JDST., vol. 5, no. 5, p84-p102, 2022
https://doi.org/10.46713/jdst.05.005

www.jdst.eu, ISSN 2534-9805, e-ISSN 2534-9813

Research Article

Video-based Detection Algorithms in FOLDOUT: Through-Foliage Detection in Ground-Based Border Surveillance

David Schreiber [a], *Andreas Opitz* [a]

[a] *AIT Austrian Institute of Technology GmbH, Giefinggasse 4, Vienna 1210, Austria*
https://www.ait.ac.at

ABSTRACT

The FOLDOUT project is concerned with through-foliage detection, which is an unsolved important part of border surveillance. FOLDOUT builds a system that combines various sensors and technologies to tackle this problem. This paper reviews the work done by AIT in FOLDOUT concerning visual sensors (RGB and thermal) for through-foliage object detection. Through-foliage scenarios contain an unprecedented amount of occlusion, specifically fragmented occlusion (e.g., looking through the branches of a tree). It is demonstrated that current state-of-the-art detectors based on deep learning approaches perform inadequately under moderate to heavy fragmented occlusion. Various state-of-the-art and beyond state-of-the-art detection algorithms, based on deep learning as well as on other approaches, dealt within FOLDOUT to detect objects in the case of fragmented occlusion, are presented, discussed, and compared.

ARTICLE INFO

RECEIVED: 09 Oct 2021
REVISED: 10 Nov 2021
ACCEPTED: 30 Nov 2021
ONLINE: 12 DEC 2021

KEYWORDS

AI, border surveillance, human detection, machine learning, fragmented occlusion, visual sensors

I. INTRODUCTION

Through-Foliage detection is an unsolved important part of border surveillance of remote areas between regular border crossing points. The FOLDOUT project [1] built a system that combines various sensors and technologies and fuses these into a robust detection platform [2]. In this paper, the work concerning detection

✉ Corresponding Author: Tel.: +43 505504282; E-mail: david.schreiber@ait.ac.at

algorithms on visual sensors is described. In particular, the challenge of fragmented occlusion (e.g., a person walking behind trees) is addressed. In contrast to the case of partial occlusion, detection under fragmented occlusion was not handled before in the literature of computer vision and is an unsolved challenge. Furthermore, there remains a lack of datasets and standard benchmarks on real-world surveillance problems that address natural environments such as forests and open land.

Although detectors in FOLDOUT were developed to detect several categories of objects, such as persons, vehicles, boats and animals, the focus was on detecting persons under fragmented occlusion. Consequently, in the paper we refer solely to person detection. Several state-of-the-art (SOTA) approaches based on Deep Neural Networks (DNNs) were developed to detect persons (and other objects) on videos of RGB and thermal cameras, under fragmented occlusion. Results indicate that state-of-the-art DNN detectors are unable to detect persons under moderate to severe fragmented occlusion, even when trained on data exhibiting fragmented occlusion [3]. Some improvement was gained, however, when the object detection task was relaxed to the simpler localisation task (without a bounding box and no classification), and temporal information was employed in training the DNN [4].

Within FOLDOUT, the PETS2021 workshop [5] was sponsored and organized (jointly by AIT and the university of Reading), held together with the AVSS (IEEE International Conference on Advanced Video and Signal-Based Surveillance) 2021 conference, on the topic: Through-foliage detection and tracking challenge and evaluation. The challenge comprises the publication of a novel annotated video surveillance dataset on through-foliage detection, addressing person detection and tracking in fragmented occlusion scenarios, and performance evaluation of challenge results submitted by six worldwide participants (including AIT). The evaluation shows that the problem of detection and tracking remains an open problem where difficult conditions are tackled such as in foliated environments.

AIT employed and tested video detections algorithms on various RGB and thermal sensors during the field demonstrations taking place within FOLDOUT in Bulgaria, Greece, and Finland, where actors were performing according to planned scenarios having various degrees of occlusion by vegetation. The main findings of the detection evaluation are: persons were detected successfully day and night, in ranges between 10 to 130 metres (the optics was setup to enable detections up to 150 meters); detection performance is degraded in the case of moderate to heavy fragmented occlusion, as well in case that the persons are crouching or walking/running while crouched, as typically detectors are trained on upright persons; improved detection results were obtained when fusing the visual sensors

with additional sensors (e.g., acoustic sensors), which is part of FOLDOUT as well [6].

In addition to DNN-based detection approaches, a more traditional approach was developed to detect moving objects in the case of heavy fragmented occlusion. The approach is based on background modeling [7], namely detecting moving objects by comparing the current frame of a video with the background model which is learned and adapted over time. A robust real-time background modeling algorithm was developed which is able, on one hand, to detect moving persons even when they are heavily occluded by foliage, while, on the other hand, being resilient against dynamic background such as moving foliage and lighting changes. The background-based detector was tested on RGB and thermal test videos recorded within FOLDOUT, especially comprising heavy fragmented occlusion, when only small fragments of the moving persons can be seen, and was demonstrated to compete favorably with DNN approaches.

The results indicate the inability of current state-of-the-art DNN-based detectors, to detect persons in case of considerable fragmented occlusion. Further research and development are needed to include temporal (motion) information in the architecture of DNNs, and to integrate the strength of both types of detectors (DNN-based and traditional motion-based) to achieve high true positive and low false positive rates.

II. State-of-the-Art of Deep Learning-Based Object Detection Under Partial Occlusion

Object detection is fundamental, and the longstanding computer vision problem has been a major active research area for a few decades [8]. It has been used in various computer vision applications such as surveillance, face detection, face recognition, pedestrian counting, security systems, vehicle detection, self-driving cars, etc. Some computer vision terms like object localization, classification, and recognition are interlinked to the object detection processing. Object classification defines the class (or category, e.g., a person, a car, …) of one or more objects that exist in the image and assigns the labels of the objects. Object localization is a process that locates the position of one or more objects in the image or video with the help of a bounding box. A combination of object localization and classification process is known as Object detection. A complete object recognition process takes an image as input, identifies the objects, assigns labels to the object of the

associated class, and gives the class probability of the recognized object [8] (see Figure 2, left).

Classification is used to predict the class from a given image region of interest. In this process, relevant features are extracted from the region of interest and combined to represent the object, compared with the trained model. Before the advancement of deep neural networks, traditional classifiers were employed. These classifiers were using pre-designed hand-crafted features. Traditional detectors had some limitations, such as a huge amount of detection proposals generated, from which many proposals were redundant. These redundant proposals resulted in many false positives (e.g., regions in the image not corresponding to persons which were erroneously classified as persons). A classifier needs to be trained offline on annotated datasets, i.e., manually marked regions in the training images, containing an object of interest from a class (positive examples) and regions which do not contain that such objects (negative examples). Bounding boxes are the most commonly used way to manually annotate training images defining the rectangular region of the objects in an image [8].

In the last decade or so, object detectors based on deep neural networks have become state-of-the-art. Artificial neural networks are computing systems inspired by biological neural networks. Such systems learn progressively to perform tasks by considering examples. A deep neural network is an artificial neural network with multiple layers between the input and output layers. There are different types of neural networks, but they always consist of the same components: neurons, synapses, weights, biases, and functions. One of the advantages of DNNs is their ability to extract internal representations automatically, outperforming hand-crafted features that are needed for traditional classifiers [8].

The development of deep neural networks takes a lot of time because it needs a huge amount of data (can be of the order of millions of examples) for training (setting all its internal parameters). Object detection techniques based on deep learning are divided into two general architecture types: two-stage detectors (e.g., R-CNN, Faster R-CNN) and one-stage detectors (such as YOLO and SSD). The two stage detectors (proposal generation then classification) employ two stages to detect the objects from the image, and these detectors often provide state-of-the-art results or high accuracy on available datasets. But these detectors have a lower inference speed as compared to one-stage detectors. One-stage detectors are mostly used in real-time object detection applications and provide comparable results much faster than two-stage detectors [8].

However, object detection is a challenging task due to many factors such as clutter, imaging conditions, large number of object categories and instances, and occlusion. Consequently, the ability of state-of-the-art deep learning detectors is far from that of human beings due to several factors, occlusion being one of them. Since occlusion can happen in various locations, scale, and ratio, it is very difficult to handle. With objects partially occluded in the scene (*partial occlusion*), as demonstrated in Figure 1 (right), both traditional detectors as well as deep neural network-based classifiers are less robust compared to humans. Therefore, handling of partial occlusion has been studied extensively, such as in the case of person detection [9].

Finally, the continuous development of object detection algorithms has generated the need for evaluation tools to quantify algorithm performance. Many evaluation metrics were proposed in the literature for quantifying different aspects of a detection algorithm's performance. However, the most basic and intuitive measures, which we will use in this paper, are the *true positive rate (TPR)* and *false positive rate (FPR)*. TPR is the percentage of bounding boxes of persons which are detected by the detector in the annotated test data. FPR is the percentage of erroneously detected bounding boxes where there is no person. In case of fragmented occlusion (to be discussed in the next Section), it is not always possible to mark a person via a bounding box, as the person is hardly visible in the image. In such cases, the TPR and FPR are evaluated not based on individual bounding boxes, but rather on the whole frame. A true positive is then a detection of at least one person in a frame where there is at least one person visible.

Figure 1: Examples of object detection (left) and partial occlusion (right).

III. The Challenge of Deep Learning-Based Detectors Under Fragmented Occlusion

To the best of our knowledge, none of the current SOTA deep learning-based approaches for object detection allow the detection of objects through foliage, namely detection under fragmented occlusion. Fragmented occlusion occurs, for example, when viewing objects behind tree and bush leaves. Contrary to partial occlusion, fragmented occlusion gives no clear view on minimal recognisable parts of the object used to detect the object [10]. To the best our knowledge, fragmented occlusion has not been considered for object detection previously to FOLDOUT.

A. *Beyond SOTA Deep Learning: Training Augmented with Fragmented Occlusion*

To study state-of-the-art deep learning methods in the case of fragmented occlusion, a new dataset (Figure 2) was created, recorded in a forest consisting of three videos with a total of 18,360 frames and 33,933 bounding boxes which were manually defined by human annotators. These bounding boxes were divided into four different occlusion levels including: no occlusion (level L0), slight occlusion (L1), moderate occlusion (L2), and heavy occlusion (L3). This data raises new challenges on the labelling and evaluation methodologies, which were partially answered in FOLDOUT. For example, bounding boxes are the standard annotation technique in current evaluation of detectors, but such labels are hard or infeasible to apply to datasets that contains heavy fragmented occlusion. Moreover, fragmented body parts of the person are detected (if detected at all) as separate objects and not as a single person. Furthermore, as the state-of-the-art detectors deliver bounding boxes, fragmented occlusion poses new questions on the evaluation methodology.

Next, the Microsoft COCO1 training data was augmented by occluding the ground truth masks similarly as leaves occlude people behind bushes and trees. Since the focus of the work was to the detect humans, this methodology was applied only to images containing humans. The trees used for the augmentation were generated from real images that were obtained from the recorded test data. The method generates whole artificial trees by randomly adding branches to previously manually segmented tree trunks. The branches attached to these trunks were also randomly generated by also adding several manually segmented leaves. The trees were placed in front of humans randomly. The deep learning-based Mask R-CNN detector was then trained with the augmented images.

The conclusion of this evaluation was that there is no significant difference between Mask R-CNN trained on Microsoft COCO and on the augmented dataset for L0 (no occlusion). However, clear improvement has been achieved for L1 (slight occlusion) which proves the applicability of the idea to model fragmented occlusion by the masks. Nevertheless, all approaches basically do not reach the expected robustness

and accuracy for moderate L2 and heavy L3 occlusion. One reason for this is that our current technique is not accurate enough to model fragmented occlusion. Furthermore, clear limits exist as heavy fragmented occlusion removes local spatial and structural information necessary for current approaches in object detection [3].

Figure 2: Fragmented occlusion examples. From left to right: the 4 categories of fragmented occlusion - no occlusion (level L0), slight occ. (L1), moderate occ. (L2), heavy occ. (L3) occlusion (taken from [3]).

B. Beyond SOTA Deep Learning: spatiotemporal Network Architecture

As discussed previously, object detection under moderate and heavy fragmented occlusion is an unsolved problem. The general task of object detection includes both localisation and classification of the targets. To make some progress beyond SOTA within the framework of deep learning, further work was carried out concerning the simpler task of localisation only. In this approach, the position of persons in the image was estimated as a likelihood map (heatmap) rather than using a conventional bounding box (as a bounding box also encodes the person's extent, an information often not present in fragmentally occluded images).

Current research in object recognition and object localization is mostly based on single images, hence lacks temporal information, which is crucial to overcome fragmented occlusion. Therefore, the new approach focused on training a spatiotemporal DNN, where training is performed using short image sequences rather than single images. The neural network is based on the U-net architecture [4]. For training, the well-known KTH dataset (recognition of human action) was used, augmented with a small self-made new dataset. Figure 3 shows the results of the method when trained on KTH dataset and directly applied on the new dataset.

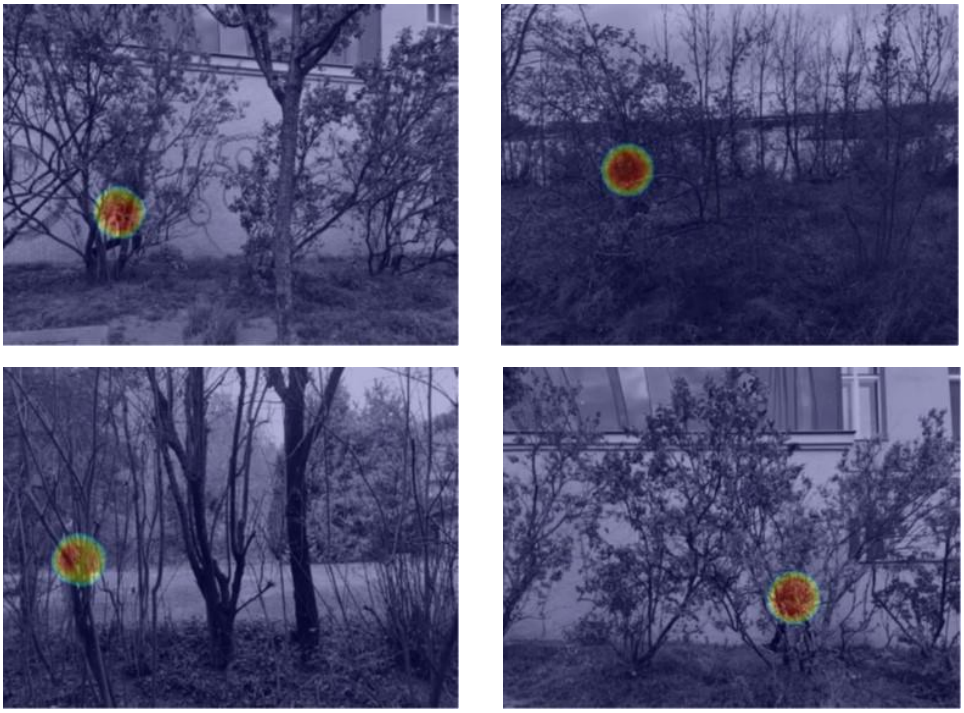

Figure 3: Qualitative results of the method on the new dataset. The method's heatmaps are shown in color (taken from [4]).

The adapted U-net spatiotemporal network works as a person localiser under fragmented occlusion on the novel test dataset with a much higher precision compared to a SOTA DNN-based detector trained on single images. The experiments show that currently, motion of the occluding vegetation degrades the performance. Furthermore, the present work is unable to cope with empty scenes (no persons present), neither with the case of multiple persons. It is also unclear how training behaves. This is left to future work [4].

IV. Evaluation of State-of-the-Art Deep Learning Detectors on FOLDOUT field demonstrations

AIT employed and tested various sensors during the field demonstrations taking place in Bulgaria (2-3.6.21 and 15-17.6.21), Greece (20-22.7.21 and 27-28.7.21) and Finland (25-27.1.22). The sensors included in the demonstrations were: (i) the AIT's Smart Sensor Technology Platform, which was developed by AIT within FOLDOUT [2], (ii) a standalone thermal camera (FLIR F-606), and (iii) a DH-SD6AL830V-HNI DAHUA sensor, with infrared illumination (NIR) for nighttime. The Smart Sensor

Technology platform, in fact, is comprised of several sensors, however, due to time constraints within FOLDOUT, the detector, which was implemented on its embedded platform, was operating only on its thermal camera (FLIR Boson). The detector implemented on the embedded platform was the DNN-based YOLOv5 [11]. The same detector was implemented also on a PC and was running on the video streams obtained from the standalone thermal camera as well as on the Dahua camera video data.

During the field demonstrations, hired actors were asked to act out scenarios having various degrees of occlusion by vegetation (Figure 4). The actors were carrying GNSS receivers, and the GNSS logs were used to monitor the relevant time intervals of the scripts as well as to compute the relative distances to the sensors. The recorded videos of the scripts together with the GNSS logs comprised the data for the evaluation. The evaluation was based on whole frames rather than on individual bounding boxes, namely, a positive example was a frame where at least one person was visible, and a true positive was the case were the detector detected at least one person in such a frame. Similarly, a negative example was a frame where no person was visible, and a false positive was the case where at least one person was detected in such a frame. The main findings of the evaluation are:

- Employing the 3 visual sensors, person detection was possible in the range of 10 – 130 meters (this range is dependent on the focal length of the sensors - chosen within FOLDOUT to enable detections up to 150 meters - and on the actors' trajectories during the scripts).
- As foreseen, detection performance of the YOLOv5 DNN-based person detector is degraded in the case of moderate to heavy fragmented occlusions. Averaging evaluation results from different sensors, time of day and weather condition, the estimated detection rate (TPR) is approximately 85% without fragmented occlusions, 60% for light occlusions, and 20% for medium occlusions, and, finally, 0% for heavy fragmented occlusions.
- Performance of the YOLO5 DNN-based person detector is reduced in cases where the persons in the scene are crouching or walking/running while bending. This is due to the fact that typically detectors are trained on upright persons.
- The DAHUA camera with NIR illumination does not fit for nighttime surveillance as it failed to illuminate the persons properly. Otherwise, the passive thermal sensors employed (inside the smart sensor as well as the standalone) were suitable for day and night detections.
- Better detection results were obtained when fusing the visual sensors with additional sensors, as was done within FOLDOUT as well [6].

Figure 4: Exemplary frames from FOLDOUT field demonstrations (in Bulgaria, Greece and Finland) using the 3 sensors and employing actors acting out various scenarios.

V. Evaluation of State-of-the-Art Deep Learning Detectors in the PETS2021 Workshop

AIT jointly with the University of Reading co-organized the PETS2021 workshop in conjunction with the AVSS (IEEE International Conference on Advanced Video and Signal-based Surveillance) 2021 conference, on the topic: Through-foliage detection and tracking challenge and evaluation. The challenge comprises the publication of a novel video surveillance annotated dataset on through-foliage detection, the defined challenges addressing person detection and tracking in fragmented occlusion scenarios, and quantitative and qualitative performance evaluation of challenge results submitted by six worldwide participants (including AIT).

PETS2021 proposes person detection and tracking challenges on environments where occlusion is highly present. The data acquisition was undertaken within FOLDOUT, mostly using the two sensors (4K-RGB and thermal) within the AIT Smart Sensor Technology Platform. Specifically, two challenges are proposed. Challenge 1 deals with the problem of person detection on three different sequences with increasing level of difficulty given by the foliage present in the image scene. Challenge 2 deals with long term tracking of a person who is several times occluded by foliage. The dataset is now publicly available for scientific research. To the best of our knowledge, a dataset dedicated to through-foliage detection was not available beforehand. In cases that fragmented occlusions do not allow to use a bounding box to annotate an occluded person, the person appearance was marked by a collection of points and/or ellipses (Figure 5).

Deep learning-based submission dominated the nature of employed methods. For this challenge, AIT submitted several state-of-the-art detectors, and a tracker. All submissions by six worldwide participants show the problem of detection and tracking remains an open problem where difficult conditions are tackled such as in foliated environments. More details can be found in [5]. In particular, the following table (Table 1) shows the detection results for the 6 state-of-the-art deep learning detectors employed by AIT for the 2 challenge sequences.

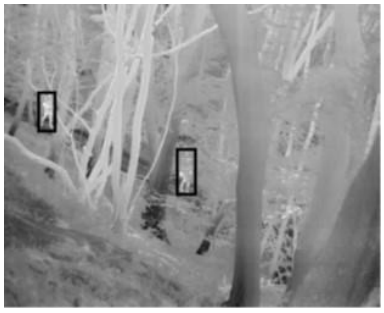

Figure 5: Exemplary images from the 2 PETS2021 challenges. In cases that the fragmented occlusion does not allow to use a bounding box or single ellipse as label (bottom-left image), the annotation was achieved by sampling the person's appearance by a set of points and/or ellipses (taken from [5]).

Table 1: AIT's detectors performance on PETS21 sequences (based on [5])

	TPR (seq. 1)	FPR (seq. 1)	TPR (seq. 2)	FPR (seq. 2)
centerNet	0.22	0.00	0.14	0.00
efficientDet	0.28	0.00	0.33	0.00
fasterRCNN	0.38	0.05	0.39	0.06
maskRCNN	0.32	0.04	0.30	0.01
SSD	0.23	0.00	0.29	0.00
YOLOv5	0.35	0.00	0.35	0.00

VI. Traditional moving object Detectors Under Fragmented Occlusion

As FOLDOUT scenarios contain an unprecedent amount of fragmented occlusion, and as SOTA deep learning detectors perform poorly on such scenarios, an alternative approach to detect through-foliage moving objects was developed as well. The approach undertaken relies on traditional background modeling. The advantage of this approach is that no prior knowledge and no offline training with annotated training data are required to detect the moving objects, rather the only assumption is that their appearance of the objects differs from the background of the scene. On the other hand, background modeling is more prone to false alarms, as no negative examples (non-person objects) are learned as in the training phase of DNNs. Additionally, background modeling-based detectors can localize moving objects but do not classify their category (e.g., a person, a car, …). Within FOLDOUT, AIT developed a robust real-time background modeling algorithm which is able, on one hand, to detect moving persons even when they are heavily occluded by foliage (fragmented occlusion), while, on the other hand, being resilient against

background appearance changes (dynamic background such as moving foliage and consequent lighting changes in the background).

Background modeling (also referred to as *background subtraction*) and the related foreground detection are important steps in video-surveillance and well-established themes in computer vision [7]. Conventional background modeling methods exploit the temporal variation of each pixel to model the background. For a recent survey regarding deep neural networks in the context of background modeling, see [12]. Deep neural networks have been demonstrated to be a powerful framework for background subtraction in video acquired by static cameras. However, deep neural networks require high computational overhead, special hardware such as expensive GPUs and extensive datasets for training. Moreover, deep neural networks outperform previous background modeling methods when they are trained with a scene-specific dataset. Furthermore, DNNs require extensive labeled datasets for training, while labeling background/foreground pixels in the case of heavy fragmented occlusions seems hardly feasible. These limitations render deep neural networks-based background modeling methods impractical for border surveillance applications such as the target of the FOLDOUT project. Consequently, traditional approaches that do not require offline training were resorted to.

The background modeling approach developed is pixel-based and comprises three modules. The core module is an extension of a previous work of AIT, published in [13], where a compressed non-parametric representation of the history of each pixel was employed, achieving ultra-real-time performance on CPU as well as on GPU. The algorithm does not require even an initial phase of observing an empty scene (background only) and can adapt over time even when initialized with a scene occupied by moving persons. The second post-processing module refines the foreground detection results of the first module by carefully removing foreground which is due to moving vegetation and corresponding lighting changes, by using motion consistency constraints. In this second phase, removing dynamical background via image filtering is avoided as much as possible, as not to eliminate the fragments of the persons observed, which can be rather tiny (consisting of several pixels only). Finally, a third module, the event detector, makes use of smoothed local spatial-temporal regions of the foreground mask to predict the presence of moving persons (generating alarms).

For the evaluation of the background-based detection algorithm, 2 video clips were chosen from the acquired FOLDOUT dataset (recordings of May 2019, in Finland), recorded by the RGB and thermal cameras within the AIT Smart Sensor Technology Platform. The scenarios depict persons walking behind tress with various degrees

of occlusion, but mainly with heavy occlusions, including frames where only small fragments of the persons are visible. The intensity of fragmented occlusion is way higher than in previous evaluations mentioned above (Sections IV and V). The 2 video clips are of length of about 2 minutes, where the frames were manually annotated as being empty of or occupied with persons.

For comparison, the DNN-based YOLOv5 detector implemented within FOLDOUT was applied to these 2 video clips. The DNN-based detector was able to classify correctly only 3% of the thermal and 14% of the RGB frames occupied with persons (TPR). The result on the RGB video is slightly better than on the thermal video since on the RGB camera a larger focal length (smaller field of view) was used, meaning more spatial resolution, and in fact avoiding some of the trees that provided more fragmented occlusions. These low detection rates were in fact achieved with maximal sensitivity, namely, accepting all detection as true regardless of how low the probability of detection was reported by the YOLOv5 detector. On the positive side, no empty frames were classified as occupied frames (no false positives). The results indicate once again the inability of current state-of-the-art DNN-based detectors to detect persons in case of severe fragmented occlusions. In contrast, the background-based detector does have a non-zero FPR, since it is not trained on positive and negative samples as in the case of a DNN-based detector. However, the background-based detector can detect significantly better moving persons under heavy fragmented occlusion. The trade-off between true and false positive rates, according to the tuning of the background-based detectors can be seen in Table 2. Figure 6 shows examples of foreground (moving objects) detection by the background modeling algorithm, for RGB and thermal cameras (output of second module of the algorithm).

Table 2: Performance of background modeling algorithm on the RGB and thermal sensors

Thermal Camera		RGB camera	
TPR	FPR	TPR	FPR
0.48	0.03	0.62	0.01
0.72	0.04	0.83	0.04
0.85	0.09	0.91	0.11

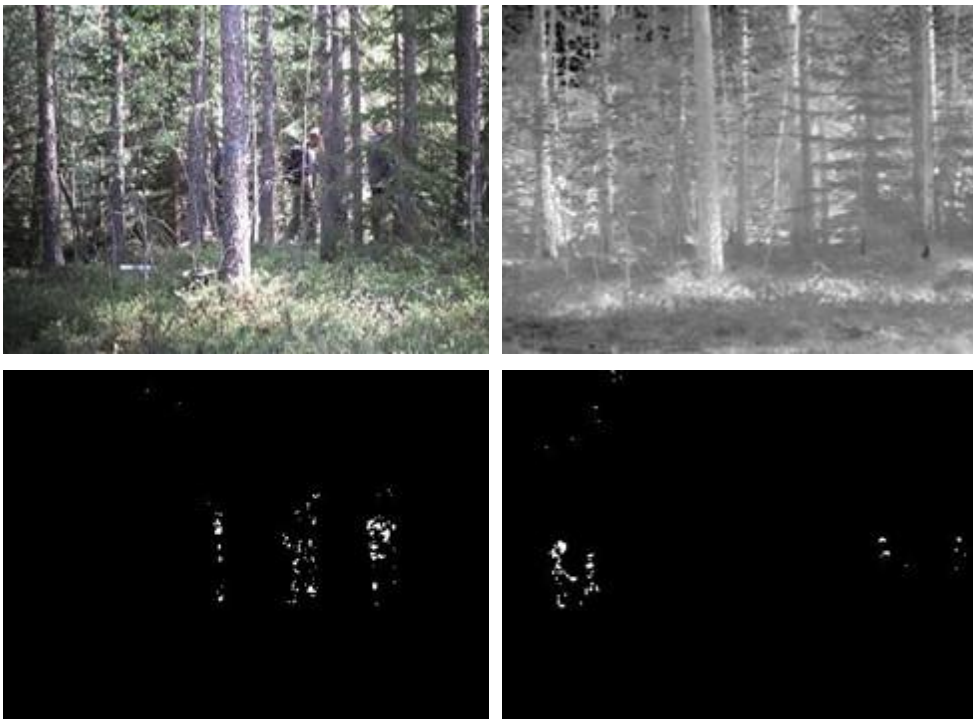

Figure 6: Top/Bottom-left: original RGB image/foreground detected. Top/Bottom-right: original thermal image/foreground detected.

VII. Conclusions

Through-foliage detection is an unsolved important part of border surveillance of remote areas between regular border crossing points. The FOLDOUT project builds a system that combines various sensors and technologies and fuses these into a robust detection platform. In this paper, the work concerning detection algorithms on visual sensors was described. In particular, the challenge of fragmented occlusion (e.g., a person walking behind trees) was addressed. To the best of our knowledge, detection under fragmented occlusion was not handled before in the literature of computer vision and is an unsolved challenge.

Several approaches based on Deep Neural Networks (DNNs) were developed to detect persons (and other objects) in videos of RGB and thermal cameras, under fragmented occlusion. Results indicate that state-of-the-art DNN detectors are unable to detect persons under moderate to severe fragmented occlusions, even when trained on data exhibiting fragmented occlusions. Some improvement was

gained, however, when the person detection task was relaxed to the simpler person localisation task, and temporal information (motion) was employed by the DNN.

Within FOLDOUT, the PETS2021 workshop was organized and sponsored, jointly with the AVSS 2021 conference, on the topic: Through-foliage detection and tracking challenge and evaluation. The challenge comprises the publication of a novel annotated video surveillance dataset on through-foliage detection, addressing person detection and tracking in fragmented occlusion scenarios, and performance evaluation of challenge results submitted by six worldwide participants (including AIT). Additionally, AIT employed and tested video detections on various RGB and thermal sensors during the field demonstrations taking place in Bulgaria, Greece, and Finland, where actors were performing according to planned scenarios having various degrees of occlusion by vegetation. Both types of evaluations show that the problem of detection and tracking remains an open problem where difficult conditions are tackled such as in foliated environments.

In addition to DNN-based detection approaches, a more traditional approach was developed to detect moving objects in the case of heavy fragmented occlusion, namely background modeling (background subtraction). A robust real-time background modeling algorithm was developed which is able, on one hand, to detect moving persons even when they are heavily occluded by foliage, while, on the other hand, being resilient against dynamic background such as moving foliage and lighting changes. The background-based detector was tested on RGB and thermal test videos recorded within FOLDOUT, especially comprising heavy fragmented occlusion, when only small fragments of the moving persons can be seen, and was demonstrated to compete favorably with DNN approaches, although, at the cost of increased false alarm's rate.

These results indicate the inability of current state-of-the-art DNN-based detectors, to detect persons in case of strong fragmented occlusions. Further research and development are needed to include temporal (motion) information in the architecture of DNNs, and to integrate the strength of both types of detectors (DNN-based and motion-based) to achieve high true positive and low false positive rates.

The applicability of a hyperspectral camera for through-foliage detection was also investigated, due to its extended spectral bandwidth. However, it was found to be

noisy compared to other visual sensors, and its data highly redundant. Several types of detectors were developed for the hyperspectral video; however, no gain was foreseen compared to other visual sensors, hence further research was carried on only for RGB and thermal sensors (More detailed can be found in [14, 15]).

ACKNOWLEDGEMENTS

The work described in this article was accomplished within the FOLDOUT project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 787021.

Bibliography

[1] "The Horizon 2020 Programme FOLDOUT project," [Online]. Available: https://foldout.eu/.

[2] C. Picus, D. Bauer, M. Hubner, A. Kriechbaum-Zabini, A. Opitz, J. Pegoraro, D. Schreiber and S. Veigl, "Novel Smart Sensor Technology Platform for Border Crossing Surveillance within FOLDOUT," *Journal of Defence & Security Technologies,* vol. 5, no. 3, 2022.

[3] J. Pegoraro and R. Pflugfelder, "The Problem of Fragmented Occlusion in Object Detection," in *Joint Austrian Computer Vision and Robotics Workshop*, 2020.

[4] R. Pflugfelder and J. & Auer, "Person Localisation under Fragmented Occlusion. .," in *17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2021.

[5] L. P. e. al., "PETS2021: Through-foliage detection and tracking challenge and evaluation," in *17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2021.

[6] L. Patino, M. Hubner, R. King, M. Litzenberger, L. Roupioz, K. Michon, Ł. Szklarski, J. Pegoraro, N. Stoianov and J. Ferryman, "Fusion of Heterogenous Sensor Data in Border Surveillance," *Sensors,* vol. 7351, no. 22, 2022.

[7] T. Bouwmans, F. Porikli and B. Höferlin, Background Modeling and Foreground Detection for Video Surveillance, New York: Chapman and Hall/CRC, 2014.

[8] J. Kaur and W. Singh, "Tools, techniques, datasets and application areas for object detection in an image: a review," .*Multimed Tools Appl,* no. https://doi.org/10.1007/s11042-022-13153-y, 2022.

[9] C. Ning, L. Menglu and Y. e. a. Hao, "Survey of pedestrian detection with occlusion," *Complex Intell. Syst.,* vol. 7, no. https://doi.org/10.1007/s40747-020-00206-8, p. 577–587, 2021.

[10] S. Ullman, L. Assif, E. Fetaya and D. & Harari, "Atoms of recognition in human and computer vision," in *PNAS Proceedings of the National Academy of Sciences of the United States of America*, 2016.

[11] G. Jocher, K. Nishimura, T. Mineeva and R. Vilariño, "YOLOv5," 2020. [Online]. Available: https://github.com/ultralytics/yolov5.

[12] J. Giraldo, T. Le and T. Bouwmans, "Deep Learning based Background Subtraction: A Systematic Survey.," in *Handbook of Pattern Recognition and Computer Vision*, World Scientific, 2020, pp. 51-73.

[13] D. Schreiber and M. Rauter, "GPU-based non-parametric background subtraction for a practical surveillance system," in *IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, 2009.

[14] A. Papp, J. Pegoraro, D. Bauer, P. Taupe, C. Wiesmeyr and A. Kriechbaum-Zabini, "Automatic Annotation of Hyperspectral Images and Spectral Signal Classification of People and Vehicles in Areas of Dense Vegetation with Deep Learning," *Remote Sensing,* vol. 12(13), 2020.

[15] D. Schreiber and A. Opitz, "A Novel Background Modeling Algorithm for Hyperspectral Ground-Based Surveillance and Through-Foliage Detection," *Sensors,* vol. 22, no. https://doi.org/10.3390/s22207720, 2022.