Note that in our stereo context the weights for the same 3D point $\mathbf{p}$ are different for correspondences of the left and right camera since their optical centers differ.

The reason for choosing this formula is a straightforward application of the theorem of intersection of parallel lines with two intersecting lines (see also figure 1c):

$$\frac{d_p}{||\mathbf{o_c}-\mathbf{RBM}(\mathbf{p})||}=\frac{d_I}{||\mathbf{o_c}-\mathbf{P}(\mathbf{RBM}(\mathbf{p}))||}.$$

Since the weight w_p is supposed to equalize the effect of d_p we need to divide by

$$d_p=d_I\cdot\frac{||\mathbf{o_c}-\mathbf{RBM}(\mathbf{p})||}{||\mathbf{o_c}-\mathbf{P}(\mathbf{RBM}(\mathbf{p}))||}$$

We can assume the image uncertainty d_I as constant and approximate $||\mathbf{o_c}-\mathbf{P}(\mathbf{RBM}(\mathbf{p}))||$ by the focal length (i.e., by a constant as well). Both constants do not influence the relative weighting of constraint equations and can therefore be neglected such when we devide by d_p we end up with equation (8).

5 Results

We applied the scaling in 3 different scenarios: motorway (figure 2A-D), lab (figure 2E) and country road (figure 2F). In the lab scenario, depth differences were rather small compared to the ego-motion while in the other the depth

[3] In addition there is also uncertainty in the calibration. However, we negelect these effects here.

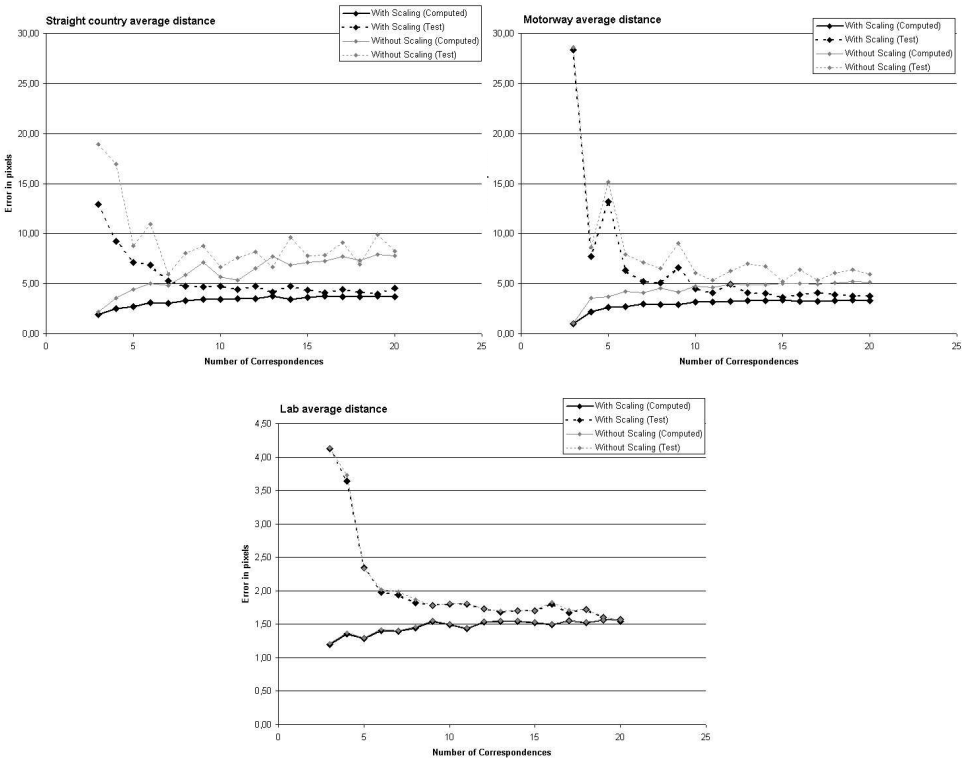

Figure 3: The average pixel distance of estimated image points depending on the number of correspondences used for computation is shown for the three scenarios: country road (top,left), motorway (top,right), and lab (bottom).

differences were rather large. From our consideration above we expect small effects for the low depth variation (lab scene) and improvement for the other two cases. For all sequences we generated 25 point correspondences manually. We computed the RBM on a subset of those (computing set). We calculated from the computing set and the set of remaining points (test set) the average pixel distance in the image plane separately.[4] The results are shown in figure 3.

Different observations are of interest. First, the average pixel error is significantly lower with our scaling compared to the non-scaling case for the motorway and the country road sequence (the error can be reduced to approximately the half). For the lab sequence there is no siginificant difference if scaling is applied or not due to the small depth variation. We can further observe that we need approximately 8 to 10 correspondences to get a good generalization. For less correspondences, we get much better results on the computing set compared to the test set.

[4]For a fixed number of correspondences we did 20 runs on different subsets.

6 Summary

We described a problem of a sucessful 3D–2D pose estimation algorithm [14, 15] when it is applied in scenarios with large depth variation. Then the image uncertainties are inhomomogenuously reflected in the Euclidian space where the constraint equations are formulated. We introduced a scaling of the constraint equations that equalizes this inhomogenity. We could show that we can reduce the error significantly in outdoor scenarios with large depth discontinuities. As expected from the motivation of the scaling method, no measurable improvement is achieved for scenes with small depth variation.

References

[1] H. Araujo, R.J. Carceroni, and C.M. Brown. A fully projective formulation to improve the accuracy of lowe's pose–estimation algorithm. *Computer Vision and Image Understanding*, 70(2):227–238, 1998.

[2] R.S. Ball. *The theory of screws*. Cambridge University Press, 1900.

[3] J. Beis and D. Lowe. Learning indexing functions for 3–d model based object recognition. *CVPR'94*, pages 275–280, 1994.

[4] C. Fagerer, D. Dickmanns, and E.D. Dickmanss. Visual grasping with long delay time of a free floating object in orbit. *Autonomous Robots*, 1(1):53–68, 1991.

[5] O. Granert. Posenschättzung kinematischer ketten. *Diploma Thesis, Universität Kiel*, 2002.

[6] W.E.L. Grimson, editor. *Object Recognition by Computer*. The MIT Press, Cambridge, MA, 1990.

[7] R.I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision.* Cambridge University Press, 2000.

[8] Selig J.M. Some remarks on the statistics of pose estimation. *Technical Report SBU-CISM-00-25, South Bank University, London*, 2000.

[9] K. Klein. *Vorlesungen über nicht–Euklidische Geometrie*. AMS Chelsea, 1927.

[10] N. Krüger, M. Ackermann, and G. Sommer. Accumulation of object representations utilizing interaction of robot action and perception. *Knowledge Based Systems*, 15:111–118, 2002.

[11] N. Krüger and F. Wörgötter. Multi-modal primitives as initiators of recurrent disambiguation processes. *Early Cognitive Vision Workshop, Isle of Skye*, 2004.

[12] N. Krüger and F. Wörgötter. Statistical and deterministic regularities: Utilisation of motion and grouping in biological and artificial visual systems. *Advances in Imaging and Electron Physics*, 131, 2004.

[13] B. Rosenhahn. *Pose Estimation Revisited (PhD Thesis).* Institut für Informatik und praktische Mathematik, Chrsitian–Albrechts–Universität Kiel, 2003.

[14] B. Rosenhahn, C. Perwass, and G. Sommer. Cvonline: Foundations about 2d-3d pose estimation. *In CVonline: On-Line Compendium of Computer Vision [Online]. R. Fisher (Ed). http://homepages.inf.ed.ac.uk/rbf/CVonline/"*., 2004.

[15] B. Rosenhahn and G. Sommer. Adaptive pose estimation for different corresponding entities. In L. van Gool, editor, *Pattern Recognition, 24th DAGM Symposium*, pages 265–273. Springer Verlag, 2002.