Midwave/Longwave Dual-Band Infrared Improves Recall in Pre-Trained YOLOv4 Small Object Detection

John R. Junger III[1], Guoliang Fan[2], and Joseph P. Havlicek[3]
[1]Artificial Intelligence Group, Camgian, USA
[2]School of Electrical and Computer Engineering, Oklahoma State University, USA
[3]School of Electrical and Computer Engineering, University of Oklahoma, USA

***Abstract*—We evaluate the object detection capabilities of deep learning based CNNs on midwave/longwave dual-band infrared (DBIR) video sequences for the first time. The characterization of CNN object detection performance on DBIR data, and in particular comparative analysis of the performance of DBIR systems relative to single-band longwave infrared (LWIR) and midwave infrared (MWIR) systems, has not been reported previously in the open literature. This is due at least in part to a general lack of labeled, publicly available DBIR data sets. In this paper, we apply a well-known, state-of-the-art CNN to DBIR data for the first time. A new labeled DBIR data set was generated comprising multiple classes of vehicles, people, airplanes, and birds. YOLOv4, pre-trained on the MS COCO dataset, was used for inference on the MWIR and LWIR channels of the DBIR sensor independently. The resulting detections from the two bands were considered both separately and jointly. The labeled objects of this DBIR data set were grouped into small, medium, and large classes. Detection performance on the medium and large objects was comparable to YOLOv4 performance reported previously in the open literature for visible wavelength objects in terms of average precision and average recall. Recall performance on small objects showed a significant size-dependent advantage for DBIR over LWIR or MWIR alone.**

***Index Terms*—dual-band infrared, small object detection, longwave infrared, midwave infrared, CNN, YOLO**

I. Introduction

Dual-band infrared (DBIR) camera systems integrate information from multiple subbands of the infrared (IR) spectrum independently to produce two output channels separated in wavelength. Commonly, and in this paper, the spectral bands of interest are the 3-5 μm midwave IR (MWIR) band and the 8-12 μm longwave IR (LWIR) band, but other configurations are possible [1]. In [2], we reported previous work where we constructed an experimental DBIR sensor using a research grade 320 $\times$ 240 MWIR/LWIR QWIP focal plane array (FPA) by QmagiQ LLC of Nashua, NH, and used it to acquire 48 mid-long DBIR video sequences depicting a rich variety of civilian objects. Fig. 1 shows a field test where this sensor was used for DBIR data collection in Santa Barbara, CA. By *research grade,* we mean that the midwave array of the detector had a swath of damaged pixels occurring in an area of size approximately 70 $\times$ 140 pixels localized primarily to the right edge of the midwave image, as also reported in [2]. A typical MWIR/LWIR frame pair acquired with this sensor is shown in Fig. 2, where the damaged pixels can be seen in the midwave image at right.

Fig. 1. Experimental dual-band MWIR/LWIR sensor deployed for field tests to acquire DBIR video sequences in Santa Barbara, CA.

Recently, we completed the arduous task of manually annotating ground truth bounding boxes and centroids for all of the objects in 43 of the 48 acquired DBIR video sequences (manual annotation of the remaining five sequences was not completed in time for the results to be included in this paper). We refer to this newly annotated collection of sequences as the OK-DBIR dataset. Our main contribution in this paper is to report detection results obtained by applying a state-of-the-art pre-trained convolutional neural network (CNN) object detector to this newly annotated dataset. As discussed below in Section V, we found that DBIR provided a significant gain in recall performance for small object detection relative to LWIR or MWIR single-band detection.

Meaningful comparison studies applying machine learn-

This work was supported in part by the U.S. Army Research Laboratory and the U.S. Army Research Office under grant W911NF-08-1-0293.

Fig. 2. Typical MWIR/LWIR frame pair acquired with the DBIR sensor of Fig. 1, cropped to a 240 $\times$ 240 pixel 1:1 aspect ratio. The 8-12 μm longwave band is shown on the left, while the 3-5 μm midwave band is shown on the right where the swath of damaged pixels is evident. Ground truth bounding boxes (blue and green) are added for one object of interest.

ing techniques to DBIR data are not available in the open literature. In [3], a NATO research team generated another DBIR MWIR/LWIR dataset in France and used generative adversarial networks to investigate super resolution. In [1], multi-layer perceptrons were used to detect military targets in DBIR sequences. However, to the best of our knowledge there are no previously published open studies reporting data on sensor performance characteristics using deep learning CNNs for object detection on DBIR datasets.

II. Background

IR camera systems play a role in a wide variety of important civilian and military applications [4]. Single-band cameras are used in the vast majority of these. In order to better understand if there are potential benefits that could be gained by replacing a single-band IR camera with a DBIR system, a study is needed to evaluate the performance of modern detection algorithms on DBIR data. The last publicly available detection studies were reported in 2001 and 2002 [1]. However, since that time there have been significant advances in detection algorithms, particularly with regards to the use of CNNs.

The training of CNN-based object detectors typically requires substantial training data comprising at least tens of thousands of labeled images, the labeling and curation of which involve a significant cost [5]. Currently, there is not any sufficiently large publicly available DBIR data set that could support training or re-training of a CNN for object detection in DBIR video sequences. Despite this lack of DBIR data, there are previous studies demonstrating relatively strong performance for CNN-based object detectors retrained using single-band IR data alone or in combination with visible spectrum data [6], [7]. New studies providing evidence of the efficacy of DBIR sensor systems could stimulate increased investment supporting the generation of larger comprehensive labeled DBIR data sets for training, which is one of the main objectives of the work we report in this paper.

Many CNNs run slower than real time, limiting their near-term applicability in practical field-deployed sensor systems [8]. You Only Look Once (YOLO) is one notable example of a state-of-the-art algorithm that does not suffer from this shortcoming [8], [9]. For this reason, YOLOv4 pre-trained on the MS COCO visual spectrum data set was selected as the initial CNN for evaluation here.

Many CNNs also do not detect small and distant targets as well as larger targets [10]. One reason for this is that small objects often lack appearance information to distinguish object from background [11]. The YOLO architecture specifies a minimum grid size as part of its approach to reducing the run time for detection [9]. This minimum grid size induces lower limits on detectable object size. While the definition of small, medium, and large targets varies somewhat in the literature, here we adopt the values provided in the MS COCO evaluation metrics [5]: small targets are those with an area under 32^2 pixels, large targets are those with an area over 96^2 pixels [12], and medium targets are those with an area falling between these two values.

Two significant characteristics needed to integrate a sensor into a practical system are probability of detection (PD) and false detection rate (FD) [13]. Ideally, the detector system has high PD and low FD. However, CNN performance is more commonly characterized in terms of recall, which is strongly related to PD, and precision, which is inversely related to FD [14]. Thus, one desires a detector that achieves high recall and high precision simultaneously. Well-known techniques such as multiple-hypothesis tracking and clutter modeling may be used to reduce the number of false positives (FP) [13]. But it must be borne in mind that reducing FP by these techniques almost universally reduces TP by some marginal and tunable amount [13]. Receiver operating characteristic (ROC) curves and/or the closely related precision-recall curves can be used to select an appropriate sensitivity for the FP rejection algorithms via detection threshold optimization [13].

CNN detection algorithms facilitate sensitivity tuning by reporting a measurement of similarity between the region being classified and internal representations of that object class, referred to here as class confidence (CC). Precision-recall curves typically show the relationship between precision and recall, parameterized by CC. In addition to detection threshold optimization, TP and FP rates can also be tuned by thresholding CC. Given these two approaches for FP reduction (concomitantly increasing precision), our focus here is simultaneously to increase recall in the IR scenario by incorporating DBIR data.

A second commonly used FP suppression algorithm is non-maximum suppression (NMS) [9], [15]. The main idea is to eliminate multiple redundant detections of a single object, which would otherwise be counted as false positives, by thresholding the intersection over union (IoU) between pairs of detections. NMS begins by selecting the detection with the highest CC and calculates the IoU of all other detections in the frame relative to the selected detection. Other detections that exceed the NMS threshold on IoU are deemed too similar to the selected detection and are removed [9], [15]. The process then repeats iteratively by selecting the next highest CC detection among those remaining in the frame. A lower

NMS threshold results in higher Precision and lower Recall.

III. Measuring Performance

The determination of whether a given detection is a TP or FP is made by thresholding the IoU between that detection and each ground truth object present in the frame. If the IoU exceeds the specified threshold, then the detection is deemed to be a TP. Alternatively, if the maximum IoU is below threshold, then the detection is deemed to be an FP. Similarly, a ground truth object is determined be to a False Negative (FN) if the maximum IoU between that ground truth object and any detection falls below threshold. The three calculated values TP, FP, and FN are then used to define precision P and recall R in the standard way according to

$$P = \frac{TP}{TP + FP} \tag{1}$$

and

$$R = \frac{TP}{TP + FN}. \tag{2}$$

IV. Experiment Design

As we mentioned in Section I, objects in 43 DBIR sequences of the OK-DBIR dataset were manually labeled with ground truth. The objects were classified as Pickup, Car, SUV, Van, Semi, Person, Motorcycle, Airplane, Fuel Truck, or Birds. In some cases, there were small, distant objects that were moving away from the sensor at the beginning of the sequence and lacked sufficient appearance information for reliable classification by the human annotator. These objects were labeled "Indeterminate." All together, the 43 labeled sequences in the data set contain 78,606 instances of 242 objects.

These data include many examples of partial and full occlusions. Many of the objects undergo gradual appearance change as they move towards or away from the sensor system. Some objects also change appearance significantly with changing attitude or direction relative to the sensor. While there are some advantages to having many sequential frames of the same object, especially in evaluating tracking algorithms, there are some disadvantages when studying CNN based detection in isolation. A principle concern is the problem of over-fitting. Given concerns about the relatively low appearance variation and low object count in this dataset (relative to, e.g., MS COCO), a decision was made against transfer learning or retraining. Thus, YOLOv4 was acquired pre-trained on MS COCO from github [8] as implemented by the original authors.

Because the OK-DBIR sequences were acquired at a sample resolution of 14 bits, some preprocessing was necessary prior to running pre-trained YOLOv4 inference. The brightest pixels of the images were clipped and histogram stretching was subsequently performed to achieve an intensity profile matching the dynamic range of typical visual spectrum images. Inference was then performed on the MWIR and LWIR channels independently. After inference, NMS was applied to suppress duplications and minor variations in the detection reports. The threshold for NMS was set at 0.99. This value was selected to remove identical or nearly identical FPs, preventing over counting due to duplicate classification for the same FP.

The IoU between detections and ground truth was evaluated against the NMS refined detections from both bands. Precision and recall were calculated for each (IoU, CC) pair in the MWIR band, in the LWIR band, and jointly for DBIR, where both IoU and CC were varied in a range from 0.05 to 0.95.

The 43 sequences of the OK-DBIR data set comprise 53,181 small object instances, 23,882 medium object instances, and 1,543 large object instances. Of the small objects, 56.4% are smaller than 10^2 pixels and 25.5% are smaller than 8^2 pixels. To evaluate the gain in recall provided by DBIR without respect to CC, the maximum recall across all CC for a given IoU was calculated for LWIR, MWIR, and jointly for DBIR, grouped separately by small, medium, and large objects.

V. Results

Maximum recall with respect to CC as a function of IoU is shown in Fig. 3 for large and medium objects and in Fig. 4 for small objects. These results show a consistent gain in maximum recall performance for DBIR relative to MWIR or LWIR alone across all object sizes and at all IoU. The maximum recall performance gain for DBIR relative to MWIR is significant in all cases tested, whereas the gain relative to LWIR shows an improvement of 3.5% for medium and large objects and 50% for small objects averaged over IoU ranging from 0.1 to 0.95 in increments of 0.05.

In [8], YOLOv4 average precision (AP) on visible spectrum data was reported ranging from 0.44 to 0.66, with 0.66 being obtained at an IoU of 0.5. At an IoU of 0.5, the DBIR AP in our study was 0.47 for medium and large objects and 0.13 for small objects, meaning that we obtained lower AP performance than that reported in [8]. Possible factors contributing to this performance loss include dissimilarity between visible spectrum and IR data, damage to the MWIR portion of the sensor under test, and dissimilarity relative to object signatures as seen in MS COCO. Average recall (AR) for YOLOv4 on MS COCO is more challenging to find; however, an AR of 0.61 is reported in [16] for retraining YOLOv4 on MS COCO with a smaller classification set.

To better understand detection performance on the smallest objects with area less than or equal to 16^2 pixels, we also performed a separate experiment with the CC threshold set at 0.3 and the IoU threshold set at 0.5. For this experiment, we calculated PD separately for all object instances with an area of i^2 pixels for $1 \leq i \leq 16$. The results are given in Fig. 5. The smallest detected object had an area of 42 pixels and there were 3,363 smaller object instances that were not detected at these threshold settings. The graph in Fig. 5 illustrates the strong relationship between object size and PD for small objects.

VI. Conclusion

Using DBIR data improves the performance of YOLOv4 trained on MS COCO for detection of small, medium, and large targets. A significant improvement in recall for small

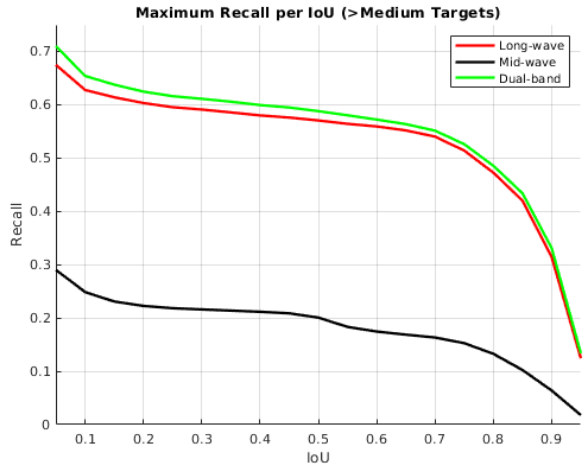

Fig. 3. Medium and large object Maximum recall per IoU for all CC. NMS thresholding set to 0.99.

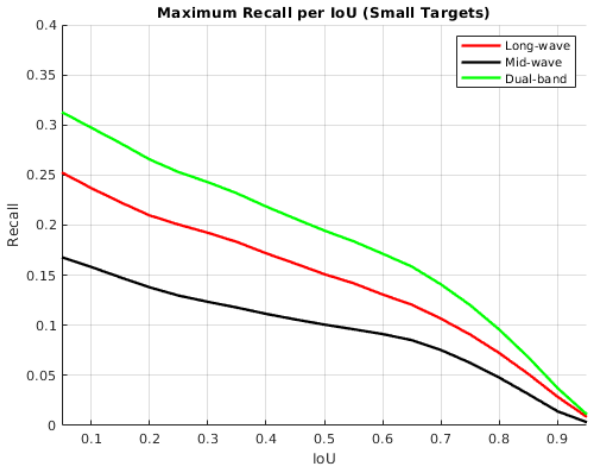

Fig. 4. Small object maximum recall per IoU for all CC. NMS thresholding set to 0.99.

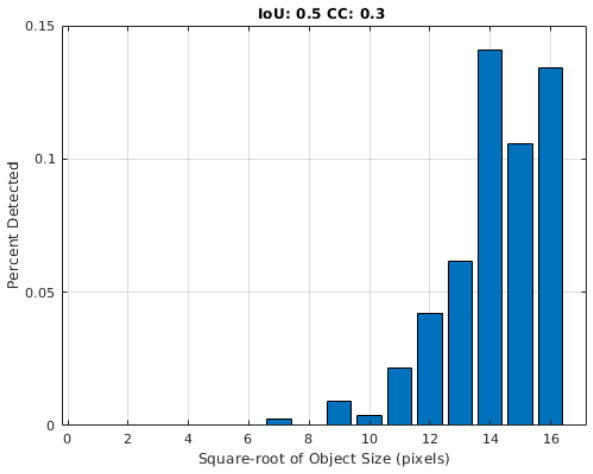

Fig. 5. Percent of small objects detected by square root of bounding-box area in pixels; MWIR and LWIR both considered.

objects, a class of objects providing the biggest opportunity for improvement, was observed. A more moderate performance gain for medium and large objects was observed. Given the low base-level of detection observed in small objects compared to medium and large objects in this study, additional work is needed to improve small object detection with CNN based object detectors. As work progresses in the field of small and distant object detection, continued evaluation of these algorithms with respect to the OK-DBIR dataset is likely to prove beneficial. The results of the experiments here suggest that there may be a practical benefit to using DBIR for CNN based object detection. It is likely that the biggest performance gains would be observed among small and distant objects. Retraining CNNs with DBIR data may be beneficial and should be studied further. However, new larger DBIR data sets would likely be needed to support those efforts

REFERENCES

[1] A. C. Goldberg, T. Fischer, and Z. I. Derzko, "Application of dual-band infrared focal plane arrays to tactical and strategic military problems," in *Infrared Tech., Applications XXVIII*, ser. Proc. SPIE, B. F. Andresen, G. F. Fulop, and M. Strojnik, Eds., vol. 4820, 2003, pp. 500–514.

[2] C. T. Nguyen, J. P. Havlicek, G. Fan, J. T. Caulfield, and M. S. Pattichis, "Robust dual-band MWIR/LWIR infrared target tracking," in *Proc. 48th Asilomar Conf. Signals, Syst., Comput.*, Nov. 2014, pp. 78–83.

[3] A. R. Weiß, U. Adomeit, P. Chevalier, S. Landeau, P. Bijl, F. Champagnat, J. Dijk, B. Göhler, S. Landini, J. P. Reynolds, and L. N. Smith, "A standard data set for performance analysis of advanced IR image processing techniques," in *Infrared Imaging Syst.: Design, Anal., Modeling, Testing XXIII*, ser. Proc. SPIE, G. C. Holst and K. A. Krapels, Eds., vol. 8355, 2012, pp. 354–363.

[4] P. David, "Multiple-sensor cueing using a heuristic search," in *Appl. Artificial Intell. IX*, ser. Proc. SPIE, M. M. Trivedi, Ed., vol. 1468, 1991, pp. 1000–1009.

[5] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollr, "Microsoft COCO: Common objects in context," 2015, arXiv preprint 1405.0312.

[6] J. Liu, S. Zhang, S. Wang, and D. N. Metaxas, "Multispectral deep neural networks for pedestrian detection," 2016, arXiv preprint 1611.02644.

[7] M. Chaverot, M. Carr, M. Jourlin, A. Bensrhair, and R. Grisel, "Object detection on thermal images: Performance of YOLOv4 trained on small datasets," in *Proc. European Symp. Artificial Neural Networks, Comput. Intell., Machine Learning*, Oct. 6-8 2021, pp. 207–212.

[8] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, arXiv preprint 2004.10934.

[9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," 2016, arXiv preprint 1506.02640.

[10] M. Kisantal, Z. Wojna, J. Murawski, J. Naruniec, and K. Cho, "Augmentation for small object detection," 2019, arXiv preprint 1902.07296.

[11] K. Tong, Y. Wu, and F. Zhou, "Recent advances in small object detection based on deep learning: A review," *Image, Vision Comput.*, vol. 97, p. 103910, 2020.

[12] H. Luo, P. Wang, H. Chen, and V. P. Kowelo, "Small object detection network based on feature information enhancement," *Computational Intell., Neuroscience*, vol. 2022, Article ID 6394823, 2022.

[13] Y. Bar-Shalom, P. K. Willett, and X. Tian, *Tracking and Data Fusion: A Handbook of Algorithms*. YBS Publishing, 2011.

[14] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," 2020, arXiv preprint 2010.16061.

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," 2016, arXiv preprint 1506.01497.

[16] J. Woo, J.-H. Baek, S.-H. Jo, S. Y. Kim, and J.-H. Jeong, "A study on object detection performance of YOLOv4 for autonomous driving of tram," *Sensors*, vol. 22, no. 22: 9026, Nov. 2022.