Image-based Visual Perception and Representation for Collision Avoidance

Cevahir Cigla
Jet Propulsion Laboratory / California Institute of Technology
cevahircigla@gmail.com

Roland Brockers
Jet Propulsion Laboratory / California Institute of Technology
brockers@jpl.nasa.gov

Larry Matthies
Jet Propulsion Laboratory / California Institute of Technology
lhm@jpl.nasa.gov

Abstract

We present a novel on-board perception system for collision avoidance by micro air vehicles (MAV). An egocentric cylindrical representation is utilized to model the world using forward-looking stereo vision. This efficient representation enables a 360° field of regard, as the vehicle moves around and disparity maps are fused temporally on the cylindrical map. For this purpose, we developed a new Gaussian Mixture Models-based disparity image fusion algorithm, with an extension to handle independently moving objects (IMO). The extension improves scene models in case of moving objects, where standard temporal fusion approaches cannot detect movers and introduce errors in world models due to the common static scene assumption. The on-board implementation of the vision pipeline provides disparity maps on a 360° egocentric cylindrical surface at 10 Hz. The perception output is used in our system by real-time motion planning with collision avoidance on the MAV.

1. Introduction

On-board obstacle detection and avoidance is essential for autonomous vehicle navigation. This is particularly challenging for small micro aerial vehicles (MAVs) that have limited payload and power budget. Vision-based approaches using compact cameras are good alternatives in this context.

There are several fundamental requirements for vision systems for obstacle avoidance. The extracted model should be sufficiently dense and accurate with a wide depth range to handle near and far objects. The model should be stable and consistent as the vehicles moves around and IMOs should be detected. Finally, a sufficiently high frame rate is required to enable real time control of the vehicle. Stereo matching is a common technique that addresses these constraints by providing dense depth maps of a scene using passive stereo cameras.

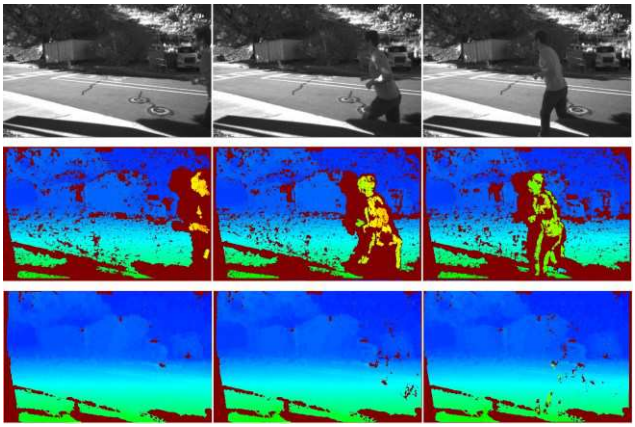

Figure 1: Top row: three frames from left camera of a stereo pair. Middle row: corresponding disparity maps via stereo matching. Last row: temporal fusion with rigid-static scene assumption.

It works well both indoors and outdoors, which is an advantage over comparably small active depth sensors. The depth range is adjustable via the baseline of the stereo cameras and the resolution of the images; fast, compact implementations of stereo matching are now available and are progressing rapidly.

Stereo matching algorithms with low computational complexity provide depth maps for each frame individually. Therefore, obstacle detection errors are inevitable, due to environmental factors and stereo matching errors. The obstacle avoidance system is prone to errors if these techniques are applied without temporal fusion of the dense depth maps. In the robotics literature, occupancy grid and voxel data structures have been standard approaches [1]-[3] for temporal fusion in 3D space. These techniques are specifically designed for generating accurate maps of the environment that could be much more complex than is necessary for obstacle avoidance. Image space representations can be an efficient alternative, as proposed in [4]-[7]. Temporal fusion in image space has potential to reduce stereo depth map errors and extend the depth range at lower computational cost [8]-[10], particularly for reactive navigation in cluttered environments.

The literature on temporal fusion of depth data usually assumes rigid and static scenes. IMOs have the potential to be invisible in fused 3D representations under these assumptions (Figure 1), and should be handled carefully for a complete and reliable collision avoidance framework. IMO handling has been given much less attention, limited mostly to feature and optical flow based approaches [11]-[13]. The sparse representations of feature-based techniques are not adequate for collision avoidance, while optical flow techniques are computationally heavy for on-board processing.

In this paper, we modify an efficient depth data fusion technique [10] for the on-board vision system of a MAV that enables live obstacle avoidance. In this set-up, forward-looking stereo cameras are used to sense the environment, while vehicle poses are estimated by visual-inertial odometry (VIO) using an IMU and images from a downward-looking camera. The scene is represented via an egocentric cylinder [7] that provides a 360^o representation of the environment with constant angular resolution. The fusion algorithm is also extended to handle moving objects with an incremental increase in computational complexity. The output of the visual perception system can be used by motion planning approaches providing online collision avoidance for MAVs in cluttered environments. To our knowledge, this is the first on-board implementation of temporally fused egocylinder representation with IMO handling.

The remainder of the paper is organized as follows. The next section summarizes prior work related to temporal fusion of disparity maps from stereo cameras. Section 3 presents the details of the visual system including proposed IMO handling on egocylinder surface, which is followed by experimental results in Section 4. Finally, we discuss conclusions in Section 5.

2. Related Work

Temporal fusion is a common way to relate frame-wise extracted depth maps in various representations of the environment. Temporal depth map consistency can be achieved by incorporating consecutive frames in a cost function with additional constraints on the estimated depth maps [14]. This also can be achieved in a multi-view framework as in [15], with complex optimizations to merge consecutive depth maps. In [16][17], cost functions are aggregated temporally as an extension to spatial aggregation in order to extract reliable depth maps for each frame. SLAM techniques [18] also provide consistency through online depth updates in the key frames, where a simple Gaussian model is utilized to model depth measurements and the depth search is limited within the standard deviation of the prior hypothesis. Recently, [19] extended SLAM approaches with introduction of stereo cameras in order to adjust scale parameter in mapping and increase the number of reliable points. However, depth maps from SLAM frameworks are still inadequate for obstacle avoidance due to the sparse representations.

Another group of methods is based on multi-view filtering techniques to improve depth map quality by removing outliers and filling holes. A visibility-based fusion method in [20] requires multiple 3D warpings as well as depth ordering that may not be applicable for on board processing. In [21], a median filter is used along consecutive frames to filter out outliers and provide smooth depth maps. [9] uses projection uncertainties in the reference view to estimate probability density functions of depth hypotheses. Recently, [10] has extended Gaussian Mixture Models for temporal depth fusion, updating depth models online with new depth observations. This decreases the memory requirement and computational complexity as well as yields more accurate results compared to recent filtering based techniques.

The most common way to merge multiple depth map observations uses 3D models such as voxels or surfaces [3][8][23][24]. The depth data is mapped to 3D coordinates to form volumetric representations of the environment that are widely utilized for generating accurate maps. Grid maps require a lot of memory and computation since the main motivation is the generation of a complete map. On the other hand, less complex and more efficient representations are available for collision avoidance based on image space representations. Recently, [5][7] proposed an efficient 2.5D image space world representation that enables fast collision checking in image space, using an egocylindrical data structure to provide 360^o representation of the environment with constant angular resolution. This approach has good potential for fast motion planning.

Research on temporal fusion has mostly focused on rigid or static scene assumptions, where moving objects are neglected. Intruders in an environment have potential to corrupt the 3D representations by violating the static scene assumption. On the other hand, missing IMOs in the scene representation may cause failures especially for reactive collision avoidance. [25] exploits two geometric constraints to detect IMOs for moving surveillance cameras based on structure consistency and plane-parallax filtering. Sparse [11][12] and dense [13] optical flows are utilized to detect objects that do not follow the scene flow. Sparse flow is insufficient, especially for close objects that have significant importance for collision avoidance. Dense flow is computationally expensive for onboard processors currently available for MAVs.

In this paper, we borrow the idea of 2.5D image space

Figure 2: Asctec Pelican quad-copter equipped with a 1.86 GHz Intel Core2Duo processor (*Asctec Mastermind*), *Odroid* XU4 flight computer, forward–looking stereo cameras and a downward looking camera for odometry.

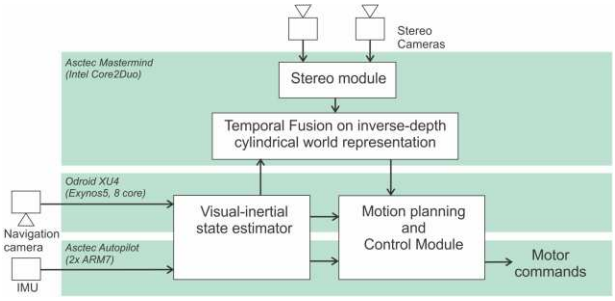

Figure 3: System architecture on-board the *Asctec Pelican*.

world representation on an egocylinder surface [5] for MAV motion planning and collision checking. We use forward-looking stereo cameras to sense the world, as well as an IMU and visual odometry [26] with downward looking imagery to estimate the pose of the MAV. Gaussian mixture models [10] are utilized to fuse frame-wise extracted disparity maps on the egocylinder representation, which is an efficient way to perform fusion in image space. In addition, we propose an IMO handling step that yields much more reliable fusion for obstacle avoidance. All vision algorithms are implemented on-board an Asctec Pelican (Figure 2) that uses a hierarchical processing architecture (Figure 3).

3. Vision System for Collision Avoidance

Gaussian Mixture Models are widely used to represent background/foreground intensity distributions for detecting moving objects in surveillance [27]. Successive intensity observations of each pixel are combined in a compact representation. Gaussian models have been used for SLAM [18] and extended to Gaussian mixtures for temporal fusion of disparity maps [10]. In both approaches, models are formed in disparity space with inverse depth representation, so that uncertainty is represented for inverse range. [18] uses Gaussian models to narrow the search range during the estimation of disparity maps for the following frame, while [10] approaches fusion as a filtering step by relating frame-wise estimated disparity maps. In that manner, the framework proposed in [10] fits our set-up with stereo matching for depth sensing. Moreover, having the background disparity models of a scene is convenient for detection of IMOs that violate the rigid-static scene assumption.

3.1. GMM-based Depth Fusion

The depth fusion is achieved by extending intensity based models by pixel position and disparity values ($\vec{x} = (u, v, d)$) in image space. Each pixel is represented as a mixture of *K* Gaussian distributions as follows:

$$\mathrm{P}(\vec{\mathrm{x}}_\mathrm{t}|\mathrm{X}_\mathrm{T}) = \sum_{\mathrm{m}=1}^{\mathrm{K}} \mathrm{W}(\mathrm{O}_\mathrm{m}, \sigma_\mathrm{m})\mathrm{N}(\vec{\mathrm{x}}_\mathrm{t}; \vec{\mu}_\mathrm{m}, \sigma_\mathrm{m}) \quad (2)$$

where $\vec{\mu}$'s are the mean and $\vec{\sigma}$'s are the variance estimates of $\vec{x}$ and S is the set of observations along *T* frames. O_m is the number of frames that corresponding mode m is observed and W is a weighting function that defines the strength of the corresponding mode. This model can directly be utilized per pixel as long as the platform does not move. For moving platforms, forward warps are required to relate corresponding pixels w.r.t. platform motion.

There are three steps for GMM based fusion: first; as the new disparities are observed for pixels with no GMMs or if there is a significant disparity difference between the models and the observation, a new mode is created by

$$N(\vec{x}; \vec{\mu}_0, \sigma_0): \begin{cases} \vec{\mu}_0 = (u, v, d) \\ \sigma_0 = \sigma_{init} \\ O_0 = 1 \end{cases} . \quad (3)$$

In (3), the triplet corresponds to the new disparity measurement and its pixel coordinates, and σ_{init} is set to a high value. The second step involves the mapping of GMMs in the previous frame to the current frame. The mapping is performed according to pose changes, for each mode. This generates groups of GMMs per pixels from different sources in the recent frame. The models from the neighboring pixels are also utilized within a specified window (i.e., 3x3) to handle holes in forward mapping.

There can be multiple hypotheses for a pixel after the mapping, as illustrated in Figure 4. The center pixel (red) gets contributions from the neighbor pixels within a window, where each circle in the disparity plot corresponds to a Gaussian distribution. In order to specify the new GMMs just before the recent disparity observation, a merge and reduction step is required. At that point, as stated in [10], there may be various alternatives to merge Gaussian distributions; however, this step is applied for each pixel and should be as simple as possible for fast operation. In that manner, the modes are grouped according to similarity of mean disparity values. Then, an averaging is performed for all parameters except standard deviation of disparity, on which the minimum σ among the grouped models is chosen. The final number of models is fixed to a predefined threshold by neglecting the modes with high standard deviations.

The final step is the model update and disparity assignment. The most recent disparity observations are compared to the GMM hypotheses per pixel individually. There is considered to be a match as long as the minimum disparity distance between the current observation and GMMs is below a threshold, T_d, (such as 3 pixels). In case of a match, the corresponding mode (M) is updated as follows:

$$\begin{aligned} \sigma_M^2 &= \alpha\sigma_M^2 + (1-\alpha)|d - \vec{\mu}_M(d)|^2 \\ \vec{\mu}_M &= \alpha\vec{\mu}_M + (1-\alpha)\vec{x} \\ O_M &= O_M + 1 \end{aligned} \quad (4)$$

where d is the newest disparity observation, $\vec{x}$ is the recent triplet representation, and α corresponds to an update rate that determines the adaptation speed of the fusion to new observations. The same equations in (4) are utilized to update the unmatched modes, where the standard deviation is updated and the number of occurrences is decremented instead. If there is no match, a new mode is generated as given in (3). Depending on the scene geometry and motion of the vehicle, some pixels may not have an observation. In that case, each mode is penalized by incrementing standard deviation and decrementing occurrence count by a forgetting factor (α_{forget}).

Each pixel is assigned a disparity value as long as the matched mode or the mode with best standard deviation (in case of no observation) satisfies the validity condition given as:

$$valid: \begin{cases} 1 & \sigma_M^2 < p\sigma_{init}^2 \; or \; O_M > T_C \\ 0 & else \end{cases} \quad (5)$$

where p is the scale factor that shows the reliability of the current mode, T_C is the occurrence threshold. Finally, a resulting disparity map is obtained by robust GMM models that have been observed for sufficient number of frames with low variation.

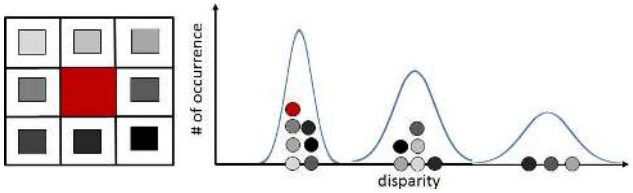

Figure 4: The center pixel (red) gets contributions within a neighborhood which forms a large number of mixture of Gaussians. Multiple models are merged for compact representation.

3.2. Independent Moving Object Handling

GMM-based temporal fusion, as with most fusion approaches, uses a rigid and static scene assumption by neglecting IMOs. As discussed previously, in dynamic scenes IMO detection is a crucial step for reliable collision avoidance. Therefore, we now extend the depth fusion framework to handle moving objects as well. Dense optical flow is not practical with current onboard computational limitations of MAVs. Instead, using GMMs enables efficiently detecting IMOs with methods similar to foreground object detection in surveillance videos. The main assumption in surveillance applications is the existence of noticeable intensity differences from the background models. Exactly the same idea can be modified by introducing disparity change with respect to background scene structure for IMO detection in temporal fusion.

We extend the parameterization of GMM-based fusion discussed in 3.1 with the addition of an intensity model (I) of the pixels. In the new model, each mode is represented by quadruple $\vec{x} = (u, v, d, I)$. Candidate moving pixels are detected in two steps. First, pixels that do not match to a background mode and have disparity values significantly larger than the background are considered as candidate moving objects. This group is classified into strong and weak candidates. The strong candidates have larger intensity differences, while the weak candidates have intensity values barely differentiated from the background. This type of classification approach helps to grow IMO regions (obtained by strong candidates) at the final step that yield more complete object detection.

Connected component analysis is performed on the strong candidates to eliminate false alarms such as small regions. At that point, a moving object is expected to have sufficiently large area (T_{area}) that it cannot be ignored for collision avoidance. Then, those regions are grown within bounding boxes and through weak pixels as long as they have connected paths in between. In that way, objects with visible disparity divergence are detected completely even though they have intensity variation within. After the detection of candidate moving pixels, disparity values for

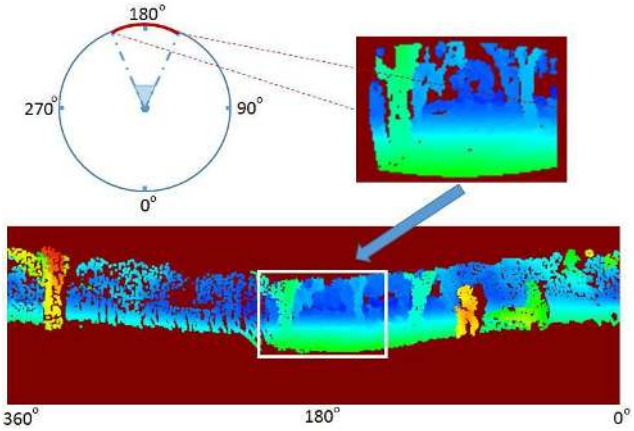

Figure 5: An egocentric cylinder is an efficient way to model the world with a 360° of Field-of-Regard (FOR). At any time, the FOV of the stereo vision system corresponds to a region centered at 180° of the egocylinder image.

these pixels are directly copied from the recent disparity observation without any update in GMMs. This does not force consistency along the moving objects, which would require object motion models. Instead, this avoids invisibility of moving objects and prevents incorrect background mode generation.

3.3. Egocylinder Representation

An egocentric cylinder surface image [7] is an efficient way to maintain a persistent 360° representation of the world. As illustrated in Figure 5 with the top view of egocylinder, the disparity map at a time instant covers a part (white rectangle) of the egocylinder corresponding to the FOV of the stereo cameras. As the vehicle moves around and covers different view angles, temporal fusion accumulates depth data to form a complete scene representation. At each time, new observations are located at the center (forward direction) of the egocylinder image.

With this representation, the update step of temporal fusion is performed on the FOV of stereo cameras (within the white square) where the most recent disparity map is observed. The remaining areas are subject to fade out (or not updated) with a speed related to the forgetting factor. The parameterization of the fusion approach enables defining the fade out rate based to the motion of the vehicle. Currently, we use a constant forgetting factor for the sake of simplicity. This representation is used by a motion planner (similar to [7]) for collision avoidance.

4. Experimental Results

Two sets of experiments were performed to test the performance of the proposed vision system as a basis for collision avoidance. The first set measured the detection performance of the IMO handling algorithms and compared resulting disparity maps to the original GMM-based temporal fusion algorithm. This used the well-known KITTI stereo benchmark [30], which specifically includes temporal stereo datasets with moving objects. The second set of experiments analyzed on onboard performance with real data captured by our MAV.

4.1. Offline Performance Evaluation

The KITTI 2015 stereo dataset provides an excellent benchmark to test the IMO handling algorithm. The dataset includes 194 different scenes with 20 consecutive stereo frames captured from a car. The ground truth disparity maps of center frames are also provided for each sequence. The center frame of each sequence also includes labeled moving objects to evaluate the performance of detection. The average distribution of static and moving regions in this dataset is 85 and 15%, respectively.

With this data, we use the Semi Global Matching algorithm [28] to extract disparity maps from stereo images for each frame independently. The vehicle poses are estimated through stereo visual odometry [29]. The parameter set for the fusion algorithm is given as follows:

Table 1: The parameter values throughout the experiments

α	α_{forget}	σ^2_{init}	p	T_d	T_C	T_{area}
0.1	0.05	20	0.1	3	5	20

IMO detection performance is measured through the object labels provided by the KITTI benchmark. The distribution of the distance (meters) of all moving objects is given in Figure 6, where the missed objects are also shown in orange color. The remaining blue color corresponds to the successfully detected objects. Detection performance improves as objects get closer to the observer. Our approach detects all of the IMOs that are closer than 9 meters, which are important for collision avoidance. The distributions of the spatial location of detected and missed moving objects in image space are also illustrated in Figure 6. Missed vehicles are generally located at the center of the image and have mostly the same moving direction with the observer. Therefore, these objects are stored and modeled as background in GMMs due to repeated observations. On the other hand, detected objects move along nearby lanes, most of which are located on the left of the observer and move in opposite direction with high probability of collision.

The average distance of detected objects is 12 meters, while missed objects are at an average distance of 25 meters and average disparity error on these objects is 1.8 pixels. Thus, missed vehicles are located at greater distances with small disparity errors in the fused maps.

Table 2: The performances of stereo matching and temporal fusion with and without IMO handling are given based on two different error statistics for static and moving regions.

Static/Moving (85/15) %	Out-3%	Avg Disp. error
SGM [28]	12.6 / 22.7	2.6 / 3.3
GMM [10]	8.4 / 61.4	1.9 / 12.1
IMO Handle	8.6 / 37.8	1.9 / 5.3

In In terms of collision avoidance, IMO detection can sense nearby moving objects that are collision risks (the left region of the histogram), while it misses distant objects with low probability of collision. This is a natural consequence of testing disparity differences: as object distance increases, the frame-to-frame disparity difference decreases.

The IMO handling step has an influence on the accuracy of fused disparity maps as well. The precision of the disparity maps is calculated based on two measures: the percentage of pixels with $\Delta d>3$ (Out-3%) compared to the ground truth disparity maps and the average disparity error. The results for the stereo matching algorithm alone [28], GMM-based fusion [10], and the IMO handling extension are given for static and moving regions in Table 2. GMM-based temporal fusion decreases the error ratio by almost 30% compared to frame independent stereo matching. As expected, the IMO handling approach has an insignificant effect in the static regions.

On the other hand, temporal fusion fails for the moving regions that violate the rigid-static scene assumptions. The average disparity error is almost 4 times larger than the initial disparity maps, indicating that the background disparity modes are assigned for those regions. The proposed IMO handling step significantly decreases the error rates of standard temporal fusion while it is still worse compared to frame-wise stereo matching for moving regions. Overall (weighted with average distributions (85-15%)), the proposed approach has the best error rates, providing a trade-off by improving disparity maps along static regions without large failures on the moving pixels.

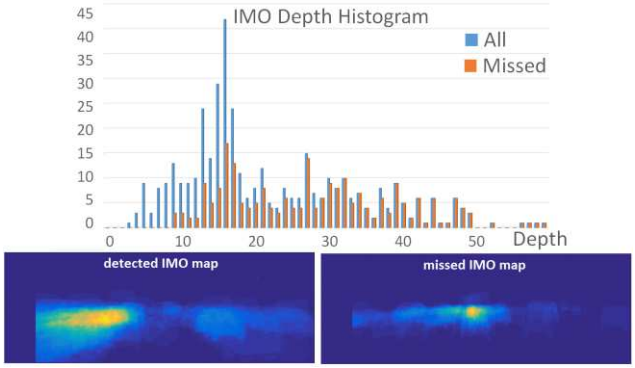

Figure 6: Top: depth distribution of moving objects in the data set (blue), missed objects (orange). Bottom: the distribution of objects masks in image space for detected and missed cases.

Table 3: Onboard computation times for each step in visual perception system.

Perception Step	Time (msec)
Stereo Matching	100
Cylindrical Mapping	14.4
GMM Forward Mapping	38.5
GMM Selection	10.6
GMM Update	3.5
IMO Handling	2.6

The output disparity maps are illustrated in Figure 7 for visual interpretation. The initial disparity maps are shown in the second row and temporal fusion results are given in the third row. The red regions in the last row belong to the IMO detection mask of the proposed algorithm. These regions are compensated by the disparity values given in the second row. The improvement is clear for the static regions, which is the result of accumulating temporal data. On the other hand, as long as the disparity and intensity differences are significant, proposed approach can detect the IMOs with sufficient object coverage.

4.2. Onboard Experiments

The *Asctec Pelican* implementation platform (Figure 2) is equipped with a 1.86 GHz Intel Core2Duo processor running the stereo vision, egocylinder, and temporal fusion modules and an Odroid XU4 processor for VIO. The forward-looking stereo cameras (752x480) are installed with a baseline of 25 cm and frame-wise stereo disparity maps are calculated by block matching over search range of 100 pixels. Temporal fusion is performed on an egocylinder image with resolution of 660x200.

The computation time of stereo matching and the steps of temporal fusion are given in Table 3. The full perception pipeline maintains a 10 Hz update rate using both cores of the Core2Duo, which enables real-time motion planning on the MAV.

Typical results of temporal fusion on the egocylinder are illustrated in Figure 8. The left stereo image, unfused disparity map, and the corresponding egocylinder images are shown for five different time instants in the scenario of moving towards an obstacle. Temporal fusion increases the density of the initial disparity maps. The world representation propagates around the egocylinder as the vehicle moves around, with new frames of stereo data being fused in the forward direction. The consistency of the model can be observed by following the same tree, as shown by the black ellipse, through the successive time instants even though it is out of sight at some point. Moreover, temporal fusion retains memory of close

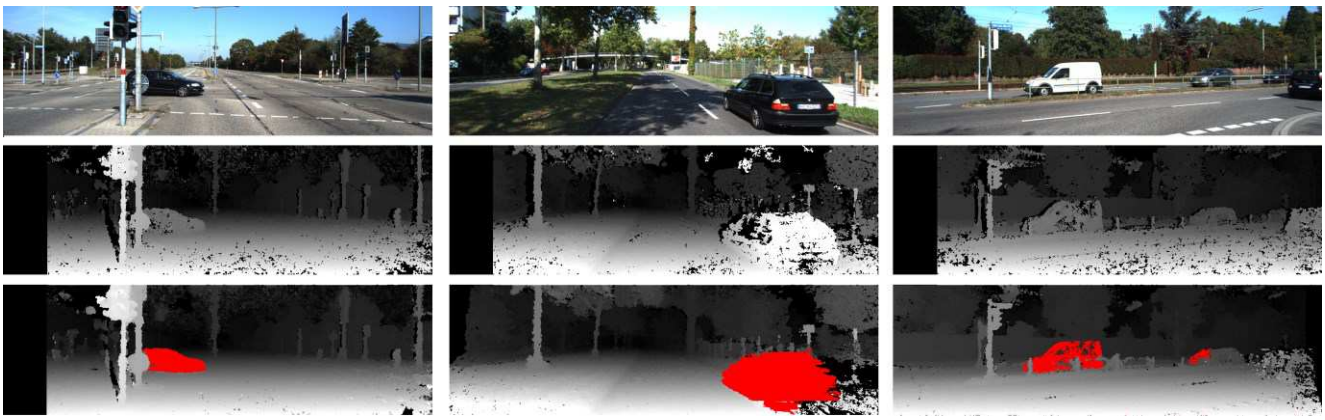

Figure 7: Top row: left color images. Middle row: corresponding unfused disparity maps (lighter pixels are closer to the camera). Last row: temporal fusion with IMO detection mask (red).

objects in the egocylinder after they disappear from the raw disparity maps because they are too close for the maximum disparity search range. The trees pass out of the FOV of the disparity maps as the vehicle approaches, while they are retained in the egocylinder representation. These characteristics benefit collision avoidance by increasing the representation range in both depth and field of regard. In both cases shown in Figure 8, collisions can be avoided by the temporally fused egocylinder representation, while it would be failure if only the frame-wise stereo disparity maps were exploited.

The proposed IMO handling approach is tested by introduction of movers in the static scenes. IMO detection is performed in the egocylinder region corresponding to stereo camera FOV with the most recent disparity observation (white rectangle in Figure 5). The egocylinder representations with and without IMO handling are illustrated in Figure 9 for two different scenes. The first row is the left image, and successive rows are for disparity maps, fusion with static assumption, the proposed IMO handling extension, and IMO detection masks. We crop the egocylinder representations for better visualization, where the corresponding angles are (120^o-240^o) and (100^o-260^o) for two different scenes consecutively. As is clearly observed, IMOs disappear under the static scene assumption; on the other hand, the proposed IMO approach detects those objects completely, improving the obstacle avoidance capability. It is also important to note that IMO handling not only detects the moving objects but also preserves the fine structure of the background model. Especially under small motion of the IMOs, due to continuous observation of the same disparity levels, these values are observed in the background model when IMO handling is not active. In the second and third time instants of scene 2, incorrect disparity assignments are observed on the left side of the tree (the third row), which are the results of fusion of repetitive regions to the background. These regions correspond to false alarms that are not desired for collision avoidance. On the other hand, this effect is removed by IMO handling and a more reliable model of the environment is provided.

5. Conclusion and Future Work

In this paper, we propose an efficient visual perception system implemented onboard for MAV collision avoidance. Forward-looking stereo cameras are used to sense the world via disparity maps that are fused temporally using an egocentric cylindrical representation yielding a 360^o scene model. We extend image based temporal depth fusion to handle independently moving objects to provide reliable perception for cluttered and dynamic environments. The proposed IMO handling step detects moving objects and improves the fused disparity maps. The onboard implementation on an *Asctec Pelican* MAV provides 10 Hz visual maps on the egocylinder that are used in live motion control for collision avoidance.

As future work, we plan to compare the proposed representation with voxel based 3D representations through the collision avoidance framework that includes motion planning for obstacle avoidance.

Acknowledgment

This work was funded by the Army Research Laboratory under the Micro Autonomous Systems & Technology Collaborative Technology Alliance program (MAST) and was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

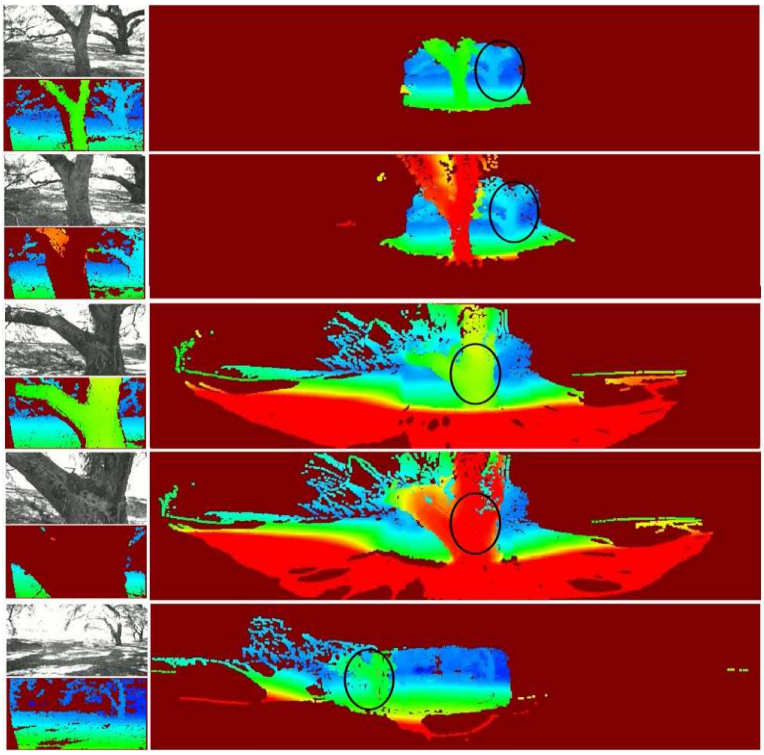

Figure 8: First column: left stereo image and disparity maps for five time instants. Second column: corresponding egocylinder representations (blue to red: far to close). The same tree is marked by the ellipse throughout the five frames.

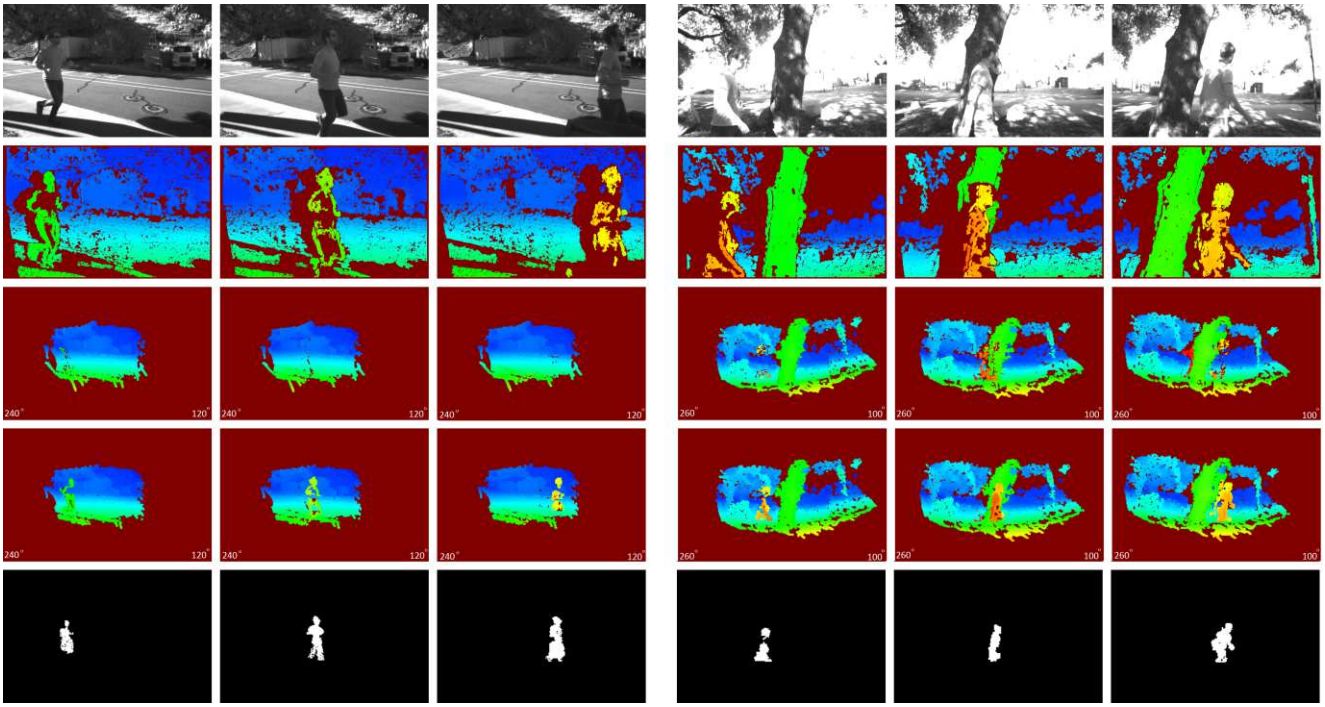

Figure 9: First row: left stereo image. Second row: unfused disparity map. Third row: cropped egocylinder image with static scene assumption (120^o-240^o for the first scene and 100^o-260^o for the second scene). Fourth row: cropped egocylinder image after IMO handling, last row: detected IMO masks.

References

[1] A. Hornung, K.M. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard, OctoMap: An Efficient Probabilistic 3D Mapping Framework Based on Octrees. *Autonomous Robots*, 2013

[2] D. Cole and P. Newman, Using Laser Range Data for 3D SLAM in Outdoor Environments. *IEEE International Conference on Robotics and Automation*, 2006

[3] I. Dryanovski, W. Morris and J. Xiao, Multi-volume Occupancy Grids: An Efficient Probabilistic 3D Mapping Model for Micro Aerial Vehicle. *International Conference on Intelligent Robotics and Systems*, 2010

[4] M. W. Otte, S. Richardson, J. Mulligan and G. Grudic, Path Planning in Image Space for Autonomous Robot Navigation in Unstructured Outdoor Environments, *Journal of Field Robotics*, 2009

[5] L. Matthies, R. Brockers, Y. Kuwata and S. Weiss, Stereo vision-based Obstacle Avoidance for Micro Air Vehicles using Disparity Space, *IEEE International Conference on Robotics and Automation*, 2014

[6] H. Oleynikova, D. Honegger and M. Pollefeys, Reactive Avoidance Using Embedded Stereo Vision for MAV Flight, *IEEE International conference on Robotics and Automation*, 2015

[7] R. Brockers, A. Fragoso, B. Rothrock, C. Lee and L. Matthies, Vision-based Obstacle Avoidance for Micro Air Vehicles using an Egocylindrical Depth Map, *International Symposium on Experimental Robotics*, 2016

[8] C. Hane, C. Zach, J. Lim, A. Ranganathan and M. Pollefeys, Stereo Depth Map Fusion for Robot Navigation. *International Conference on Intelligent Robots and Systems*, 2011.

[9] C. Unger, E. Wahl, P. Strum and S. Ilic, Probabilistic Disparity Fusion for Real-time Motion Stereo. *Machine Vision and Applications*, Vol 25, 2011.

[10] C. Cigla, R. Brockers and L. Matthies, Gaussian Mixture Models for Temporal Depth Fusion, *IEEE Winter Conference on Applications of Computer Vision*, 2017

[11] P. Lenz, J. Ziegler, A. Geiger and m. Roser, Sparse Scene Flow Segmentation for Moving Object Detection in Urban Environments, *IEEE Intelligent Vehicles Symposium*, 2011

[12] D. Zhou, V. Fremont, B. Quost and B. Wang, On Modeling Ego-Motion Uncertainty for Moving Object Detection from a Mobile Platform, *IEEE Intelligent Vehicles Symposium*, 2014

[13] A. Talukder and L. Matthies, Real-time Detection of Moving Objects from Moving Vehicles using Dense Stereo and Optical Flow, *IEEE International Conference on Intelligent Robots and Systems*, 2004

[14] G. Zhang, J. Jia, T. T. Wong and H. Bao, Consistent Depth Maps Recovery from a Video Sequence. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 31(6), June 2009

[15] M. Pizzoli, C. Forster and D. Scaramuzza, REMODE: Probabilistic, Monocular Dense Reconstruction in Real Time. *IEEE International Conference on Robotics and Automation* 2014.

[16] C. Richardt, D. Orr, I. Davies, A. Criminisi and N. A. Dodgson, Real-time Spatiotemporal Stereo Matching Using the Dual-Cross-Bilateral Grid, *European conference on Computer vision*, 2010

[17] A. Hosni, C. Rhemann, M. Bleyer, M. Gelautz, Temporally Consistent Disparity and Optical Flow via Efficient Spatio-Temporal Filtering. *Pacific-Rim Synmposium on Image and Video Technology*, 2011.

[18] J. Engel, J. Strum and D. Cremers, Semi-Dense Visual Odometry for a Monocular Camera. *IEEE International Conference on Computer Vision*, 2013

[19] J. Engel, J. Stueckler and D. Cremers, Large-Scale Direct SLAM with Stereo Cameras. *International Conference on Intelligent Robots and Systems*, 2015

[20] P. Merrel et al. Real-time Visibility-based Fusion of Depth Maps. *IEEE International Conference on Computer Vision* 2007

[21] S. Matyunin, D. Vatolin and M. Smirnov, Fast Temporal Filtering of Depth Maps. *International Conference on Computer Graphics, Visualization and Computer vision*, 2011.

[22] C. Unger, E. Wahl, P. Strum and S. Ilic, Probabilistic Disparity Fusion for Real-time Motion Stereo. *Machine Vision and Applications*, Vol 25, 2011.

[23] D. Droeschel, M. Nieuwenhuisen, M. Beul, D. Holz, J. Stucker and S. Behnke, Multi-layered Mapping and Navigation for Autonomous Micro Air Vehicles, *Journal of Field Robotics*, 2015

[24] S. Shen, N. Michael and V. Kumar, 3d Indoor Exploration with a Computationally Constrained MAV, *Robotics: science and Systems*, 2003

[25] J. Kang, I. Cohen, G. Medioni and C. Yuan, Detection and Tracking of Moving Objects from a Moving Platform in Presence of Strong Parallax, *IEEE International Conference on Computer Vision*, 2005

[26] C. Forster and M. Pizzoli and D. Scaramuzza, SVO: Fast Semi-Direct Monocular Visual Odometry, *IEEE International Conference on Robotics and Automation, 2014*

[27] C. Stauffer and W. Grimson, Adaptive Background Mixture Models for Real-time Tracking. *International Conference on Computer vision and Pattern Recognition*, 1999.

[28] H. Hirschmuller, Accurate and Efficient Stereo Processing by Semi-Global Matching and Mutual Information. *International Conference on Computer Vision and Pattern Recognition*, 2005

[29] B. Kitt, A. Geiger and H. Lategahn. Visual Odometry based on Stereo Image Sequences with RANSAC-based Outlier Rejection Scheme. *Intelligent Vehicle Symposium*, 2010.

[30] A. Geiger, P. Lenz and R. Urtasun, Are we ready for Autonomous Driving? The KITTI Benchmark Suite. *Conference on Computer Vision and Pattern Recignition*, 2012