Visual Drone Detection and Tracking for Autonomous Operation from Maritime Vessel

Timothy, Halleux
Department of Mechanics
Royal Military Academy
Brussels, Belgium
Timothy.Halleux@mil.be

Tien-Thanh, Nguyen
Department of Mechanics
Royal Military Academy
Brussels, Belgium
* TienThanh.Nguyen@mil.be

Charles, Hamesse
Department of Mathematics
Royal Military Academy
Brussels, Belgium
Charles.Hamesse@mil.be

Geert, De Cubber
Department of Mechanics
Royal Military Academy
Brussels, Belgium
Geert.DeCubber@mil.be

Bart, Janssens
Department of Mechanics
Royal Military Academy
Brussels, Belgium
Bart.Janssens@mil.be

***Abstract* — To allow incorporation of autonomous Unmanned Aerial Vehicles (UAV's/drones) into maritime military operations, it is critical to be able to accurately localize the UAV with respect to the moving maritime vessel during the take-off and landing phases. This work addresses the study and implementation of a visual detection, tracking and three-dimensional positioning method for a specific drone from a moving maritime vessel. The YOLOv5 detector and the OceanPlus tracker have been trained on a custom dataset with good performance in accuracy and processing time. The drone's position with respect to the vessel is estimated by applying stereo triangulation to the centres of the bounding boxes returned by the object detectors and trackers. The performance of the proposed positioning method was evaluated in a realistic simulated environment in the Unreal Game Engine. The proposed method allows detection, tracking, and positioning of a target drone at ranges exceeding 100m while achieving positioning errors below 10cm during landing phases.**

Keywords — Object Detection, Tracking, UAV, synthetic data

I. Introduction

Unmanned Aerial Vehicles (UAV's), also referred to as drones, are ever more being employed during maritime military operations. They offer advantages for over sea reconnaissance, tactical planning and situations in which human lives would have to be put at risk. Advances in technology have already rendered in flight operation of UAV's fairly easy. Thanks to global navigation satellite systems and software aided flight control, pilots can be highly assisted during flight maneuvers and under certain conditions even be completely replaced, making the system fully autonomous during flight. The takeoff and landing phases however remain critical, especially for maritime operations in which UAV's are operated from moving platforms characterized by complex superstructures. Being able to localize the UAV with respect to the platform during these phases is necessary to allow development of autonomous systems which are able to plan trajectories and avoid collisions with surrounding obstacles. This work addresses a visual object detection and tracking algorithm for a specific drone from a moving maritime vessel and proposes a positioning technique using stereo camera triangulation.

A. Related Methods

For the task of drone detection and positioning, a multitude of techniques involving various types of sensors have been addressed in literature. The available methods can be split into two categories: Detection and positioning methods not requiring the installation of any additional systems on the drone itself will further be called "passive" methods, whereas detection and positioning methods requiring the installation of additional hardware on board of the drone will be called "active" methods, as they require the drone to actively cooperate with the system to position it.

Global Navigation Satellite Systems (GNSS) with enhanced accuracy by differential technique Real-Time Kinematic (RTK) is one of the common "active" positioning methods for drone, can achieve position error less than 10cm. However, it relies entirely on the availability of GNSS signals and is vulnerable to jamming and spoofing. Other active methods such as radio frequency and acoustics also provides good accuracy and even specific solution for drone landing. But they are required additional hardware on both drone and maritime vessel which reduce their versatility.

Passive methods such as RADAR, LIDAR, visual and IR optical sensors do not require the drone to be cooperative in the positioning process. However, to achieve the accuracy level for guiding the drone during landing phase, they require improvement in hardware and software for drone detection and position estimation or sensor fusion technique for multimodal detecting system to provide better performance. This paper focuses on visual detecting and tracking method for a specific drone during the landing phase on a maritime vessel using deep learning algorithms.

II. Detection And Tracking

A. Detection

Object detection is a computer vision task to locate and identify target objects in an image, which has been greatly improved in performance with the rapid development of deep learning networks. Recent surveys of object detection techniques based on deep learned features have been provided by [1] and [2]. In [3] and [4], the focus is set especially on the task of drone detection. In general, the techniques can be divided into two main categories: One-stage and two-stage based detectors. The two-stage detector splits the object detection task into image classification and object localization. Examples are: RCNN, fast-, faster and mask-RCNN or FPN. These algorithms can produce high detection accuracy (bounding box tightness), however at the cost of detection speed therefore they are not yet suitable for real-time applications. For example, faster-RCNN provides 5 Frames Per Second (FPS) on a K40 GPU [5].

This research was funded by Royal Higher Institute for Defense with the collaboration between the Royal Military Academy and the Belgian Navy.

One-stage object detectors can generate class probabilities and object coordinates by performing a single pass through a deep CNN that provides all information at once. Examples for one-stage detectors are SSD, RetinaNet, EfficientDet and the family of YOLO algorithms [6]. One-stage detector are faster and can achieve good accuracy for our application. At the time of writing, newly developed YOLOv5 [7] has been reported to have best performance in accuracy and framerate [8]. Therefore, all network sizes of the YOLOv5 algorithm (YOLOv5s, -m, -l, -x) are selected and will be evaluated further in section III.

B. *Tracking*

Object tracking is a computer vision task to track the movement of target objects in a sequence of images. For drone tracking application, only single-target object tracking is considered. The Visual Object Tracking (VOT) Challenges [9] were introduced in 2013, providing datasets and clearly defined evaluation methods with available toolkit for researchers to evaluate their trackers. To select tracking algorithms for our drone tracking application, the best performance algorithms of VOT2020 were listed up, and ranked. As the selected tracker will need to operate in real-time on limited hardware, we focused our comparison on the results of short-term trackers in the real-time challenge. Providing best performance in accuracy and robustness and being able to achieve real-time tracking using limited computing resource, the following trackers have been selected for further evaluation: GOTURN [10], Ocean, OceanPlus, OceanPlus Online [11] and AlphaRefine [12].

III. Algorithm Evaluation

A. *Dataset*

A custom dataset of our research subject drone – a DJI Matrice M300 – was acquired with the raw dataset taken from multiple drone footages from different cameras, from various view angles, and different types of background and illumination conditions during our field tests at the Damage Control Center Military Domain in Beernem (Belgium).

The raw images then are annotated and processed according to the defined format of evaluated algorithms. A Python tool for automatic video annotation and processing based on video object tracking and manual evaluation is developed to generate the annotated images and ground truth bounding box data from the recorded videos (see Figure 2). It guarantees proper distribution of the bounding box sizes and positions in the dataset as shown in Figure 1. This dataset was augmented then randomly split into training, validation and test sets and used to train and evaluate multiple algorithms for object detection and tracking.

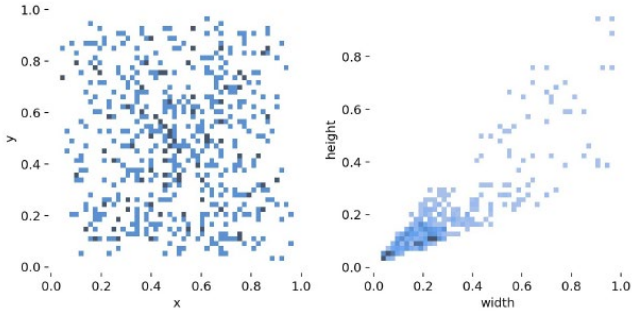

Figure 1: Bounding box sizes and positions after processing

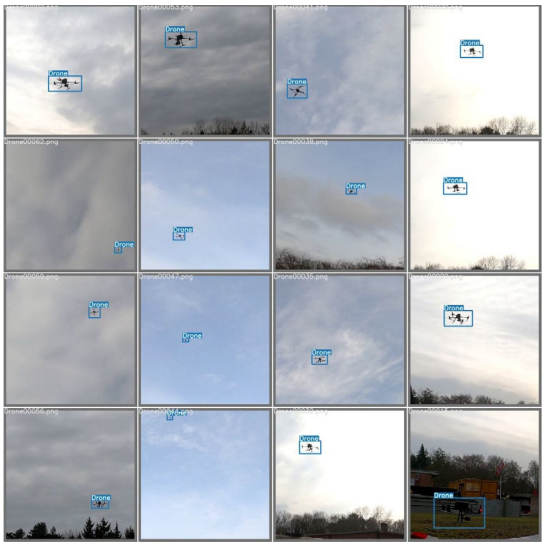

Figure 2: Example images from the raw dataset

B. *Performance metrics*

For object detection, the most used metric is the Average Precision (AP) which is almost equivalent to Area Under the Precision (P) – Recall (R) Curve (AUC). It is a combined measure, reflecting the performance of a detector in both precision and recall. Intersection over Union (IoU) measures the overlap between the predicted and the ground truth bounding box. The AP can be evaluated and averaged for different levels of IoU (0.5 to 0.95 with steps of 0.05 for instance which is indicated as AP@50:5:95) or for a single value (e.g., AP50). The mean Average Precision (mAP) is obtained by averaging all AP values obtained for different object classes.

For object tracking, [13] suggests using these metrics: Accuracy (A), Robustness (R) and Expected Average Overlap (EAO) to evaluate the performance of visual object tracking algorithms. Accuracy (A) is defined as the average overlap between the predicted and ground truth bounding boxes during successful tracking periods. Robustness (R) is defined as the number of times the tracker failed, i.e., drifted from the target, and had to be reinitialized. A failure is detected when IoU drops to zero. And EOA is an estimator of the average overlap a tracker is expected to attain on a large collection of short- term sequences with the same visual properties as the given dataset [13].

The speed evaluation of both detectors and trackers was done in limited computing power hardware: ThinkPad P50 with an Intel i7-6700HQ CPU, 2GB VRAM Quadro M1000M GPU and CUDA 10.2. Inference time and Non-max suppression process (NMS) time were used for evaluating YOLOv5 detectors. Frame rate (FPS) was considered for evaluating the speed of different trackers

C. *Detection Algorithms Evaluation*

Four network sizes of the YOLOv5 algorithm (YOLOv5s, -m, -l, -x) are compared in both accuracy performance and speed on our custom dataset. Table 1 shows that Precision equals almost 100% for all models, while Recall varies between 89% and 98%. Increasing the network size doesn't increase the performance mAP50 or mAP@50:5:95 much.

On other hand, Table 2 shows that the inference time is significantly increased when using larger network sizes.

TABLE 1:COMPARISION OF PERFORMANCE OF DIFFERENT YOLOV5 NETWORK SIZES ON OUR CUSTOM DRONE DETECTION DATASET

Model size	*F1-score*	*Precision*	*Recall*	*mAP @50%IoU*	*mAP @50:5:95%IoU*
S	*0.97*	*1*	*0.89*	*0.90*	*0.65*
M	*0.99*	*1*	*0.97*	*0.97*	*0.69*
L	*0.99*	*0.98*	*0.95*	*0.95*	*0.70*
X	*0.97*	*0.99*	*0.98*	*0.99*	*0.68*

Because of the insignificant improvement in performance and significantly longer inference time of the large network sizes (L and X), the smaller size networks (S and M) are selected for the drone detection task.

TABLE 2: COMPARISON OF MODEL SPEED OF DIFFERENT YOLOV5 NETWORK SIZES ON OUR CUSTOM DRONE DETECTION DATASET (IMAGE RESOLUTION = 640X640, BATCH SIZE = 1 ON THE M1000M GPU)

Model size	*Inference Time (ms)*	*NMS Time (ms)*	*Total FPS*
S	41.1	0.7	23.9
M	103.7	*0.8*	9.6
L	*195.2*	*0.8*	*5.1*
X	*362.2*	*2.1*	*2.7*

D. Tracking Algorithms Evaluation

To identify the most suited tracking algorithm for our application, five tracking algorithms (GOTURN, Ocean, OceanPlus, OceanPlus Online and AlphaRefine) have been evaluated in performance and speed. On performance, Table 3 presents the evaluation results on the VOT2022 dataset of the five trackers in both baseline and real-time tracking evaluation. The real-time evaluation assumes an image stream at 30 FPS. Since our application focus more on real-time aspect, Figure 3 shows the real-time evaluation of the accuracy-robustness plot and Figure 4 shows the EAO curve in function of tracked frames for each tracker. OceanPlus tracker achieves in overall best result in real-time performance evaluation.

On the speed aspect, shown in Table 4, OceanPlus and its online version achieved second best in speed with 10.77 and 10.49 FPS respectively, after Ocean with 14.02 FPS.

With good result in both tracking performance and speed, OceanPlus was chosen as the tracking algorithm to be further implemented for our application. Figure 5 illustrates an example of a tracking sequence using OceanPlus tracker.

TABLE 3: TRACKER EVALUATION RESULTS ON THE VOT2020 DATASET. THE FIRST, SECOND AND THIRD BEST SCORES PER METRIC ARE RESPECTIVELY COLORED IN RED, GREEN, AND BLUE.

	baseline			realtime			unsupervised
	EAO analysis	AR analysis		EAO analysis	AR analysis		Average accuracy
Trackers	EAO	A	R	EAO	A	R	AUC
GOTURN	0.114	0.419	0.341	0.120	0.391	0.365	0.145
Ocean	0.285	0.484	0.723	0.257	0.455	0.688	0.347
OceanPlus	0.430	0.693	0.754	0.348	0.636	0.672	0.533
OceanPlus_Online	0.430	0.693	0.754	0.344	0.632	0.666	0.533
AlphaRef	0.484	0.755	0.783	0.299	0.604	0.635	0.585

TABLE 4: TRACKING SPEED RESULTS (ON THINKPAD P50 & QUADRO M1000M GPU)

Tracker	Frame Rate (FPS)
GOTURN	13.67
Ocean	14.02
OceanPlus	10.77
OceanPlus Online	10.49
AlphaRefine	5.48

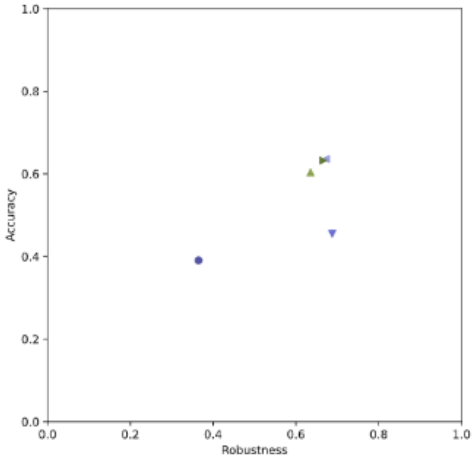

Figure 3: Real-time evaluation accuracy-robustness plot.

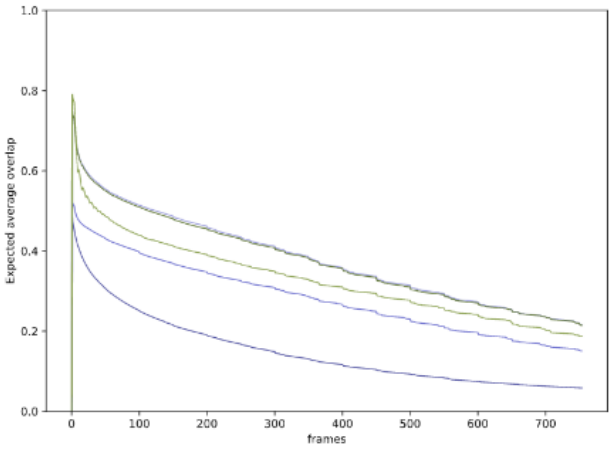

Figure 4: Real-time evaluation EAO curve in function of tracked frames

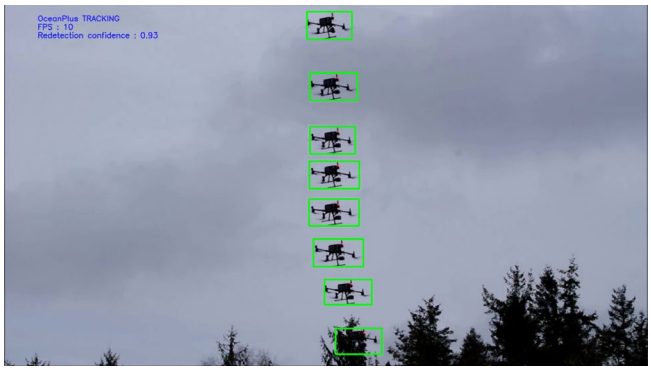

Figure 5: Example of a tracking sequence

IV. IMPLEMENTATION AND VALIDATION

This Section discusses the integration of YOLOv5 and OceanPlus as one detection and tracking pipeline. A simple position estimation method is also introduced and validated using realistic Maritime Environment Simulation. The simulation, based on the work of [14], created using Unreal Game Engine. It was used to acquire the ground truth position data of the simulated drone with respect to the stereo camera.

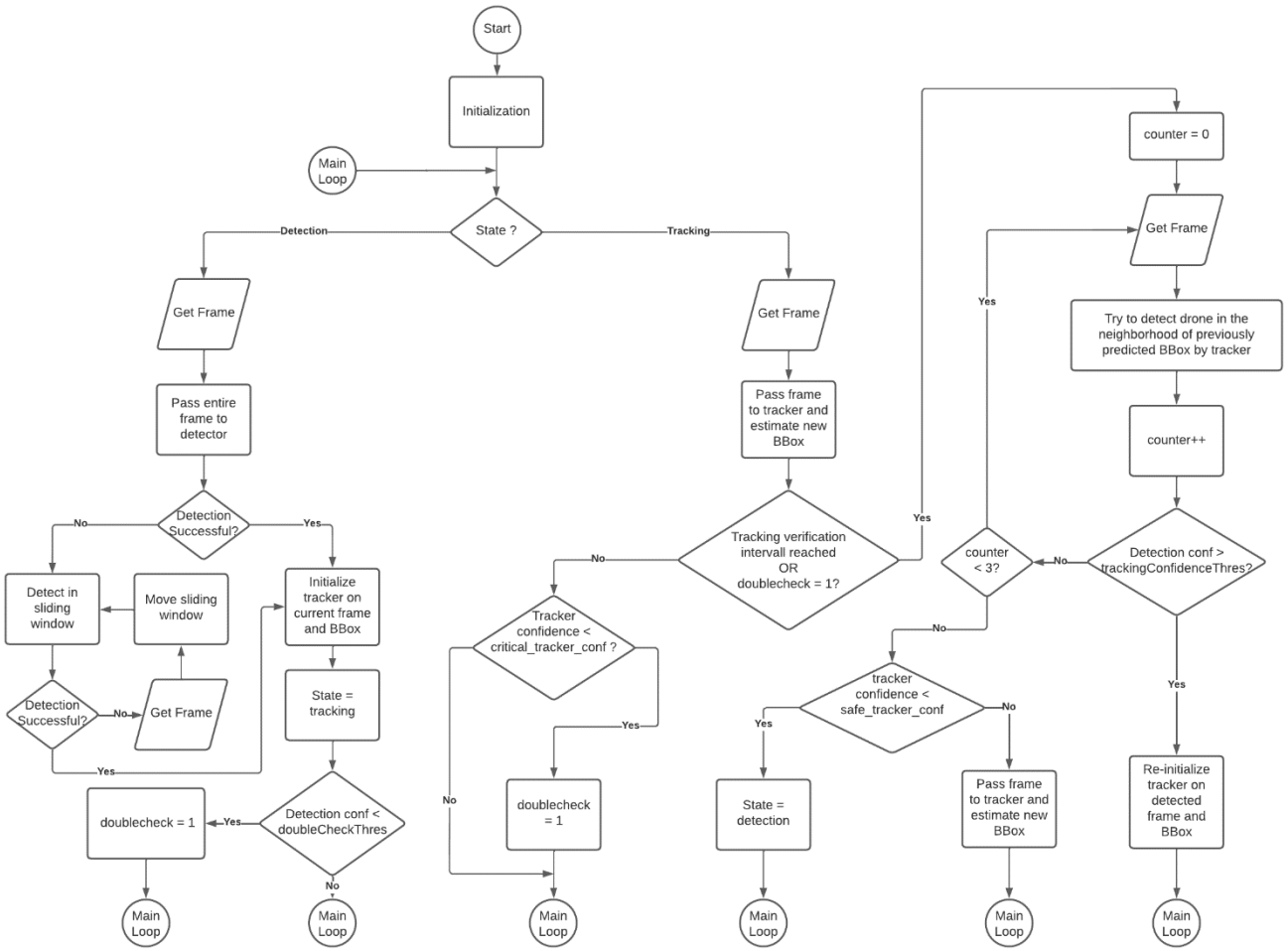

Figure 6: Flowchart illustrating the detector and tracker fusion methodology

A. Detection and Tracking Fusion Implementation

The full detection and tracking pipeline starts by initializing a series of parameters for the two algorithms and their networks are loaded into computer memory. First, YOLOv5 detector attempts on an entire frame of which the resolution has been scaled down to fit the specified YOLO input image size. This allows to detect the drone if it is very close to the camera. If no drone is detected, the detection process in the following frames is continued in a smaller window, keeping original resolution, and sliding over the entire image. Performing detection on small images allows to limit detection delay, which provides more recent bounding boxes to the tracker. If a drone is detected, the tracker is initialized on the current frame and given bounding box. The system state is switched to tracking. During tracking process, the detection continues to run in the background on the window contains the current tracking bounding box to double check the result of the tracking algorithm. In case of tracking failure, the detector will attempt to re-detect the drone and re-initiate the tracking process. A flowchart of the detection and tracking pipeline is illustrated in the Figure 6.

B. Position Estimation Methodology

To estimate the position of the target drone, 3D triangulation method is used with a stereo camera set-up. Once the drone's center has been detected in both video feeds of the stereo camera (see Figure 9), a maximum likelihood 3D position can be triangulated using the camera projection matrices and the 2D point correspondence in both images.

A MATLAB script was written which will take the bounding boxes result of the detection and tracking pipeline on two images of the stereo camera and their intrinsic and extrinsic parameters to calculate the 3D location of the target drone with respect to the stereo camera.

C. Maritime Environment Simulation

A realistic maritime environment simulation was created using Unreal Game Engine, which includes an accurate 3D model of the DJI Matrice 300 drone, a ship model with buoyancy movement, realistic sea environment with waves and sky containing clouds. One simulated stereo camera is placed on the afterdeck of the ship, with a baseline of 2m long. The camera resolution can be selected between 4K (4096x3072) and HD (1280x960), this option is used to analyze the relation between resolution and detection range. The origin of the coordinate system is the optical center of the right camera (Figure 7). The position of the drone with respect to this origin is recorded as the ground truth for the position estimation evaluation.

The simulation includes two programmed drone trajectories (see Figure 8) and at two velocities: 5m/s and 1m/s. Figure 7 and 9 show the field of view of both cameras of the stereo setup, and the bounding box result of the detection and tracking pipeline.

Figure 7: Afterdeck area of the ship model, with stereo camera (blue - on the top image) and their respective images (bottom left and right)

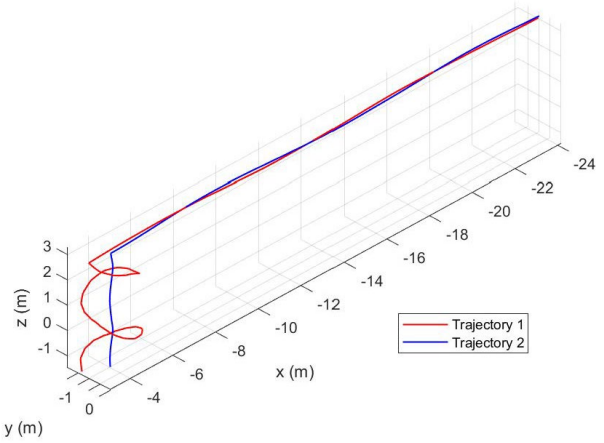

Figure 8: Two programmed drone trajectories in the simulated environment – the origin of the coordinate system is the optical center of the right camera

Figure 9: Example of the drone center estimation

D. Position Estimation Results

Absolute Trajectory Errors (ATE) is the accuracy metric commonly used in SLAM (Simultaneous Localization and Mapping) for the evaluation of a reconstructed trajectory based on a ground truth trajectory [15]. The ATE computes for every triangulated point the Euclidean distance to the corresponding ground truth point (see Figure 10). The mean and standard deviation of the ATE over all points of the trajectory are used as performance metrics to evaluate the quality of the reconstructed trajectory.

Comparison between different trajectories and drone velocities: Mean and Standard Deviation of the ATE for trajectory 1 and 2 in two different drone velocities (1m/s and 5m/s) using HD stereo camera is shown in Table 5. The ATE does not change much from different trajectories, however, the drone speed is found to have direct impact on the systems accuracy due to tracking failure at higher drone velocity.

TABLE 5: MEAN AND STANDARD DEVIATION OF THE ATE FOR DIFFERENT TRAJECTORIES AND DRONE VELOCITIES.

	Trajectory 1		**Trajectory 2**	
	v = 5m/s	*v = 1m/s*	*v = 5m/s*	*v = 1m/s*
ATE Mean (cm)	17.20	4.57	14.44	4.03
ATE Standard Deviation (cm)	5.26	1.71	5.14	1.86

Detection range: the resolution of the camera and the size of the sliding window used during the detection phase can affect the detection range. Using HD camera with large sliding window of 640x640 pixels provides the detection range around 9.5m. Reducing the sliding window to 384x384 pixels increases the detection range to 21m. Using 4K camera with sliding window of 640x640 pixels can achieve the detection range above 100m.

Tracking failure: in the simulation scenarios, tracking failure occurs during the end of the trajectory, just before touchdown on the deck. Lighting and background conditions make it difficult to distinguish between the drone and the background. (See Figure 10). This can be improved by improving lighting condition, background contrast.

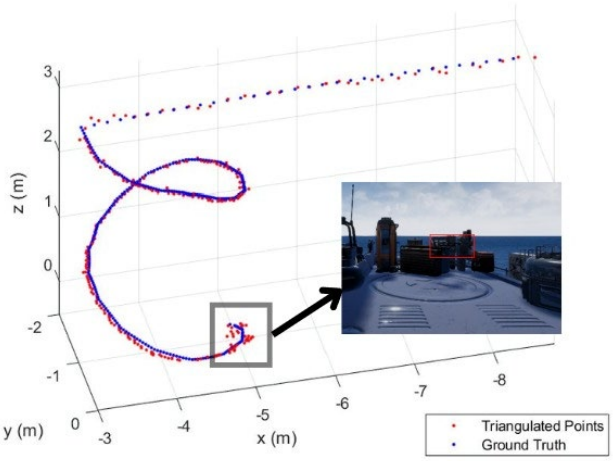

Figure 10: The triangulated and corresponding ground truth points for trajectory 1. Accuracy decreases at the end of the trajectory because tracking failure occurs due to difficulty in distinguishing drone and background

Accuracy of Position Estimation: depends on the camera resolution, distance of the target drone with respect to the camera, the relative velocity of the drone and the ship and configuration of the detection algorithm. In our analysis, the trajectory 1 with drone velocity (v = 1m/s), HD camera, and sliding window of 384x384 pixels during the detection phase are used. Figure 11 shows the histogram and Cumulative Distribution Function of the ATE of drone's position estimation during the approach and touchdown phase – less than 10m away from the camera. The largest error found in this case is below 22cm with over 90% below 10cm. The main contribution to the ATE are the estimation error in the x-axis, in which the drone performs its main movement, and is the optical axis of the right camera (see Figure 7). It follows by the errors in the z-axis. The errors in the y-axis are very low (see Figure 12). Figure 13 presents the total ATE and ATE from different axes in function of tracking frame number with the drone from far distance till touchdown on the ship deck. At the large distance, the ATE is high. And it will drop down

significantly when it gets closer to the camera, which is the most critical part of the landing phase.

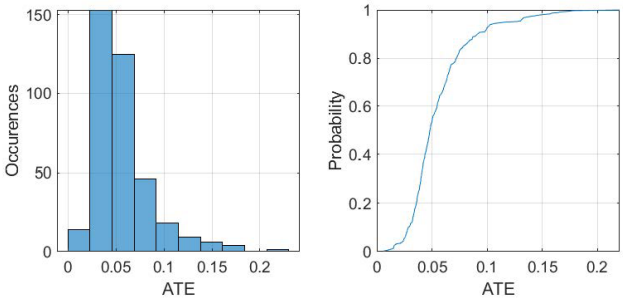

Figure 11: Absolute Trajectory Error histogram and Cumulative Distribution Function during landing phase in trajectory 1

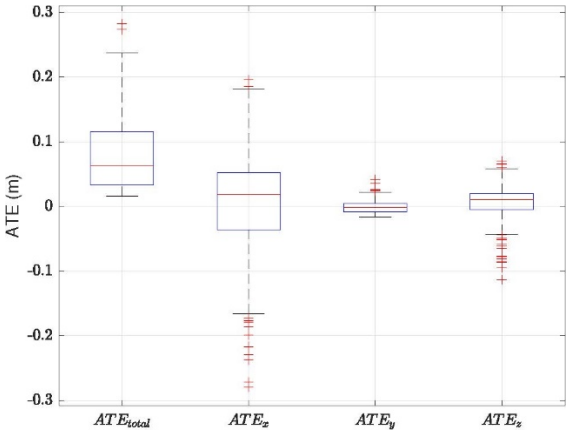

Figure 12: Box plots of total ATE and ATE in different axes during landing phase in trajectory 1

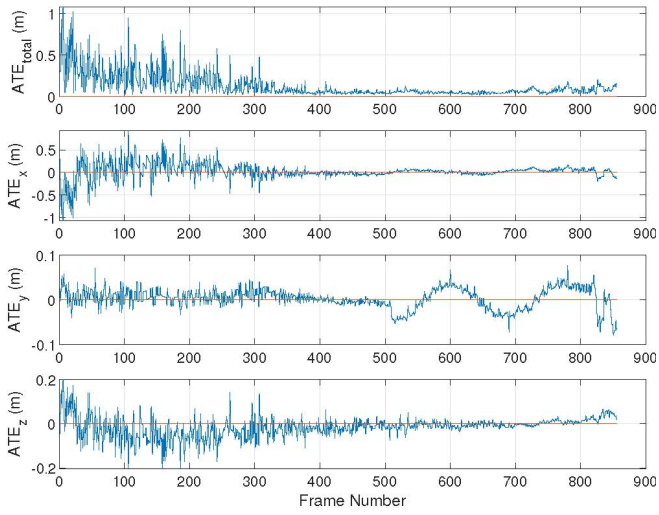

Figure 13: Total ATE and ATE from different axes in function of frame number for trajectory 1 – from starting tracking to the drone touchdown

V. Discussion And Conclusion

The study and implementation of deep learning method for visual detection and tracking of a subject drone in maritime environment are introduced in this paper. A video data set of the subject drone was acquired during our field tests. The image data was extracted, processed, and annotated. Multiple state of the art detection and tracking algorithms were evaluated to select the most suitable detector and tracker to be used in the task, taking into count the limited computing resource. The selected algorithms: YOLOv5 detector and OceanPlus tracker were trained with our custom dataset and integrated as one single detection and tracking pipeline. A simple and effective method to estimate the position of the subject drone with respect to the tracking camera was introduced and validated in maritime environment simulation on Unreal Game Engine with ground truth data for quantitative analysis. Our proposed method can detect and track the target drone from up to 100m with 4K stereo camera and can estimate the position of the drone with less than 10cm error during the critical landing phase.

Due to the nature of visual system, our method depends on lighting conditions and visibility of the target drone, therefore, can't be used in all conditions. However, because of its cost-effective solution, it offers possibilities to combine with other positioning methods to increase the accuracy and reliability of the drone positioning system.

Future work will focus on implementation of our method on the real-life tests, integration with other methods to provide better positioning estimation accuracy and redundancy, and establishing a communication channel with the drone to guide it during its autonomous landing task.

Acknowledgment

We thank two drone pilots: Fabian Filée and Alexander Borghgraef for their participation in our field tests. Without their support, establishing the dataset would not have been possible. We also thank personnel in Damage Control Center Military Domain in Beernem (Belgium) and colleagues in Department of Communications, Information, Systems and Sensors (CISS) – Royal Military Belgium for organizing the field tests in February 2021.

References

[1] Y. Xiao, Z. Tian, J. Yu, Y. Zhang, S. Liu, S. Du, and X. Lan, "A review of object detection based on deep learning," Multimedia Tools and Applications, vol. 79, June 2020, pp. 23729–23791.

[2] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 159, Jan. 2020, pp. 296–307.

[3] S. Samaras, E. Diamantidou, D. Ataloglou, N. Sakellariou, A. Vafeiadis, V. Magoulianitis, A. Lalas, A. Dimou, D. Zarpalas, K. Votis, P. Daras, and D. Tzovaras, "Deep learning on multi sensor data for counter UAV applications—a systematic review," Sensors, vol. 19, Nov. 2019, p. 4837.

[4] U. Seidaliyeva, D. Akhmetov, L. Ilipbayeva, and E. T. Matson, "Real-time and accurate drone detection in a video with a static background," Sensors, vol. 20, July 2020, p. 3856.

[5] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," CoRR, vol. abs/1506.01497, 2015.

[6] Y. Zhang, X. Li, F. Wang, B. Wei and L. Li, "A Comprehensive Review of One-stage Networks for Object Detection," 2021 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), 2021.

[7] G. Jocher, "ultralytics/yolov5." https://github.com/ultralytics/yolov5. Accessed: 02.04.2021.

[8] M. Karthi, V. Muthulakshmi, R. Priscilla, P. Praveen and K. Vanisri, "Evolution of YOLO-V5 Algorithm for Object Detection: Automated Detection of Library Books and Performace validation of Dataset," 2021 IEEE ICSES, 2021, pp. 1-6

[9] M. Kristan, J. Matas, A. Leonardis, T. Vojir, R. Pflugfelder, G. Fernandez, G. Nebehay, F. Porikli, and L. Čehovin, "A novel performance evaluation methodology for single-target trackers," IEEE

Transactions on Pattern Analysis and Machine Intelligence, vol. 38, pp. 2137–2155, Nov 2016

[10] D. Held, S. Thrun, S. Savarese, “Learning to Track at 100 FPS with Deep Regression Networks,” CoRR, vol. abs/1604.01802, 2016.

[11] Z. Zhang, H. Peng, J. Fu, B. Li, and W. Hu, “Ocean: Object-aware anchor-free tracking,” European Conference on Computer Vision (ECCV), 2020

[12] B. Yan, X. Zhang, D. Wang, H. Lu and X. Yang, “Alpha-refine: Boosting tracking performance by precise bounding box estimation,” IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 5289-5298.

[13] L. Cehovin, A. Leonardis, and M. Kristan, “Visual object tracking performance measures revisited,” CoRR, vol. abs/1502.05803, 2015.

[14] C. Hamesse, H. Luong, and R. Haelterman, “Evaluating the impact of head motion on monocular visual odometry with synthetic data,” in Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP), Vol 5, Online, 2022, pp. 836–843.

[15] D. Prokhorov, D. Zhukov, O. Barinova, K. Anton, and A. Vorontsova, “Measuring robustness of visual SLAM,” 16th International Conference on Machine Vision Applications (MVA), 2019.