UAV AI-based Visual Detection for Remote Archaeological Heritage Site Surveillance

Bernardo Teixeira, André Moura, José Antunes, André Dias, Hugo Silva
INESCTEC Porto, Portugal
bernardo.g.teixeira@inesctec.pt, andre.f.moura@inesctec.pt, jose.f.silva@inesctec.pt, andre.dias@inesctec.pt, hugo.m.silva@inesctec.pt

***Abstract*—Surveillance of all archaeological and culturally relevant sites can be a tremendous endeavour for most local authorities, which are frequently understaffed to prevent vandalism or plunder activities in or around all those sites which are under their purview.**

As such, the need for employing robust, persistent autonomous robotic solutions in the scope of preemptive surveillance is becoming very evident. The use of UAVs, equipped with different types of sensor payloads, is key to combat the destruction of our common cultural heritage, as they can provide constant surveillance at a fraction of the cost of a human resource.

The SHIELD project framework allowed for the collection of a large set of aerial surveillance footage consisting of both RGB and Thermal imagery that will form the basis for the development of future object detection algorithms. In this paper, we layout the dataset construction and present a first approach to the aerial detection of people and vehicles using a learning-based object detector, as well as a visual tracking algorithm.

***Index Terms*—Deep Learning, Robotic Perception, Object Detection**

I. Introduction

Archaeological sites in remote areas, even those classified as UNESCO world heritage sites, are subject to abandonment and at mercy of vandalism and/or plunder activities [1]. To cope with this problem, law enforcement agencies require the use of more automated means of surveillance, that can preemptively detect and alert to the presence of intruders and/or illegal activities. This is particularly relevant at night, which is when most vandalism and plundering activities occur and human resources are most scarcely available. The development of such systems is thus an urgent necessity around globe, especially in countries where the amount of archaeological and/our culturally relevant sites is too large to have permanent physical human surveillance.

Unmanned Aerial Vehicles are becoming ubiquitous in today's aerial surveillance applications. They can provide an enhanced area coverage compared to ground surveillance systems and therefore are a key component nowadays in law enforcement applications.

Following the framework of the SHIELD project [1], a UAV (i.e. a drone) equipped with a gimbal with visual and infrared imaging capabilities, was developed as a potential solution to provide surveillance capabilities to remote archaeological sites in Cyprus. The drone used AI-based visual detection solutions for detecting and classifying intruders, e.g., persons, cars, and possible plunder activities such as night excavations.

Fig. 1. Shield surveillance UAV.

Our AI-based approach utilizes a learning-based object detection scheme and object tracking framework. This research topic is currently a very active topic in the AI and robotics communities and different solutions have surfaced in the past years.

II. State-of-the-Art

Before deep learning took off in 2013, almost all object detection was done through classical vision-based and machine learning techniques. Common ones included SIFT (scale-invariant feature transforms) [2], and histogram of oriented gradients [3].

A. Object Detection

Those early approaches would detect a number of common salient features across the image, such as corners and edges and classify their most prominent clusters using some sort of ML algorith such as logistic regression, color histograms, or random forests. In the present, deep learning-based techniques vastly outperform these, and are commonplace in most high-performance object detection systems.

Deep learning-based approaches use neural network architectures like RetinaNet [4], YOLO (You Only Look Once)[5], SSD (Single Shot Multibox detector) [6], which are single-stage object detectors. Region proposals algorithms, or two stage detectors like R-CNN[7] or Fast-RCNN [8], work by first extracting ROIs (Regions of interest), and then classify and regress their class labels.

B. Object Tracking

Object tracking aims at estimating not only the bounding boxes but also the individual identity of each object in video sequence. It takes in a set of initial object detections, develops a visual model for the objects, and tracks the objects as they move around. Furthermore, object tracking enables us to assign a unique ID to each tracked object, making it possible for us to count unique objects in a video.

Simple Online And Realtime Tracking (SORT)[9] is a seminal example of an object tracking algorithm. SORT uses the position and size of the bounding boxes for both motion estimation and data association through a sequence of frames.The IOU metric and the Hungarian algorithm [10] are utilized for choosing the optimum box association.

Despite achieving overall good performance in most scenarios, in terms of tracking precision and accuracy, SORT suffers from a high number of identity switches and can struggle in to deal with for example occlusions. To overcome these limitations DeepSORT[11] replaces the association metric with a more informed metric that combines motion and appearance information.

III. Dataset Construction

In MONET [12], the construction of a large LWIR dataset was presented as a tool to study the problem of object localisation and behaviour understanding of targets undergoing large-scale variations and being recorded from different and moving viewpoints. We are proposing an extension of the current available data to include also RGB imagery in addition to LWIR data and robotic system metadata , properly synchronized and independently annotated. Excluding night-time scenarios, where visual manual annotation is not feasible, a total of roughly 12k RGB-LWIR image pairs are made available to aid the development of future multimodal perception algorithms. We refer to the MONET paper for more information about data structure and formats.

Fig. 2. Example of annotated RBG image: class people in green, class vehicle in red

IV. Implementation

In Robotic perception development, is is critical to employ strategies that amount to a good performance/timeliness balance. It is of no use to have a high accuracy system that is not able to deliver predictions at a frame-rate that properly accompanies the robot dynamics. Furthermore it is also worth taking into account that robotic systems typically are not equipped with powerful GPU hardware and thus we are limited in choosing the object detection framework to employ.

A. Object Detection

Taking all the aforementioned information into regard, we chose the YoloV7 [13] framework as our object detector. As was to be expected, due to the singular nature of the data, captured at high altitude with varying moving viewpoints, the pretrained model on ImageNet was not enough for our use case. As such, we performed a transfer-learning strategy, finetuning the model to our data type, using a train-test-split across our dataset of 50-40-10 %.

In addition, a data augmentation pipeline was introduced in the framework, making use of rotations, translations, shifts and the introduction of gaussian blur samples. All in all, the volume of training data was augmented fivefold, adding robustness to the visual detection system.

It is also worth noting that we have two different scenarios in our data, namely ”dirtroad” and ”runway” data. This data was evenly split across our training split.

Fig. 3. Illustration of model prediction of vehicle instance

Evaluation of the system in the testing split reached a 82% Mean Average Precision result, with almost the error coming from missed or erroneous people detections, with close to perfect vehicle detections. The most reasonable explanation is that people bounding boxes are mostly small to tiny targets representing just a handful of pixels, and thus prone to higher percentual errors when calcuting the bounding box error rate.

B. Object Tracking

At this stage, the next logical step to construct a functional object tracker was to develop a bounding box association scheme, that can track object detections through time and keep unique object identifiers. To achieve this, a version of the Hungarian Algorithm Association was employed. The Hungarian algorithm works by minimizing a cost matrix between subsequent image’s object detections. Typically, this matrix is constructed by computing the IoU cost for all predicted

TABLE I
BENCHMARK AVERAGE PROCESSSING SPEED IN MILLISECONDS (MS)

	Inference	Non-Maximum Supression	Topic Publication	Total
Ryzen 7	294.061 ±12.551	0.285 ±0.451	1.232 ±0.257	295.579
RTX 3060	13.130 ±0.002	0.266 ±0.230	1.130 ±0.380	14.525
Nvidia Jetson (15W)	134.503 ±2.894	1.993 ±1.452	2.791 ±2.120	139.256
Nvidia Jetson (30W)	84.720 ±5.997	1.372 ±1.038	2.044 ±1.534	88.137
Nvidia Jetson (50W)	37.270 ±1.562	1.169 ±1.051	3.240 ±1.601	41.680
Nvidia Jetson (MaxN)	26.050 ±1.295	0.869 ±0.756	2.483 ±1.268	29.402

matches and minimizing to find the beast association between detections. Detections are only accepted as matches if their IoU cost is bigger than a threshold that can be tuned to specific use cases.

In our implementation, this cost matrix consist not only of an IoU cost but a coupound sum between 3 similarity cost functions:

1) IoU cost: The standard formulation typically employed.
2) Linear Sanchez-Matilla cost [14]
3) Exponential Yu Cost [15]

The combination of the three provides a more smooth and robust association, that will in term contribute to better results and less dropped trackers. ID switches also occur less frequently.

The tracking system is still able to achieve real-time performance in a normal computer, running smoothly over video sequences, which was always an essential requisite for the system. The system keeps track of individual ID's until they are not present in the image (see illustration on fig.4).

Fig. 4. Example ID instances in people detections

C. Real-time performance on embedded devices

In order to establish the suitability of the developed algorithms for real time archaeological surveillance , an assessment of the real-time performance was needed. Accounting for the computing requirements needed for Deep Learning-based perception, the adoption of a GPU embedded device became necessary. The Nvidia Jetson Orin platform was chosen as the subject of our experiments, as it is currently the standard for real-time performance on embedded GPU devices for Deep Learning apllications.

In order to understand our requirements for the integration of the embedded GPU platform, we ran a hardware-on-the-loop simulation using the Nvidia Jetson Orin Development Kit and the real-world data collected on our field mission. Our object detection pipeline was deployed on ROS and processing speeds were logged throughout the pipeline as can be observed in table I.

The Nvidia Jetson Orin Development Kit has 4 different power settings, which allows us to understand how to project the requirements for future integration in Robotic applications. Table I shows the real-time performance of the object Detection algorithm, comparing between laptop hardware (both CPU and discrete GPU) and the different Jetson power settings. Significant performance gains can be observed relative to CPU-only implementation, thus reinforcing the idea that embedded dedicated GPU's massively improve on-board processing for Deep Learning real-time inference. This conclusion holds truth even for lower power settings, including 15W PoE connection.

V. CONCLUSIONS

Even though the algorithm is its early stages of development and shall still be finetuned to unlock greater levels of performance, we show the applicability and suitability of visual object detection and tracking of people and vehicle targets in aerial imagery captured via UAV.

We show that even though the detection of such small targets at such a great altitude is not an easy feat for an object detection algorithm, we are able to achieve an accuracy of roughly 82 % using a finetuned YOLO algorithm.

From there, a Hungarian algorithm approach was leveraged for the task of association of visual detections and tracking over time.

VI. FUTURE WORK

Implementing data fusion from different sensors, in this case making use of both RGB and LWIR collected data, could massively improve the accuracy and robustness of our object detection algorithms. Which is particularly beneficial for the case of UAVs equipped with different types sensors, where diverse information about the environment is being captured simultaneously and synchronously, enabling a more comprehensive environment perception and thus object detection. We intent to pursue such a strategy in the near future, so as to enhance the perception capabilities of the SHIELD surveillance UAV.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Rui Mendes and Francisco Neves for their laborious effort in collaborating with the annotation process of the dataset. This paper is funded by the European Union's Joint Programming Initiative – Cultural Heritage, Conservation, protection and use Joint Call - JPICH-0085, and by the Portuguese Science Foundation under the grant JPICH/0001/2019. The authors also acknowledge DEEPFIELD project, funded by the European Commission under the H2020 EU framework programme for research and Innovation with project h2020-WIDESPREAD-2018-3,857339. Corresponding author funded through national funds under Fundação para a Ciência e Tecnologia (FCT) Ph.D. Grant 2020.05052.BD.

REFERENCES

[1] URL: https://www.inesctec.pt/en/projects/shield#technical_sheet.

[2] David G Lowe. "Distinctive image features from scale-invariant keypoints". In: *International journal of computer vision* 60 (2004), pp. 91–110.

[3] N. Dalal and B. Triggs. "Histograms of oriented gradients for human detection". In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol. 1. 2005, 886–893 vol. 1. DOI: 10.1109/CVPR.2005.177.

[4] Tsung-Yi Lin et al. "Focal loss for dense object detection". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2980–2988.

[5] Joseph Redmon et al. "You only look once: Unified, real-time object detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 779–788.

[6] Wei Liu et al. "Ssd: Single shot multibox detector". In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer. 2016, pp. 21–37.

[7] Ross Girshick et al. "Rich feature hierarchies for accurate object detection and semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 580–587.

[8] Ross Girshick. "Fast r-cnn". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1440–1448.

[9] Alex Bewley et al. "Simple online and realtime tracking". In: *2016 IEEE international conference on image processing (ICIP)*. IEEE. 2016, pp. 3464–3468.

[10] Harold W Kuhn. "The Hungarian method for the assignment problem". In: *Naval research logistics quarterly* 2.1-2 (1955), pp. 83–97.

[11] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. "Simple online and realtime tracking with a deep association metric". In: *2017 IEEE international conference on image processing (ICIP)*. IEEE. 2017, pp. 3645–3649.

[12] Luigi Riz et al. "The MONET dataset: Multimodal drone thermal dataset recorded in rural scenarios". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 2545–2553.

[13] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 7464–7475.

[14] Ricardo Sanchez-Matilla, Fabio Poiesi, and Andrea Cavallaro. "Online multi-target tracking with strong and weak detections". In: *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II 14*. Springer. 2016, pp. 84–99.

[15] Fengwei Yu et al. "Poi: Multiple object tracking with high performance detection and appearance feature". In: *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II 14*. Springer. 2016, pp. 36–42.