Semi-Dense Visual Odometry for Monocular Navigation in Cluttered Environment

Shreyansh Daftry, Debadeepta Dey, Harsimrat Sandhawalia, Sam Zeng, J. Andrew Bagnell and Martial Hebert

***Abstract*— Recently, there have been numerous advances in the development of biologically inspired lightweight Micro Aerial Vehicles (MAVs). Due to payload and power constraints it is necessary for such systems to have autonomous navigation and flight capabilites in highly dense and cluttered environments using only passive sensors such as cameras. This is a challenging problem, given they have to operate in highly variable illumination conditions and be responsive to large environmental variations. In this paper we describe a scale-aware monocular vision based semi-dense direct depth perception system that enables robust autonomous navigation of small agile MAVs at low altitude through natural forest environments. We also show qualitative results in an outdoor dense forest area.**

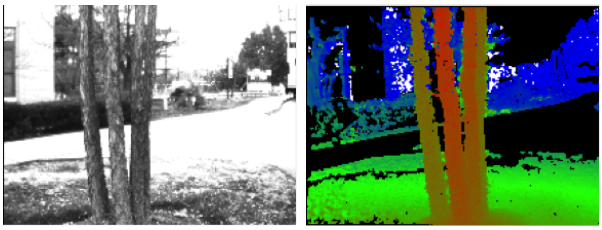
I. Introduction

In recent years we have seen small and agile Micro Aerial Vehicles (MAVs) make themselves useful in a number of search and rescue, surveillance and mapping applications. The most important benefit of using such lightweight MAVs is that it is allows the capability to fly at high speeds in cluttered and space-constrained environments. While autonomous operations and obstacle avoidance of MAVs has been well studied in general, most of these approaches uses laser range finders (lidar) [1] or Kinect cameras (RGB-D sensors) [2]. For agile MAVs with very limited payload and power, it is not feasible to carry such active sensors.

We develop and demonstrate a system that allow the MAVs to autonomously navigate through a cluttered natural environment using only passive monocular vision. In previous work we have demonstrated an imitation learning framework to learn a purely reactive controller [3] and non-linear regression based depth prediction method [4] to learn a more deliberative planning and control technique. However, these data driven approaches are either myopic in nature or dependent on pre-trained models that do not generalize well to previously unseen scenerios.

In this work, we propose the use of geometry-driven depth estimation. Such techniques are agnostic towards obstacle type, size and shape. The proposed approach does not rely on availability of training data and operate directly on raw pixel values and thus alleviate the bottle-necks described above while producing a high resolution depth map for the scene. Feature-based methods, are computationally expensive to be used onboard in real-time. In contrast, we utilize recent advancements in direct methods for visual odometery [5]. However, a critical drawback of such methods for use in obstacle avoidance applications is that the true scale of the environment cannot be recovered by monocular vision alone. We thus adapt them for autonomous navigation in cluttered environments.

All members are members or alumni of the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA. Email: daftry@cmu.edu, {debadeep, hsandhawalia, samlzeng, dbagnell, hebert}@ri.cmu.edu



(a) (b)

Fig. 1. (a) Input image frame and (b) the estimated semi-dense depth map in an outdoor setting. Note: Red is near and Blue is far.

II. Approach

A. Hardware and Software Overview

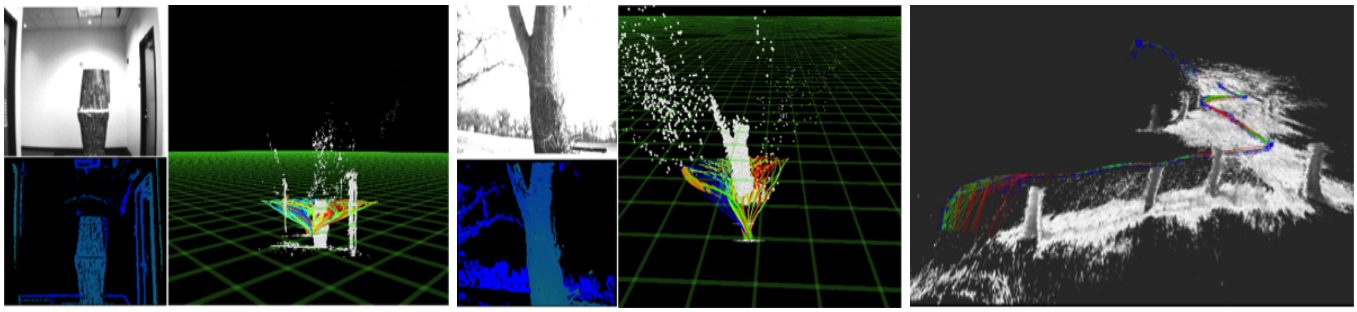
We have use a modified version of the 3DR ArduCopter with an onboard quad-core ARM processor and a Microstrain 3DM-GX3-25 IMU. Onboard, there are two monocular cameras: one PlayStation Eye camera facing downward for real time pose estimation and one high-dynamic range Point-Grey Chameleon camera for monocular depth estimation. An image stream from the front facing camera is streamed to the base station where the perception module is used to estimate depth maps; the planning module then finds the best trajectory to follow, to minimize probability of collision and transmits it to the onboard computer where the trajectory following module runs a pure pursuit controller to do trajectory tracking.

B. Semi-Dense Depth Map Estimation

In this section we present a summary of the semi-dense visual odometry approach following the method of Engel *et al.* [5]. Given a semi-dense inverse depth map for the current image, the camera pose of the new frames is estimated using direct image alignment: given the current map $\{I_M, D_M, V_M\}$, the relative pose $\xi \in SE(3)$ of a new frame I is obtained by directly minimizing the photometric error

$$E(\xi) := \sum_{x \in \Omega_{D_M}} \|I_M(x) - I(w(x, D_m(x), \xi))\|_\delta$$

where $w : \Omega_{D_M} \times \mathbf{R} \times SE(3) \rightarrow \omega$ projects a point from the reference frame image into the new frame. D and V denote mean and variance of the inverse depth, and $\|\cdot\|_\delta$ is the



(a) (b) (c)

Fig. 2. The estimated depth map is projected into the 3D space and the planner selects the optimal trajectory (show in orange) to avoid the fixed target obstacle in (a) Indoor setting with fake trees and (b) Outdoor cluttered setting. (c) The resulting 3D map generated of the outdoor test-site.

Huber norm to account for outliers. The minimum is computed using iteratively re-weighted Levenberg-Marquardt minimization [6].

The depth measurements are obtained by a recently proposed probabilistic approach for adaptive-baseline stereo [5]. This method explicitly takes into account the knowledge that in video, small baseline frames occurs before large baseline frames. Accordingly, a subset of pixels is selected for which the disparity is sufficiently large and for each selected pixel a suitable reference frame is selected.

The inverse depth map is propagated from frame to frame, once the pose of the following frame has been determined and refined with new stereo depth measurements. Based on the inverse depth estimate d_0 for the pixel, the correspoding 3D point is calculated and projected into the new frame and assigned to the closest integer pixel position providing the new inverse depth estimate d_1. We assume the camera rotation to be small, thus the new inverse depth map can be approximated by

$$d_1(d_0) = (d_0^{-1} - t_z)^{-1}$$

where t_z is the camera translation along the optical axis. Now, for each frame, after the depth map has been updated, a regularization step is performed by assigning each inverse depth value the average of the surrounding inverse depths, weighted by their respective inverse variance (σ^2). An example of the obtained depth estimates has been shown in Figure 1 Note: In order to prevent sharp edges, which can be critical in detecting trees, we only perform this step if two adjacent depth values are statistically similar i.e. their variances are within 2σ.

C. Scale Estimation

Scale ambiguity is inherent to all monocular visual odometry based methods. This is not critical in visual mapping tasks, where the extrenal scale can be obtained using fiducial markers [7] as a post processing step. However, for osbtacle avoidance in real-time, it is required to recover the current scale robustly so that the distance to the object is known precisely. We resolve the absolute scale $\lambda \in \mathbf{R}^+$ by leveraging motion estimation from a highly accurate metric sensor onboard. We measure, at regular intervals (operating at 15 hz), the 3-dimensional distance travelled accoriding to the visual odometry $\mathbf{x}_i \in \mathbf{R}^3$ and the metric sensors $\mathbf{y}_i \in \mathbf{R}^3$. Given such sample pairs $(\mathbf{x}_i, \mathbf{y}_i)$, we obtain a scale $\lambda(t_i) \in \mathbf{R}$ as the running arithmetic average of the qoutients $\frac{\|\mathbf{x}_i\|}{\|\mathbf{y}_i\|}$ over a small window size. We further pass the obtained set of scale measurements through a low-pass filter in order to avoid erroneous measurements due to sensor noise and thus converge to a true scale λ.

D. Planning and Control

Once the planner module receives a scaled-depth map from the perception module, the local point cloud is updated using the current pose of the MAV. A trajectory library of 78 trajectories of length 5 meters is budgeted and picked from a much larger library of 2401 trajectories using the maximum dispersion algorithm by Green et al. [8]. Further, for each of the budgeted trajectories a score value for every point in the point cloud, is calculated by the method proposed by Dey *et al.* [4] and the optimal trajectory to follow is selected. The control module takes in the coordinates of the trajectory to follow and the current pose of the vehicle from the optical flow-based pose estimation system to successfully track it using a pure pursuit strategy [9].

III. Discussion

In this section, we show the qualitative assesment of our approach. First, an indoor setup with artifially created trees (Fig. 2(a)) was created as a test-bed to benchmark accuracy of depth estimation with respect to a laser rangefinder as ground truth. Once the sanity of depth estimates was verified in an indoor setting, we performed experiments with the whole system in a densely cluttered forest area, while restraining the MAV with a light-weight tether. The snapshot of an output instance as well as the 3D map generated has been shown in Figure 2(b-c).

References

[1] A. Bachrach, R. He, and N. Roy, "Autonomous flight in unknown indoor environments," *International Journal of Micro Air Vehicles*, vol. 1, no. 4, pp. 217–228, 2009.

[2] A. Bachrach, S. Prentice, R. He, P. Henry, A. S. Huang, M. Krainin, D. Maturana, D. Fox, and N. Roy, "Estimation, planning, and mapping for autonomous flight using an rgb-d camera in gps-denied environments," *The International Journal of Robotics Research*, vol. 31, no. 11, pp. 1320–1343, 2012.

[3] S. Ross, N. Melik-Barkhudarov, K. S. Shankar, A. Wendel, D. Dey, J. A. Bagnell, and M. Hebert, "Learning monocular reactive uav control in cluttered natural environments," in *IEEE International Conference on Robotics and Vision (ICRA)*, 2013.

[4] D. Dey, K. S. Shankar, S. Zeng, R. Mehta, M. T. Agcayazi, C. Eriksen, S. Daftry, M. Hebert, and J. A. Bagnell, "Vision and learning for deliberative monocular cluttered flight," in *In Proceedings of the International Conference on Field and Service Robotics (FSR)*, 2015.

[5] J. Engel, J. Sturm, and D. Cremers, "Semi-dense visual odometry for a monocular camera," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1449–1456.

[6] J. Engel, T. Schöps, and D. Cremers, "Lsd-slam: Large-scale direct monocular slam," in *Computer Vision–ECCV 2014*. Springer, 2014, pp. 834–849.

[7] S. Daftry, C. Hoppe, and H. Bischof, "Building with drones: Accurate 3d facade reconstruction using mavs," in *IEEE International Conference on Robotics and Vision (ICRA)*, 2015.

[8] C. Green and A. Kelly, "Optimal sampling in the space of paths: Preliminary results," Robotics Institute, Pittsburgh, PA, Tech. Rep. CMU-RI-TR-06-51, November 2006.

[9] R. C. Coulter, "Implementation of the pure pursuit path tracking algorithm," DTIC Document, Tech. Rep., 1992.