$SE\left(3\right)$ Multimotion Estimation Through Occlusion

Kevin M. Judd[1] and Jonathan D. Gammell[1]

***Abstract*— Visual motion estimation is an integral and well-studied challenge in autonomous navigation. It is significantly more challenging in highly dynamic environments with multiple moving objects. This paper introduces an approach to this *multimotion estimation problem* capable of estimating the full $SE\left(3\right)$ trajectory of every motion in a scene, even when motions become occluded. The Multimotion Visual Odometry (MVO) pipeline employs multilabeling techniques and continuous motion models to estimate all motions simultaneously, including the camera egomotion. Motion closure is used to recognize when trajectories become unoccluded, and the motion models are used to interpolate the occluded estimates. The estimation performance of the pipeline is demonstrated on real-world trajectory data from the Oxford Multimotion Dataset.**

I. Introduction

Safely navigating through dynamic environments is important in robotics and consistently estimating multiple, continuous motions from incomplete observations is integral to this task. Visual odometry (VO) is widely used to estimate the egomotion of a camera by isolating the static parts of a scene. Recent work has addressed the *multimotion estimation problem* by focusing on the dynamic regions of a scene that VO rejects. A rigid-motion assumption is often used to simplify the problem as common objects (e.g., vehicles) tend to move rigidly, and more complex dynamic objects (e.g., humans) can be treated as piecewise-rigid motions.

Aspects of the rigid multimotion estimation problem have been addressed by a variety of techniques, including factorization and model selection. Factorization techniques use matrix decomposition to determine the motion and shape of each dynamic object [1]. This factorization usually requires points to be tracked for the entirety of the estimation window, which is difficult in complex scenes due to motion blur or lighting changes that deteriorate the quality of measurements.

Ozden et al. [2] consider many practical challenges in multimotion estimation, such as incomplete feature tracks, and propose a model selection framework that relies on separate egomotion estimation. While this technique explicitly models the merging and splitting of motions, it does not address *occlusions*, where objects temporarily obscure each other or leave the field of view of the sensor. Motion estimation systems must be robust to observation dropouts as highly dynamic scenes tend to include significant amounts of occlusion.

Tracking occluded objects in dynamic scenes is a principle problem in multiple object tracking (MOT). A variety of specific, appearance-based object models are used to detect targets in each frame and techniques focus on accurately associating present and past detections [4]. Partially occluded objects can be tracked using these models by inferring the position of the entire object from the portions that are visible [5]. The motions of fully occluded objects are often predicted using motion models [6]. These object models are difficult to generalize, so tracking techniques are often designed for specific applications and employ constrained motion models. These assumptions limit their ability to track general objects and estimate the full $SE\left(3\right)$ pose of each object.

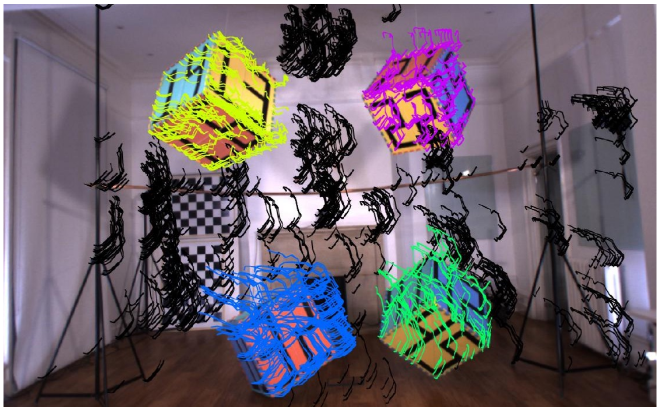


Fig. 1. Motion segmentation produced by our multimotion visual odometry (MVO) system. The egomotion of the camera is estimated from the static points in the scene shown in black. The other colors represent the segmentation of the other estimated motions in the scene.

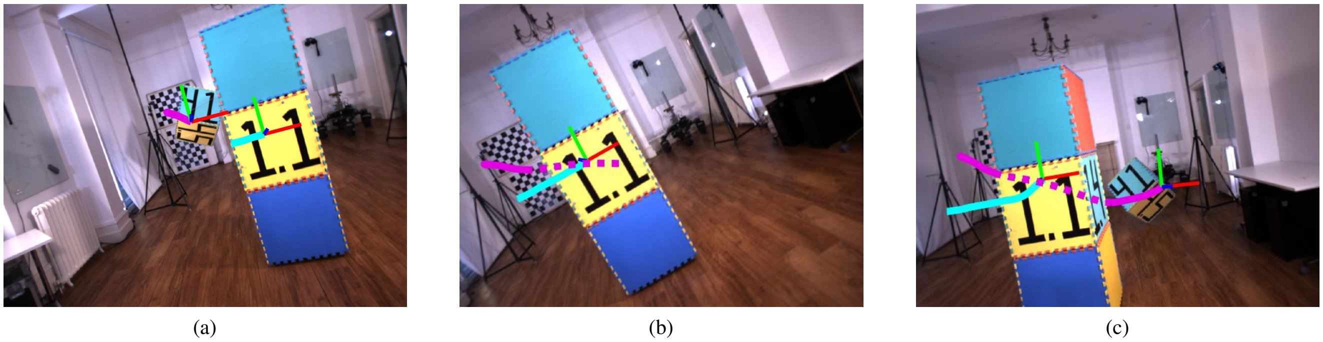
Our multimotion visual odometry (MVO) pipeline [7] addresses the multimotion estimation problem by applying multimodel fitting techniques to the traditional VO pipeline. MVO simultaneously estimates the full $SE\left(3\right)$ trajectory of every motion in a scene, including the egomotion. The original pipeline relies on direct observations and is therefore unable to handle significant observation dropouts.

This paper demonstrates how MVO can be extended to estimate multiple motions through occlusion by exploiting a physically founded motion model. A white-noise-on-acceleration (i.e., locally constant-velocity) model is used to extrapolate motion estimates until the object becomes visible. These estimates are used in *motion closure* to recover tracking when objects reappear in the predicted location, after which the occluded estimates can be interpolated. The full $SE\left(3\right)$ trajectory of every motion in the scene is estimated at all times, including when previously observed motions are occluded, and performance is demonstrated on ground-truth data from the Oxford Multimotion Dataset [3].

II. Multimotion Visual Odometry

The MVO pipeline extends VO to *multimodel* segmentation and estimation. As with traditional stereo VO pipelines, a set of tracklets is generated by matching salient image points

[1] K. M. Judd and J. D. Gammell are with the Estimation, Search, and Planning (ESP) research group at the Oxford Robotics Institute (ORI), University of Oxford, United Kingdom. {kjudd, gammell}@robots.ox.ac.uk



(a) (b) (c)

Fig. 2. Trajectory estimates produced by our occlusion-aware multimotion visual odometry (MVO) system before (a), during (b), and after (c) an occlusion in the `occlusion_2_unconstrained` segment of the Oxford Multimotion Dataset [3]. The trajectory estimate of the swinging block (4, magenta) is extrapolated using the constant-velocity motion model (dashed line) when the block is occluded by the moving tower (1, cyan) in (b). When the block becomes unoccluded in (c), it is rediscovered through motion closure and the estimates are interpolated to match the directly estimated trajectory (solid line).

across rectified stereo image pairs and temporally across consecutive stereo frames. The motion segmentation and estimation are then cast as a multilabeling problem where a label represents a motion hypothesis calculated from a group of tracklets. The labeling is found using CORAL [8], a convex optimization approach to the multilabeling problem.

The tracklets are embedded within a geometric graph that encodes spatial proximity over multiple frames. This graph forms the basis of both label assignment and generation. Labels are assigned based on the reprojection residual of the associated trajectory, as well as a local smoothness regularization. Disconnected subgraphs within label supports are used to estimate new trajectories through a multiframe RANSAC procedure. Redundant and oversegmented labels are later merged.

The algorithm iterates this process until the labeling converges. All motion hypotheses up to this point are treated as egocentric and potentially belonging to the static portions of the scene (i.e., the camera's egomotion). In a final step, a label is selected to represent the motion of the camera and a full-batch estimation of each trajectory is performed in a geocentric frame. The geocentric frame is chosen over an egocentric frame as it is more appropriate for the constant-velocity prior. This is because two frames, each moving with constant velocities relative to a static reference frame (e.g., the Earth), do not move with constant velocity relative to *each other*.

If a previously estimated motion is not found in the current frame, its estimate is extrapolated using the white-noise-on-acceleration motion prior described by Anderson et al. [9]. In practice, the prior penalizes the trajectory's deviation from a locally constant body-centric velocity. The prior is physically motivated, as objects tend to move smoothly through their environment.

We apply motion closure to determine if a recently discovered trajectory is similar to an extrapolated motion in both location and velocity. Trajectories belonging to the same motion on either side of an occlusion can be linked, and the occluded estimates can be corrected via interpolation. More detailed explanations of MVO are available in [7] and [10].

III. Discussion and Conclusion

This abstract introduces how the MVO pipeline can be extended to address the challenges posed by occlusions in highly dynamic environments. The pipeline uses a white-noise-on-acceleration motion model to extrapolate occluded trajectories until they are observed again. A motion-based similarity threshold incorporating both position and velocity can then be used to determine if a newly discovered motion belongs to the same occluded object.

The MVO pipeline performance was demonstrated on a challenging segment from the Oxford Multimotion Dataset [3] exhibiting significant occlusion and highly dynamic $SE(3)$ motions (Fig. 2). The system performs comparably to similarly defined visual odometry systems used solely for egomotion estimation while estimating all motions in the scene. This rigid-estimation approach can be applied to other problems, such as autonomous driving and human tracking. The motion of vehicles, cyclists, and pedestrians can generally be approximated as rigid or piecewise-rigid and they often interact in complex, dynamic environments.

Current and future work focuses on applying more accurate motion models, as well as extensions to other sensor modalities, such as RGB-D and event cameras.

References

[1] J. P. Costeira and T. Kanade, "A multibody factorization method for independently moving objects," *IJCV*, 29(3):159–179, 1998.

[2] K. E. Ozden, K. Schindler, and L. V. Gool, "Multibody structure-from-motion in practice," *PAMI*, 32(6):1134–1141, 2010.

[3] K. M. Judd and J. D. Gammell, "The Oxford multimotion dataset: Multiple SE(3) motions with ground truth," *RA-L*, 4(2):800–807, 2019.

[4] D. Reid, "An algorithm for tracking multiple targets," *TAC*, 24(6):843–854, 1979.

[5] G. Shu, A. Dehghan, O. Oreifej, E. Hand, and M. Shah, "Part-based multiple-person tracking with partial occlusion handling," in *CVPR*, pp. 1815–1821, 2012.

[6] D. Mitzel, E. Horbert, A. Ess, and B. Leibe, "Multi-person tracking with sparse detection and continuous segmentation," in *ECCV*, pp. 397–410, 2010.

[7] K. M. Judd, J. D. Gammell, and P. Newman, "Multimotion visual odometry (MVO): Simultaneous estimation of camera and third-party motions," in *ICRA*, pp. 3949–3956, 2018.

[8] P. Amayo, P. Piniés, L. M. Paz, and P. Newman, "Geometric Multi-Model Fitting with a Convex Relaxation Algorithm," in *CVPR*, pp. 8138–8146, 2018.

[9] S. Anderson and T. D. Barfoot, "Full STEAM ahead: Exactly sparse Gaussian process regression for batch continuous-time trajectory estimation on SE(3)," in *ICRA*, pp. 157–164, 2015.

[10] K. Judd and J. Gammell, "Occlusion-robust MVO: Multimotion estimation through occlusion via motion closure," 2019, arXiv: 1905.05121 [cs.RO].