In: SPIE Defense + Commercial Sensing, 30 April - 4 May 2023, Orlando, Florida, Volume 12536
DOI: 10.1117/12.2665984

CNN-fusion architecture with visual and thermographic images for object detection

Ndidiamaka Adiuku[1*], Nicolas P. Avdelidis[1], Gilbert Tang[1], Angelos Plastropoulos[1], Suresh Perinpanayagam[2],

[1] Integrated Vehicle Health Management Centre, School of Aerospace, Transport and Manufacturing, Cranfield University, UK.

[2] University of York, UK.

ABSTRACT

Mobile robots performing aircraft visual inspection play a vital role in the future automated aircraft maintenance, repair and overhaul (MRO) operations. Autonomous navigation requires understanding the surroundings to automate and enhance the visual inspection process. The current state of neural network (NN) based obstacle detection and collision avoidance techniques are suitable for well-structured objects. However, their ability to distinguish between solid obstacles and low-density moving objects is limited, and their performance degrades in low-light scenarios. Thermal images can be used to complement the low-light visual image limitations in many applications, including inspections. This work proposes a Convolutional Neural Network (CNN) fusion architecture that enables the adaptive fusion of visual and thermographic images. The aim is to enhance autonomous robotic systems' perception and collision avoidance in dynamic environments. The model has been tested with RGB and thermographic images acquired in Cranfield's University hangar, which hosts a Boeing 737-400 and TUI hangar. The experimental results prove that the fusion-based CNN framework increases object detection accuracy compared to conventional models.

Keywords: Deep learning, Object Detection, Image Fusion, aircraft inspection.

1. INTRODUCTION

The increasing potential for autonomous mobile robots (AMR) to achieve intelligent operations has attracted continuous research interest in different domains like in aviation industry. One of the major application areas is visual aircraft inspection for capturing aircraft surface defects during maintenance, repairs and Overhaul (MRO) operations in a hangar environment. The busy hangar environment is inconsistent in structure, with objects of varying shapes, sizes, colours, and intensities that contribute to difficulties in the robot's real-time motion control while detecting and avoiding obstacles in a navigation task. AMR requires high detection accuracy in distinguishing solid and low-density moving objects for safe navigation in different lighting or environmental conditions. Significant improvement has been achieved by enhancing the robot's perception [1], which is heavily dependent on sensing devices and background information which can be ambiguous.

Thermal and RGB cameras are vision-based sensors used for object detection in autonomous mobile robot navigation tasks and have different capabilities based on light variations. RGB camera images provide more detailed information in good light conditions, but shadows and illumination intensity impair performance. On the other hand, thermal images are good in spotting objects of even small shapes in visually degraded and high-contrast conditions. Still, image information usually lacks detailed features like colour, which compromises application in different scenes and situations.

* Correspondance e-mail: n.p.adiuku@cranfield.ac.uk

The fusion of visible and thermal images [2] has been seen to improve the perception of object features in complex environments and has increased the robustness of the information in object detection and avoidance tasks.
Image fusion-based methods are critical concepts that have been widely developed in the literature to improve the visual quality of the fused image [3] and enhance object detection in different applications with more accurate information and clarity. This includes using traditional methods like deep neural network-based approach [4], fuzzy logic [5], multiscale and sparse representation [6]. These algorithms use a handcrafted feature extraction method to fuse image information, which is challenging to implement, limited for shallow objects [7] and results in poor-quality image classification.

This research aims to develop an efficient and more robust object detection technique for mobile robots to respond timely to obstacles and navigate safely in unstructured and dynamic scenarios using deep learning on fused image data. Image-based fusion for object detection has been widely studied using deep learning and has been applied in different use cases with a reasonable success rate [8]. Convolutional neural network (CNN) is one of the variants of deep learning models and is powerful for feature extraction from image data. Recent improvement with CNN architecture has been proposed to extract relevant object features and classify them accordingly to enhance object detection like ResNet, Alex Net, FusionGAN [9], NestFuse [10], DenseFuse with encoding and decoding networks [11]. These CNN networks utilise dense blocks, which compromises computational requirements and limits the real-time application of the models. To address this drawback, we propose an adaptive YOLOv5 model [12] based on fused RGB and thermal images using a pre-processing module and attention mechanism to improve object detection accuracy. In addition, the model will incorporate custom augmentation techniques and a convolutional block Attention module (CBAM) to take advantage of the image sources' feature dissimilarity to improve object recognition and avoidance. The goal is to create more informative images that increase object visibility and improve scene understanding for accurate object detection and avoidance in challenging cases.

The main contribution of this work can be summarised as follows:

1. We prepared a multisensory dataset of objects of 12 classes from the mobile robot perspective in hangar environment with 600 aligned RGB and thermal image frames.
2. The input module is designed with pre-processing fusion techniques that transform, scale and fuses custom RGB and thermal images to enhance object feature extraction for improved accuracy in object detection.
3. To improve YOLOv5m network with CBAM module for relevant feature map extraction and network layers compression to achieve lightweight model with reasonable computing resources.
4. Performance evaluation using public and custom dataset on proposed fuse-YOLO and YOLOv5 models

The rest of this paper is organised as follows: In section 2, we summarised some recent literature in YOLOv5 for autonomous vehicle object detection. Section 3 describes the YOLOv5 architecture and proposed fuse-YOLO model. Section 4 shows the experiment process and result. Section 5 presents the conclusions of this research.

2. RELATED WORK

There has been some significant research on object detection recently with multi-source images, showing improvement and robust environmental adaptation capability. Currently, traditional and deep-learning detection methods are two major categories of object detection that have been widely studied. Conventional methods like the sliding window approach [13], support vector machine (SVM) [14] and template matching-based algorithm [15] involve manually designed feature extraction procedures for object detection and classification in an image. This method is challenging to implement, computationally inefficient and poorly performed with small and diverse objects. Therefore, the automatic and hierarchical approach to feature extraction in convolutional neural network (CNN) algorithms [16] has gaed more research attention over traditional methods, especially in solving the problems mentioned earlier.

The most popular CNN architectures for object detection are based on the single-stage and two-stage detectors network. Faster R-CNN [17] is among the two-stage variants that was improved by using region of interest (ROI) and region proposal network (RPN) pooling [18], which generate sparse related bounding boxes for target objects and then classified and regressed to increase detection accuracy. However, this still is limited in achieving the real-time improved inference speed that is required for application in autonomous vehicles. On the other hand, Yolov5 is among the one-stage detection methods that use a grid framework to predict the position and classification of multiple objects in the frame for better detection accuracy. Different improvement has been made in yolov5 in [19] the method of structural adjustment of the model elements was used to improve small object detection and inference time, [20] authors expanded cross-stage-partial-connections (CSP) module to enhance the use of shallow features.

Many more researchers have proposed better models using fused visible and thermal. Yunfan et al. in [21] use a multi-layered CNN architecture that combines RGB and thermal images for improved pedestrian detection. Wagner et al. in [22 used a CNN-based detection model to train the KAIST dataset described for pedestrian detection using merged FIR and RGB images. Osman et al. [23] trained the YOLOv3 model with a merged image spectrum but requires more images to optimise the performance of the network. In [24], a multispectral model based on Yolov4 with combined RGB and thermal images was use to demonstrate high detection accuracy and adaptation capability in changing scene scenarios. More improvement was seen in [25] where the authors used illumination-aware deep neural networks with both visual and thermal images to enhance object detection performance. Early fusion enhancement was used in [26] by using a CNN-based fusion module to extract useful features from the RGB and thermal images towards better object performance. Based on these studies, we have proposed fuse-YOLO that leverages thermal and RGB features and CBAM module for the yolov5n network to improve object detection accuracy. This will enhance relevant object feature extraction, lightweight capability and computational suitability for mobile robot navigation tasks.

3. THERMAL IMAGE ENHANCEMENT FOR ROBOTS OBJECT DETECTION

In this paper, we focus on improving the performance of an existing deep neural network capable of recognising different types of objects under varied lighting situations by supplementing thermal images with RGB images through a fusion network. Thermal imaging is among the sensing mechanisms that can be utilised to improve the navigation of autonomous mobile robots, particularly under challenging conditions like low-light conditions. It can detect the temperature differences of objects in complex scenes, which can assist robots in detecting obstacles that are not obvious in RGB images. In this work, autonomous mobile robots' navigation will be enhanced to provide adequate safety and motion control using thermal imaging through improved object recognition and avoidance algorithm for aircraft inspection operations. The image representation below in Figure 1. shows a resulting fused image from thermal image enhancement with RGB images which is the input to the YOLOv5 object detection algorithm proposed.

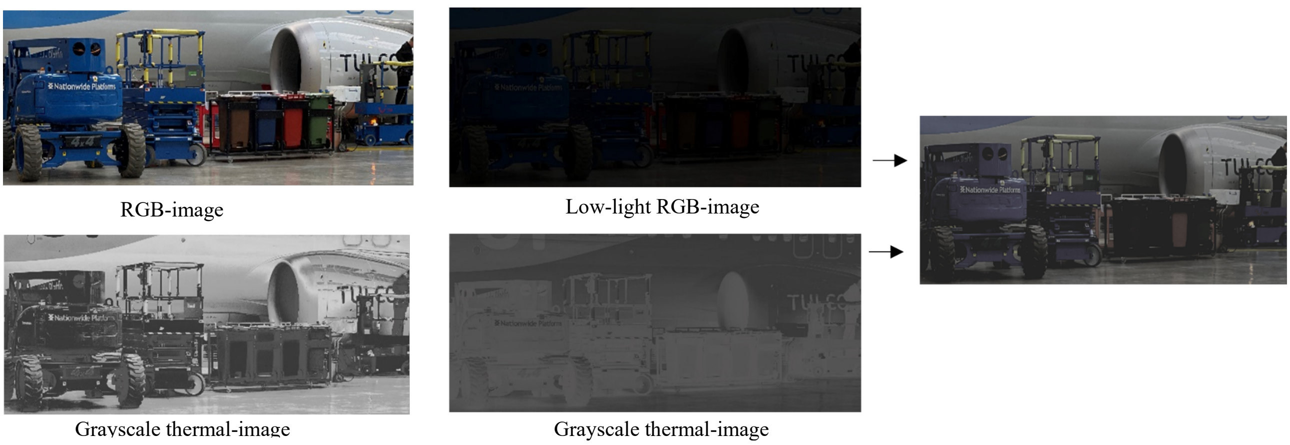

Figure 1. Visual representation of RGB and thermal images at different light variations and resulting fused image.

3.1. Overview Of Yolov5 Algorithm

YOLOv5 is a popular CNN-based object detection framework that uses a single-stage architecture and grid of cells to train the detector to operate precisely and with more visibility enabling the system to identify any object in complex environments. It is easy to deploy and train. The architecture comprises the backbone based on the cross-age partial connection (CSP-Darknet-53) model [27] that connects convolutional layers to extract object features from input images at a reduced computational cost. The neck structure is at the middle phase and uses Path Aggregation Network (PANet) [28] to up-sample, down-sample, fuse features maps, and extract the multiscale object. This is done at different levels of the backbone region and helps the model to generalise on unknown data. It provides an improvement in the detection and localisation accuracy of objects of interest in an image. The YOLOv5 detection head uses a dynamic anchor assignment approach to cluster the ground-truth objects by adjusting the sizes and shapes from input images. This will generate multi-dimensional bounding box information, including probabilities of the objects class, confidence and box coordinates.

The Yolov5 model is configured with a yaml file that is used to define the parameters that make up the model architecture [19] and has a variety of pretrained models represented as YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l and YOLOv5x where the difference between them lies on inference time and size of each model in terms of the depth and width of the network [29] The general Yolov5 model [30] consist of series of convolutional layers with different kernel sizes, batch normalisation [31] and ReLU activation functions [32] to optimise feature extraction and detection accuracy. The YOLOv5m model performed better with our custom dataset and was used as the baseline model in this paper.

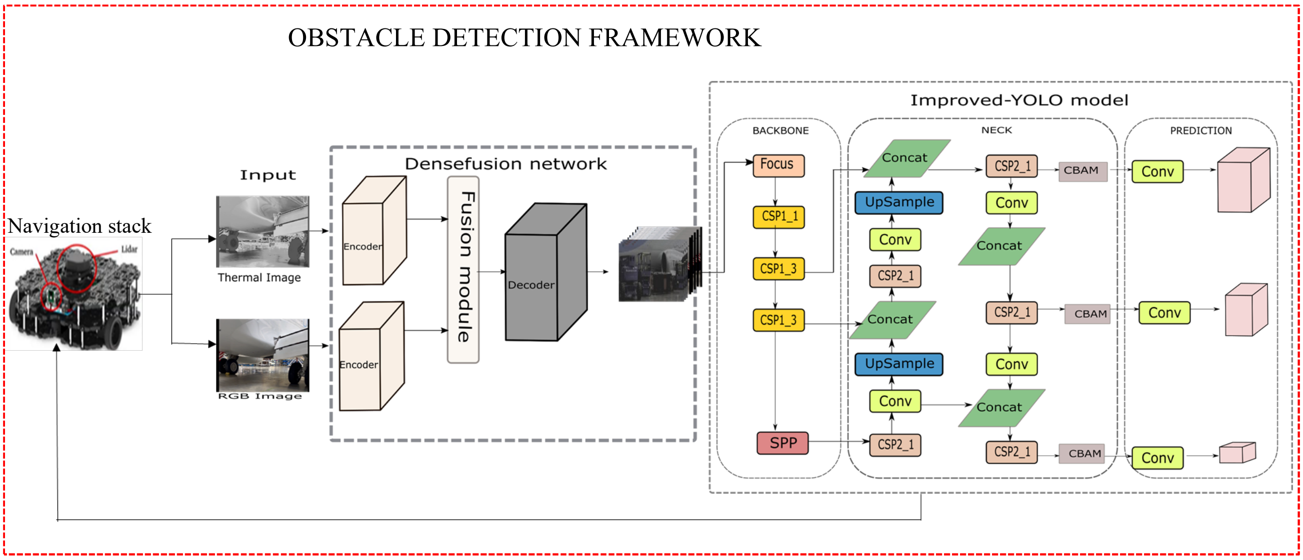

Figure 2. The Fuse-YOLO robot navigation architecture for object detection.

3.2. Proposed Object Detection Model.

The proposed model enhances thermal images by fusing RGB and thermal images to improve the effectiveness of existing deep neural network models in object detection for AMR navigation. Thermal images have lower resolution, lack good texture, and have blurred visual effects, which affects the detection of low-intensity and small objects. The challenges reduce object detection performance in many networks and have recently seen reasonable improvement through data augmentation and fusion networks. Image fusion significantly improves image quality and clarity compared to a single image and Densefuse [11] among other fusion methods, has shown great enhancement in object detection. The network uses encoder-fusion-decoder architecture with convolutional layers that extract rough features and dense blocks that extract high-level depth features from the input. The feature maps from the RGB and thermal pictures are concatenated and passed via a dense block module during the fusion process. The fused feature maps are then fed through a decoding network, which reconstructs the fused image using upsampling layers and convolutional layers to improve visual quality. The thermal images were converted to grayscale to reduce computational demand. The reconstructed fused images were integrated to the backbone region of the YOLOv5 model to improve object detection.
At the input level, the model structure was optimised by redesigning the YOLOv5 input structure to take a four-channel image stream and removed the mosaic data augmentation network, which was added to enrich the datasets but impact negatively in generalisation performance as small objects in the datasets become smaller and difficult to recognise. A

redefined data augmentation structure was introduced using different transformation such as random image cropping, flipping, rotation, and contrasts adjustment for both visible, thermal images and fused images. The overall Fuse-YOLO architecture presented in Figure 2 shows the structure of the thermal image enhancement process for an improved robot navigation experience. The fused input images are fed into the backbone network for feature extraction, then to the neck region that performs cross-stage feature fusion using an attention mechanism to enhance the distinctiveness of relevant features.

The convolutional block Attention module (CBAM) [33] shown in Figure 3 is an attention mechanism and was integrated at the neck layer to adaptively refine the layered feature maps. This contains the channel attention module (CAM) and the spatial attention module (SAM). The network structure was redesigned to improve less effective features by assigning different weights to the feature maps. The channel attention module compresses feature maps using global average and global maximum pooling, while the spatial attention module uses mean and maximum pooling to produce relevant feature maps [34]. This helps to focus and extract useful features for the target object. Optimisation parameters were fine tune accordingly and the final multiscale output was extracted using a fully connected SoftMax layer to output detected feature maps and applied anchor boxes with object class probability and confidence scores.
The fuse-YOLO was pre-trained on MS COCO [35] datasets and the experiment was optimised with different hyperparameters for the relevant output. This process was repeated, observing whether specific parameters complemented or diminished each other and adding more complex combinations progressively.

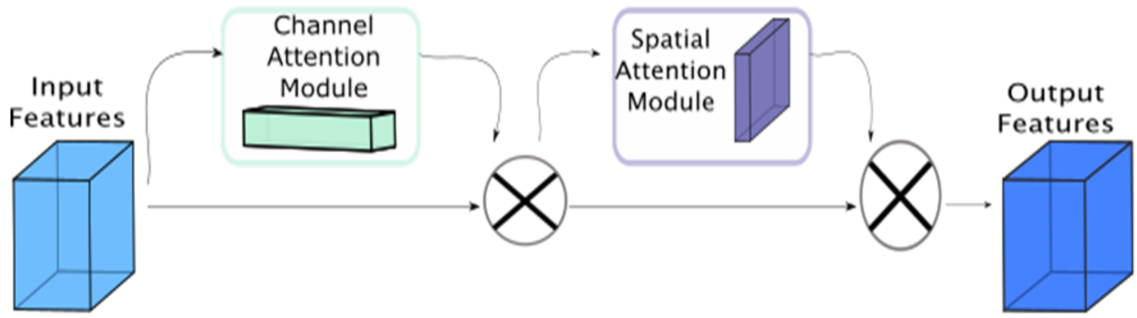

Figure 3. The structure of CBAM

3.3. Data Preparation:

Our custom dataset was generated from Cranfield University DARTeC hangar that hosts Boeing B737 aircraft and TUI hangar [38] with standard RGB and FLIR T560 thermal camera. The datasets consist of 365 pairs of thermal and RGB images, manually annotated based on 13 categories of objects. This includes crowded image with instance of objects found in maintenance, repair and overhaul (MRO) hangar environments such as Chairs, airstairs, person, waste-packs, cable, materials, stands, toolbox, airplane-body, ground vehicle, trolley, electricals and tire.
We use customised data augmentation techniques to increase the diversity of the training data, which was relatively small and as well introduce LLVIP public dataset [36] obtained for object detection in low light conditions. The LLVIP dataset consisted of 15,488 thermal-visible images and was added to increase the number of training samples, assist the model in learning to recognise a broader range of objects and handle various scenarios. With a more extensive and more diverse dataset, the fuse-YOLO model was able to accurately detect and classify objects in the broader range of environments. The dataset was divided into training, validation and testing sets in a ratio of 80:10:10. The custom training dataset contained 290 images, and the test and validation datasets contained 37 images, respectively. The LLVIP custom training set includes 12680, val. is 1586 and test is 15853.
MS COCO [35] benchmark dataset with 200,000 RGB images and 80 annotation class that contains everyday scene was used to pre-train the Yolov5 model.

4. EXPERIMENT AND RESULT

4.1. Model Training:

To train our models and guide our studies, we used a dataset of bounding box annotated objects from the perspective of an autonomous mobile robot navigation for aircraft visual inspection in a hangar environment. The software configuration of the proposed method includes PyTorch framework, python 3.9, and CUDA version 11, running on Windows 10 operating system (OS). The OS runs on Intel(R) Core(TM) i7-10750H CPU working at 2.60GHz with 32G memory.
The datasets are of 12 categories of objects with many qualities like size and positions, trained on the Fuse-YOLO algorithm. This was performed in a batch format that is randomly selected and consists of 16 pairs of images per the training datasets. The thermal grayscale image and the RGB image three channel inputs were used, an example is shown in Figure 1. Data augmentation technique was used at the pre-processing stage to capture a greater variety of object angles. The image input is first sampled independently and trained using the custom visible and thermal images. Next, the fused datasets were trained.

The network was initiated using pre-trained model on MS COCO public dataset [35]. During training, we used the Adam optimiser with momentum at 0.9 and learning rate at 0.005 for 80 epochs and reduced to 0.001 when the model became stable for 20 epochs. First, we trained only the RGB and as well the THERMAL dataset, then the THERMAL and RGB dataset together for 100 total epochs set for each training stage. We also conducted an experiment on the LLVIP object detection public dataset to verify the performance of the model, and this was compared with the classical yolov5m model as detailed in Table 1 below. The best weight of each trained network was saved and used to test and detect on the test datasets.

Table 1 Comparative performance of YOLOv5m and Fuse-YOLO on four datasets based on **(mAP)$_{0.5}$**.

Datasets	YOLOv5m (%)	Fuse-YOLO(%)
RGB-Image	58.2	61.8
Thermal-Image	47.4	49.1
|RGB_Thermal	62.5	64.2
LLVIP	73.3	75.6

4.2. Evaluation Metrics

To evaluate the performance of the proposed fuse-Yolov5 model, we used the custom and public LLVIP datasets to compare the detection accuracy and computational efficiency. We will go over some of the evaluation metrics that we employed in our experiment that are widely used to evaluate the performance of object localisation, and classification model include precision and recall, which measure the percentage of true positive detections made by the algorithm and the percentage of true positive detections made by all ground truth objects, respectively. Next is the Intersection over union (IoU) concept that measures the ratio of overlap between the predicted boundaries and ground truth bounding boxes to the region of the union. It measures the correctness of object localisation and is used as a threshold value for accurate measurement with mean Average Precision (mAP) [37]. The mAP represents the standard metric for object detection accuracy evaluation of a detection model. The speed with FPS (Frames per second) is another evaluation metric that is very relevant to real-time applications. See the below formulas representing the evaluation metrics.

$$\mathbf{Precision} = \frac{TP}{TP+FP} \qquad (1)$$

$$\textbf{Recall} = \frac{TP}{TP+FN} \quad (2)$$

$$\boldsymbol{mAP} = \frac{1}{N}\sum AP \quad (3)$$

$$\boldsymbol{AP} = \int_0^1 Pre\,(Rec)dRec \quad (4)$$

$$\boldsymbol{IoU} = \frac{target \cap prediction}{target \cup prediction} \quad (5)$$

Where, TP = True Positive is the number of correct objects detected, FP = False Positive is number of falsely detected objects, FN = False Negative is a number of missed objects, and AP = Average precision (average accuracy of each type of object).

4.3. Results

The YOLO models and fuse-YOLO were trained and tested on four different datasets to verify the performance. The results were evaluated, and performance was compared with standard YOLOv5 models. We used the same datasets and evaluation metrics, including precision, recall and mAP, where mAP IoU threshold was set to 0.5 for each category and then averaged for all the object categories on the classical yolov5(m, l, & x) models. The result of the model's performance on each set of the dataset was evaluated and their performance was recorded as shown in Table 1. including the performance from LLVIP datasets at different illumination conditions.

The Fuse-YOLO improved mean average precision (mAP) by 1.5% on the custom dataset, and 2.3% on the LLVIP and the precision dimension showed slight improvement. The general accuracy was minimal by several per cent in some categories, especially for objects like stands that represent different frames of stands, while others object classes showed significant improvement. To further evaluate the performance of our model, we compare the performance of the datasets using conventional yolov5m, yolov5l and yolov5x models on the fused datasets, as shown in Table 2. The proposed model performed better in accuracy than the YOLOv5 classical models with slight variations. Fuse-YOLO detected varying objects through the fusion network shown in Figure 5, though more improvement is required in the model to increase the bounding box confidence score. The overall experiment results, as graphically represented in Figure 4, showed that the fuse-YOLO proposed approach could significantly increase the identification and detection of objects in complex settings, proving that the method can improve obstacle avoidance in robot navigation, as is the focus of this study.

Table 2: Result Comparison between YOLOv5 models and fuse-YOLO on RGB_thermal.

Models	mAP(0.5)%	mAP(0.5-0.9)%	P(%)	R(%)
Yolo5m	65.8	48.8	72.9	64.6
Yolov5l	63.9	48.8	67.5	66.9
Yolov5x	65.1	50.8	72.5	63.4
Fuse-YOLO	68.1	47.9	63.4	74.9

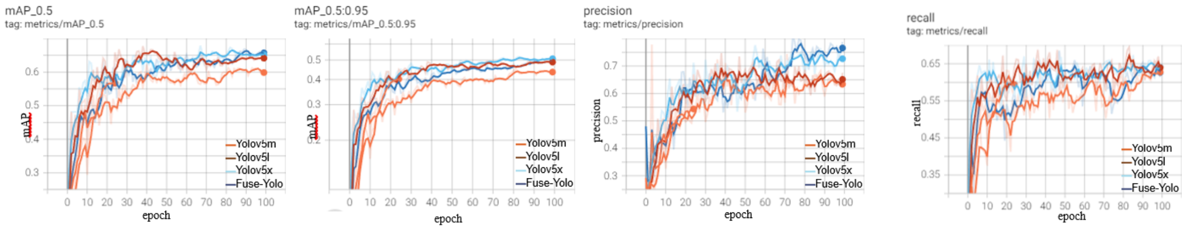

Figure 4. Performance comparison of the yolov5 models and Fuse-YOLO

Thermal image RGB image RGB_Thermal image

Figure 5. Object detection results on the model test using RGB_Thermal image, thermal and RGB images

CONCLUSIONS

The motivation of this work is to develop a system that improves the accurate detection of objects of different structures in light variations, and this is a significant challenge in the design of autonomous mobile robots for aircraft visual inspections. The research proposes a multi-spectral convolutional neural network called fuse-YOLO that combines the features information of RGB and thermal images together. The fuse-yolov5 model was developed based on Yolov5m network and can detect objects varying at different light variations compared to the baseline model. The work significantly improved real-time object detection accuracy by combining rgb and thermal images to help identify features under bright and low light intensity, respectively.

The result of the experiment demonstrated the efficiency of the algorithm and was verified based on the custom and LLVIP public datasets. Compared with the standard Yolov5 network, mAP overall value was improved by approximately 2% and most object classes AP were slightly improved.

Our model shows great potential in identifying objects even in low light states and facilitates application in optimising robot obstacle avoidance and motion control decision in complex environments.

Further work will be done by training the images independently and fusing the output at feature fusion or decision fusion level to improve model performance and generalisation capability.

REFEERENCE

[1] A. Chtourou, P. Merdrignac, and O. Shagdar, “Collective Perception service for Connected Vehicles and Roadside Infrastructure,” *IEEE Vehicular Technology Conference*, vol. 2021-April, Apr. 2021, doi: 10.1109/VTC2021-SPRING51267.2021.9448753.

[2] B. Khalid, A. M. Khan, M. U. Akram, and S. Batool, “Person Detection by Fusion of Visible and Thermal Images Using Convolutional Neural Network,” *2019 2nd International Conference on Communication, Computing and Digital Systems, C-CODE 2019*, pp. 143–148, Apr. 2019, doi: 10.1109/C-CODE.2019.8680991.

[3] H. Li, X.-J. Wu, and J. Kittler, “Infrared and Visible Image Fusion using a Deep Learning Framework,” *Proceedings - International Conference on Pattern Recognition*, vol. 2018-August, pp. 2705–2710, Apr. 2018, doi: 10.1109/ICPR.2018.8546006.

[4] S. Rajkumar and P. V. S. S. R. C. Mouli, “Infrared and Visible Image Fusion Using Entropy and Neuro-Fuzzy Concepts,” *Advances in Intelligent Systems and Computing*, vol. 248 VOLUME I, pp. 93–100, 2014, doi: 10.1007/978-3-319-03107-1_11/COVER.

[5] J. Saeedi and K. Faez, “Infrared and visible image fusion using fuzzy logic and population-based optimisation,” *Appl Soft Comput*, vol. 12, no. 3, pp. 1041–1054, Mar. 2012, doi: 10.1016/J.ASOC.2011.11.020.

[6] W. Ding, D. Bi, L. He, and Z. Fan, “Infrared and visible image fusion method based on sparse features,” *Infrared Phys Technol*, vol. 92, pp. 372–380, Aug. 2018, doi: 10.1016/J.INFRARED.2018.06.029.

[7] “IEEE Xplore Full-Text PDF:” https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7965570 (accessed Mar. 16, 2023).

[8] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, “THIS PAPER HAS BEEN ACCEPTED BY IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS FOR PUBLICATION 1 Object Detection with Deep Learning: A Review”.

[9] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, “FusionGAN: A generative adversarial network for infrared and visible image fusion,” *Information Fusion*, vol. 48, pp. 11–26, Aug. 2019, doi: 10.1016/J.INFFUS.2018.09.004.

[10] H. Li, X.-J. Wu, and T. Durrani, “NestFuse: An Infrared and Visible Image Fusion Architecture based on Nest Connection and Spatial/Channel Attention Models,” *IEEE Trans Instrum Meas*, vol. 69, no. 12, pp. 9645–9656, Jul. 2020, doi: 10.1109/TIM.2020.3005230.

[11] H. Li and X.-J. Wu, “DenseFuse: A Fusion Approach to Infrared and Visible Images,” *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2614–2623, Apr. 2018, doi: 10.1109/TIP.2018.2887342.

[12] “(15) (PDF) A comparative study of YOLOv5 models performance for image localization and classification.” https://www.researchgate.net/publication/363824867_A_comparative_study_of_YOLOv5_models_performance_for_image_localization_and_classification (accessed Apr. 05, 2023).

[13] S. M. Pan and D. H. Madill, “Generalised sliding window algorithm with applications to frame synchronisation,” *Proceedings - IEEE Military Communications Conference MILCOM*, vol. 3, pp. 796–800, 1996, doi: 10.1109/MILCOM.1996.571384.

[14] “Using Machine Learning to Determine Fold Class and Secondary Structure Content from Raman Optical Activity and Raman Vibrational Spectroscopy Myra Kinalwa-Nalule”.

[15] [object Object], “Template Matching Advances and Applications in Image Analysis”.

[16] J. Wang, L. Song, Z. Li, H. Sun, J. Sun, and N. Zheng, “End-to-End Object Detection with Fully Convolutional Network,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 15844–15853, Dec. 2020, doi: 10.1109/CVPR46437.2021.01559.

[17] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” *Adv Neural Inf Process Syst*, vol. 28, 2015, Accessed: Mar. 13, 2023. [Online]. Available: https://github.com/

[18] X. Chen and A. Gupta, “An Implementation of Faster RCNN with Study for Region Sampling,” Feb. 2017, Accessed: Mar. 20, 2023. [Online]. Available: https://arxiv.org/abs/1702.02138v2

[19] A. Benjumea, I. Teeti, F. Cuzzolin, and A. Bradley, “YOLO-Z: Improving small object detection in YOLOv5 for autonomous vehicles”.

[20] S. Li, Y. Li, Y. Li, M. Li, and X. Xu, “YOLO-FIRI: Improved YOLOv5 for Infrared Image Object Detection,” *IEEE Access*, vol. 9, pp. 141861–141875, 2021, doi: 10.1109/ACCESS.2021.3120870.

[21] Y. Chen, H. Xie, and H. Shin, “Multi-layer fusion techniques using a CNN for multispectral pedestrian detection,” *IET Computer Vision*, vol. 12, no. 8, pp. 1179–1187, Dec. 2018, doi: 10.1049/IET-CVI.2018.5315.
[22] “(16) (PDF) Multispectral Pedestrian Detection using Deep Fusion Convolutional Neural Networks.” https://www.researchgate.net/publication/302514661_Multispectral_Pedestrian_Detection_using_Deep_Fusion_Convolutional_Neural_Networks (accessed Mar. 23, 2023).
[23] M. O. Gani, S. Kuiry, A. Das, M. Nasipuri, and N. Das, “Multispectral Object Detection with Deep Learning,” *Communications in Computer and Information Science*, vol. 1406 CCIS, pp. 105–117, 2021, doi: 10.1007/978-3-030-75529-4_9.
[24] K. Roszyk, M. R. Nowicki, and P. Skrzypczyński, “Adopting the YOLOv4 Architecture for Low-Latency Multispectral Pedestrian Detection in Autonomous Driving,” *Sensors 2022, Vol. 22, Page 1082*, vol. 22, no. 3, p. 1082, Jan. 2022, doi: 10.3390/S22031082.
[25] D. Guan, Y. Cao, J. Yang, Y. Cao, and M. Y. Yang, “Fusion of Multispectral Data Through Illumination-aware Deep Neural Networks for Pedestrian Detection,” *Information Fusion*, vol. 50, pp. 148–157, Feb. 2018, doi: 10.1016/j.inffus.2018.11.017.
[26] C. Li, D. Song, R. Tong, and M. Tang, “Illumination-aware faster R-CNN for robust multispectral pedestrian detection,” *Pattern Recognit*, vol. 85, pp. 161–171, Jan. 2019, doi: 10.1016/J.PATCOG.2018.08.005.
[27] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, “YOLOv4: Optimal Speed and Accuracy of Object Detection,” Apr. 2020, Accessed: Apr. 02, 2023. [Online]. Available: https://arxiv.org/abs/2004.10934v1
[28] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, “Path Aggregation Network for Instance Segmentation,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 8759–8768, Mar. 2018, doi: 10.1109/CVPR.2018.00913.
[29] J. Yao, J. Qi, J. Zhang, H. Shao, J. Yang, and X. Li, “A Real-Time Detection Algorithm for Kiwifruit Defects Based on YOLOv5,” *Electronics 2021, Vol. 10, Page 1711*, vol. 10, no. 14, p. 1711, Jul. 2021, doi: 10.3390/ELECTRONICS10141711.
[30] “(18) (PDF) A comparative study of YOLOv5 models performance for image localization and classification.” https://www.researchgate.net/publication/363824867_A_comparative_study_of_YOLOv5_models_performance_for_image_localization_and_classification (accessed Apr. 03, 2023).
[31] N. Bjorck, C. P. Gomes, B. Selman, and K. Q. Weinberger, “Understanding Batch Normalisation,” *Adv Neural Inf Process Syst*, vol. 31, 2018.
[32] A. M. Fred Agarap, “Deep Learning using Rectified Linear Units (ReLU),” Mar. 2018, Accessed: Apr. 03, 2023. [Online]. Available: https://arxiv.org/abs/1803.08375v2
[33] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional Block Attention Module.” pp. 3–19, 2018.
[34] L. Miao, N. Li, M. Zhou, H. Zhou Lize Miao, and H. Zhou, “CBAM-Yolov5: improved Yolov5 based on attention model for infrared ship detection,” *https://doi.org/10.1117/12.2631130*, vol. 12168, pp. 564–571, Mar. 2022, doi: 10.1117/12.2631130.
[35] T.-Y. Lin *et al.*, “Microsoft COCO: Common Objects in Context,” *Proceedings - International Conference on Image Processing, ICIP*, pp. 3076–3080, 2014, doi: 10.1109/ICIP46576.2022.9897741.
[36] X. Jia, C. Zhu, M. Li, W. Tang, and W. Zhou, “LLVIP: A Visible-infrared Paired Dataset for Low-light Vision”, Accessed: Mar. 30, 2023. [Online]. Available: https://bupt-ai-cz.github.io/
[37] P. Henderson and V. Ferrari, “End-to-end training of object class detectors for mean average precision,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10115 LNCS, pp. 198–213, Jul. 2016, doi: 10.1007/978-3-319-54193-8_13.
[38] TUI©. Available online: https://www.tuigroup.com/en-en (accessed on 12 April 2023).

2023-06-12

CNN-fusion architecture with visual and thermographic images for object detection

Adiuku, Amaka

SPIE

Adiuku N, Avdelidis NP, Tang G, etal., (2023) CNN-fusion architecture with visual and thermographic images for object detection. In: SPIE Defense + Commercial Sensing, 30 April - 4 May 2023, Orlando, Florida, Volume 12536, Thermosense: Thermal Infrared Applications XLV

https://doi.org/10.1117/12.2665984