Epipolar-Based Stereo Tracking without explicit 3D Reconstruction

Andre Gaschler Darius Burschka
Department of Computer Science
Technische Universität München
München, Germany
{*gaschler|burschka*}*@cs.tum.edu*

Gregory Hager
Computational Interaction and Robotics Laboratory
Johns Hopkins University
Baltimore, MD 21218
hager@cs.jhu.edu

***Abstract*—We present a general framework for tracking image regions in two views simultaneously based on sum-of-squared differences (SSD) minimization. Our method allows for motion models up to affine transformations. Contrary to earlier approaches, we incorporate the well-known epipolar constraints directly into the SSD optimization process. Since the epipolar geometry can be computed from the image directly, no prior calibration is necessary. Our algorithm has been tested in different applications including camera localization, wide-baseline stereo, object tracking and medical imaging. We show experimental results on robustness and accuracy compared to the known ground truth given by a conventional tracking device.**

***Keywords*-stereo tracking; epipolar geometry;**

I. Introduction

The problem of tracking image patches in binocular stereo videos arises in a variety of applications. Clearly, the use of multiple cameras should improve tracking performance as more information (i.e. two images) is available, and the well-known epipolar constraints must be satisfied on the image [1]. A well known approach to model moving image patches in a stereo environment is to employ 3D coordinates. However, modeling an image patch with 3D coordinates leads to a number of non-linearities, at least for perspective cameras [2]. Moreover, 3D tracking introduces the additional potential inaccuracies due to the effect of calibration errors on the 3D reconstruction. Waxman and Duncan [3] fuse stereo and motion together using the differences in the optical flow in a binocular setup. They describe an elegant method to recover rigid body motions from two rectified views, which is a special case of non-calibrated camera setups.

In contrast to 3D approaches, patches on the image plane can be modeled directly in 2D using epipolar constraints. A simple special case is a rectified stereo setup where corresponding image points lie on the same scan line. Ni and Dellaert [4] exploit rectified geometry to set up a coupled tracking process for a visual odometry application where small image patches share the same y-coordinate. Similarly, Stoyanov et al. [5] incorporate the scan line constraint to improve robustness in surgical applications where tools often occlude the field of view. Other groups also use epipolar constraints to discard poorly tracked features: Yang and Zhang [6] use this approach for their head pose tracker. The same goes for Argyros and Lourakis' [7] hand tracker and Kanbara's et al. [8] augmented reality system. Finally, Ramey et al. [9] directly parameterize their surface tracker and solve for disparity. That way, they can employ epipolar constraints in the case of non-verged cameras.

Our work extends these results to *general epipolar constraints without prior rectification* up to the affine motion model. Epipolar constraints couple the targets in both images without explicit 3D reconstruction. Since the processing stays in the image domain, no additional errors due to calibration inaccuracies or non-linearities due to perspective projection are introduced into the system.

II. Stereo Region Tracking Approach

Our method is derived from the well known SSD tracking [10] [11], as firstly described by Lucas and Kanade [12]. For a single view, motion parameters μ are obtained minimizing the objective function $O(\mu) = \|I(\mu) - I_0\|^2$, where $I(\mu)$ is the warped image patch and I_0 the template image. In the case of stereo images, we make use of the redundant information between left and right view. Integrating the epipolar constraints into the objective function, we obtain a stereo-coupled tracking algorithm. Denoting the two parameter vectors μ_l and μ_r with n entries for each view, we can describe k constraints with the k by $2n$ constraint matrix C. Then, the objective function is given by

$$\begin{aligned} \min \quad O(\delta\mu) \quad &= \|M_l\delta\mu_l + I_l(\mu_l) - I_0\|^2 \\ &\quad + \|M_r\delta\mu_r + I_r(\mu_r) - I_0\|^2 \\ \text{s.t.} \quad C \quad &\begin{bmatrix} \mu_l + \delta\mu_l \\ \mu_r + \delta\mu_r \end{bmatrix} = 0. \end{aligned} \tag{1}$$

M denotes the Jacobian of the image with respect to the parameter vector. Introducing the Lagrange multipliers λ, we can solve for the objective function. With $\nabla O(\delta\mu_l, \delta\mu_r, \lambda) = 0$, we obtain the linear system suitable

for iterative solution:

$$\begin{bmatrix} {M_l}^T M_l & 0 & C^T \\ 0 & {M_r}^T M_r & \\ C & & 0 \end{bmatrix} \begin{bmatrix} \delta\mu_l \\ \delta\mu_r \\ \lambda \end{bmatrix} = - \begin{bmatrix} {M_l}^T \left(I_l(\mu_l) - I_0\right) \\ {M_r}^T \left(I_r(\mu_r) - I_0\right) \\ 0 \end{bmatrix} \quad (2)$$

A. Stereo Constraints for non-calibrated Views

In rectified stereo views, corresponding points in the left and right image share the same y-coordinate. From this section on, we assume the affine motion model. Let A_l and A_r be affine transformation matrices whose variable entries mirror the parameter vectors μ_l and μ_r. For rectified stereo views, we easily verify three linear constraints $a_{l21} = a_{r21}$, $a_{l22} = a_{r22}$ and $a_{l23} = a_{r23}$. Translation and similarity motion models are only special cases of the affine model and therefore possess linear constraints analogously.

Provided that the fundamental matrix is known, stereo constraints can also be set up for non-calibrated views. We will show this procedure for the affine transformation—translation and similarity transformation are special cases of the affine transformation and again analogous.

We pick three points in the template image: The template image center $P_1 = (0\ 0\ 1)^T$ and two infinitesimal close points $P_2 = (d\ 0\ 1)^T$ and $P_3 = (0\ d\ 1)^T$. Given the fundamental matrix F and the affine transformations A_l and A_r, these three points will fulfill the following stereo constraint:

$$\left(\left(A_l + \delta A_l\right) P\right)^T F \left(A_r + \delta A_r\right) P = 0 \quad (3)$$

F may be given as a result of any standard algorithm described in [1]. Note that A_l and A_r are the affine transformations for the previous camera frame. Therefore, the magnitude of the subsequent parameter updates δA is usually small. Expanding the above equation, we can therefore ignore higher order terms of δA and d. Thus, we obtain three purely linear constraints

$$\begin{aligned} \left(\delta A_l P\right)^T F\ A_r P + \left(A_l P\right)^T F\ \delta A_r P \\ = - \left(A_l P\right)^T F A_r P. \end{aligned} \quad (4)$$

Since the points P are chosen infinitesimally close to the center of the template, the constraints are in fact a local linearization of the epipolar geometry. Note that A_l, A_r are known from the previous iteration and F is a constant. Thus, we can calculate a constraint matrix C connecting the components of the parameter updates δA_l and δA_r as needed for the iterative solution in (2).

B. Direct Solution of the Constrained Problem

Considering equation 2, we wonder about the high dimensionality of the linear system. Since the constraints reduce the dimensionality of possible solutions, we might rather expect a $2n - k$ linear system.

For that, we introduce the transformed parameter vector $x \in R^{2n-k}$ and the $2n$ by $2n - k$ transformation matrix T. T should be chosen in such a way that all $\mu = Tx$ fulfill the constraints. Using this parameter transformation we can reformulate the objective function to

$$\left\|M\delta\mu + I(\mu) - I_0\right\|^2 = \left\|MTx + I(\mu) - I_0\right\|^2. \quad (5)$$

That way, the constraints are always ensured by $C\mu = CTx = 0$. Plus, the dimensionality of the solution of x becomes $2n - k$:

$$\left(T^T M^T M T\right) x = -T^T M^T \left(I(\mu) - I_0\right) \quad (6)$$

Solving the linear system, we obtain the linear approximation step

$$\delta\mu = -T\ \left(T^T M^T M T\right)^{-1} T^T M^T \left(I(\mu) - I_0\right). \quad (7)$$

Obviously, Tx yields only allowed parameters.

As a more general solution, we define T to be the *basis of the null space* of C. That way, we theoretically verify $C\mu = 0$ if and only if $\mu = Tx$. Thus, we can always construct the transformation matrix T from a given constraint matrix C and vice versa, as C is a basis of the null space of T^T. In the special case of rectified views, T becomes a constant and the direct solution is further simplified. This special case is identical to the algorithm in [4] and very similar to that in [13].

An intuitive explanation of our reduced parameter set is the following: Due to the epipolar geometry, any transformation perpendicular to the epipolar lines has to occur in both views. Translation, scaling and skew perpendicular to the epipolar lines give 3 independent parameters. Similarly, translation, scaling and skew parallel to the epipolar lines result in 3 parameters for each view. Thus, we can model the affine motions with 9 parameters in total.

C. Practical Implementation

In practical applications, neither rectification nor the fundamental matrix are perfectly accurate. For that reason, the constraints given above do not hold perfectly. It is therefore beneficial to set up *soft constraints*:

$$\begin{aligned} \min\ O\left(\delta\mu\right) &= \left\|M_l\delta\mu_l + I_l(\mu_l) - I_0\right\|^2 \\ &+ \left\|M_r\delta\mu_r + I_r(\mu_r) - I_0\right\|^2 \\ &+ \lambda \left\|C\left(\mu + \delta\mu\right)\right\|^2 \end{aligned} \quad (8)$$

where $\lambda > 0$ is a design parameter.

Analogously solving for $\nabla O(\delta\mu) = 0$ results in the linear system

$$\left[\begin{bmatrix} {M_l}^T M_l & 0 \\ 0 & {M_r}^T M_r \end{bmatrix} + \lambda^2 C^T C\right] \begin{bmatrix} \delta\mu_l \\ \delta\mu_r \end{bmatrix} = - \begin{bmatrix} {M_l}^T \left(I_l(\mu_l) - I_0\right) \\ {M_r}^T \left(I_r(\mu_r) - I_0\right) \end{bmatrix} - \lambda^2 C^T C\ \mu \quad (9)$$

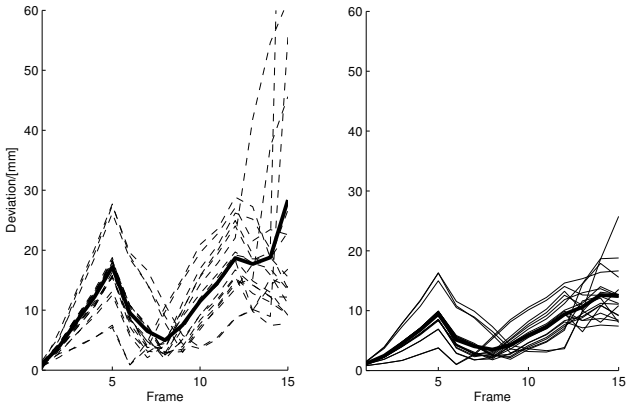

Figure 1. Spatial deviation of tracked image patches. Left: Unconstrained tracker. Right: Epipolar constraints enforced. Bold lines show the average.

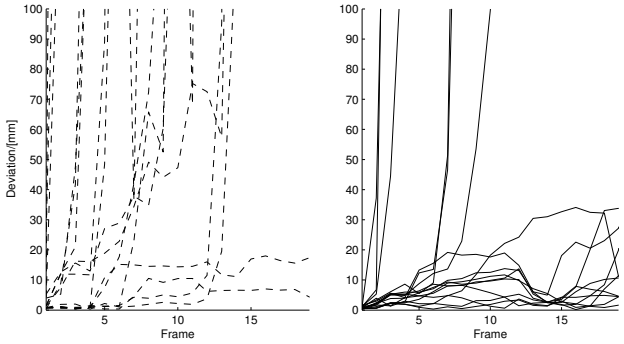

Figure 2. Spatial deviation of tracked image patches in a wide baseline setup. Left: Unconstrained tracker. Right: Epipolar constraints enforced.

In contrast to the hard constraints, the soft constraint matrix C needs to be scaled (by λ) depending on how precisely the conditions should be met. In our experiments, we used scale factors at the order of $\lambda = 10^{-1}N$, where N is the number of pixels.
In order to improve computational performance, we also made use of factorizing the Jacobian M into a constant matrix M_0 and a time-varying matrix Σ, as described by [10].

III. Experiments and Applications

The algorithm presented was implemented in Matlab and C++. The latter uses the Intel Integrated Performance Primitives (IPP) libraries, OpenCV 1.0 as well as CISST Stereo Vision. With the C++ version, we could show that integrating epipolar constraints in the tracking algorithm does not slow down performance. The bottleneck of the tracking process is warping and interpolation.

A. Binocular Tracking for Localization

For many applications, image features need to be tracked accurately in order to obtain camera motion. In the first experiment, we set up a stereo rig of two Point Grey High Definition cameras. Different planar images were moved in front of the cameras. The image set contained both blurry and well-textured standard test images. The motion of the images was captured by an NDI Optotrak motion capture system. Using the explicit $AX = XB$ solution by Park and Martin [14] and the optimization method by Strobl and Hirzinger [15], we calibrated the moving image plane up to spatial deviation of 0.8 mm and an angular error of 0.6 degrees. Thus, we compared the visual tracking results to a known ground truth given by the motion capture system.

Figure 1 shows the accuracy of our proposed tracking algorithm (on the right) in comparison to regular unconstrained tracking (left). The difference in accuracy can be explained as follows: The regular tracking algorithm gives adequate results on pixel level in each camera image. However, it does not directly estimate disparity. In contrast to that, our stereo tracking algorithm directly solves for disparity. Thus, the integration of epipolar constraints allows a far better estimation of disparity and therefore z-coordinates. That way, the overall spatial deviation in our algorithm is superior to the regular approach.

B. Stereo Tracking in Wide Baseline Camera Setups

In the second experiment, we set up a verged stereo system with a wide baseline of 749.2 mm. A planar image was moved in the common field of view. We could register the poses of the image plane up to a residual RMS error of 1.16 mm. Then, we ran the tracking algorithms on a number of image patches on the moving plane. In the case of the conventional tracking algorithm, almost all image patches were lost within a short sequence (see Figure 2). Integrating epipolar constraints leads to a substantially higher robustness. This result can be explained by the significantly different perspectives. The image patch shows much greater distortion in one view than in the other. Coupling both views therefore allows for higher robustness.

C. Tracking of Biomedical Surfaces

In a third experiment, we used binocular video data from a minimally-invasive prostatectomy. Since the surgery was conducted with a da Vinci tele-manipulation system (Intuitive Surgical, Sunnyvale, CA), a stereo endoscopic view was readily available. Biomedical surfaces usually pose certain difficulties on the tracking process: First, the surface is constantly moving, deforming and poorly textured. Second, illumination leads to specularities in the endoscopic view. Third, occlusions by surgical tools are frequent.

In this experiment, ground truth is not known. However, we could compare the convergence of random image patches. Rectangular image regions were randomly picked and visually tracked for a number of frames. Regions with

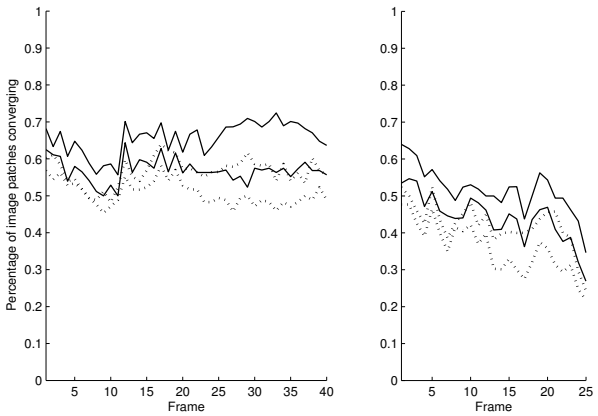

Figure 3. Random patches being tracked up to deviation of 1 pixel (lower lines) and 5 pixels (upper lines) in two video sequences. Dotted lines: Unconstrained tracker. Solid lines: Epipolar constraints enforced.

high residuals were automatically discarded, ambigious regions were manually classified. After that, the percentage of inliers up to a certain deviation was calculated. As shown in Figure 3, our algorithm has slightly higher robustness. This result can be explained by the high number of specularities in the endoscopic view. Since specularities appear differently in the respective cameras, a stereo tracking process can handle them more easily. Poor image conditions in one view can partly be compensated for by the other.

IV. Conclusion

This paper has proposed a stereo tracking approach based on epipolar constraints. We have used epipolar geometry to directly couple affine region tracking in both image views. In contrast to conventional 3D tracking, we work directly in the image space and avoid the need for 3D reconstruction. Our method achieves robust registration of region features in stereo views, even for wide-baseline setups. We have shown high accuracy compared to the ground truth given by a conventional tracking device.

Acknowledgment

Andre Gaschler was supported by Stiftung Familie Klee Advancement Award from Stiftung Familie Klee, Germany.

References

[1] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge University Press New York, NY, USA, 2003.

[2] L. Shapiro, A. Zisserman, and M. Brady, "3D motion recovery via affine epipolar geometry," *International Journal of Computer Vision*, vol. 16, no. 2, pp. 147–182, 1995.

[3] A. Waxman and J. Duncan, "Binocular image flows: steps toward stereo-motion fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, p. 729, 1986.

[4] K. Ni and F. Dellaert, "Stereo tracking and three-point/one-point algorithms-a robust approach in visual odometry," in *IEEE International Conference on Image Processing*, 2006, pp. 2777–2780.

[5] D. Stoyanov, G. Mylonas, F. Deligianni, A. Darzi, and G. Yang, "Soft-tissue motion tracking and structure estimation for robotic assisted mis procedures," *Lecture Notes in Computer Science*, vol. 3750, p. 139, 2005.

[6] R. Yang and Z. Zhang, "Model-based head pose tracking with stereovision," in *Proc. Fifth IEEE International Conference on Automatic Face and Gesture Recognition (FG2002)*, pp. 255–260.

[7] A. Argyros and M. Lourakis, "Binocular hand tracking and reconstruction based on 2D shape matching," in *Pattern Recognition. 18th International Conference on*, vol. 1, 2006, pp. 207–210.

[8] M. Kanbara, N. Yokoya, and H. Takemura, "Registration for stereo vision-based augmented reality based on extendible tracking of markers and natural features," in *Pattern Recognition. 16th International Conference on*, vol. 2, 2002, pp. 1045–1048.

[9] N. Ramey, J. Corso, W. Lau, D. Burschka, and G. Hager, "Real time 3d surface tracking and its applications," in *Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04) Volume 3*, 2004, p. 34.

[10] G. Hager and P. Belhumeur, "Efficient region tracking with parametric models of geometry and illumination," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 10, pp. 1025–1039, 1998.

[11] S. Baker and I. Matthews, "Lucas-kanade 20 years on: A unifying framework," *International Journal of Computer Vision*, vol. 56, no. 3, pp. 221–255, 2004.

[12] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *International joint conference on artificial intelligence*, vol. 3, 1981, p. 3.

[13] J. Morat, F. Devernay, and S. Cornou, "Tracking with stereo-vision system for low speed following applications," in *Intelligent Vehicles Symposium, 2007 IEEE*, june 2007, pp. 955–961.

[14] F. Park and B. Martin, "Robot sensor calibration: solving AX= XB on the Euclidean group," *IEEE Transactions on Robotics and Automation*, vol. 10, no. 5, pp. 717–721, 1994.

[15] K. Strobl and G. Hirzinger, "Optimal hand-eye calibration," in *Proceedings of the IEEE/RSJ International Conference on Robots and Intelligent Systems*, 2006, pp. 4647–4653.