Stereo 3D Object Trajectory Reconstruction

Sebastian Bullinger, Christoph Bodensteiner, Michael Arens
Fraunhofer IOSB
Ettlingen, Germany
<firstname>.<lastname>@iosb.fraunhofer.de

Rainer Stiefelhagen
Karlsruhe Institute of Technology
Karlsruhe, Germany
rainer.stiefelhagen@kit.edu

Abstract

We present a method to reconstruct the three-dimensional trajectory of a moving instance of a known object category using stereo video data. We track the two-dimensional shape of objects on pixel level exploiting instance-aware semantic segmentation techniques and optical flow cues. We apply Structure from Motion (SfM) techniques to object and background images to determine for each frame initial camera poses relative to object instances and background structures. We refine the initial SfM results by integrating stereo camera constraints exploiting factor graphs. We compute the object trajectory by combining object and background camera pose information. In contrast to stereo matching methods, our approach leverages temporal adjacent views for object point triangulation. As opposed to monocular trajectory reconstruction approaches, our method shows no degenerated cases. We evaluate our approach using publicly available video data of vehicles in urban scenes.

1. Introduction

1.1. Trajectory Reconstruction

The reconstruction of three-dimensional object motion trajectories is important for autonomous systems and surveillance applications. There are different platforms like drones or wearable systems where one wants to achieve this task with a minimal number of devices in order to reduce weight or lower production costs. We propose an approach to reconstruct three-dimensional object motion trajectories using two cameras as sensors. These results are essential for applications like environment perception and geo-registration of three-dimensional object trajectories. 3D stereo measurement precision deteriorates quickly with camera distance [20] due to limited camera baselines. We tackle this problem by combining temporal adjacent views using Structure from Motion techniques. Even small object rotations may result in big camera baseline differences.

In many scenes objects cover only a minority of pixels. This increases the difficulty of reconstructing object motion trajectories using image data. In such cases current state-of-the-art Structure from Motion (SfM) approaches [18,22] treat moving object observations most likely as outliers and reconstruct background structures instead. Previous works, e.g. [15,16], detect moving objects by applying motion segmentation or keypoint tracking. Recent progress in instance-aware semantic segmentation [7,11] and optical flow [13,12] techniques allow for object tracking on pixel level [2] and handle stationary objects naturally. We extend the approach in [2] to track objects on pixel level in stereo video data. Stereo object tracking allows us to use [22] and [18] for object and background reconstruction. We refine the reconstruction results by incorporating stereo constraints using GTSAM [6]. GTSAM provides functionality to model reconstruction problems with factor graphs. The incorporation of stereo constraints removes the scale ambiguity between object and background reconstruction and allows us to compute consistent object motion trajectories.

1.2. Related Work

Semantic segmentation or scene parsing is the task of providing semantic information at pixel-level. Shelhamer et al. [23] applied Fully Convolutional Networks for semantic segmentation, which are trained end-to-end. Recently, [7,11] proposed instance-aware semantic segmentation approaches. The field of Structure from Motion (SfM) can be divided into iterative and global approaches. Iterative or sequential SfM methods [18,22,25] are more likely to find reasonable solutions than global SfM approaches [18,25]. However, the latter are less prone to drift. GTSAM [6] allows to model and to optimize SfM problems using factor graphs, but does not provide functionality to perform data association and initialization. [3] analyze the importance of initialization techniques for Simultaneous Localization and Mapping (SLAM) using GTSAM. We perform data association and initialization using state-of-the-art SfM libraries [18,22]. Previous works [4,24] exploit specific camera
poses to reconstruct object trajectories in monocular video data. These approaches are specifically defined for driving scenarios. [8] reconstruct vehicle shapes and trajectories in stereo video data using off-the-shelf ego-motion and stereo reconstruction algorithms. [19] combine object proposals, stereo, visual odometry and scene flow to compute three-dimensional vehicle tracks in traffic scenes. The object trajectory reconstructions in [8] and [19] are limited by the stereo camera baseline.

1.3. Contribution

The core contributions of this work are as follows. (1) We present a new framework to reconstruct the three-dimensional trajectory of moving instances of known object categories in stereo video data leveraging state-of-the-art semantic segmentation and structure from motion approaches. (2) We propose a novel approach to track the two-dimensional shape of objects on pixel level in stereo video data. (3) We present a novel method to compute object motion trajectories consistent to image observations and background structures using state-of-the-art SfM techniques for data initialization and factor graphs for refinement exploiting stereo constraints. (4) Opposed to stereo matching methods, our approach leverages views from different time steps for object point triangulation. (5) We demonstrate the usefulness of our method by showing qualitative results of reconstructed object motion trajectories.

2. Object Motion Trajectory Reconstruction

The pipeline of our approach is shown in Fig. 1. The input is an ordered stereo image sequence. We track two-dimensional object shapes on pixel level across video sequences exploiting instance-aware semantic segmentation [17] to identify object shapes and optical flow [13] to associate extracted object shapes in corresponding stereo images and subsequent frames. Without loss of generality, we describe motion trajectory reconstructions of single objects. We apply SfM [18, 22] to object and background images as shown in Fig. 1. Object images denote pictures containing only color information of single object instances. Similarly, background images show only environment structures. We combine information of object and background SfM reconstructions to determine consistent object motion trajectories. We use GTSAM [6] to refine object and background reconstructions and resolve the scale ambiguity using stereo camera baseline constraints.

The point triangulation of stereo matching or stereo correspondence [21] methods are limited by the baseline of corresponding the stereo camera [20]. In contrast, SfM allows to triangulate 3D points by exploiting information of subsequent frames. Since already small object rotations may result in big camera baseline changes, our method is not necessarily limited by the stereo camera baseline. In contrast to stereo correspondence methods, the proposed approach builds object models reflecting the information of each frame. To build an object model with stereo matching techniques requires additional steps to fuse the 3D points of subsequent frames. The presented method does not require a calibration of the stereo camera.

2.1. Stereo Online Multiple Object Tracking

The proposed Stereo Multiple Object Tracking (MOT) approach extends the monocular tracking algorithm presented in [2] and is depicted in Fig. 2. [2] allows to track the two-dimensional shape of objects of known categories across video sequences on pixel level. We use optical flow matches to associate instance-aware semantic segmentations between subsequent frames to maintain the tracker state. In contrast to motion model based tracking methods, this approach allows to naturally associate objects between left and right images of stereo cameras. [2] uses the Kuhn-Munkres algorithm [14] to solve the assignment problem, i.e. to determine object associations of objects between image pairs. The assignment problem consists of finding a maximum weight matching in a weighted two-dimensional (or bipartite) graph. This problem translates in the stereo MOT case to a four-dimensional matching problem, because the object instances in the left image \( I_{i,t} \) and the right image \( I_{i,r} \) at time \( i \) as well as the object instances in the left image \( I_{i+1,t} \) and the right image \( I_{i+1,r} \) at time \( i + 1 \) must be associated. Let \( OF_{i,t,r} \) and \( OF_{i+1,t,r} \) denote the optical flow between image \( I_{i,t} \) and \( I_{i,r} \) as well as \( I_{i+1,t} \) and \( I_{i+1,r} \). We do not solve the associations of \( I_{i,t}, I_{i+1,t}, I_{i,r} \) and \( I_{i+1,r} \) simultaneously, since (a) the four-dimensional matching problem is NP-complete and (b) the simultaneous
determination of two subsequent stereo image pairs requires the computation of three optical flow fields in addition to $O_{F_{i+1,l}}$ and $O_{F_{i,ln}}$. Instead, we track object instances in the left images $I_{i,l}$ and $I_{i+1,l}$ using the object affinity matrix presented [2] as input for the Kuhn-Munkres algorithm. Concretely, the affinity matrix is defined according to equation (1)

$$A_t = \begin{bmatrix} O_{1,1} & \cdots & O_{1,v} & \cdots & O_{1,n_v} \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
O_{u,1} & \cdots & O_{u,v} & \cdots & O_{u,n_v} \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
O_{n_u,1} & \cdots & O_{n_u,v} & \cdots & O_{n_u,n_v} \end{bmatrix}.$$  

(1)

Here, $O_{u,v}$ denotes the overlap of prediction $u$ in $P_{i,ln}$ and detection $v$ in $D_{i+1,l}$ (see Fig. 2). Let $n_u$ denote the number of predictions in $P_{i,ln}$ and $n_v$ denote the number of detections in $D_{i+1,l}$. This affinity measure reflects locality and visual similarity. The tracker state $T_{i+1,l}$ contains only tracks of object instances in images corresponding to the left camera. We use the optical flow between left and right images $O_{F_{i+1,lr}}$ to associate the tracker state of left images $T_{i+1,l}$ with objects visible in the corresponding right image. The association between predictions $P_{i+1,lr}$ and detections $D_{i+1,r}$ in the right images are also solved using the affinity matrix of [2] as input for [14]. In this case $O_{u,v}$ denotes the overlap of prediction $u$ in $P_{i+1,lr}$ and detection $v$ in $D_{i+1,r}$. $n_u$ denotes the number of predictions in $P_{i+1,lr}$ and $n_v$ de-
notes the number of detections in $D_{i+1,r}$.

### 2.2. Object Motion Trajectory Computation

We follow the pipeline outlined in Fig. 1 and apply SfM simultaneously to object and background images. We denote corresponding reconstruction results with $s_f m^{(o)}$ and $s_f m^{(b)}$. Each object image has a corresponding background image, i.e., the background image extracted from the same input frame. We consider only object-background-image-pairs, which are part of $s_f m^{(o)}$ and $s_f m^{(b)}$. Reconstructed cameras without corresponding object or background camera are removed from the reconstruction.

Let $o_j^{(o)}$ denote the 3D points contained in $s_f m^{(o)}$. The superscript $o$ in $o_j^{(o)}$ describes the corresponding coordinate frame. The variable $j$ denotes the index of the points in the object point cloud. We combine information of object-background-image-pairs to define object motion trajectories parameterized by a single parameter. The object reconstruction $s_f m^{(o)}$ contains object point positions $o_j^{(o)}$ as well as corresponding camera centers $c_i^{(o)}$ and rotations $R_i^{(o)}$. We convert the object points $o_j^{(o)}$ defined in the coordinate frame system (CFS) of the object reconstruction to points in the camera CFS $o_j^{(i)}$ of camera $i$ using $o_j^{(i)} = R_i^{(o)} \cdot (o_j^{(o)} - c_i^{(o)})$. We use the camera center $c_i^{(b)}$ and the corresponding rotation $R_i^{(b)}$ contained in the background reconstruction $s_f m^{(b)}$ to transform object points in camera coordinates to the background CFS using equation (2).

$$o_j^{(b)} = c_j^{(b)} + R_j^{(b)}^T \cdot o_j^{(i)}$$ (2)

The naive combination of object and background reconstruction results in inconsistent object motion trajectories due to the scale ambiguity of SfM. We adjust the scale between object and background reconstruction using the baseline of the stereo cameras in object and background reconstructions as reference. Reconstructions of dynamic objects using state-of-the-art SfM tools contain occasionally badly registered cameras and incorrectly triangulated object points (see Fig. 3). Reasons for these are small object sizes, changing illumination and reflecting surfaces. Incorrectly estimated camera baselines hamper the correct estimation of the scale ratio between object and background reconstruction.

We leverage factor graphs to model stereo camera constraints and to refine the previously computed SfM reconstructions. For each triangulated point we search for corresponding stereo feature observations, i.e., pairs of feature observations which appear in the left and the right image of the same time step. Stereo image rectification preprocessing allows us to assume that the stereo feature observation positions should show (almost) the same $y$ coordinate. Since the feature observations are computed for each image independently we use only stereo feature observations with an $y$ pixel difference smaller than three pixels. We average the $y$ coordinate to define the final stereo constraint. The resulting reconstructions show consistent camera stereo baselines. Note that GTSAM does not provide functionality to perform data association and initialization. Fig. 3 shows a comparison of initial and refined reconstructions.

We can recover the full object motion trajectory computing equation (2) for each object-background-image-pair. We use $o_j^{(b)}$ of all cameras and object points as object motion trajectory representation.

### 3. Experiments and Evaluation

Due to the lack of suitable benchmark datasets, we show qualitative results using publicly available video data. For object tracking we evaluated for instance-aware semantic segmentation and for optical flow computations. We observed that and achieved the best segmentation and optical flow results. computes more stable optical flow vectors for moving objects than. We considered the following SfM pipelines for object and background reconstructions: Colmap, OpenMVG, Theia, and VisualSfM. Our object trajectory reconstruction pipeline uses Colmap for object and OpenMVG for background reconstructions. Colmap and Open-
Figure 4: Vehicle trajectory reconstruction using three sequences (stuttgart01-stuttgart03) contained in the Cityscape dataset \cite{Pirk2016} and one sequence (2011_09_26_drive_0013) of the KITTI dataset \cite{Geiger2013}. Object segmentations and object reconstructions are exemplarily shown for one of the vehicles visible in the scene. The reconstructed cameras are shown in red. The vehicle trajectories are colored in green and blue. The figure is best viewed in color.
MVG created the most reliable object and background reconstructions in our experiments.

4. Conclusions

This paper presents a pipeline to reconstruct the three-dimensional trajectory of moving objects using stereo video data. We presented a novel approach to track objects on pixel level across stereo video sequences. This allows us to apply state-of-the-art SFM techniques simultaneously to different objects. We demonstrate how to resolve the scale ambiguity of object and background sfm reconstructions leveraging stereo constraints. In contrast to previously published methods, our approach w.r.t decreasing object sizes.

References

[1] S. Baker, S. Roth, D. Scharstein, M. J. Black, J. P. Lewis, and R. Szeliski. A database and evaluation methodology for optical flow. In 2007 IEEE 11th International Conference on Computer Vision, pages 1–8, 2007.

[2] S. Bullinger, C. Bodensteiner, and M. Arens. Instance flow based online multiple object tracking. In IEEE International Conference on Image Processing (ICIP), 2017.

[3] L. Carlone, R. Tron, K. Daniilidis, and F. Dellaert. Initialization techniques for 3d SLAM: A survey on rotation estimation and its use in pose graph optimization. In IEEE International Conference on Robotics and Automation, ICRA 2015, Seattle, WA, USA, 26-30 May, 2015, pages 4597–4604, 2015.

[4] F. Chhaya, N. D. Reddy, S. Upadhyay, V. Chari, M. Z. Zia, and K. M. Krishna. Monocular reconstruction of vehicles: Combining SLAM with shape priors. In 2016 IEEE International Conference on Robotics and Automation, ICRA 2016, Stockholm, Sweden, May 16-21, 2016, 2016.

[5] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[6] F. Daellert. Factor graphs and gtsam: A hands-on introduction. Technical report, GT-RIM-CP&R-2012-002, 2012.

[7] J. Dai, K. He, and J. Sun. Instance-aware semantic segmentation via multi-task network cascades. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[8] F. Engelmann, J. Stückler, and B. Leibe. SAMP: shape and motion priors for 4d vehicle reconstruction. In IEEE Winter Conference on Applications of Computer Vision, WACV, 2017.

[9] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. Int. J. Rob. Res., 32(11), 2013.

[10] R. I. Hartley and A. Zisserman. Multiple View Geometry in Computer Vision. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[11] K. He, G. Gkioxari, P. Dollar, and R. Girshick. Mask r-cnn. In The IEEE International Conference on Computer Vision (ICCV), Oct 2017.

[12] Y. Hu, R. Song, and Y. Li. Efficient coarse-to-fine patchmatch for large displacement optical flow. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[13] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[14] H. W. Kuhn. The hungarian method for the assignment problem. Naval Research Logistics Quarterly, 2:83–97, 1955.

[15] A. Kundu, K. M. Krishna, and C. V. Jawahar. Realtime multi-body visual slam with a smoothly moving monocular camera. In ICCV, 2011.

[16] K. Lebeda, S. Hadfield, and R. Bowden. 2D or not 2D: Bridging the gap between tracking and structure from motion. In Proceedings, Asian Conference on Computer Vision (ACCV), 2014.

[17] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei. Fully convolutional instance-aware semantic segmentation. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017.

[18] P. Moulon, P. Monasse, R. Marlet, and Others. Openmvg. an open multiple view geometry library., 2013.

[19] A. Ösep, W. Mehner, M. Mathias, and B. Leibe. Combined image- and world-space tracking in traffic scenes. In ICRA, 2017.

[20] P. Pinggera, D. Pfeiffer, U. Franke, and R. Mester. Know your limits: Accuracy of long range stereoscopic object measurements in practice. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, Computer Vision – ECCV 2014, 2014.

[21] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International Journal of Computer Vision, 2002.

[22] J. L. Schönberger and J.-M. Frahm. Structure-from-motion revisited. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[23] E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell., 39(4):640–651, 2017.

[24] S. Song, M. Chandraker, and C. C. Guest. High accuracy monocular SFM and scale correction for autonomous driving. IEEE Trans. Pattern Anal. Mach. Intell., 38(4), 2016.

[25] C. Sweeney, Theia Multiview Geometry Library: Tutorial & Reference. University of California Santa Barbara., 2014.

[26] C. Wu. Visualsfm: A visual structure from motion system, 2011.