

Featureless 6 DoF Pose Refinement from Stereo Images

Wolfgang Sepp Gerd Hirzinger

Institute of Robotics and Mechatronics,
German Aerospace Center (DLR),
82234 Wessling, Germany

E-mail: wolfgang.sepp@dlr.de

Abstract

We present a pose estimation method from an initial unreliable guess using calibrated stereo images. The approach does not rely on a priori known salient features on the surface. The stereo images are brought in congruence without computing a disparity map like in standard stereo algorithms. Instead, the pose parameters of the object are varied to match the stereo images on the known surface shape. Our approach takes into account the aliasing effects introduced due to irregular sub-sampling and is not limited to simple geometric surfaces.

1. Introduction

The problem of pose estimation arises in many tasks such as grasping objects, head tracking, camera calibration and multi-modal surface registration. While the task seems to be solved for 2-D motion [5] research is still ongoing in the field of 3-D tracking in multiple images. In the following, only exemplary approaches on this topic are outlined.

Existing solutions are mainly based on the localization of a priori known artificial or known natural landmarks, which in general is not the case. Many approaches rely on the correspondence between edges on the object and edges in the perceived intensity image [9] which are principally not given for free form surfaces and for richly textured objects. Other approaches work on range images computed from stereo algorithm [6] with all uncertainties and problems associated to surface reconstruction methods. Also optical flow coupled to 3-D surface models is used for pose tracking [1].

Only few approaches exist, which do not match the surface model with a 3-D point cloud but which match the surface texture in two or more calibrated images. Diehl et. al [4] developed a fast method for tracking of 4 DoF motion

of planar objects. Cernuschi-Frias et al. presented in [3] a model for estimating the parameters of simple parameterized surface models. The approach is based on an orthographic imaging model and the method has been evaluated only for up to 4 DoF geometric surfaces. Heipke proposed in [7] a model for the joined optimization of surface height and image orientation whereas the approach has been shown only for surface reconstruction. La Cascia et al. [2] have developed an algorithm which determines a heads pose by minimizing the residual error of the surface texture seen from a single camera to a reference image. The near real-time capability of the approach is given by the linearization of the residual function which is valid only for small variations from the reference pose.

We present a passive method for pose estimation from an initial guess which is capable to determine all 6 DoF pose parameters irrespective of a reference pose and a reference texture. Here, we derive an estimation model based on the minimization of the dissimilarity between the assumed surface textures which accounts also for aliasing effects and for substantial perspective distortions. Therefore, the method is applicable to a vast variety in object positions and in single camera positions. Unlike previous approaches, we do not restrict the surface patch to simple geometric forms but allow any free form surface where sampled surface points are arranged on a two dimensional grid.

The potential applications of the approach comprise, but are not limited to, 6 DoF tracking of rigid objects, accurate texture mapping and accurate localization of known objects.

Sect. 2 starts with the description of the pose estimation model. Experiments with known ground-truth are reported in Sect. 3. And finally, Sect. 4 contains some concluding remarks.

2. Estimation Model

First, let $I_i : \mathbf{x} \in \{0, \dots, N_I\} \times \{0, \dots, M_I\} \mapsto \mathbb{N}$ be the image of size $N_I \times M_I$ taken from the calibrated camera

i. The pose estimation model requires a three dimensional surface patch s given in parametric form:

$$s : \mathbf{u} \in [0, N_s] \times [0, M_s] \mapsto \mathbf{w} \in \mathbb{R}^3 \quad (1)$$

which resembles a finite regular grid of size $N_s \times M_s$ over an arbitrary surface. A point (w_x, w_y, w_z) in \mathbb{R}^3 is mapped into the image plane according to the perspective projection

$$q((w_x, w_y, w_z)^T) = \frac{C}{-w_z}(w_x, w_y)^T \quad (2)$$

where C denotes the camera aperture constant. In the following we consider the projection of a point \mathbf{w} in space to the image given the pose parameters $\mathbf{r} \in \mathbb{R}^6$:

$$p(\mathbf{w}, \mathbf{r}) = q(R(\mathbf{r}) \mathbf{w} + t(\mathbf{r})) \quad (3)$$

where $R(\mathbf{r})$ is the 3×3 rotation matrix and $t(\mathbf{r})$ defines the translation vector for the pose \mathbf{r} . In our experiments, we choose the object centered yaw-pitch-roll rotation and Cartesian translation for the object pose parameters. For convenience, we define a function which maps a surface grid point \mathbf{u} directly to an image coordinate given the pose \mathbf{r} :

$$m_{\mathbf{r}}(\mathbf{u}) = p(s(\mathbf{u}), \mathbf{r}) . \quad (4)$$

Next, the object intensity for the surface grid point \mathbf{u} is determined for the current pose assumption \mathbf{r} and for each camera. An overall surface texture is computed by space variant re-sampling and smoothing of the camera image i according to

$$J_i(\mathbf{u}, \mathbf{r}) = \sum_{\mathbf{x} \in \{0..N_I\} \times \{0..M_I\}} I_i(\mathbf{x}) G_{\sigma}(m_{\mathbf{r}}^{-1}(\mathbf{x}) - \mathbf{u}) \left| \frac{\partial m_{\mathbf{r}}^{-1}(\mathbf{x})}{\partial \mathbf{x}} \right| \quad (5)$$

where $G_{\sigma}(\cdot)$ is a Gaussian function defined over \mathbb{R}^2 with a variance of σ . This term is responsible for the space variant smoothing of each pixel due to the irregular sampling of the camera image. The expression $|\partial m_{\mathbf{r}}^{-1}(\mathbf{x}) / \partial \mathbf{x}|$ is the functional determinant of $m_{\mathbf{r}}^{-1}(\cdot)$ and assures that each pixel is weighted according to the corresponding patch size on the surface grid.

Equation 4 contains the inverse of $m_{\mathbf{r}}(\cdot)$ which is not given in analytic form for arbitrary surfaces. Generally, the inverse can be numerically found through iterative approximation. In the case of the functional determinant we are interested in the gradient of the inverse function which is solved following the rules for inverse functions:

$$\left| \frac{\partial m_{\mathbf{r}}^{-1}(\mathbf{x})}{\partial \mathbf{x}} \right| = \left| \frac{\partial m_{\mathbf{r}}(\mathbf{v})}{\partial \mathbf{v}} \right|^{-1} \Bigg|_{\mathbf{v}=m_{\mathbf{r}}^{-1}(\mathbf{x})} . \quad (6)$$

The textures computed for each camera are pairwise compared to evaluate the current pose estimation. To be

concrete, let \mathbf{r}_i be the fixed pose parameters of a common base to the optical center of camera i . Then, the texture error function is the sum of squared differences of the individual surface textures which reads in the special case of two cameras

$$E(\mathbf{r}) = \sum_{\mathbf{u} \in \{0..N_s\} \times \{0..M_s\}} (J_0(\mathbf{u}, \mathbf{r} \circ \mathbf{r}_0) - J_1(\mathbf{u}, \mathbf{r} \circ \mathbf{r}_1))^2 \quad (7)$$

for the pose estimation \mathbf{r} . Here, the operator \circ denotes the concatenation of two inhomogeneous coordinate transformations. To be precise, the integration in Equation 5 has to be confined to those pixels which are within the relevant area of $G_{\sigma}(\cdot)$. Additionally, the summation in Equation 7 has to be restricted to inner surface grid points because otherwise the smoothing operation with $G_{\sigma}(\cdot)$ would expect surface points outside of the sampled regions.

Finally, the error function is minimized with a minimization technique which employs the error gradient. To be concise, we discuss only selected components of the gradient. Take a look at the gradient for the surface texture, which reads

$$\begin{aligned} \frac{\partial}{\partial \mathbf{r}} J_i(\mathbf{u}, \mathbf{r}) = & \sum_{\mathbf{x} \in \{0..N_I\} \times \{0..M_I\}} I_i(\mathbf{x}) \left\{ \frac{1}{|\partial m_{\mathbf{r}} / \partial \mathbf{v}|} \frac{\partial}{\partial \mathbf{r}} G_{\sigma}(m_{\mathbf{r}}^{-1}(\mathbf{x}) - \mathbf{u}) \right. \\ & \left. + G_{\sigma}(m_{\mathbf{r}}^{-1}(\mathbf{x}) - \mathbf{u}) \frac{\partial}{\partial \mathbf{r}} \frac{1}{|\partial m_{\mathbf{r}} / \partial \mathbf{v}|} \right\} . \quad (8) \end{aligned}$$

Note that the partial derivative consists of two additive components which may work in an antagonistic way. Never the less, we expect the overall gradient to point in the opposed direction toward the correct pose.

According to the chain rules, the derivative of $m_{\mathbf{r}}^{-1}$ has to be computed for Equation 8. The non inverse function reads for a surface point \mathbf{u} and a pixel coordinate \mathbf{x}

$$p(s(\mathbf{u}), \mathbf{r}) = \mathbf{x} . \quad (9)$$

Since we are interested only in the inverse for a fixed coordinate \mathbf{x} under pose \mathbf{r} , the derivative of Equation 9 for \mathbf{r} is

$$\partial_1 p \partial_{\mathbf{u}} s \partial_{\mathbf{r}} \mathbf{u} + \partial_2 p = 0 \quad , \quad (10)$$

where ∂_1 denotes the partial derivative for the first function argument and ∂_2 respectively. This formula can be solved for $\partial_{\mathbf{r}} \mathbf{u}$

$$\partial_{\mathbf{r}} \mathbf{u} = -(\partial_1 p \partial s)^{-1} \partial_2 p \quad (11)$$

which is exactly the derivative of $m_{\mathbf{r}}^{-1}$. The components of the error gradient not reported here are straightforward to compute.

3. Experiments

We consider in the following a pyramidal surface with a texture as depicted in Fig. 1. The surface patch has a base-

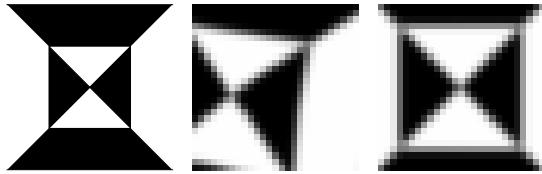


Figure 1. Left: surface texture of the overall pyramid; Middle: surface texture perceived from the right camera under the initial pose assumption; Right: surface texture after pose refinement

line of 30 cm and a height of 7 cm and is sampled on a grid of size 31×31 . The sampled surface points are interpolated with cubic b-splines to get a continuous representation.

The setting consists of two fronto-parallel cameras with an horizontal aperture of 58.25 degree and with a baseline of 20 cm. The pyramid is translated to $(0, 0, -60)$ cm and points toward the observers.

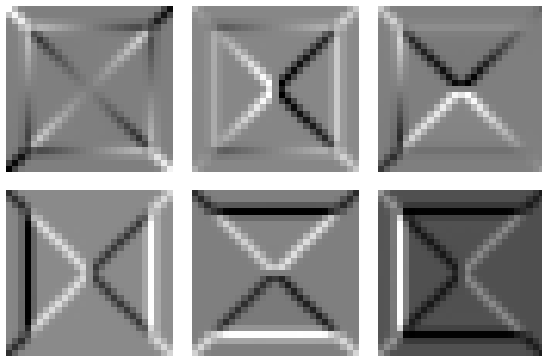


Figure 2. Partial derivatives of the right camera texture for rotation roll, pitch, yaw (top row from left to right) and for translation x,y,z (bottom row from left to right). The intensities are scaled to full contrast.

The partial derivatives (Equation 8) of the surface texture for a single camera are shown in Fig. 2 at the ground-truth pose. While the term $\frac{\partial}{\partial r} |\partial m_r / \partial \mathbf{v}|$ determines the change in weight of single image pixels, the contrasts in the figure are mainly produced by $\frac{\partial}{\partial r} G_\sigma(\cdot)$. Herein, the derivative of the inverse of the mapping function given in Equation 11 plays the key role, as it describes the flow of texture on the surface for small changes in the parameters (see Fig. 3).

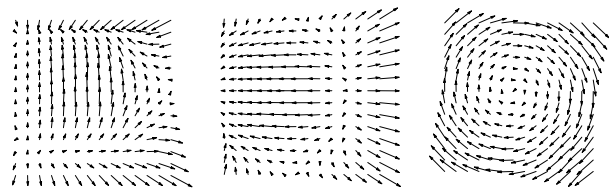


Figure 3. Texture flow according to the derivative of the inverse mapping function (Eqn. 4) for rotation yaw, pitch, roll (from left to right).

The confidence of the pose estimation is determined by the shape of the error function in the neighborhood of the ground-truth pose. The error plots of Fig. 4 show big differences in the concise location of the global minimum for the setting. While the translation along the x-axis and along the

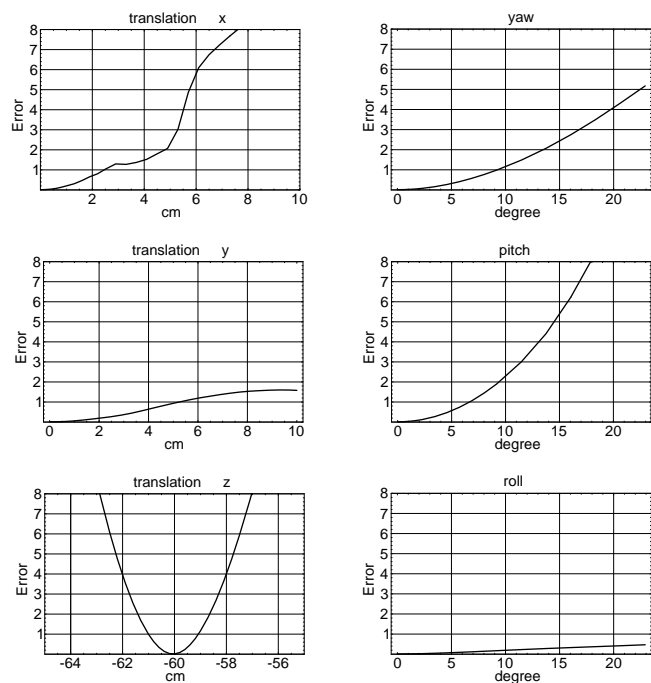


Figure 4. Error for translation along the x-, y-, z-axis (left column) and for rotation yaw, pitch, roll (right column)

z-axis can be reliably estimated, the translation along the y-axis shows no steep changes in the error. The rotational parameters exhibit the same characteristic.

The ability of the estimation model to determine the correct pose is exemplified at the above mentioned setting. Here, we employ the Levenberg-Marquardt optimization scheme for the minimization of the dissimilarity

function. Under the initial assumption, that the object were placed at $(5, 3, -64)$ cm with a yaw-pitch-roll rotation of $(8.59^\circ, 8.59^\circ, 8.59^\circ)$, the algorithm converges to the pose $(0.16, 0.01, -60.05)$ cm and a rotation of $(0.00^\circ, 0.42^\circ, 0.00^\circ)$, which is in the proximity of the ground-truth pose. Fig. 5 documents the evolution of the error.

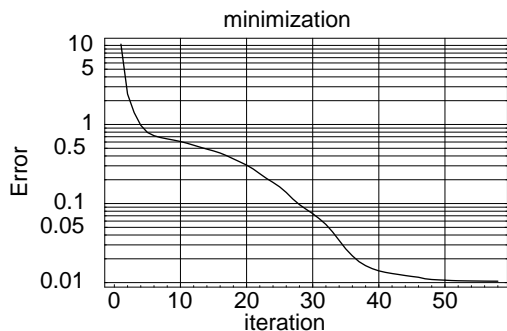


Figure 5. Evolution of the dissimilarity measure during parameter optimization with a Levenberg-Marquardt minimization scheme.

ror. The surface texture computed independently from both cameras show no perceivable difference (see also Fig. 1). The computational complexity depends on the number of pixels which have to be processed for the textures. In above example each iteration takes 7 minutes on a Sun UltraSPARC-III workstation at 750 Mhz.

4. Conclusion

We present a pose refinement model from stereo images which does not depend on prior established point or feature correspondences. The approach does not rely on known texture features of the surface but is solely based upon the surface shape. It is robust against illumination changes on diffuse (Lambertian) materials, such as for example shadows, because the surface points are perceived with the same intensity at all camera positions.

The here presented approach fulfills the Nyquist criteria for sub-sampling which eliminates false matches in the case of surface textures with high frequency components. Therefore, the method should also be robust against noise, which in general shows same frequency characteristics.

A further advantage is the comparison of texture on variable grids in image space. Hence, differences in image resolutions can be compensated, for instance, when the object is close to one camera. We pay attention to this close-up configuration by not approximating the perspective camera projection.

The approach has inherent limitations, though in practice not very restrictive. It is limited to those surface parts which

can be expressed as a two dimensional manifold. Also, the surface texture should allow no disambiguities in the objects pose. Since the approach relies on sub-sampling in the image, it will fail for small or distant objects.

Several extensions to the approach are possible. First, the model can be simply extended to more than two cameras, which increases the reliability of the pose estimations. Second, a multi-grid minimization technique can be implemented [8] extending the radius of convergence and improving the computational speed. For this purpose, only the grid size and therefore the integration scale has to be varied.

The results reported here demonstrate the functionality of the algorithm but also unveil a high computational complexity and the small confidence for rotational parameters in this setting. Our next steps comprise the evaluation of surface and texture dependency of the approach as well as field test with non-synthetic images. We are also committed to improve the time efficiency of the current implementation.

5. Acknowledgment

We would like to thank Dr. Ulrich Hillenbrand for his fruitful comments.

References

- [1] S. Basu, I. Essa, and A. Pentland. Motion regularization for model-based head tracking. In *ICPR96*, page C8A.3, 1996.
- [2] M. L. Cascia, S. Sclaroff, and V. Athitsos. Fast, reliable head tracking under varying illumination: an approach based on registration of texture-mapped 3d models. *IEEE Trans. on PAMI*, 22(4):322–336, 2000.
- [3] B. Cernuschi-Frias, D. B. Cooper, Y.-P. Hung, and P. N. Belhumeur. Toward a model-based bayesian theory for estimating and recognizing parameterized 3-d objects using two or more images taken from different positions. *IEEE Trans. on PAMI*, 11(10):1028–1052, 1989.
- [4] N. Diehl and H. Burkhardt. Planar motion estimation with a fast converging algorithm. In *Proc. 8th Int. Conf. on Pattern Recognition*, pages 1099–1102, 1986.
- [5] G. Hager and P. Belhumeur. Efficient region tracking with parametric models of geometry and illumination. *PAMI*, 20(10):1025–1039, October 1998.
- [6] M. Harville, A. Rahimi, T. Darrell, G. Gordon, and J. Woodfill. 3d pose tracking with linear depth and brightness constraints. In *ICCV99*, pages 206–213, 1999.
- [7] C. Heipke. A global approach for least-squares image matching and surface reconstruction in object space. *Photogrammetric Engineering & Remote Sensing*, 58(3):317–323, 1992.
- [8] G. Wei, W. Brauer, and G. Hirzinger. Intensity-based and gradient-based stereo matching using hierarchical gaussian basis functions. *Trans. on PAMI*, 20(11):1143–1160, 1998.
- [9] P. Wunsch and G. Hirzinger. Real-time visual tracking of 3-d objects with dynamic handling of occlusions. In *Proc. IEEE Int. Conf. on Robotics and Automation*, 1997.