Onboard Stereo Vision for Drone Pursuit or Sense and Avoid

Cevahir Cigla *
Aselsan Inc.
ccigla@aselsan.com.tr

Rohan Thakker *
Jet Propulsion Laboratory / California Institute of Technology
rohan.a.thakker@jpl.nasa.gov

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

Abstract

We describe a new, on-board, short range perception system that enables micro aerial vehicles (MAVs) to detect, track, and follow or avoid nearby drones (within 2-20 meters) in GPS-denied environments. Each vehicle is able to sense its neighborhood and adapt its motion accordingly without use of centralized reasoning or inter-vehicle communication. To enable a lightweight, low power solution, on-board stereo cameras are used for detection and tracking with depth images, while a downward-looking camera and an inertial measurement unit are used to estimate the position of the observer without use of GPS. We illustrate the robustness and accuracy of this approach through real-time, outdoor leader-follower experiments with three quadrotors. Our experiments show that state-of-art trackers are far less robust in detection against cluttered background. This demonstrates that stereo vision is a highly effective approach to perception for safe navigation of multiple MAVs in close proximity.

1. Introduction

Many potential applications of micro aerial vehicles (MAVs), including mapping, reconnaissance, and delivery, require capabilities for autonomous localization, collision avoidance, and 3D reconstruction [1], [2], [3], As these technological capabilities mature, applications involving multiple MAVs are becoming possible, including team reconnaissance, cooperative lift, and counter-UAV operations. These applications require detecting, tracking, and maneuvering each MAV relative to other nearby MAVs; this has not been demonstrated previously in a self-contained system that performs all sensing and computing needed for operating each MAV on-board that MAV.

Most research for multi-MAV applications focuses on controlling multiple vehicles to fly in harmony [4], [5], using external motion capture (mocap) systems, e.g. Vicon cameras, to accurately estimate the positions of all MAVs. Such work usually involves communication between vehicles or between a central control station and all vehicles. Many real-world application scenarios cannot use mocap systems, centralized reasoning, or even inter-vehicle communication. Previous studies of airborne sense and avoid (SAA) systems mostly address large scale drones [6] and planes using relatively large, heavy, power-hungry sensors or sensor combinations, including radar, due to the need for long range, day/night, all-weather operation [7], [8]. Lighter weight vision-based approaches are also used [9],[10], with more limited performance.

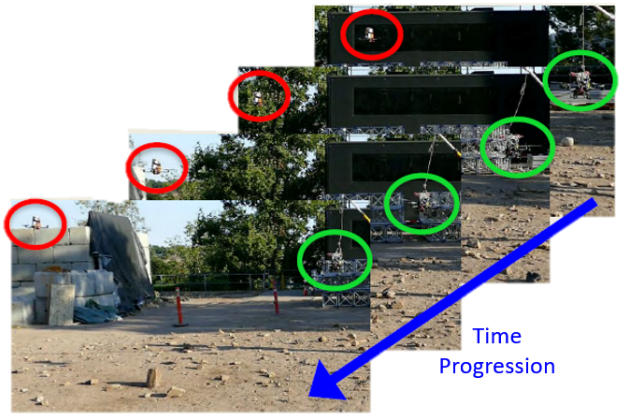


Figure 1: Follower drone (green) autonomously tracks a manually joy-sticked leader drone (red)

Short range SAA systems for MAVs have received little attention. Recently, a cascade of detectors was used with a single camera to detect nearby MAVs [11]. This approach may fail in outdoor applications with cluttered backgrounds, where the target can be invisible due to texture similarities. Other state-of-the-art video object trackers can also fail due to texture/color ambiguity, as seen in experiments we report here. In [12], miniature radar is used on a

*Cigla Cevahir and Rohan Thakker contributed equally to this work.

quadrotor to detect nearby flying objects based on Doppler signature. This active approach consumes more power than vision-based alternatives and has a very limited field of regard with poor angular separation capability. Acoustic vector sensors have been used on fixed wing MAVs [13] to detect civil aircraft; this was promising for detecting loud targets, but it has limitations for position sensing.

This paper addresses the need for on-board SAA and leader follower capability in small MAVs, e.g. under 1 kg. Detection is performed by segmentation of depth images produced by a lightweight, low power, passive stereo vision system and tracking is performed by frame to frame motion prediction and depth blob association. A single downward-looking camera and an IMU are use for ego-motion estimation without assistance from GPS or other external navigation aids. This enables SAA or pursuit with a self-contained, on-board sensing and computing system, without reliance on centralized reasoning or inter-vehicle communication.

Section 2 reviews related work on SAA and leader-follower applications for UAVs and MAVS. Our approach to detection and tracking is described in section 3. Experimental evaluation was done with real-time, outdoor testing of a self-contained perception and navigation system on a quadrotor; experiments were most practical to do for a leader-follower scenario where the lead vehicle was teleoperated and the follower was autonomous. Section 4 describes these experiments and compares performance of our system with other tracking algorithms. Section 5 discusses conclusions and future research directions.

2. Related Work

Sense and avoid technology has been studied for large aerial vehicles (UAVs and airplanes) to support ground traffic control [6]. GPS and radar sensors are the most endeavored technologies to provide long distance collision protection. Radar cannot be used on MAVs due to size, weight, and power constraints. Whereas, GPS cannot provide submeter accuracy that is required for small vehicles and also cannot be used in GPS-denied environments. At that point, vision based approaches are good alternatives for sense and avoid due to their light-weight and low-power characteristics. Commonly used monocular vision based approaches [9], [14] use texture based classification and/or motion compensated frame differencing to detect moving targets. [11] adapts monocular vision for MAVs by Histogram of Oriented Gradients (HOG) based ensemble learners trained for each frame. The Hungarian algorithm [15] is used for data association and tracking of the candidate regions in consecutive frames. Monocular vision has the potential to fail easily due to texture ambiguity and fast vehicle motion, especially against cluttered background which is commonly encountered in outdoor applications.

Drawback of monocular vision can be handled by use of stereo cameras, that could provide 3D map of an environment by stereo matching [16]. Flying vehicles can easily be separated from the background by observing the depth images. In the last decade, significant amount of research [17],[18] has been devoted for stereo matching through evaluation among common benchmarks [16],[19]. Extraction of dense depth images is still an open problem due to its high computational complexity and errors in the matching process.

The position of an observer MAV is important to track nearby vehicles and keep them in field-of-view. An alternative to GPS is Simultaneous localization and mapping (SLAM) [20] techniques which can localize a vehicle accurately. A downward looking camera is used to estimate the motion of the vehicle by tracking visual features on the ground. Vision based odometry estimation may fail in untextured environments or under severe lighting changes. Hybrid techniques have also been proposed that fuse IMU data with visual odometry in an SSF framework [21] providing much better accuracy and robustness compared to visual SLAM approaches.

Video object tracking (VOT) is a common tool used in surveillance applications. The recent VOT benchmarks and challenges [22] provide a framework to compare algorithms which highly impacted the tracking research. DSST [23] and KCF [24] are state-of-art correlation based trackers that provide high accuracy with low computation. Use of multiple features (HOG) is adapted to fast FFT based kernel correlation. In recent years significant amount of research has been devoted to improve the performance of correlation filters. SAMF [25] extends KCF to adapt to changes in scale by a brute force scale-search and performs a histogram based modification to handle visual appearance changes. It is one of the best performing trackers in the experiments of [26], [27]. Recently, context aware background modeling (SAMF CA)[28] has been adapted for fast correlation tracking that improves accuracy and handles background change with incremental increase in computation. DS-KCF [29] extends correlation based tracking to RGB-D data that balances efficiency and accuracy in RGBD tracking evaluation [30].

Two new benchmarks [26],[31] have been released to address video object tracking on unmanned aerial vehicle. Both studies have the same conclusion that current state-of-the-art video object trackers perform poorly on the videos captured through MAVs compared to common VOT challenges. Especially in the scenario where one MAV is tracked by another one. As stated in [26], all of the trackers fail due to abrupt relative motion and visual appearance changes. It is important to note that most of the state-of-the-art trackers and the VOT benchmarks are developed for monocular videos. These benchmarks have a diverse set

of videos that have various content characteristics which makes it is difficult to expect good performance for specific applications that are not referenced in the challenges adequately. This includes, drone pursuit or sense and avoid related applications, which require depth information. These applications have not received much attention compared to monocular object tracking.

Hence, in this work we propose a stereo vision based drone tracking algorithm to sense the surrounding and detect/track nearby flying MAVs that is much more robust than state-of-art VOT algorithms.

3. Algorithm

In this study, we perform detection and tracking of nearby micro aerial vehicles in disparity space. The key idea is that flying drones are separate objects in 3D space that can be differentiated from their surroundings by blob detection in disparity images. The flowchart of the algorithm is shown in Fig. 2. Once the blobs are detected for each frame, various approaches can be utilized to relate them among consecutive frames which is discussed in the tracking section.

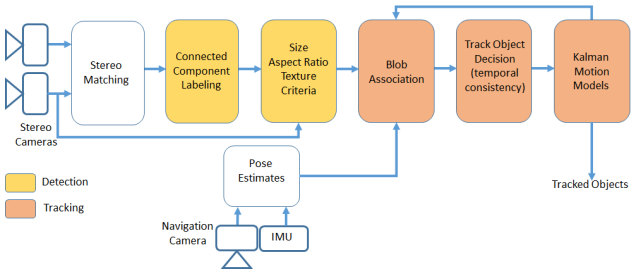


Figure 2: The flowchart of nearby drone detection and tracking

3.1. Detection

Depth images provide significant information about the neighborhood of a moving object. This can be used to sense, avoid and track nearby moving objects. As shown in Fig. 3, a quadrotor can be observed as a compact blob in a depth image extracted from stereo images. In this work, due to limitation on processing power and real-time constraints, block matching algorithm is executed on intensity measurements. Even for the state-of-the-art matching algorithms, errors in depth image are common due to the presence of texture-less regions and changes in lighting that have uneven effects on the scene. These errors may generate false blobs that do not correspond to actual drones. Hence post-processing and noise removal steps are required to detect target drones reliably.

Connected component labeling is the first step to group neighboring pixels that are located at similar distances from the observer. Each label is parameterized by mean 3D coordinates, volume, diameter and aspect ratio based on the depth data and also the texture in image along the connected components. During this step, connected components with very low spatial distances are merged together to account for disjoint parts in disparity image that correspond to the same object which may be observed due to errors in stereo matching. An estimate of the size and speed capability of the nearby drones are known for the leader-follower application. Thus, the connected components that do not satisfy the size criteria, such as having a much larger/smaller volume, diameter and aspect ratio compared to the expected values are eliminated. A typical elimination result is illustrated in Fig. 3 where the small and large connected component labels are eliminated, this results in a more clear candidate map. The larger volume regions correspond to background such as floor, buildings, clouds or trees; whereas smaller regions correspond to errors in depth images especially along texture-less regions. In addition to size/shape criteria, average texture within the blobs is also utilized to eliminate false candidates that could meet the size criteria. The texture of a blob S is measured by the average edge power along horizontal and vertical axes as follows:

$$S_{texture} = \sqrt{\sum_{i \in S} I_x^2(i) + I_y^2(i)} \tag{1}$$

As a result, the detection step provides candidate blobs that are easily differentiated from their surroundings according to the depth image extracted through stereo matching. The candidates can be target drones or noisy blobs; hence, they are further analyzed in the tracking step according to their motion characteristics.

3.2. Tracking

Detection of candidate blobs is applied for each frame independently and we need to relate these candidates to extract the motion models of the nearby drones. Use of depth image enables the estimation of 3D motion models including speed and position in world coordinates. It is important to note that the observer is also moving in the leader-follower scenario, hence we use pose estimates from visual odometry to account for the ego motion of the stereo camera. At each frame, the 3D coordinates of the candidate targets are mapped to the world coordinates by the use of position and orientation of the observer. This mapping enables us to estimate the position and speed of the target by using Kalman Filter with different motion models such as constant velocity, constant acceleration or jerk. During the experiments it has been observed that, targets do not follow specified motion models perfectly. This is mainly due to presence of noise in the estimate of center positions of the target that arises from errors in depth images that

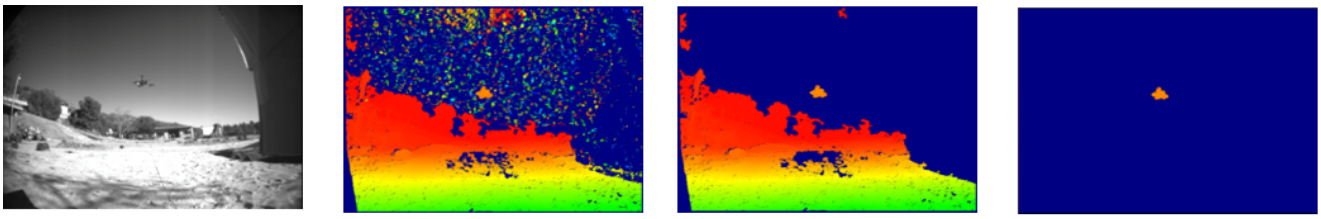


Figure 3: Left camera image, disparity image after stereo matching, blobs that are larger than a predefined threshold and blobs within expected size.

causes abrupt changes in blob size. The blob size enlarges or shrinks independent of the relative vehicle motion which makes it difficult to assign proper motion models. Thus, complex motion models did not show any significant improvement and constant velocity model provided the best performance.

For each candidate blob, we implement a Kalman Filter that uses these motion model to predict the 3D position in the next frame. Then, a search window is defined around the predicted position of the target. In this window, the blob with closest spatial distance and highest similarity in terms of size and shape is assigned as the preceding blob of the tracked target. Thus, distance and size similarity are utilized to relate blobs in consecutive frames. Each blob is assigned to the closest target and tracked along previous frames that are within the search window as shown in Fig. 4. At this point, some targets may not be assigned to the blobs in the most recent frame. These targets are kept with no observation update in the Kalman Filter for certain number of frames. The position and speed of the tracked blobs are updated with the most recent observation.

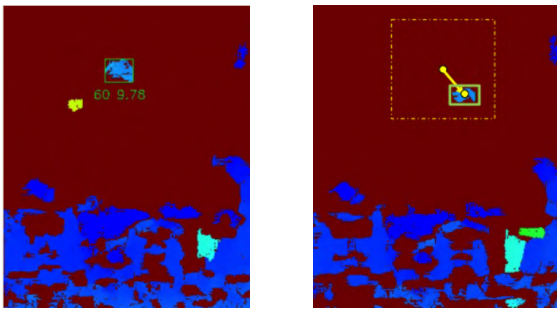


Figure 4: Tracking example for a target that has been tracked for 60 frames and have 9.78m distance to the observer

The association based tracking relates the blobs in consecutive frames, however it does not guarantee that those blobs correspond to an actual nearby flying target. As expected, an actual target continues its motion for a period of time, this characteristic can be used to eliminate false tracked blobs, that have short track history. Hence, the blobs that are tracked for a predefined number of frames are considered as the target. This approach enables us to track multiple targets which requires an additional task of identification of the target that is supposed to be followed by the observer. The number of tracked frames, the distance, and the location of the targets can be considered as the criteria for target selection. In this work, the target which is non-stationary and has been tracked for the largest number of frames compared to the other candidates is chosen as the target to be followed.

As a common fact in tracking systems, the targets may disappear for a period of time due to an occluding structure, low contrast or errors in depth image. In order to keep these candidates under consideration during the coast mode, the search window is enlarged gradually and the unmatched blobs are not dropped from the track list. The duration of coast mode is a design parameter of a system which is the range of 2 to 5 seconds for our application. During this period, if the target is matched to a blob, track continues and the search window is set to the initial size until the coast mode is activated again. The track is dropped if no match is found within this time interval.

4. Experiments

To demonstrate the accuracy and robustness of our approach we conduct a leader-follower experiment on 3 different target drones in various outdoor environments. We also show a quantitative comparison of the accuracy and robustness our tracker with the state-of-art VOT algorithms on data logs of the leader-follower experiment.

4.1. Leader-Follower

In the leader-follower experiment the leader drone is joysticked manually by a human pilot and the follower drone navigates autonomously. Fig. 5 shows the pipeline of our implementation of the leader-follower application. For the follower drone, we used the Asctec Pelican drone mounted with an Intel NUC, Odroid XU-4, a downward facing camera and a forward facing pair of stereo cameras with a 20cm baseline.

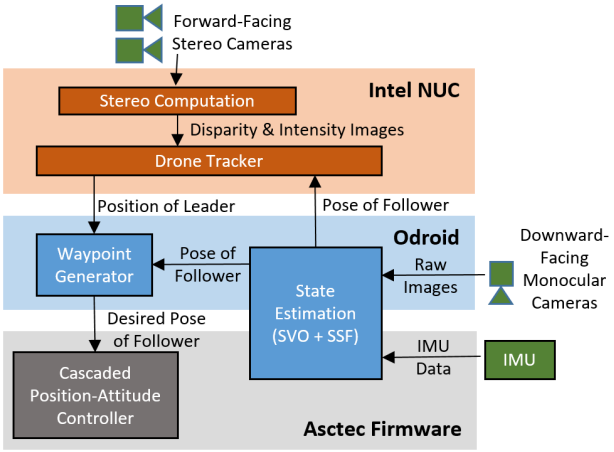


Figure 5: Pipeline of leader follower application

Table 1: Computation time distribution for the leader-follower application

	SVO	Stereo Matching	Tracking
Time(msec)	16.6	100	15

Images from the stereo cameras are received on the Intel NUC and a disparity image is computed. Our tracker computes the position of the leader drone using intensity and disparity images along with pose estimates of the follower drone that are obtained from the state estimation module based on Single Sensor Fusion (SSF) [21] and Semi-Direct Visual Odometry (SVO) [20] algorithms. The way-point generator uses position estimates of the leader and the follower to compute the desired position of the follower which is then sent to the AscTech firmware. Finally, the AscTech firmware runs a cascaded position and attitude control loop that takes the follower drone to the desired pose.

The leader-follower behavior is shown in Fig. 1. Fig. 6 shows the images acquired by the follower drone while the leader-follower behavior was being executed and the complete video can been viewed in [32]. This shows that using the proposed approach the follower successfully tracks the leader and also avoids collisions with it.

In this experiment, disparity images are calculated at a resolution of 752x480 pixels with a search range of 75 disparities. The timing details are given in Table 1, where SVO runs on Odroid XU-4, stereo matching runs on the Intel NUC with our drone tracking algorithm. The stereo matching runs at 10 fps which is followed by a rather fast tracking algorithm that results in overall 7 fps for sense and avoid. The way-point generator has negligible computation time compared to other processes.

4.2. Quantitative Comparison of Trackers

In this section, we compare the proposed tracker with the state-of-art Visual object trackers DS-KCF [29], SAMF [25] and SAMFCA [28] using stereo data collected by the follower drone during execution of the leader-follower experiment with three different target drones. Deep learning based trackers are not considered due to lack of large labeled dataset that would be required for training. We execute the tracking algorithms off-line and compare their performance with our tracker using manually generated ground truth (GT) bounding boxes on four different sequences as shown in Fig. 6 and [32].

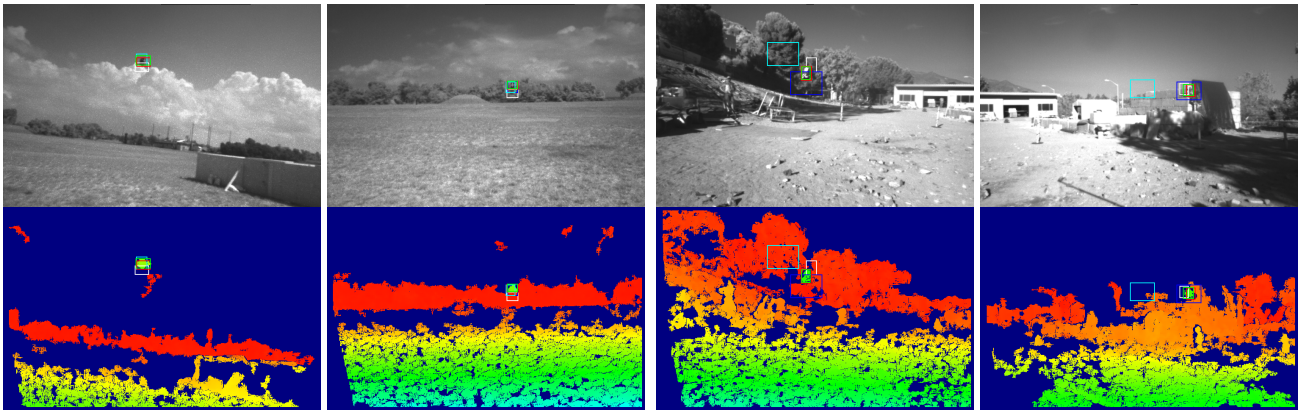
DS-KCF adapts FFT based correlation to 3D, while SAMFCA incorporates nearby background pixels as negative sample regions. Each tracker adapts the scale to track targets that have large variations in the distance to the observer. The target models in these trackers are defined by multi-channel HOG feature representation based on depth images. We use existing open source implementations of these trackers and run them on the disparity images obtained after stereo matching. We also tested using the intensity images but this approach mostly failed due to background mixing and lack of texture differentiation.

Unlike other trackers, our tracker uses visual inertial odometry data to incorporate the ego motion of the camera. Hence, for a fair comparison, at each frame we re-initialize the search window of the trackers around the position of the target obtained from ground truth. New position of the target is detected at the region with maximum correlation score in the search window. Finally, the target representation model is updated according to the newly detected position and bounding box. Note that this helps to evaluate these trackers independent of the ego motion and to focus only on their matching capabilities. However, this also incorporates the motion of the target drone, hence our results show the upper bound on the performance of other trackers and their actual performance may be worse.

The following well-known quantitative metrics are utilized to compare the performance of the trackers:

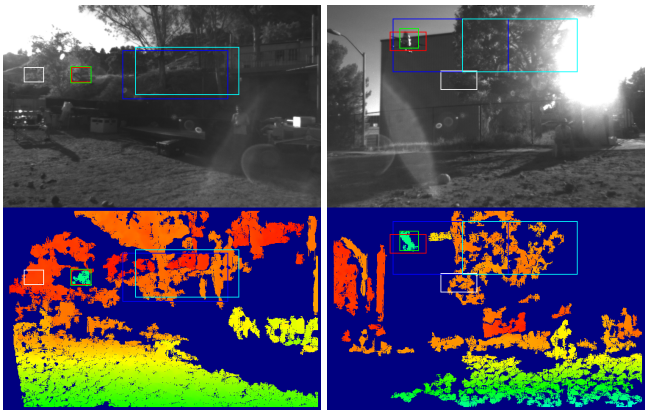
1. **Area of Overlap**: of the bounding boxes generated by the tracker and ground truth (normalized by the maximum of the two areas).
2. **Center Distance**: between bounding boxes of the tracker and ground truth.
3. **Failure Rate**: ratio of number of frames with zero area of overlap to total number of frames in the sequence. This metric captures the false positive detections.

Fig. 7 shows a comparison of the robustness (failure rate) and accuracy (area of overlap and center distance) of the trackers. The results show that our algorithm has better accuracy and a significantly lower failure rate; hence is robust to cluttered background unlike the state-of-art trackers. Note that the performance of the state-of-art trackers

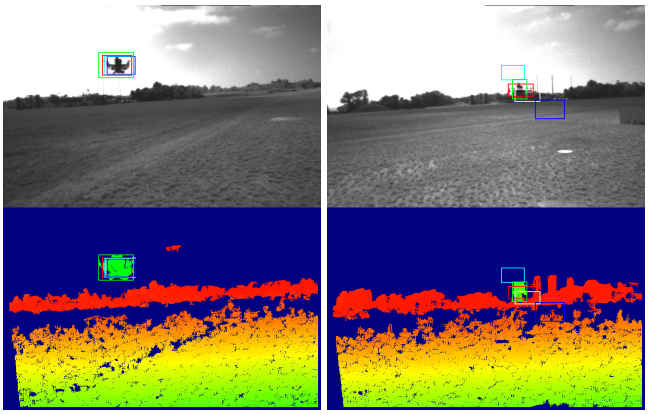


(a) Sequence s1 with custom ∼250mm frame target drone

(b) Sequence s2 with Asctec Hummingbird as target drone



(c) Sequence s3 with Asctec Hummingbird as target drone



(d) Sequence s4 with Asctec Pelican as target drone

Figure 6: Comparison of trackers using stereo data from the leader-follower experiment shows that state-of-art algorithms perform well against sky but are not robust to cluttered background. The algorithms compared are DS-KCF (white), SAMF (cyan), SAMFCA (blue), proposed (green) and manually labeled ground truth (red).

is worse in sequence 2 and 3 which involves more cluttered background.

Another major reason for failure of the correlation based trackers is abrupt shape change of the target in the depth image. Some typical shapes of the targets in disparity images of consecutive frames are shown in Fig. 8, where the corresponding response maps between the two models are given in the last column. The shape and size of the targets change significantly which distorts the desired correlation distribution around the actual correspondences. Besides, in some cases, the target is not observed in the disparity images; this enforces the tracker to match a background region. It is difficult to determine if there is a match or not depending on the correlation score of these trackers, hence they have larger number of false positives. Since they are sensitive to shape and appearance change, low correlation scores can be observed even though there is a good match.

Hence, it can be concluded that the state-of-art trackers do not produce reliable results on depth images for this specific application. This is also consistent with the observations in [26].

5. Conclusion and Future Work

Real-world applications of MAV SAA and pursuit require fully self-contained perception systems on each MAV, and some applications either cannot use inter-vehicle communication (counter-UAS) or may prefer to minimize communication. We have described such a system that uses passive stereo vision for onboard detection and tracking of nearby MAVs. Detection is done by segmenting depth images; tracking is done by Kalman filter-based prediction plus association of detections between frames. With the resolution (752x480) and baseline (20 cm) of the existing cameras, detection range up to 20 m has been shown. Stereo vision observes all elements of the state vector of target MAVs, whereas monocular vision approaches cannot do that without additional information, such as known size of the target, or constant target velocity plus maneuvering

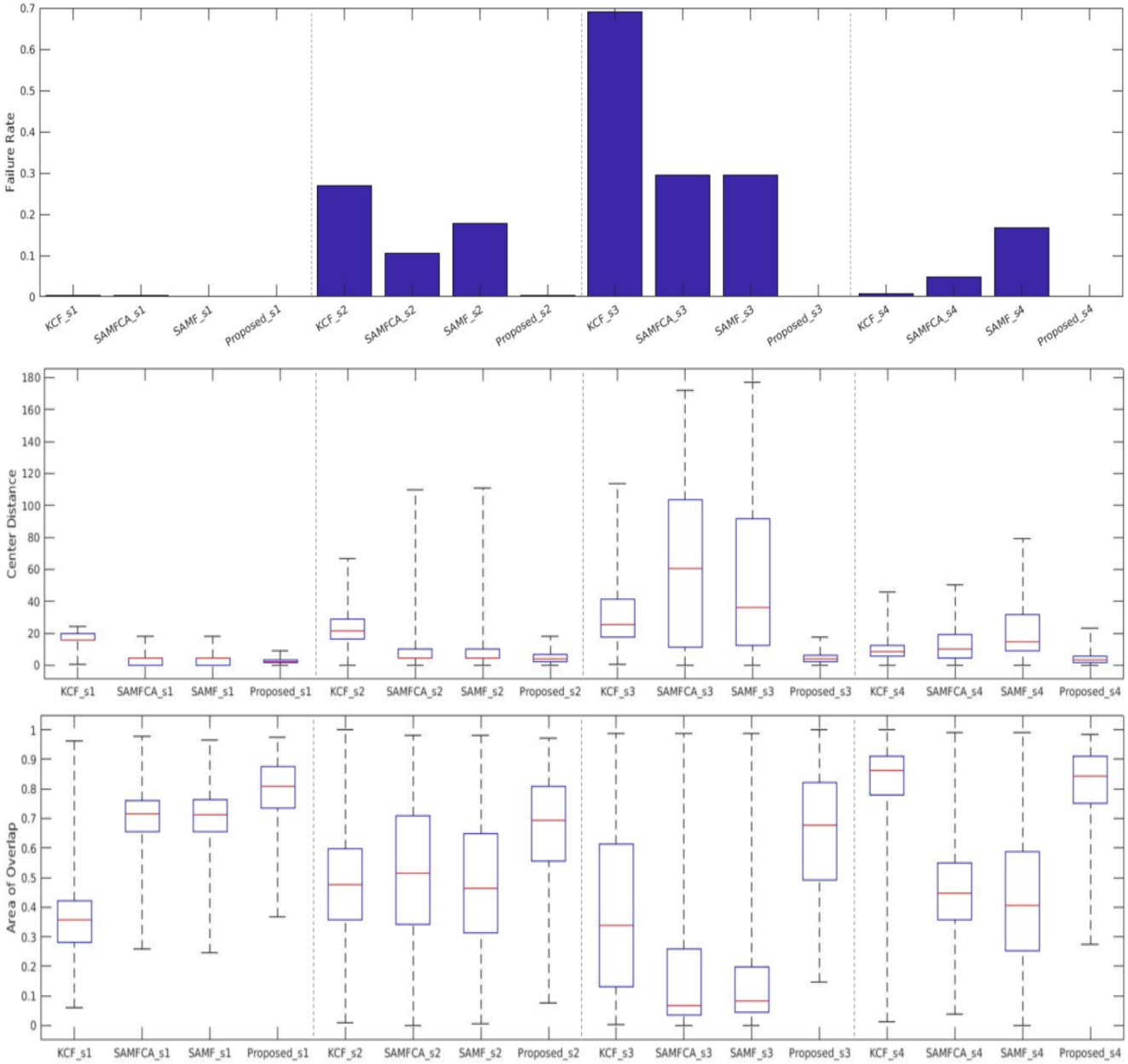


Figure 7: Comparison of robustness (failure rate) and accuracy (center distance and area of overlap) of the trackers on 4 sequences shows that our tracker is robust with significantly lower failure rate and has better accuracy than the state-of-art trackers.

of the observer MAV.

We conducted outdoor experiments with a leader-follower behavior in several challenging conditions, including cluttered backgrounds, large frame-to-frame motion of the target, and flying toward the sun. Stereo-based tracking was very robust and was able to reacquire the target reliably after a loss of track. Using data logs, we compared this to existing implementations of several state-of-the-art trackers, which were less reliable and less accurate. Higher resolution is achievable with ASIC implementations of stereo algorithms, which would enable either longer range or shorter baselines. Pushing detection and tracking beyond the range limits of stereo is still of interest, integrating stereo and monocular tracking is an important topic for future work.

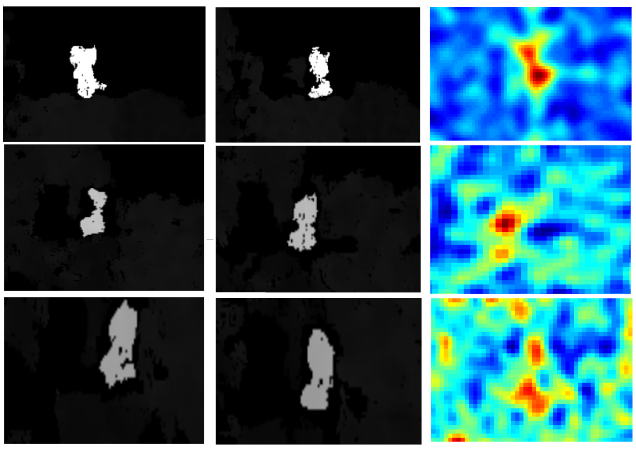


Figure 8: Typical correlation response maps on dynamically changing depth data. Shape changes degrade the crispness of correlation scores.

6. Acknowledgement

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

The authors would like to thank Dr. Roland Brockers for helping with system implementation and testing, Dr. Jeff Delaune and Mr. Christian Brommer for helping with the state estimation module and the leader-follower experiments.

References

[1] C. Forster, M. Pizzoli, and D. Scaramuzza. Appearance-based active, monocular, dense reconstruction for micro aerial vehicles. In *Proceedings of Robotics: Science and Systems*, Berkeley, USA, July 2014.

[2] L. Matthies, R. Brockers, Y. Kuwata, and S. Weiss. Stereo vision-based obstacle avoidance for micro air vehicles using disparity space. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pages 3242–3249. IEEE, 2014.

[3] R. Brockers, A. Fragoso, B. Rothrock, C. Lee, and L. Matthies. Vision-based obstacle avoidance for micro air vehicles using an egocylindrical depth map. In *International Symposium on Experimental Robotics*, pages 505–514. Springer, 2016.

[4] M. Saska, J. Chudoba, L. Preucil, J. Thomas, G. Loianno, A. Tresnak, V. Vonasek, and V. Kumar. Autonomous deployment of swarms of micro-aerial vehicles in cooperative surveillance. *IEEE Proceedings of 2014 International Conference on Unmanned Aircraft Systems (ICUAS)*, 1:584–595, 2014.

[5] Y. Mulgaonkar, G. Cross, and V. Kumar. Design of small, safe and robust quadrotor swarms. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 2208 – 2215, Seattle, WA, USA, May 2015.

[6] E. Barmpounakis, E. Vlahogianni, and J. Golias. Unmanned aerial aircraft systems for transportation engineering: Current practice and future challenges. *International Journal of Transportation Science and Technology*, 5:111 – 122, 2016.

[7] Y. Kwag and C. Chung. Uav based collision avoidance radar sensor. *IEEE Int. Geo-Science Remote Sensing Symposium*, pages 639–642, 01 2007.

[8] G. Fasano, D. Accardo, A. Moccia, and L. Paparone. Airborne multi-sensor tracking for autonomous collision avoidance. *IEEE Int. Conference on Information Fusion*, pages 1 – 7, 08 2006.

[9] Y. Wu, Y. Sui, and G. Wang. Vision-based real-time aerial object localization and tracking for UAV sensing system. *CoRR*, abs/1703.06527, 2017.

[10] R. Carnie, R. Walker, and P. Corke. Image processing algorithms for uav sense and avoid. *IEEE International Conference on Robotics and Automation*, pages 2848 – 2853, 06 2006.

[11] S. Raj, R. Adriaan, R. Artem, L. Vincent, G. Denis, F. Pascal, and M. Alcherio. Vision-Based Unmanned Aerial Vehicle Detection and Tracking for Sense and Avoid Systems. *2016 Ieee/Rsj International Conference On Intelligent Robots And Systems (Iros 2016)*, pages 1556–1561, 2016.

[12] A. Moses, M. Rutherford, and K. Valavanis. Radar-based detection and identification for miniature air vehicles. In *Control Applications (CCA), 2011 IEEE International Conference on*, pages 933–940. IEEE, 2011.

[13] E. Tijs, G. Croon, J. Wind, B. Remes, C. Wagter, H. Bree, and R. Ruijsink. Hear-and-avoid for micro air vehicles. In *Proceedings of the International Micro Air Vehicle Conference and Competitions (IMAV), Braunschweig, Germany*, volume 69, 2010.

[14] C. Fu, A. Carrio, M. Olivares-Mendez, R. Suarez-Fernandez, and P. Campoy. Robust real-time vision-based aircraft tracking from unmanned aerial vehicles. In *2014 IEEE International Conference on Robotics and Automation, ICRA 2014, Hong Kong, China, May 31 - June 7, 2014*, pages 5441–5446, 2014.

[15] H. Kuhn and B. Yaw. The hungarian method for the assignment problem. *Naval Res. Logist. Quart*, page 8397, 1955.

[16] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comput. Vision*, 47(1-3):7–42, 2002.

[17] H. Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions Pattern Analysis Machine Intelligence*, 30(2):328–341, 2008.

[18] A. Geiger, M. Roser, and R. Urtasun. Efficient large-scale stereo matching. In *Asian Conference on Computer Vision (ACCV)*, 2010.

[19] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[20] C. Forster, M. Pizzoli, and D. Scaramuzza. Svo: Fast semi-direct monocular visual odometry. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pages 15–22. IEEE, 2014.

[21] S. Weiss, M. Achtelik, S. Lynen, M. Chli, and R. Siegwart. Real-time onboard visual-inertial state estimation and self-calibration of mavs in unknown environments. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 957–964. IEEE, 2012.

[22] M. Kristan et al. The visual object tracking vot 2015 challenge results. In *Visual Object Tracking Workshop 2015 at ICCV2015*, Dec 2015.

[23] M. Danelljan, G. Häger, F. Shahbaz Khan, and M. Felsberg. Accurate scale estimation for robust visual tracking. *Proceedings of the British Machine Vision Conference*, 2014.

[24] J. Henriques, R. Caseiro, P. Martins, and J. Batista. High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, abs/1404.7584, 2014.

[25] Y. Li and J. Zhu. *A scale adaptive kernel correlation filter tracker with feature integration*, pages 254–265. Springer International Publishing, 2014.

[26] M. Mueller, N. Smith, and B. Ghanem. A benchmark and simulator for uav tracking. In *Proc. of the European Conference on Computer Vision (ECCV)*, 2016.

[27] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. Torr. Staple: Complementary learners for real-time tracking. June 2016.

[28] M. Mueller, N. Smith, and B. Ghanem. Context-aware correlation filter tracking. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[29] S. Hannuna, M. Camplani, J. Hall, M. Mirmehdi, D. Damen, T. Burghardt, A. Paiement, and L. Tao. Ds-kcf: a real-time tracker for rgb-d data. *Journal of Real-Time Image Processing*, 2016.

[30] S. Song and J. Xiao. Tracking revisited using rgbd camera: Unified benchmark and baselines. *Proceedings of 14th IEEE International Conference on Computer Vision*, 2013.

[31] S. Li and D. Yeung. Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models. In *AAAI*, 2017.

[32] Video sequences of the leader-follower experiment. `https://tinyurl.com/ybet6jdr`.