SHIELD: Safeguard Heritage In Endangered Looted Districts

Mohamed Lamine Mekhalfi, Nicola Saljoughi, Davide Boscaini, Fabio Poiesi
Technologies of Vision, Fondazione Bruno Kessler
<mmekhalfi, nsaljoughi, dboscaini, poiesi>@fbk.eu

Abstract

Unmanned aerial surveillance is key for the protection of archaeological sites against looting activities. Automated surveillance is a valuable decision-making means that can become of great support to public authorities. The European project Safeguard Heritage In Endangered Looted Districts (SHIELD) aims to developing a fully autonomous surveillance system that involves a drone capable of taking off and surveying a scene of interest (i.e., detecting/tracking objects, monitoring suspicious actions), as well as landing and charging at its smart helipad. In this extended abstract we focus on SHIELD perception capabilities, in particular on the the object detection module that is based on the state-of-the-art CenterNet algorithm. Preliminary results on thermal infrared data, from the BIRDSAI dataset, show promising results that could serve as a build-up of future improvements[†].

1 Introduction

Archaeological sites are an important asset that is often subject to natural and/or human-made deterioration. SHIELD is a project that aims at designing and building an artificially intelligent Unmanned Aerial System (UAS) to patrol archaeological and heritage sites in order to identify looting (e.g., illegal excavations) activities in real-time. This UAS will be designed so as to meet several criteria. For instance, it should be able to automatically take off and land to carry out scheduled surveying missions and to recharge its batteries when they are down. Thus, SHIELD involves Sensors and Imaging for the automatic acquisition part and Artificial Intelligence (AI) for the object detection/tracking part.

SHIELD's AI modules mainly exploits deep learning solutions to detect objects like humans and vehicles. Typical detectors from the literature make use of region proposals or anchor boxes to locate and determine objects. Recent trends address this downsides by modelling objects as a set of keypoints, such as CornerNet [Law and Deng, 2018], which detects corners of an object's bounding box, and ExtremeNet which detects keypoints at all four angles [Zhou *et al.*, 2019b], supplemented with a central keypoint. Despite their on-par performance with previous pipelines, they require post-keypoint grouping operations as they detect multiple keypoints. CenterNet is another appearance-based detector that relaxes the problem to a single keypoint per object (i.e., located at the centre of the bounding box), discarding exhaustive grouping stages [Zhou *et al.*, 2019a].

In view of object detection in images acquired via an unmanned aerial vehicle (UAV), the resolution of acquisition sensors is a strong factor to take into account, which was studied in previous works [Seifert *et al.*, 2019]. Nevertheless, even when high resolution sensors are accessible, the altitude of the UAV on which the sensor is mounted has a major influence on object size [Seifert *et al.*, 2019]. Camera orientation is another component that not only affects object size, but may also reveals a largely distinct appearance of the target. Furthermore, previous works have demonstrated plausible performance on object detection and tracking in the case of still cameras, where the above challenges have been tackled to some extent [Zhu *et al.*, 2020]. However, when the acquisition cameras are in motion, traditional background-subtraction techniques remain limited due to many challenges such as camouflage, dynamic background, shadows, motion blur, illumination changes, amongst others [Chapel and Bouwmans, 2020]. The influence of these changes may multiply if the frame rate of the acquired videos is low. For instance, the difference in illumination between two consecutive frames is typically smaller at high frame rates.

In this regard, SHIELD attempts to tackle the above challenges through a multimodal framework that incorporates UAV flight information (e.g., altitude, orientation, viewpoint) along with optical images for robust object detection. Further, since looting behaviours may take place during daytime and/or nighttime, it is necessary to avail data that was acquired in both scenarios. Evidently, daytime object detection and tracking can be addressed via RGB streams. Nighttime object detection, however, normally requires the adoption of thermal sensors, especially in poor lighting conditions. This is particularly challenging for SHIELD due to the scarcity of public datasets that satisfy the above conditions (i.e., RGB and thermal images along with flight metadata). The following subsection conducts a brief narrative of existing datasets.

[†]The SHIELD project (http://shield.cyi.ac.cy) is funded by the European Union's Joint Programming Initiative – Cultural Heritage, Conservation, Protection and Use joint call.

1.1 UAV datasets

Several UAV datasets can be found in the literature. For instance, The CARPK [Hsieh *et al.*, 2017] dataset contains about 90,000 cars from 4 parking lots with drone-view at an altitude of approximately 40 meters. The Campus [Robicquet *et al.*, 2016] dataset collects images and videos of various types of agents that navigate in a university campus, and is constituted by 100 sequences containing about 900,000 annotated bounding boxes. As in the CARPK dataset, camera view is fixed and the frames have been acquired at an altitude of 80 meters. UAV 123 [Mueller *et al.*, 2016] is a dataset for low altitude UAV target tracking, which consists of 123 video sequences acquired with varying flight attitude, ranging between 5 and 25 meters. VisDrone [Zhu *et al.*, 2020] dataset consists of 400 videos formed by 265,228 frames and 10,209 static images covering a wide variety of locations, environment, objects, density and in presence of challenging weather and lighting conditions. The video sequences are about 400 for a total of more than 2.6 million annotated bounding boxes. DTB70 [Kiani Galoogahi *et al.*, 2017] is constructed through the collection of 70 video sequences mostly focusing on people and cars. UAVDT [Du *et al.*, 2018] contains 10 video sequences that make up to 80,000 frames with varying flying attitude, weather, light and environment. The acquisition altitude varies between 10 and more than 70 meters. However, the flight metadata are rather sparse (e.g., the flight altitude is expressed by a hot encoding that falls within three altitude classes, namely low, medium and high altitude). The AU-AIR [Bozcan and Kayacan, 2020] dataset consists of about 32,000 annotated RGB video frames captured by means of a drone flying at low altitude, where the frames are labelled with time, GPS, IMU, altitude and linear velocities of the UAV. By contrast to UAVDT, the metadata in AU-AIR consist in real values, which suggests a more robust modelling of object detection via drones.

With regards to UAV thermal image data, however, to the best of our knowledge there exists only the BIRDSAI dataset [Bondi *et al.*, 2020] so far, which provides both real and synthetic thermal infrared data of animals and humans in an outdoor environment in Southern Africa. Nevertheless, BIRDSAI does not provide flight metadata.

2 Methodology

The methodology of SHIELD involves the development of an autonomous, flexible, sustainable and scalable, aerial surveillance system that can identify illegal excavation in real-time in order to enable public authorities and police to take proper actions. In particular, SHIELD is designed to identify and characterise illegal excavations based on imaging techniques which will be used by competent local authorities for a first early response. This system will be deployed via a portable helipad, powered by a solar-power based renewable energy supply, where the drone will be (i) stored, (ii) able to take off for scheduled missions, (iii) land, (iv) automatically recharge and (v) download the data collected during the survey. The project involves Sensors and Imaging, Machine Learning (ML) and Artificial Intelligence (AI) techniques. The main challenges characterising this surveillance scenario are:

(a) variable target appearance as the camera moves with the drone in a 3D space;
(b) low-resolution images that are captured by the thermal camera;
(c) discrimination of the targets from background due to heat emitted by the ground (e.g. warm terrain/rocks);
(d) real-time requirement necessary to timely trigger alarms;
(e) low-power consumption required by the drone for on-board data processing;
(f) comprehensive data collection and annotation to train and validate AI algorithms.

The technical aspects we deem important to investigate involve the generalisation of the AI algorithms to different scenarios/domains, such as flight configurations (camera orientation, altitude), acquisition sensors (RGB, thermal, zoom) and scenes (background structure, target types). Because we use data-driven methods, diversity of the collected data is key. Ideally, data should cover all possible cases, scenarios, variability of the target appearance. However, edge cases are highly likely to occur.

For object detection, we opt for CenterNet [Zhou *et al.*, 2019a] as motivated above. In practice, CenterNet feeds a given query image to a fully convolutional network to produce a heatmap, where the peaks pertain to object centers and the features at each peak predict object bounding box size. Thus, given an image $I \in R^{W \times H \times 3}$, for each ground-truth keypoint $p \in R^2$ of class c, a low-resolution equivalent $\tilde{p} = \frac{p}{r}$ where r is the output stride, is envisioned. Afterwards, a heatmap $Y \in [0,1]^{\frac{W}{r} \times \frac{H}{r} \times C}$, where C is the number of object classes, is obtained by splatting all ground-truth keypoints via a Gaussian kernel. The training loss follows a pixel-wise logistic regression:

$$L_k = -\frac{1}{N} \sum_{xyc} \begin{cases} (1-\hat{Y}_{xyc})^{\alpha} log(\hat{Y}_{xyc}), \\ \text{if } Y_{xyc} = 1 \\ (1-Y_{xyc})^{\beta} (\hat{Y}_{xyc})^{\alpha} log(1-\hat{Y}_{xyc}), \\ \text{otherwise} \end{cases} \tag{1}$$

where α and β are hyper-parameters of the focal loss [Zhou *et al.*, 2019a], and N is the number of keypoints in the image. To compensate the discretization error due to the output stride, a local offset $\hat{O} \in R^{\frac{W}{r} \times \frac{H}{r} \times 2}$ is predicted for each keypoint according to an L_1 loss:

$$L_{\text{off}} = \frac{1}{N} \sum_{p} \left| \hat{O}_{\tilde{p}} - (\frac{p}{R} - \tilde{p}) \right|. \tag{2}$$

To predict object size, a single size prediction $\hat{S} \in R^{\frac{W}{r} \times \frac{H}{r} \times 2}$ for all object classes is used, an L_1 loss is calculated as

$$L_s = \frac{1}{N} \sum_{k=1}^{N} \left| \hat{S}_{pk} - s_k \right|, \tag{3}$$

where s_k is the object size. The final detection loss is

$$L_{\text{det}} = L_k + L_{\text{off}} + \lambda_s L_s, \tag{4}$$

Table 1: Preliminary results on the BIRDSAI dataset [Bondi *et al.*, 2020]. (left) human. (right) animals. Green bounding boxes are the ground truth, blue bounding boxes are the detector estimations.

P	R	FS	AP (Animal)	AP (Human)	mAP
0.73	0.64	0.68	72.77	6.08	39.42

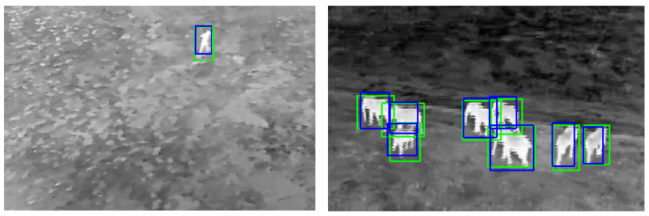

Figure 1: Detection examples from BIRDSAI dataset. Left: Human class. Right: Animal class. Green and blue bounding boxes pertain to the ground-truth and inference, respectively.

where λ_s is set to 0.1. It is to note that a single network is used to predict the aforementioned keypoints, offset and size. To predict bounding box locations, the coordinates of the 100 peaks across the heatmap that are greater or equal to their immediate neighbors are used, and the keypoint responses within the heatmap are considered as a measure of their detection confidence.

3 Preliminary results

We provide preliminary results obtained by training CenterNet on the BIRDSAI dataset [Bondi *et al.*, 2020], which amounts to 100K Animal bounding boxes and 34K Human bounding boxes. The images were captured throughout protected regions in the countries of South Africa, Malawi, and Zimbabwe using a battery-powered fixed-wing UAV. All flights took place at night and the altitude ranged from approximately 60 to 120m, and flight speed ranged from 12 to 16 m/s depending on conditions such as wind.

We adopt the pre-trained Resnet-50 as backbone. We use the Adam optimiser and an initial learning rate of 2e-04, which is scheduled following a cosine annealing. The threshold of intersection over union is set to 0.2 and that of the detection confidence is set to 0.3.

Table 1 shows the results in terms of Precision (P), Recall (R), F-Score (FS), Average Precision (AP) per class, and Mean Average Precision (mAP).

The results indicate a plausible performance in terms of FS and mAP. In terms of AP, however, the network scores far better on the Animal class w.r.t Human class, which may be traced back to two reasons. The first one is the high imbalance of the dataset, where the number of Animal samples largely outnumbers that of Human ones. The second one refers to the size of Animal objects which is mostly larger than that of Human. A detection example is given in Figure. 1.

4 Conclusions

This paper briefly described the SHIELD project and laid out preliminary results. Although the obtained object detection scores are encouraging, we believe that further improvement is possible. The current research line of SHIELD investigates the use of flight metadata as a supplementary information, besides the input images, in order to render CenterNet more robust against altitude, viewpoint, and orientation changes.

References

[Bondi *et al.*, 2020] E. Bondi, R. Jain, P. Aggrawal, S. Anand, R. Hannaford, A. Kapoor, J. Piavis, S. Shah, L. Joppa, B. Dilkina, and M. Tambe. BIRDSAI: A dataset for detection and tracking in aerial thermal infrared videos. In *Proc. of the IEEE WACV*, 2020.

[Bozcan and Kayacan, 2020] I. Bozcan and E. Kayacan. AU-AIR: A Multi-modal Unmanned Aerial Vehicle Dataset for Low Altitude Traffic Surveillance. *arXiv:2001.11737*, 2020.

[Chapel and Bouwmans, 2020] M.N. Chapel and T. Bouwmans. Moving objects detection with a moving camera: A comprehensive review. *Computer Science Review*, 2020.

[Du *et al.*, 2018] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, and Q. Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proc. of the IEEE ICCV*, 2018.

[Hsieh *et al.*, 2017] M.R. Hsieh, Y.L. Lin, and W.H. Hsu. Drone-based object counting by spatially regularized regional proposal network. In *Proc. of IEEE ICCV*, 2017.

[Kiani Galoogahi *et al.*, 2017] H. Kiani Galoogahi, A. Fagg, C. Huang, D. Ramanan, and S. Lucey. Need for speed: A benchmark for higher frame rate object tracking. In *Proc. of IEEE ICCV*, 2017.

[Law and Deng, 2018] H. Law and J. Deng. CornerNet: Detecting objects as paired keypoints. In *Proc. of IEEE CVPR*, 2018.

[Mueller *et al.*, 2016] M. Mueller, N. Smith, and B. Ghanem. A benchmark and simulator for uav tracking. In *Proc. of ECCV*, 2016.

[Robicquet *et al.*, 2016] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *Proc. of ECCV*, 2016.

[Seifert *et al.*, 2019] E. Seifert, S. Seifert, H. Vogt, D. Drew, J. Van Aardt, A. Kunneke, and T. Seifert. Influence of drone altitude, image overlap, and optical sensor resolution on multi-view reconstruction of forest images. *Remote Sensing*, 2019.

[Zhou *et al.*, 2019a] X. Zhou, D. Wang, and P. Krahenbuhl. Objects as points. *arXiv:1904.07850*, 2019.

[Zhou *et al.*, 2019b] X. Zhou, J. Zhuo, and P. Krahenbuhl. Bottom-up object detection by grouping extreme and center points. In *Proc. of IEEE CVPR*, 2019.

[Zhu *et al.*, 2020] P. Zhu, L. Wen, D. Du, X. Bian, Q. Hu, and H. Ling. Vision meets drones: Past, present and future. *arXiv:2001.06303*, 2020.