

A Direct Method for Real-Time Tracking in 3-D under Variable Illumination

Wolfgang Sepp

Institute of Robotics and Mechatronics,
German Aerospace Center (DLR),
82234 Wessling, Germany
wolfgang.sepp@dlr.de

Abstract

3-D tracking of free-moving objects has to deal with brightness variations pronounced by the shape of the tracked surface. Pixel-based tracking techniques, though versatile, are particularly affected by such variations. Here, we evaluate two illumination-adaptive methods for a novel efficient pixel-based 3-D tracking approach. Brightness adaption by means of an illumination basis is compared to with a template update strategy with respect to both robustness and accuracy on tracking in 6 degrees-of-freedom.

1. Introduction

Numerous applications in the fields of robotics, augmented reality, and human-machine interfaces demand solutions for object related pose tracking. Vision-based tracking systems are particularly interesting owing to their low cost and high accuracy. The most popular systems are based on infrared reflecting markers, which can be tracked very robustly. However, these systems require some kind of augmentation of the target object and require dedicated infrared cameras.

On the other side, passive visual tracking methods offer a natural way of tracking objects. Feature-based tracking techniques based on 3-D edges and corners are prevalently stable under changes of illumination but cannot be applied to free-form surfaces. These techniques are outperformed by pixel-based tracking based on template matching, and therefore are not limited to certain shapes. However, latter techniques are affected by illumination changes between the templates.

This paper compares two adaptive methods for direct (pixel-based) tracking of free-moving objects in 6 Degrees-of-Freedom (DoF) with illumination changes. Here, an exact registration and not simply a rough match between target and object template is desired. Thus, accurate models of illumination are required. The first method considered extends tracking to more parameters including the coefficients of an illumination base. The second method updates the tracked 3-D texture template over time. These methods are evaluated with regard to object velocity and accuracy of the tracked pose.

1.1. Previous Work

Sum-of-squared differences (SSD) alignment reaches back to the work of Lucas and Kanade [9] upon which many 2-D tracking algorithms are based. The work did also consider contrast and brightness adjustment within the SSD description of the problem. Extending this tracking method to 3-D requires an efficient formulation of the problem which proves to be increasingly demanding toward full 6-DoF motion.

Diehl et al. [7] developed a fast method for tracking 4-DoF motion of planar objects where illumination effects are neglected. If a reference texture is registered to the model, then planar tracking can be extended to more DoF. Baker et al. [1] solve the 8-D homography while Buenaposada et al. [4] compute the 6-DoF motion explicitly without taking changes in illumination into account.

La Cascia et al. [5] use numerical difference decomposition to minimize the residual error of the projected surface texture from the current view to the reference view. Variations in illumination are compensated by using illumination templates along the lines of Hager and Belhumeur [8].

Some efforts have been undertaken to upgrade tracking of primitive planar surfaces to more general surfaces. Cernuschi-Frias et al. [6] presented an estimation model for simple parameterized surfaces by matching two views on the surface. The approach is based on an orthographic imaging model and has been evaluated for up to 4-DoF geometric surfaces and without handling of illumination changes. The approach of Sepp et al. [12] iteratively estimates the pose of 3-D free-form surface patches in stereo images. A small stereo baseline ensures that illumination changes on the surface, including shadows, need not be modeled explicitly. The computational expense of the method however does not fit the requirements for tracking at frame rate. Recently, Ramey et al. [11] were able to track the coefficients of a b-spline 2 1/2-D surface in stereo images. They used the zero-mean SSD as comparison measure to gain robustness against brightness variations. Promising results have been achieved by Belhumeur et al. [2] whose approach establishes the basis vectors for an optical-flow (pose) subspace and an illumination subspace. The coefficients of these vectors are mapped to 6-DoF object motion under the orthographic projection model. The method, however, has the drawback of a long training session.

Recently Sepp et al. [13] presented an approach capable of tracking arbitrary 3-D surfaces in 6-DoF under full perspective projection in real-time. Yet, this approach has not been evaluated under illumination changes. In this paper we empirically evaluate two illumination-adaptive methods for a similar approach.

2. Direct method for tracking in 3-D

The 3-D surface patch to be tracked is modeled as an arbitrary set of points $X = \{\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_N\} \subset \mathbb{R}^3$. No assumption is made about the topology of the points such as for instance a 2-D grid. Therefore, no constraints other than visibility are imposed on the surface. The rigid body transformation of a point $\mathbf{x} \in X$ is described by

$$m(\mathbf{x}, \mu) = R(\mu) \mathbf{x} + t(\mu) \tag{1}$$

for a pose $\mu \in \mathbb{R}^6$ and the associated 3-D rotation $R(\mu)$ and translation $t(\mu)$. A point in camera frame is mapped to the image under the full perspective projection

$$p(\mathbf{x}) = \left(\frac{\mathbf{k}_1^T \cdot \mathbf{x}}{\mathbf{k}_3^T \cdot \mathbf{x}}, \frac{\mathbf{k}_2^T \cdot \mathbf{x}}{\mathbf{k}_3^T \cdot \mathbf{x}} \right)^T, \quad K = \begin{pmatrix} \mathbf{k}_1^T \\ \mathbf{k}_2^T \\ \mathbf{k}_3^T \end{pmatrix} \quad (2)$$

where $K \in \mathbb{R}^{3 \times 3}$ is the matrix of intrinsic camera parameters. Let $I(\mathbf{u})$ be a brightness value of the *current* image of a live stream at position $\mathbf{u} \in \mathbb{R}^2$ and let $T(\mathbf{u})$ be the brightness value for the *reference* image. In the following ${}^{\text{PI}}I(\mathbf{x}) \equiv I(p(\mathbf{x}))$ and ${}^{\text{P}0}T(\mathbf{x}) \equiv T(p(m(\mathbf{x}, \mu^0)))$.

With these definitions, tracking is formulated as minimization problem $\hat{\delta\mu}^* = \arg \min_{\delta\mu} O(\delta\mu)$ of a least-squares, *compositional* objective function

$$O(\delta\mu) = \sum_{\mathbf{x} \in X} [{}^{\text{PI}}I(m(m(\mathbf{x}, \delta\mu), \hat{\mu})) - {}^{\text{P}0}T(\mathbf{x})]^2. \quad (3)$$

This error function measures the dissimilarity between the surface texture in the current view under the chained poses $\delta\mu, \hat{\mu}$ and the texture in the reference view under the initially registered pose μ^0 . The pose variation $\hat{\delta\mu}^*$ that minimizes (3) gives the pose estimation for the current image I , that is

$$\hat{\mu}^* = \hat{\mu} \circ \hat{\delta\mu}^* \quad \text{according to} \quad m(\mathbf{x}, \hat{\mu}^*) = m(m(\mathbf{x}, \hat{\delta\mu}^*), \hat{\mu}) \quad . \quad (4)$$

The above objective function is minimized with a Gauss-Newton approximation to the Hessian by repeatedly solving the linear equation system

$$\begin{aligned} \sum_{\mathbf{x} \in X} [\partial_{\delta\mu} {}^{\text{PI}}I]^T [\partial_{\delta\mu} {}^{\text{PI}}I] \Big|_{\delta\mu=0, \hat{\mu}} \delta\hat{\mu} = \\ - \sum_{\mathbf{x} \in X} [\partial_{\delta\mu} {}^{\text{PI}}I]^T \Big|_{\delta\mu=0, \hat{\mu}} [{}^{\text{PI}}I|_{\delta\mu=0, \hat{\mu}} - {}^{\text{P}0}T] \end{aligned} \quad (5)$$

for the pose variation $\hat{\delta\mu}$. In practice (5) is not efficient since the image Jacobian has to be recomputed for every frame in the live-stream. Here, the computational expense can be lowered by taking advantage of an approximative image constancy assumption¹ in 3-D at the optimal pose $\hat{\mu} \circ \hat{\delta\mu}^*$, that is

$${}^{\text{PI}}I(m(m(\mathbf{x}, \hat{\delta\mu}^*), \hat{\mu})) = {}^{\text{P}0}T(\mathbf{x}) \quad . \quad (6)$$

The spatial gradient remains constant under this approximation, that is

$$\partial_1 {}^{\text{PI}}I \Big|_{\hat{\delta\mu}^*, \hat{\mu}} \cdot \partial_1 m \Big|_{\hat{\delta\mu}^*, \hat{\mu}} \cdot \partial_1 m \Big|_{\hat{\delta\mu}^*} = \partial_1 {}^{\text{P}0}T, \quad (7)$$

¹ The extended image constancy assumption holds if the surface normal is *parallel* to the camera ray under the current and the initial rigid body transformation.

where operator ∂_i denotes the derivative of the following function with respect to their i -th argument. Thus, the image Jacobian of (5) simplifies to

$$\begin{aligned}\partial_{\delta\mu}{}^{\text{PI}}\Big|_{\hat{\delta\mu}^*, \hat{\mu}} &= \partial_1{}^{\text{PI}}\Big|_{\hat{\delta\mu}^*, \hat{\mu}} \cdot \partial_1 m\Big|_{\hat{\delta\mu}^*, \hat{\mu}} \cdot \partial_2 m\Big|_{\hat{\delta\mu}^*} \\ &= \partial_1{}^{\text{P}0}T \cdot \partial_1 m^{-1}\Big|_{\hat{\delta\mu}^*} \cdot \partial_2 m\Big|_{\hat{\delta\mu}^*} .\end{aligned}\quad (8)$$

Since the Jacobian is evaluated at $\delta\mu = 0$, it can be further simplified to

$$\partial_{\delta\mu}{}^{\text{PI}}(m(m(\mathbf{x}, \delta\mu), \hat{\mu})) = \partial_1{}^{\text{P}0}T(\mathbf{x}) \cdot \partial_2 m(\mathbf{x}, 0) \quad (9)$$

under the assumption that $m(\mathbf{x}, 0)$ equals the identity transformation. Hence, the image Jacobian and Hessian are constant in the compositional framework for every live-stream image.

While the image constancy assumption holds in 2-D leading to the inverse compositional method of Baker and Matthews [1] this equality is only approximately satisfied for general surfaces in 3-D.

3. Illumination Adaptive Methods

The objective function (3) does not consider illumination changes of the surface texture due to a moving object or moving light source. In the following, two methods are considered to cope with these effects.

3.1. Illumination Subspace

Belhumeur and Kriegman [3] proved that illumination variation form a convex polyhedral cone in \mathbb{R}^N . That is, an image and its illumination changes can be reconstructed by a linear combination of orthogonal image vectors $\mathbf{B}_1, \mathbf{B}_2, \dots, \mathbf{B}_M$. So, illumination compensation is added to the objective function (3) in the form

$$O(\delta\mu) = \sum_{\mathbf{x} \in X} [\text{PI}(m(m(\mathbf{x}, \delta\mu), \hat{\mu})) + \mathbf{x}B\lambda - \text{P}0T(\mathbf{x})]^2 \quad , \quad (10)$$

where $\mathbf{x}B = (\mathbf{x}B_1, \mathbf{x}B_2, \dots, \mathbf{x}B_M)$ is a row vector of brightness values of the illumination base for model point \mathbf{x} and λ is the parameter column vector of coefficients to this illumination base. The overall parameters are determined in accordance with [8] by the solution of the linear equation system

$$\begin{aligned}\sum_{\mathbf{x} \in X} [\partial_{\delta\mu}{}^{\text{PI}} , \mathbf{x}B]^T [\partial_{\delta\mu}{}^{\text{PI}} , \mathbf{x}B] \Big|_{\delta\mu=0, \hat{\mu}} \begin{pmatrix} \hat{\delta\mu} \\ \hat{\lambda} \end{pmatrix} &= \\ - \sum_{\mathbf{x} \in X} [\partial_{\delta\mu}{}^{\text{PI}} , \mathbf{x}B]^T \Big|_{\delta\mu=0, \hat{\mu}} [\text{PI}|_{\delta\mu=0, \hat{\mu}} - \text{P}0T] & ,\end{aligned}\quad (11)$$

for pose variation $\hat{\delta\mu}$ and illumination coefficients $\hat{\lambda}$. Note, that omitting illumination compensation in the image constancy assumption allows for the efficient minimization techniques of Sect. 2.

3.2. Template Update Method

Matthews et al. [10] developed a strategy for updating a tracked template without drifts between the updated template and the original one. At the i -th image of the live-stream, minimization starts at the previous pose estimation $\hat{\mu}_{i-1}^*$ with an updated template ${}^{p_0}T_i$ which reads for compositional tracking

$$\hat{\delta\mu}_i = \arg \min_{\delta\mu} \sum_{\mathbf{x} \in X} [P_i(m(m(\mathbf{x}, \delta\mu), \hat{\mu}_{i-1}^*)) - {}^{p_0}T_i(\mathbf{x}))^2] \quad . \quad (12)$$

Subsequently, minimization is continued with the reference template ${}^{p_0}T_0$

$$\hat{\delta\mu}_i^* = \arg \min_{\delta\mu} \sum_{\mathbf{x} \in X} [P_i(m(m(\mathbf{x}, \delta\mu), \hat{\mu}_{i-1}^* \circ \delta\mu_i)) - {}^{p_0}T_0(\mathbf{x}))^2] \quad . \quad (13)$$

The final pose estimation reads $\hat{\mu}_i^* = \hat{\mu}_{i-1}^* \circ \hat{\delta\mu}_i$ and the template is updated following the rule

$${}^{p_0}T_{i+1}(\mathbf{x}) = \begin{cases} P_i(m(\mathbf{x}, \hat{\mu}_i^*)) & : \|\hat{\delta\mu}_i^*\| < \epsilon \\ {}^{p_0}T_i(\mathbf{x}) & : \text{else} \end{cases} \quad (14)$$

Thus, the template is updated only when subsequent minimization with the reference template leads to the same minimum. Here, this strategy is used to update the brightness appearance of the tracked surface in order to account for changing illumination conditions.

4. Evaluation

In the following, experiments are performed on a standard Pentium Xeon 1.7GHz. Video images are gathered with a interlaced camera at PAL resolution and $56^\circ \times 48^\circ$ horizontal and vertical apertures. The internal parameters of the camera together with the distortion coefficients for a 3rd degree polynomial distortion model are determined offline.

The test set consists of two objects. The first surface patch is part of the label of an ordinary 1.5l soda bottle. The patch is modeled as a 83.17° segment of a cylindrical body of radius 4.6cm. Sampling at intervals of 1mm produces

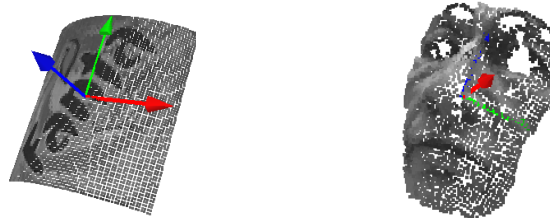


Fig. 1. Texture-registered 3-D point cloud of the objects *bottle* and *sculpture*.



Fig. 2. Screenshots of the tracked sequences *bottle* and *sculpture*. The region of the tracked 3-D model points is manually outlined for better visualization.

4624 surface points. The second surface is the face of a sculpture. By using a 3-D digitizing system, 3668 3-D points of this object are acquired.

The reference textures of the objects are acquired prior tracking. For the first object, the reference image is manually registered to the corresponding 3-D model (see Fig. 1). The point cloud of the sculpture is automatically registered with the image using the same 3-D digitizing system as mentioned above.

A video stream of the objects is recorded at 25Hz starting from the vicinity of the reference pose (see Fig. 2). Four different object velocities are simulated by sub-sampling the sequence with different step sizes. In order to cope with real-world (computing) constraints the number of total minimization steps is limited to 22. In the case of the update strategy, the first 14 steps are performed with the updated template. The dimensionality of the illumination subspace is set to 2 for sequence *bottle* and to 3 for the sequence *sculpture*.

4.1. Robustness

The first experiment evaluates the gain in robustness of the two methods for brightness adaption. Increasing the target velocity implicitly accelerates the brightness variation on the surface. Thus, improved robustness would result in a persistent sequence of the tracked target. Table 1 reports the number of frames

bottle	vel 1	vel 2	vel 3	vel 4	sculpture	vel 1	vel 2	vel 3	vel 4
compositional	719	319	137	1	compositional	903	451	287	205
+ill. subspace	528	258	137	1	+ill. subspace	903	431	284	100
+template update	719	350	137	102	+template update	903	451	286	215
# of frames	719	359	239	228	# of frames	903	451	301	225

Table 1. Robustness under different velocities of sequences *bottle* and *sculpture*. The tables shows the number of successfully tracked frames.

successfully tracked until the object was lost. The use of the illumination subspace degrades the compositional approach for sequences with fair illumination changes while the template update strategy clearly improves overall robustness.

4.2. Accuracy

The second experiment compares both methods for brightness adaption regarding their accuracy. Retro-reflecting markers are attached to the *bottle* object and tracked with a commercially available system. The missing coordinate transformation of marker frame to object frame and world frame to camera frame are estimated offline by means of least-squares fitting over a tracked sequence.

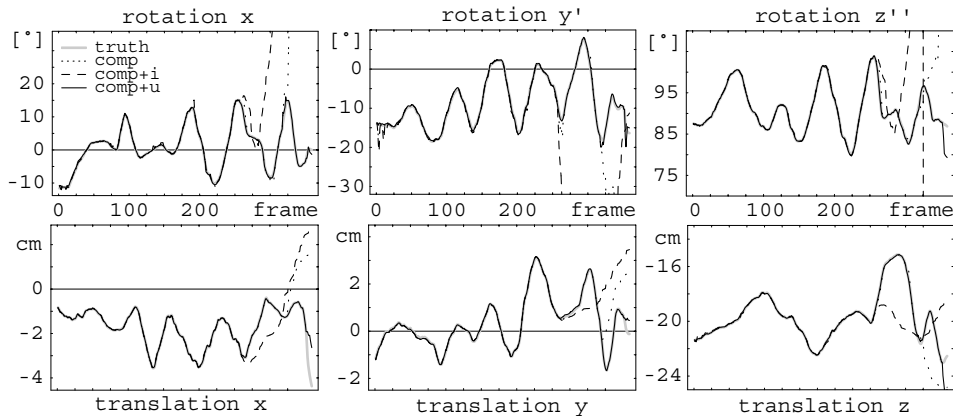


Fig. 3. Cartesian 6-DoF trajectories of sequence *bottle* at velocity 2 for tracking with the illumination subspace method or the template update strategy. The trajectories match the ground truth until the target was lost.

Figure 3 shows the trajectories of the tracked sequence at velocity 2 while Table 2 reports the standard deviations to the ground-truth trajectories. The standard deviations and the displayed trajectories show the superiority of the template update strategy for illumination adaption.

5. Summary and Conclusions

Tracking in the 2-D image plane has been recently extended to 3-D with 6 degrees-of-freedom. Here, we proposed a novel method for efficient 3-D tracking and evaluated two adaptation methods for brightness variations.

The experiments qualitatively and quantitatively showed that extending tracking to more parameters, e.g. for an illumination subspace, degrades the tracked 6-DoF poses. On the other hand, the template update strategy of Matthews et al. [10] increases both, robustness in tracking and accuracy in the trajectories and is therefore the method of choice for tracking objects in 6-DoF with the proposed compositional approach.

bottle	σ_{tx} [°]	σ_{ty} [°]	σ_{tz} [°]	σ_{tx} [mm]	σ_{ty} [mm]	σ_{tz} [mm]
compositional	0.29	0.58	0.12	0.17	0.32	0.38
+ill. subspace	0.35	0.75	0.16	0.19	0.32	0.48
+template update	0.22	0.25	0.10	0.15	0.32	0.28

Table 2. Accuracy over the first 250 frames in sequence *bottle* with velocity 2.

Acknowledgment

The author would like to thank Bernhard Thaler for his support on the evaluation. The work was partly supported by the FP-6 IP *SMErobot* no. 011838-2.

References

1. Simon Baker and Iain Matthews. Lucas-kanade 20 years on: A unifying framework. *International Journal of Computer Vision*, 56(3):221 – 255, March 2004.
2. Peter N. Belhumeur and Gregory D. Hager. Tracking in 3d: Image variability decomposition for recovering object pose and illumination. *Pattern Analysis & Applications*, 2:82–91, 1999.
3. Peter N. Belhumeur and David J. Kriegman. What is the set of images of an object under all possible illumination conditions. *International Journal of Computer Vision*, 28(3):245–260, July 1998.
4. José M. Buenaposada and Luis Baumela. Real-time tracking and estimation of plane pose. In *Proc. ICPR*, volume II, pages 697–700, August 2002. Quebec, Canada.
5. Marco La Cascia, Stan Sclaroff, and Vassilis 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.
6. Bruno Cernuschi-Frias, David B. Cooper, Yi-Ping Hung, and Peter N. Belhumer. 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.
7. Norbert Diehl and Hans Burkhardt. Planar motion estimation with a fast converging algorithm. In *Proc. 8th ICPR*, pages 1099–1102, 1986.
8. Gregory D. Hager and Peter N. Belhumeur. Efficient region tracking with parametric models of geometry and illumination. *IEEE Trans. on PAMI*, 20(10):1025–1039, October 1998.
9. Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision (darpa). In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 674–679, 1981.
10. Iain Matthews, Takahiro Ishikawa, and Simon Baker. The template update problem. In *Proc. of the British Machine Vision Conference*, September 2003.
11. Nicholas A. Ramey, Jason J. Corso, William W. Lau, Darius Burschka, and Gregory D. Hager. Real time 3d surface tracking and its applications. In *Proc. of Workshop on Real-time 3D Sensors and Their Use (at CVPR 2004)*, 2004.
12. Wolfgang Sepp and Gerd Hirzinger. Featureless 6dof pose refinement from stereo images. In *Proc. ICPR*, volume IV, pages 17–20, August 2002. Quebec, Canada.
13. Wolfgang Sepp and Gerd Hirzinger. Real-time texture-based 3-d tracking. In *Proc. 25th Pattern Recognition Symp., DAGM’03*, Sept. 2003. Magdeburg, Germany.