A UAV-based Infrared Small Target Detection System for Search and Rescue Missions

Victor J. Hansen
Dept. of Science and Industry Systems,
University of South-Eastern Norway
Kongsberg, Norway
email: vjhansen7@gmail.com

Antonio L. L. Ramos
Dept. of Science and Industry Systems,
University of South-Eastern Norway
Kongsberg, Norway
email: antonior@usn.no

José A. Apolinário Jr.
Dept. of Electrical Engineering,
Military Institute of Engineering
Rio de Janeiro, Brazil
email: apolin@ime.eb.br

***Abstract*—Using infrared imaging cameras mounted on unmanned aerial vehicles to assist in search and rescue missions by gathering and processing images can substantially improve the chances of survival of missing people. Indeed, infrared imaging cameras are well suited to support the detection of heat signatures in dark and cloudy conditions. The critical point herein is detecting heat signatures emitted by the human body. This stresses feasibility of application of infrared small target detection for search and rescue missions in wide areas. This paper presents and discusses a deep learning and a low-rank and sparse matrix decomposition approaches for infrared small target detection. Further, a framework tailored for unmanned aerial vehicles is developed. The proposed infrared small target detection system is capable of detecting heat signatures in images with complex backgrounds. Experimental results demonstrate that an infrared small target detection method based on deep learning is a valuable supporting system in search and rescue missions.**

Keywords*—*UAV; infrared imaging; object detection; computer vision; machine learning.

I. Introduction

In recent years, several missing people have been located by Unmanned Aerial Vehicles (UAVs) equipped with Infra-Red (IR) imaging cameras [1][2]. Indeed, autonomous UAVs are becoming very popular in applications such as surveillance and search and rescue missions. Generally, search and rescue missions are limited by time, area of coverage, costs, and availability of UAV pilots. This work considers a system suited for multi-rotor UAVs, as well as other vertical take-off and landing UAVs. In comparison to helicopters, UAVs require less traffic management, and are less expensive to operate and easy to deploy. In addition, they offer a high degree of design flexibility and can be equipped with a wide range of sensors. Moreover, UAVs can access confined spaces that helicopters cannot, including areas deemed hazardous to humans.

The use of automated object detection in search and rescue missions can reduce human errors, which are likely to occur in cases where the operator has to monitor a video stream for hours. The obvious consequence of this in the case of search and rescue missions is failure in locating the missing person. Compared to automated object detection, human operators have the advantage of understanding the context of a video and recognizing where to search based on experience. However, a challenging task for a human is to detect crucial details in a high-definition video, because the human eye could potentially be focusing only on a small section of the video frame.

The task of detecting missing victims is time-critical and particularly challenging since parts of the victim might be exceedingly small, and sometimes blend in with the surroundings. In natural disaster scenarios, using autonomous UAVs to discover any human activity can save lives. Object detection techniques and IR imaging are useful in automatizing the process of detecting humans in adverse conditions, thereby increasing the likelihood of survival. The combination of IR and color imaging can provide a relatively short search time in remote areas since IR small targets will appear as brighter than the local background and therefore distinguishable from the surroundings. IR imaging [3][4] is used in civilian and military applications owing to its ability to operate in dark and low-visibility conditions, such as cloudy and smoke-covered areas, making IR imaging suitable for detecting humans in low light situations. However, IR imaging cameras are not as effective in supporting search and rescue teams in finding missing people in warm areas because the heat from the surroundings might mask the heat signature of the target. This issue becomes exceptionally challenging when the target is covered by objects with thermal radiation shielding properties, or when multiple interfering heat sources are present.

In aerial IR images [3][5], IR targets occupy just a few pixels on the imaging plane due to the long imaging distance. An IR small target [6] can be defined as an object having a total size of less than 0.15% of an image. As a result, the target's thermal radiation is likely to appear weak. This makes the target difficult to recognize, as it lacks obvious shape, size, and texture characteristics. Mid-wave IR and long-wave IR cameras [7] are popularly known as thermal imaging cameras because they are capable of detecting radiation emitted by objects with a low surface temperature, typically around $25\,^{\circ}\mathrm{C}$. These cameras detect IR radiation and produce a thermal image that can be used to determine surface temperatures. Thus, there is no need for an external light source to detect an object. However, high-resolution IR imaging cameras are prohibitively expensive and not widely available to the public. Commercially available thermal imaging cameras typically generate low-resolution images, which make them inept at detecting small objects.

The field of IR small target detection has been dominated by model-driven methods. One of the highest performing [6] non-learning model-driven methods using low rank and sparse matrix decomposition is the Infrared Patch-Image (IPI) model [8]. Low-rank and sparse matrix decomposition methods [9] try to separate an image into a foreground component $\mathbf{S}$ and a background component $\mathbf{L}$. A sparse matrix [10] has a considerable majority of elements equal to zero. Thus, the IR small targets are often the non-zero values in the sparse matrix, making them easily identifiable. Conversely, a low-rank matrix [10] has a small number of linearly independent rows and columns compared to the matrix's size. The background patch of an IR image has a low rank, and the IR small targets of $\mathbf{S}$ are sparse when compared to $\mathbf{L}$.

Convolutional Neural Networks (CNNs), specifically Feature Pyramid Networks (FPNs) [11], outperform non-learning model-driven methods [5], indicating that learning from data can lead to high accuracy in IR small target detection. But, most CNNs learn high-level features by downsampling feature maps. As a result, the IR small targets become engulfed by the background features in the deepest layers. To ensure adequate detection results, a specialized network design is required [5][6]. Using a pre-trained network for the task of IR small target detection is not advised [3], but rather to train the CNN's weights from scratch using only IR small target images.

Meta-architectures, such as Faster R-CNN [12] and YOLO [13] only use the last layer's feature map to localize objects and make predictions. These models are ineffective at localizing small objects due to the absence of low-level features [14] at the last layer. The problem of detecting small objects can be alleviated by using a more fitting feature extractor (e.g., ResNet) [15]. According to [16], the main drawback of employing a CNN for the task of IR small target detection is that feature learning will become particularly challenging, as an IR small target generally lacks any prominent shape. Further, extracting features from low resolution images is difficult, and the IR small targets may disappear in the deep layers of a network due to their small size. Dai et al. [5] state that a high-resolution prediction map is crucial for detecting IR small targets, and thus propose the Attentional Local Contrast Network (ALCNet). ALCNet achieves better results than the completely data- and model-driven methods on the SIRST (Single-frame Infra-Red Small Target) dataset [17], indicating that when detecting IR small targets, one should prioritize combining CNNs with domain-specific knowledge, e.g., methods for measuring local contrast. To conserve small targets and extract feature maps, ALCNet employs a modified ResNet as its feature extractor. Further, Wang et al. [3] suggests restricting the number of downsampling operations in the feature extractor, thus gaining a sufficiently large feature map which conserves features of the IR small targets.

Having a deep network is desirable, as a deeper network can learn more features. However, as the network gets deeper, there may be instances where the accuracy saturates and then rapidly decreases. This is known as degradation [18]. Moreover, a deeper network leads to more parameters, which results in a more resource intensive model. As a solution to this problem, ResNet [19] introduced the residual block. The residual block takes the output $\mathcal{F}(\mathbf{x})$ of one or more layers and combines it with a shortcut connection containing the value $\mathbf{x}$ which is feeding those layers. Since the residual block prevents degradation, the network's depth can increase, and the accuracy will improve over time. Results from [19] demonstrate that the effect of the residual connections increases proportionally with the number of layers.

The remainder of this paper is organized as follows. Section II outlines the methods employed in this work and the rationale behind the selection of these methods. Further, Section II describes the evaluation metrics and outlines the testing process. Section III summarizes the test results and discusses the significance of the results. Section IV provides a discussion of the proposed system. Finally, Section V summarizes the performance results of the proposed system, and presents conclusions and future research opportunities on the topic.

II. METHODOLOGY

In this work, two different methodologies for detection of IR small targets are proposed and tested, namely a data-driven CNN-based method and a model-driven method using low-rank and sparse matrix decomposition. These are discussed next, beginning with some general considerations on the dataset.

A. IR Small Target Dataset Analysis

The Single-frame IR Small Target (SIRST) dataset [17], which contains 427 short-wave IR and mid-wave IR images, is used for training and testing of the proposed methods. The dataset sample size was augmented to improve the training of the model. The main reason is that scarcity or low variance in the training dataset will result in a model that performs poorly on new data. Certain targets are difficult for humans to discover as they require one to perform a focused and thorough search discriminate whether they are a target or just noise. Therefore, the classification task of IR small target detection is binary [6]. Moreover, as most of the targets in the images lack any definite features, they are all placed into a general class called ”Target”.

B. Data-driven Approach

Numerous CNN models are available, and it is difficult to differentiate between them. A model with a good trade-off between accuracy and speed (i.e., inference time for a single image) is desired. Based on suggestions from the literature review, the final choice fell on CenterNet ResNet50 V1 FPN (512×512) from the *TensorFlow 2 Detection Model Zoo* [20].

CenterNet [21] is a keypoint-based object detector, which means that it represents an object as a single point in the center of a generated bounding box. Other object properties, such as size and dimension, are obtained by moving from the center location towards the bounding box's outline. First, an input image is fed into a feature extractor (e.g., ResNet) in order to create a key-point heatmap. Peaks in the heatmap are mapped as object center points. An object's bounding box size is inferred from its center point.

1) Modified ResNet: The ResNet50 V1 FPN (512×512) is used as the object detector's feature extractor. The SIRST dataset contains images that are smaller than the original 512×512 pixels input size to the network. Thus, instead of upscaling the input images to 512×512 pixels, which would distort them, they are resized to 224×224 pixels. This is performed for the original ResNet50 as well.

We followed the general consensus reflected in most current research in the field that the downsampling operations of the feature extractor should be reduced to improve the detection of small objects. To achieve satisfactory results in terms of accuracy, the depth of the ResNet is maintained at 50 layers. The downsampling is reduced by changing the stride from 2 to 1 in the first convolutional layer of the original ResNet50. The output shape of the modified ResNet's last convolutional layer has an output shape of 14×14, whereas the output shape of the original ResNet's last convolutional layer has an output shape of 7×7.

2) Training the Data-driven Method: Training CNNs relies a great deal on matrix multiplications. GPUs (Graphics Processing Units) are well-suited for this type of computation, as their architecture allows for $100\times$ greater speed than CPUs (Central Processing Units) at this task [22].

The data-driven models are trained from scratch. The learning rate determines how fast the network learns. Goodfellow et al. [23] states that a high learning rate increases the training loss, while a low learning rate increases the risk of a slow training process, which, potentially, could become stuck at a high training loss. The original hyperparameters listed in Table I were used for training the data-driven methods, as these hyperparameters are commonly fine-tuned by the model's developers.

TABLE I: PIPELINE VALUES USED FOR TRAINING THE DATA-DRIVEN METHODS.

Pipeline values	
Warmup learning rate	2.5×10^{-4}
Base learning rate	0.001
Batch size	64 / 32
Warmup steps	5000

The batch size [23] is the number of training-samples from the dataset used in a single forward-pass. Typically, the batch size is less than the total number of training samples in the dataset. A large batch size consumes more memory. The data-driven method based on the original ResNet uses a batch size of 64. The modified ResNet uses a batch size of 32 due to the reduced downsampling which requires additional GPU memory. The training is stopped when the loss is stagnating. The training of the modified ResNet50 was stopped when the total training loss reached approximately 0.3, requiring significantly more steps than the original ResNet50, which had a training loss of approximately 0.15.

C. Model-driven Approach

The IPI model proposed by Gao et al. [8] is capable of producing accurate results even when confronted with complex scenes [9]. However, background edges, corners, or blobs infiltrate the sparse matrix, resulting in multiple discrepancies that the IPI model could treat as targets. In the IPI model, image patches from an IR image are rearranged using a sliding window to form a data matrix $\mathbf{D}$. The data matrix is then decomposed into a low-rank matrix $\mathbf{L}$ and a sparse matrix $\mathbf{S}$ using the Robust Principal Component Analysis (RPCA) algorithm in conjunction with the Principal Component Pursuit (PCP) [24]. Continuing with the IPI model, $\mathbf{D}$ can be decomposed into three components:

$$\mathbf{D} = \mathbf{L} + \mathbf{S} + \mathbf{N}, \tag{1}$$

where $\mathbf{N}$ is the noise. PCP can recover $\mathbf{L}$ and $\mathbf{S}$ from $\mathbf{D}$ by solving the following optimization problem [8]:

$$\min_{\mathbf{L},\mathbf{S}} \left(\|\mathbf{L}\|_* + \lambda \|\mathbf{S}\|_1 + \frac{1}{2\mu} \|\mathbf{D} - \mathbf{L} - \mathbf{S}\|_F^2 \right), \tag{2}$$

where μ and λ are positive-valued parameters.

The problem stated in Equation (2) can be solved through the Accelerated Proximal Gradient (APG). Solving RPCA-PCP via APG requires a significant amount of time to converge for a single IR small target image. Fortunately, however, several algorithms are available in the literature that solve the PCP. The proposed model-driven method is based on the IPI model [8] and RPCA-PCP via the Inexact Augmented Lagrangian Method (IALM) [25]. IALM is at least five times faster than APG and has a higher precision [26].

D. Evaluation Metrics

In the context of search and rescue missions, missed detection is more costly than false alarms, and this should be taken into account in the performance evaluation of the system. For the model-driven method, the sparse matrix $\mathbf{S}$ is the prediction. The center of a predicted target will be the location of pixels with a value higher than a certain threshold. The accuracy of the model-driven method is measured by checking if the center of the ground truth target intersects with the center of the predicted target. Additionally, the True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN) outcomes are recorded after testing each method. To deem a prediction to be TP, the predicted location must be within proximity of the ground truth location. This includes situations where the predicted bounding box and the ground truth bounding box are proper subsets of one another.

An applicable evaluation metric is the F-score given by

$$F_\beta = \frac{(1+\beta^2)(PPV)(TPR)}{\beta^2 PPV + TPR}, \tag{3}$$

where PPV is the Positive Predictive Value, also known as *precision*, and TPR is the True Positive Rate, also known as *recall*. Precision is a measure of how many of the predicted targets correspond to ground truth targets, and recall is the number of ground truth targets detected. For the proposed system, it is preferable to select a $\beta = 2$ in Equation (3), as the recall is more critical for evaluating the proposed methods.

Another applicable metric is the Matthews Correlation Coefficient (MCC), which returns a value in the range $[-1, 1]$.

The modified SIRST dataset is unbalanced, as it contains 268 positive samples plus 210 negative samples. Chicco and Jurman [27] recommend using the MCC rather than $F_{\beta=1}$ when evaluating predictions from a binary classifier, as F_1 can produce inaccurate results when applied to unbalanced datasets. MCC can resolve this issue by assimilating the imbalance. The proposed methods should have a high recall, a high F_2, and a high MCC score.

E. Testing

The modified SIRST dataset was used for experimental evaluation of the model- and data-driven approaches. The dataset has 210 negative images and 214 positive images containing 268 IR small targets.

The model-driven methods MD-v1 and MD-v2 are evaluated by adjusting the parameters listed in Table II. The *tolerance* ϵ_1 is required by the stopping criterion. If the value of the stopping criterion is below ϵ_1 the solution of RPCA-PCP via IALM has converged. The *iteration* parameter is used to forcibly stop the IALM if it has not converged. Further, adjacent pixels with a value above the *threshold* produce an IR small target. A high threshold removes false positives, but it could also exclude true positive predictions.

TABLE II: MODEL-DRIVEN METHODS AND THEIR PARAMETERS.

Abbreviation	Tolerance (ϵ_1)	Iterations	Stride	Patch size	Threshold
MD-v1	0.1	500	20	80	150
MD-v2	0.01	1000	20	80	150

According to [8], a patch size of 80×80 pixels, and a sliding step or stride of 14 in the sliding window produces acceptable results. However, a stride of 20 was selected for MD-v1 and MD-v2 as this decreases the required computational time. Furthermore, if the patch size exceeds 80×80, performance degrades.

The score threshold S_{th} is adjusted when evaluating the data-driven methods. The score threshold discards predictions which have a confidence score less than S_{th}. Table III contains abbreviations used for the various data-driven methods.

TABLE III: ABBREVIATIONS FOR DATA-DRIVEN METHODS.

Abbreviation	Stride	Batch size	Score threshold (S_{th})
DD-v1-03	2	64	0.3
DD-v1-05	2	64	0.5
DD-v2-03	1	32	0.3
DD-v2-05	1	32	0.5

III. Results

All results from evaluating the data-driven and model-driven methods on the modified SIRST dataset are shown in Table IV. The DD-v1 methods are clearly the fastest methods, with an average time of 0.2 seconds per image. A set of predictions performed by the proposed methods are shown in Figure 1. None of the methods are able to detect all five IR small targets in Figure 1a. However, four IR small targets were detected by MD-v2 and DD-v1-03, as shown in Figure 1c and 1e, respectively.

A. Analysis of the Model-driven Methods

As expected, and demonstrated in Table IV, the model-driven methods are inaccurate when compared to the data-driven methods. MD-v1 ($F_2 = 0.604$, $MCC = 0.224$, $PPV = 0.585$) outperforms MD-v2 ($F_2 = 0.586$, $MCC = -0.024$, $PPV = 0.345$) in terms of F_2, MCC and precision. In comparison to MD-v1, MD-v2 performs a meticulous decomposition, which may account for the low precision value, i.e., the large share of FP predictions.

MD-v2 with a reduced patch size and stride extracts excess noise from the image. In addition, a low patch size and stride results in a longer processing time.

To summarize, the MD-v1 and MD-v2 cannot compete with the data-driven methods in terms of accuracy. Also, MD-v1 and MD-v2 have an excessive computational time, requiring approximately 5 seconds per image. The lengthy computation time is primarily caused by the sliding window and singular value decomposition used for solving RPCA-PCP via IALM. The model-driven methods are, however, effective at identifying targets in complex environments.

B. Analysis of the Data-driven Methods

As illustrated in Table IV, all data-driven methods have a high precision, with the best scores going to DD-v2-05 ($PPV = 0.990$). A S_{th} of 0.3 results in a high recall, MCC and F_2. A $S_{th} < 0.3$ will introduce additional FP predictions. A S_{th} equal to 0.5 discards a portion of the false predictions, however, this also reduces the TP predictions. Further, an even higher S_{th} increases the amount of FN predictions. This is not desirable, as the system should aim at detecting all potential targets.

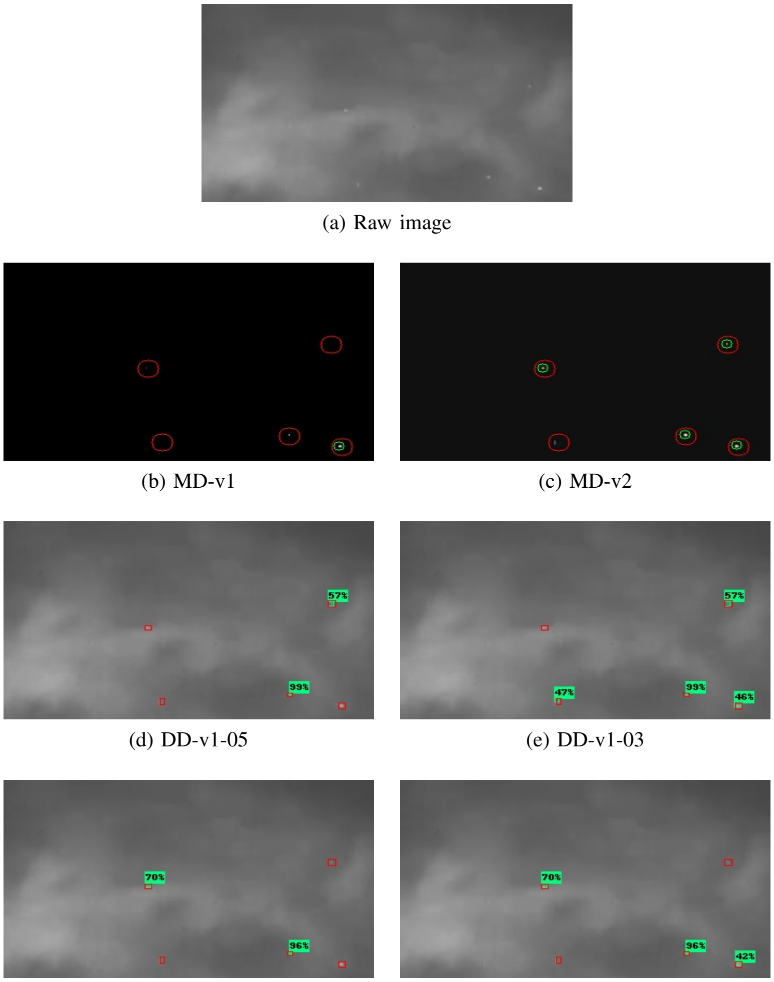
DD-v1-03 ($F_2 = 0.908$, $MCC = 0.817$, $TPR = 0.904$) and DD-v2-03 ($F_2 = 0.900$, $MCC = 0.842$, $TPR = 0.885$) are the most accurate methods. DD-v2-03 has a marginally lower F_2 than DD-v1-03. DD-v1-03 has the highest recall and F_2, and is the most appropriate approach for IR small target detection when considering the average time required to process a single image. DD-v1-03 processes a single image in approximately 0.2 seconds. However, this is not comparable to running object detection on a continuous video stream. The processing speed would, however, be higher if the methods were deployed on a GPU-equipped machine.

There is no statistically significant difference between the modified (stride = 1) and original (stride = 2) ResNet50 in terms of performance. This might originate from the modified ResNet50 not being downsampled sufficiently, or alternatively, the original ResNet50 already had suitable feature map sizes. Further reduction of the downsampling operations results in a slower system. As shown in Table IV, the modified ResNet models (i.e., DD-v2-03 and DD-v2-05) run slower due to the increased parameter count caused by the reduced downsampling.

The modified ResNet was trained with a batch size of 32, which is likely to be the reason why the training process is slower than the training of the original ResNet. A low batch size should result in a model that generalizes well

TABLE IV: RESULTS FROM EVALUATING THE DATA-DRIVEN AND MODEL-DRIVEN METHODS.

Metric	**MD-v1**	**MD-v2**	**DD-v2-03**	**DD-v2-05**	**DD-v1-03**	**DD-v1-05**
Recall	0.610	0.711	0.885	0.722	**0.904**	0.800
Precision	0.585	0.345	0.966	**0.990**	0.926	0.963
MCC	0.224	-0.024	**0.842**	0.720	0.817	0.760
F_2	0.604	0.586	0.900	0.763	**0.908**	0.828
Avg. time [s]	4.98	5.04	0.85	0.8	**0.2**	**0.2**



(a) Raw image

(b) MD-v1

(c) MD-v2

(d) DD-v1-05

(e) DD-v1-03

(f) DD-v2-05

(g) DD-v2-03

Figure 1: Predictions performed by proposed methods. Red boxes represent the ground truth targets. Green boxes represent predictions. (a) raw image from the dataset. (b) obtained using MD-v1. (c) obtained using MD-v2. (d) obtained using DD-v1-05. (e) obtained using DD-v1-03. (f) obtained using DD-v2-05. (g) obtained using DD-v2-03.

to previously unobserved data. Yet, there are no significant differences between using a batch size of 64 or 32 in terms of accuracy. The high accuracy of the data-driven methods could be the result of using the CenterNet meta-architecture with the ResNet feature extractor. CenterNet appears to perform well at the task of IR small target detection as it extracts peaks from keypoint heatmaps generated by ResNet.

IV. Discussion

What should the system do when a target or multiple targets are detected? Further, how can the location of a target be determined? Another challenge is to define how the system should behave in response to previously predicted targets. The system could register specific GPS positions. Moreover, the system should ignore heat signatures coming from the ground crew. Computational cost of CNNs results in slow inference on computationally constrained devices. Due to physical constraints, the system proposed in this work will have limited onboard computational power. To accelerate intensive tasks, edge computing can be used. Edge computing requires data transfer from the UAV to the *edge*, where required computations are carried out, then the results are relayed back to the UAV. This is a challenging task, and will likely result in unacceptable latency [28]. With edge computing, a collection of computing devices brings the capability to solve computationally intensive tasks closer to the UAVs, thereby reducing latency [28]. However, the computers located at the edge may be insufficient to perform real-time (e.g., more than 20 frames per second) inference due to restricted memory and processing power.

V. Conclusion and Future Work

This work investigates the detection of IR small targets. The results of using autonomous UAVs, IR imaging, and object detection to assist search and rescue missions are promising. In particular, a model-driven approach based on low-rank and sparse matrix decomposition which employs RPCA-PCP via IALM, and a deep learning-based data-driven approach using CenterNet with ResNet proved to be suitable choices towards solving this problem. Despite the limitations of the dataset, experimental results indicate that the proposed system is effective at detecting IR small targets. As expected, the data-driven approach outperformed the model-driven approach. Although accurate, however, the data-driven methods are slow. Training the data-driven methods on additional IR small target samples will further improve their accuracy.

The results of this work establish unequivocally that CNN-based object detection methods are accurate at IR small target detection. Conclusively, the proposed system can make a substantial impact by assisting search and rescue missions. Several areas are worth investigating further. The IR small target detection system remains incomplete. Validation of the proposed system in real-world circumstances should be given considerable attention. Furthermore, target tracking methods should be researched as they could increase the system's ability to locate missing victims and allow the system to focus on a single target if necessary.

Acknowledgment

This study was financed in part by the Norwegian Agency for International Cooperation and Quality Enhancement in Higher Education (Diku), project number UTF-2018-CAPES-Diku/10002, and in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES), finance code n° 23038.018065/2018-17.

References

[1] DJI. (2020, 12) DJI counts more than 500 people rescued by drones around the world. https://www.dji.com/newsroom/news/dji-counts-more-than-500-people-rescued-by-drones-around-the-world, retrieved: 2021.10.27.

[2] J. Frantzen. (2020, 4) Missing elderly woman found by drone pilot: Drones are saving lives in Norway (*Savnet eldre kvinne funnet av dronepilot: Nå redder droner liv i Norge*). https://www.uasnorway.no/savnet-funnet-av-dronepilot-na-redder-droner-liv-ogsa-i-norge, retrieved: 2021.10.27.

[3] K. Wang, S. Li, S. Niu, and K. Zhang, "Detection of infrared small targets using feature fusion convolutional network," *IEEE Access*, vol. 7, pp. 146 081–146 092, 2019.

[4] H. Zhang *et al.*, "A novel infrared video surveillance system using deep learning based techniques," *Multimedia tools and applications*, vol. 77, no. 20, pp. 26 657–26 676, 2018.

[5] Y. Dai, Y. Wu, F. Zhou, and K. Barnard, "Attentional local contrast networks for infrared small target detection," *CoRR*, vol. abs/2012.08573, 2020. [Online]. Available: http://arxiv.org/abs/2012.08573

[6] Y. Dai, Y. Wu, F. Zhou, and K. Barnard, "Asymmetric contextual modulation for infrared small target detection," in *IEEE Winter Conference on Applications of Computer Vision, WACV 2021*, 01 2021, pp. 949–958.

[7] M. Vollmer and K. Möllmann, *Infrared Thermal Imaging: Fundamentals, Research and Applications*. Wiley, 02 2018.

[8] C. Gao *et al.*, "Infrared patch-image model for small target detection in a single image," *Image Processing, IEEE Transactions on*, vol. 22, no. 12, pp. 4996–5009, 2013.

[9] H. Wang, M. Shi, and H. Li, "Infrared dim and small target detection based on two-stage u-skip context aggregation network with a missed-detection-and-false-alarm combination loss," *Multimedia Tools and Applications*, vol. 79, no. 47, pp. 35 383–35 404, 2020.

[10] S. L. Brunton and J. N. Kutz, *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge University Press, 2019.

[11] T.-Y. Lin *et al.*, "Feature pyramid networks for object detection," *CoRR*, vol. abs/1612.03144, 2017. [Online]. Available: http://arxiv.org/abs/1612.03144

[12] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *CoRR*, vol. abs/1506.01497, 2016. [Online]. Available: http://arxiv.org/abs/1506.01497

[13] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You Only Look Once: Unified, real-time object detection," *CoRR*, vol. abs/1506.02640, 2016. [Online]. Available: http://arxiv.org/abs/1506.02640

[14] G. Chen *et al.*, "A survey of the four pillars for small object detection: Multiscale representation, contextual information, super-resolution, and region proposal," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, pp. 1–18, 2020.

[15] Z.-Q. Zhao, P. Zheng, S. tao Xu, and X. Wu, "Object detection with deep learning: A review," *CoRR*, vol. abs/1807.05511, 2019. [Online]. Available: https://arxiv.org/abs/1807.05511

[16] H. Fang, M. Chen, X. Liu, and S. Yao, "Infrared small target detection with total variation and reweighted ℓ_1 regularization," *Mathematical Problems in Engineering*, vol. 2020, pp. 1–19, 1 2020.

[17] Y. Dai. (2021) Yimiandai/sirst. https://github.com/YimianDai/sirst, retrieved: 2021.10.27.

[18] Y. Xiao *et al.*, "A review of object detection based on deep learning," *Multimedia Tools and Applications*, vol. 79, no. 33, pp. 23 729–23 791, 9 2020.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: https://arxiv.org/abs/1512.03385

[20] TensorFlow. (2020, 9) TensorFlow 2 Detection Model Zoo. https://github.com/tensorflow/models/blob/master/research/object_detection/g3doc/tf2_detection_zoo.md", retrieved: 2021.10.27.

[21] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *CoRR*, vol. abs/1904.07850, 2019. [Online]. Available: https://arxiv.org/abs/1904.07850

[22] B. Pang, E. Nijkamp, and Y. N. Wu, "Deep learning with TensorFlow: A review," *Journal of Educational and Behavioral Statistics*, vol. 45, no. 2, pp. 227–248, 2020.

[23] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org

[24] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *CoRR*, vol. abs/0912.3599, 2009. [Online]. Available: https://arxiv.org/abs/0912.3599

[25] Z. Lin, M. Chen, and Y. Ma, "The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices," *CoRR*, vol. abs/1009.5055, 2010. [Online]. Available: https://arxiv.org/abs/1009.5055

[26] T. Bouwmans, A. Sobral, S. Javed, S. K. Jung, and E.-H. Zahzah, "Decomposition into low-rank plus additive matrices for background/-foreground separation: A review for a comparative evaluation with a large-scale dataset," *Computer Science Review*, vol. 23, p. 1–71, 2 2017.

[27] D. Chicco and G. Jurman, "The advantages of the Matthews Correlation Coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, p. 6, 2020.

[28] J. Chen and X. Ran, "Deep learning with edge computing: A review," *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1655–1674, 2019.