6-2022

HUMAN DETECTION FROM AERIAL IMAGERY FOR SEARCH AND RESCUE OPERATIONS

Namat Ahmad Bachir
United Arab Emirates University

جامعة الإمارات العربية المتحدة
United Arab Emirates University

MASTER THESIS NO. 2022:32

College of Engineering

Department of Electrical and Communication Engineering

HUMAN DETECTION FROM AERIAL IMAGERY FOR SEARCH AND RESCUE OPERATIONS

Namat Ahmad Bachir

June 2022

United Arab Emirates University

College of Engineering

Department of Electrical and Communication Engineering

HUMAN DETECTION FROM AERIAL IMAGERY FOR SEARCH AND RESCUE OPERATIONS

Namat Ahmad Bachir

This thesis is submitted in partial fulfilment of the requirements for the degree of Master of Science in Electrical Engineering

June 2022

United Arab Emirates University Master Thesis
2022: 32

Cover: Detection of Human from UAV imagery

(Photo: By of Jason J. Hatfield / iStock (drone))

Print: University Print Service, UAEU 2022

Declaration of Original Work

I, Namat Ahmad, the undersigned, a graduate student at the United Arab Emirates University (UAEU), and the author of this thesis entitled "*Human Detection from Aerial Imagery for Search and Rescue Operations*", hereby, solemnly declare that this is the original research work done by me under the supervision under supervision of Dr. Qurban A Memon, in the College of Engineering at UAEU. This work has not previously formed the basis for the award of any academic degree, diploma or a similar title at this or any other university. Any materials borrowed from other sources (whether published or unpublished) and relied upon or included in my thesis have been properly cited and acknowledged in accordance with appropriate academic conventions. I further declare that there is no potential conflict of interest with respect to the research, data collection, authorship, presentation and/or publication of this thesis.

Student's Signature:

Date: 19-5-2022

Approval of the Master Thesis

This Master Thesis is approved by the following Examining Committee Members:

1) Advisor (Committee Chair): Qurban A Memon

   Title: Associate Professor

   Department of Engineering

   College of Electrical Engineering

   Signature _____ Qurban Ali __ Date 18-06-2022

2) Member: Dr. Addy Wahyudie

   Title: Associate Professor

   Department of Electrical and Computer Engineering

   College of Engineering

   Signature addy wahyudie Date 18-06-2022

3) Member (External Examiner): Dr. M. Yakoob Siyal

   Title: Associate Professor

   Department of Electrical and Computer Engineering

   Institution: Nanyang Technological University, Singapore

   Signature ____________ Date 18-06-2022

This Master Thesis is accepted by:

Acting Dean of the College of Engineering: Professor Mohamed H. Al-Marzouqi

Signature Mohamed AlMarzouqi

Date July 01, 2022

Dean of the College of Graduate Studies: Professor Ali Al-Marzouqi

Signature Ali Hassan

Date July 01, 2022

Abstract

Mountain recreation is becoming more popular with mountaineering, rock climbing, skiing, mountain biking, hiking, and mushroom picking among the most popular sports. Despite this tendency, there is currently limited research available explaining the rise in search and rescue (SAR), as well as the injuries and illnesses that entail SAR aid in tourist-friendly mountain and desert areas. The objective of a search and rescue operation is to scan the farthest area feasible in the shortest amount of time in order to locate a lost or wounded individual. In the past decade, several new and spectacular uses for drones, including search and rescue, surveillance, traffic monitoring, and weather monitoring, have been created and deployed. Current advancements in drone technology have resulted in major modifications that enable drones to conduct a vast array of tasks with an increasing degree of complexity. Missions such as search and rescue and forest surveillance need a vast camera coverage, making drones an ideal tool for performing complex tasks.

Meantime, the rising prevalence of deep learning applications in computer vision offers exceptional insight into these research areas. In search and rescue operations, the main object is the human being; however, recordings from a bird's eye perspective are not embedded or incorporated in the large datasets that are used to train these cutting-edge detectors. To attain the best potential detection accuracy of the model, the dataset, which is to be employed for training, should contain conditions comparable to those that are encountered where model is to be tested. Hence, it is required to train the model with data, which is obtained using a bird's eye perspective. A recent dataset (SARD) has been used to detect a person's presence in mountain spots. The research conducted in this work proposes a method for identifying the presence of human's mountain setting utilizing an algorithm for human detection with a deep learning framework. Even if the individual is partially veiled, a trained deep learning system can recognize from a variety of perspectives. Existing well-known detectors such as YOLO-v4, RetinaNet, Faster R-CNN, and Cascade R-CNN have been investigated in previous research on various datasets to simulate rescue scenes. Although those algorithms achieve good recall, the other recent detector such as YOLO-v5 may be investigated for comparative performance. Thus, in this research, YOLO-v5 is trained on SARD dataset to validate its speed and accuracy, as well as the small number of false detections. It turns out that it achieves the highest mean average accuracy of 96.9% compared with other detectors. Experimental results using YOLO-v5 conducted on SARD dataset are presented for comparison.

Keywords: Search and Rescue, Aerial Imagery, UAV, YOLO-v5.

Title and Abstract (in Arabic)

استخدام الشبكات العصبية للكشف عن الاشخاص في عمليات البحث والإنقاذ

الملخص

أصبح الترفيه الجبلي أكثر شيوعًا ، حيث أصبح تسلق الجبال وتسلق الصخور والتزلج وركوب الدراجات في الجبال والمشي لمسافات طويلة وقطف الفطر من بين أكثر الرياضات شعبية. على الرغم من هذا الاتجاه ، إلا أن هناك أبحاثًا محدودة متاحة حاليًا تشرح الزيادة في عمليات البحث والإنقاذ (SAR)، فضلاً عن الإصابات والأمراض التي تنطوي على مساعدة البحث والإنقاذ في المناطق الجبلية والصحراوية الصديقة للسياح. الهدف من عملية البحث والإنقاذ هو البحث في أبعد منطقة ممكنة في أقصر فترة زمنية لتحديد مكان الشخص المفقود أو المصاب. في العقد الماضي ، تم إنشاء ونشر العديد من الاستخدامات الجديدة والمذهلة للطائرات بدون طيار ، بما في ذلك البحث والإنقاذ والمراقبة ومراقبة حركة المرور ومراقبة الطقس. أدت التطورات الحالية في تكنولوجيا الطائرات بدون طيار إلى تعديلات كبيرة تمكن الطائرات بدون طيار من إجراء مجموعة واسعة من المهام بدرجة متزايدة من التعقيد. تحتاج المهام مثل البحث والإنقاذ ومراقبة الغابات إلى تغطية واسعة للكاميرا ، مما يجعل الطائرات بدون طيار أداة مثالية لأداء المهام المعقدة.

في غضون ذلك ، يوفر الانتشار المتزايد لتطبيقات التعلم العميق في رؤية الكمبيوتر نظرة ثاقبة استثنائية في مجالات البحث هذه. في عمليات البحث والإنقاذ ، يكون الكائن الأساسي هو الشخص ؛ ومع ذلك ، لا يتم تضمين التسجيلات من منظور عين الطائر في مجموعات البيانات الكبيرة المستخدمة لتدريب هذه الكواشف المتطورة. لتحقيق أفضل دقة محتملة لنموذج الكشف ، يجب أن تحتوي مجموعة البيانات التي يتم تدريب النموذج عليها على شروط مماثلة لتلك التي تمت مواجهتها أثناء اختبار النموذج. وبالتالي ، من الضروري تدريب النموذج بالبيانات التي تم الحصول عليها من منظور عين الطائر. تم استخدام مجموعة بيانات حديثة (SARD) لاكتشاف وجود شخص في المناطق الجبلية. يقترح البحث الذي تم إجراؤه في هذا العمل طريقة لتحديد وجود بيئة جبلية للإنسان باستخدام خوارزمية لاكتشاف الكائن البشري وإطار عمل التعلم العميق. حتى لو كان الفرد محجبا جزئيا ، يمكن لنظام التعلم العميق المدرب التعرف من مجموعة متنوعة من وجهات النظر. تم فحص أحدث أجهزة الكشف الحالية مثل Faster R-CNN و YOLO-v4 و RetinaNet و Cascade R-CNN في بحث سابق على مجموعات بيانات مختلفة لمحاكاة مشاهد الإنقاذ. على الرغم من أن هذه الخوارزميات تحقق استرجاعًا جيدًا ، إلا أنه قد يتم فحص الكاشف الحديث الآخر مثل YOLO-v5 من أجل الأداء

المقارن. وبالتالي ، تم تدريب YOLO-v5 على مجموعة بيانات الزراعة والتنمية الريفية المستدامتين في هذا البحث للتحقق من سرعته ودقته ، بالإضافة إلى العدد الصغير من الاكتشافات الخاطئة. اتضح أنها تحقق أعلى متوسط دقة متوسط 96.9٪ مقارنة بأجهزة الكشف الأخرى. تم عرض النتائج التجريبية باستخدام YOLO-v5 التي أجريت على مجموعة SARD للمقارنة.

مفاهيم البحث الرئيسية: البحث والإنقاذ ، الصور الجوية ، الطائرات بدون طيار ، الشبكة يولو – ف5

Acknowledgements

My thanks go to Dr. Qurban A Memon who introduced me to the exciting field of computer vision and whose endless encouragement and continuous support motivated me to this area of research and most other studies in which I have been involved.

I would like to thank my committee for their guidance, support, and assistance throughout my preparation of this thesis. I would like to thank the chair and all members of the department of electrical engineering at the United Arab Emirates University for assisting me all over my studies and research.

Special thanks go to my parents, husband, brothers, and sisters who helped me along the way. In addition, special thanks are extended to the Bachir family for their assistance and friendship.

Dedication

To my beloved parents and family

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AP	Average Precision
CNN	Convolutional neural networks
FN	False Negatives
FP	FP: False Positives
TN	True Negatives
TP	True Positives
SSD	Single Shot Multi-Box Detector
SAR	Search and Rescue operations
IoU	Intersection of Union
mAP	Mean Average Precision
UAV	Unmanned Aerial Vehicle
YOLO	You Only Look Once

Chapter 1

Chapter 1: Introduction

In this chapter, background of search and rescue operations is detailed along with problems associated with it. At the end, the problem statement is developed, which is addressed in this thesis.

1.1 Overview

One of the most critical computer vision tasks is object detection. The problem becomes more challenging when the object of interest is small, due to a limited resolution and information. In this study, we focus on small object detection in search and rescue (SAR) operations where human is the target. Search operations are typically applied where activities such as lost in desert, hiking, mountain biking, paragliding, free climbing, and rafting are carried out. As a result, the necessity to safeguard persons in harsh environment and difficult-to-reach regions such as forests, mountains, desert, canyons, and caverns is rising. Due to the nature of these environments and lack of physical and mental preparation involving such activities, a rising number of injuries such as sliding, burying, and so on are occurring. The main goal of such operations is to locate missing people who are injured possibly. In addition to the possibility of injury, hikers face risks related to their expertise in dealing with potential situations. Sudden weather changes, insufficient preparation, inappropriate equipment or clothing, non-compliance to warnings and timely information or overestimation of one's talents or understanding can all lead to emergencies. Disorientation and sickness are typical causes for missing people reports. The likelihood of a missing person's survival drops over time, while the search area expands exponentially [1].

Search and rescue teams are always working to enhance and upgrade their daily routine operations by creating strategies for promptly locating people who have been lost in such environments. Traditionally, dogs are trained and employed to locate such lost individuals; however, the number of dogs required, and the time needed for such tasks is prohibitive. Using humans to aid in such search is an alternate method of locating individuals, albeit it is time-consuming and needs the coordination of numerous volunteers and experts [2].

To increase the survival rates, innovative technologies such as unmanned aerial vehicles (UAV) have been adopted, which can assist in identifying missing people faster as well as decreasing the cost of search and rescue operations. According to [3], the number of people estimated by about 59 have been rescued, until 2017, using drones in difficult-to-reach environments in 18 separate cases around the world.

1.2 Statement of the problem

There is a necessity to initiate a search and rescue sequence for providing required healthcare to individuals, who participate in different adventure related sports or involve in tourism that requires to stay in desert areas, mountains and other different-to-reach regions. The purpose of a SAR operation is to search as much of the land as possible in the shortest amount of time to locate a missing or wounded individual. Nowadays, unmanned aerial vehicle (UAV) has been used extensively as a significant resource in search operations where the landscape is scanned and photographed. Detection of persons or any other objects using the unmanned aerial vehicle (UAV) captured imagery is a challengeable task due to position variations, weather conditions, humans' relatively small size, and camouflaged environment. Therefore, an automatic detection system of people or objects is deemed necessary.

To facilitate and increase detection accuracy of objects and accelerate image processing, several image-processing algorithms have been developed and reported in literature. The latest detectors reported in literature are based on either deep learning networks, for example YOLO, RetinaNet, Faster R-CNN, and Cascade R-CNN or hybrid ones. The datasets for SAR operations are limited due to lack of similar conditions typically found in real life scenarios. Most of these detectors' performance evaluations also include speed for real timeliness, accuracy and number of false detections during testing. In SAR situation, the main object under search is a human being, however, seen or recorded from above. Typically, such camera views or recordings are not found in the large datasets, which are used to train well-known object detectors. To obtain the maximum possible accuracy for object detection, the dataset on which the model is to be trained should have similar environmental parameters that are under investigation. Thus, it becomes essential to train the model using views or recordings done from above.

Chapter 2

Chapter 2: Literature Review

In this chapter, survey of literature is presented that is relevant to this topic and the research works that have been published in the same direction.

2.1 Object detection challenges

Recently, extensive research has been conducted to provide several algorithms for object detection in aerial images. Low visibility owing to different elevations, the variations in position and scale, object of interest, disguised surroundings with trees, bushes and trees, and high-resolution aerial photos all play a role in object detection in aerial images [4]. Detection of objects from aerial images is still deemed a tough challenge [5], even though typical ground imagery has generated acceptable results in detection of objects. One crucial challenge is to save individuals in search and rescue (SAR) tasks using aerial imagery without incurring any losses. In practice, to locate missing people, search and rescue efforts must be carried out as promptly as feasible. It can be quite costly, and it necessitates a variety of actions such as deploying big groups of people, sniffing dogs, and a large number of air and ground vehicles such as helicopters and cars. The National Police Air Service of UK, for example, registered almost about 17,000 mission hours by 2016-17, with each operation hour costing above £2800 [6].

2.2 Object detection technologies

Humans can be recognized using several technologies such as machine learning approaches [7-9] or using thermal imaging techniques [10], which avoid the large costs and time commitments associated with traditional SAR methods. Machine learning is a technique that is utilized for a variety of road and air applications, such as mapping a specific location with drones, autonomous driving and detecting persons in search and rescue missions. A lot of aerial data is needed to train and recognize persons with machine learning-based human detection.

The alternative method of human detection from aerial imagery is the use of thermal cameras fitted in UAVs to obtain real time feedback. Since thermal infrared cameras have relatively fewer pixels, it turns out that reception of images becomes quite easy. However, the deployment of thermal cameras is not always possible because detection of people using thermal cameras is not reliable in certain weather conditions. In cold terrain, the temperature of the human body is higher than that of the environment, so humans appear clearer and bright using thermal cameras. While in tropical terrain or in summer, the human body temperature is much lower than that of the environment, so it becomes a challenge to detect persons in such environments. The use of thermal infrared cameras [11] for detection of objects have the limitation of requiring the person/object carry bulky equipment. The SARD dataset [12] used to train the model does not include any thermal imaging aerial photos, which is another reason not to include thermal imaging in the research. By combining thermal and visual imaging, Authors in [13] were able to create a real-time human and vehicle detection system. Thermal and visual imaging were also employed by authors in [10] to locate people in various stances on the ground in video sequences.

2.3 Human detection in SAR operations

To recognize people from aerial photos taken using drones, researchers [14] used image segmentation, contrast augmentation, and SSD detector. They also created synthetic datasets representing search and rescue operations with the use of a 3D game editor. The authors in [15] combined the proposed model with support vector machine (SVM) technique to search using UAVs for persons caught in an avalanche. In a similar task [16], the focus is detecting individuals at sea. Th imagery is taken by aerial vehicles equipped with multispectral camera, and MobileNet architecture. In another work, using a dataset involving GPS location calculation, the authors [18] constructed a system for detecting persons and recognizing actions. In another example [18], GPS signal is demonstrated in search and rescue operations. This is envisioned considering that the person has a mobile device turned on, and thus the person's location is calculated by combining the GSM signal strength with UAV’s GPS position. In [19], the Tiny YOLO-v3 architecture was used to create a model to identify a person in water. The model is trained on well-known MSCOCO dataset, which was developed in HD resolution (with a GoPro camera) by a UAV.

The authors [20] propose a real-time approach to recognize and track ocean surface objects. In another work [21], the authors present a technique for classifying drone imagery and object detection utilizing semi-supervised and supervised machine learning approaches, as well as a hardware and software architecture proposed for a UAV platform.

Unmanned Aerial Vehicles (UAVs) are aircraft that are piloted by a computer system or a person from a distance [22] which can be deployed for object detection tasks of SAR operations in remote areas. SAR operations have benefited greatly in recent years from aerial imagery captured by unmanned aerial vehicles (UAVs) to explore harsh, inaccessible or difficult-to-reach remote places like mountainous regions or dense forests or woodlands [23].

Several deep learning methods have been created and documented in the literature to make object detection in SAR operations easier and faster, as well as to speed up image processing. Many deep learning architectures have been divided into two frameworks based on the ROI (region of interest) [24]. ROI pooling is followed by object detection and bounding box regression in two-stage detectors, whereas end-to-end detection is done without explicitly extracting object proposals in one-stage detectors. In terms of localization and classification accuracy, two-stage detectors such as Faster RCNN [25], RetinaNet [27], and cascade R-CNN [27], are usually more accurate. They are, however, slower to process than single-stage detectors such as YOLO [28] and SSD [29]. Therefore, single stage detectors are commonly employed in real-time object detection [30]. Some of the research papers [31-32] combined handcrafted methods with hybrid deep learning methods, whereas others [33-35] employed pure convolution neural networks (CNNs).

For training and testing, most CNNs require a fixed small input size that limits network width, depth, and resolution of the image. Because the photographs recorded by drones are typically of high-resolution, it becomes one of the main issues with aerial datasets. Given the limitations of aspect ratio and scaling, some Google researchers created EfficientDET [36-37], a one-stage object detector that proved to be much more efficient than two-stage object detectors. The authors presented a scaling factor that can scale all width, depth, and resolution dimensions evenly. The EfficientNet-B7 [38] obtains a state-of-the-art 84.4 percent accuracy on ImageNet and an average precision of 52.2 percent on MSCOCO test-dev [39], thus, proposing a new coefficient and employing Bi-directional Feature Pyramid Network [36].

Transfer learning is also a machine learning approach that allows to reuse a network that has already been trained on a specific dataset. This technique is especially beneficial for remote sensing images, and when enough data is not available. By using transfer learning, the authors of [40] suggested a deep network model for classification of search and rescue images, which do not require a larger labeled dataset. To speed up the classification of remote sensing data, transfer learning was employed in [41]. According to the authors, transfer learning on bigger, generic natural image datasets outperforms transfer learning that works on tiny remotely sensed datasets. The authors in [42] employed transfer learning in a region-based convolutional network dubbed as a Double Multi-scale Feature Pyramid Network, in which intrinsic multi-scale pyramidal features with low-resolution features are merged with high-resolution features using transfer learning. Using transfer learning in [43], the authors devised a deep neural network model for search and rescue image categorization that does not require dataset that is labelled.

Transfer learning has showed considerable promise in surmounting difficulties faced by a lack of data for training deep learning frameworks, as evidenced by earlier works, and indicates a step toward strong machine learning. Labeled-image databases are commonly utilized for deep neural network training and testing and have been attempted in a research work [40]. In water rescue operations [44], the authors suggested a novel deep architecture that consists of an ensemble of different deep network classifiers coordinated by the fusion module. Individual models have been combined to improve detection results. For example, Faster R-CNN with Feature Pyramid Network with ResNet backbone (50 and 101 layers deep) have been implemented, but there is an evident drawback in the form of increased computing effort. This can be reduced in various ways, such as by sharing feature maps among detection heads with the same backbone. Aside from improved performance, it is a simpler training procedure because each model is trained separately and individually, and with flexibility in the object detector selection, which may be changed by merely a simple weight optimization step. Faster-R-CNN algorithm was used in a web application [45], where authors analyzed raw and processed images. The system outperformed an expert in recall, but the experts outperformed the algorithm in accuracy when analyzing images that had previously been processed and labeled.

The reliability of existing state-of-the-art detectors such as YOLOv4, RetinaNet, Faster R-CNN, and Cascade R-CNN has been examined in [22] using a VisDrone benchmark and custom-made dataset built to mimic rescue scenes. The detection outcomes were compared after the models were trained on selected datasets. The YOLO-v4 detector was chosen for further investigation due to its great speed and accuracy, as well as the low number of false detections. The research [46] offered a novel approach for detecting persons from aerial imagery for search and rescue missions. This approach describes how to train HERIDAL's existing high-resolution database. To tackle person detection challenge, the EfficientDET deep network is trained using a newly built database. In comparison to all current methods, the proposed method attained the accuracy of 93.29 percent mAP. The primary purpose of SAR operations is to improve detection outcomes, particularly in lowering the number of false positive detections and, as a result, improving precision. Thus, the research in [47] offered a strategy in which image sequences are used as the system's input. The proposed method revolves on the idea that an object/person detected in numerous consecutive images is more likely to be a real positive detection, whereas objects/persons detected in only one consecutive image are more likely to be false positives. Multiple neural networks were employed in both the region level and during classification stages of this method.

Images collected during the search and rescue mission are typically processed on board the UAV or provided to a third party for further analysis. However, due to the UAV's limited computational resources, processing high-resolution photos requires a high level of computational complexity, which is challenging to achieve. Images should not be compressed when being transferred from the UAV to the ground station because compression results in information loss, which might have undesirable consequences in that it might need additional processing to detect a very small object of interest. Though lossless compression solutions exist, they necessitate a substantial amount of computational resource on the UAV to run a compression algorithm on high-resolution images, which is often not practical and is very time-consuming. As a result, it is more practical for a UAV to send original photos to a ground station. With this mind, the authors [48] describe an efficient approach for transmitting high-resolution images.

2.4 Weaknesses in current approaches

The deep learning methods are frequently adjusted in numerous ways to be able to conduct small object detection and thus produce noteworthy results to address this problem. One of the obstacles in object recognition tasks from aerial imagery is in recognizing the object from multiple aerial perspectives and angles, in varied positions, when it is partially blocked or obstructed or even in motion. If an object gets rotated, translated, scaled or partially hidden from view, a human is likely to recognize it in an image. However, because of the way computers address this problem, this task is significantly more difficult with computer vision systems. Furthermore, human detection from aerial imagery may result in a large number of false positives. Because the images collected during search and rescue operations are sequential, the continuity of the detected object in several images could be utilized to reduce false positive detections. Unfortunately, only a small body of information exists on the application of deep learning and computer vision techniques to this type of problem [49]. Furthermore, most of the current works employ images obtained by drones to show everyday-life scenes that are unrealistic from a SAR aspect. It is challengeable to have representative datasets for developing such detection methods.

The first weakness in existing methods is the performance, like Faster R-CNN detector based on two-phase region-based detection has lower performance than YOLO detector. In [22], an example of overlapping of persons' detection in an image was recognized by several detectors, but most of them exhibit occlusion issues and failed to detect a person crouching behind a moving figure. Another example described in [50] used salient detection approach that began by narrowing the search space by using a visual attention algorithm to detect the image's most salient or significant segments. This approach used pre-trained and fine-tuned convolutional neural networks (Faster-R-CNN) with a detection rate of 88.9%.

The second weakness in existing methods is the speed of training and testing. One of the most prominent methods, Faster R-CNN, provides a good speed value of 0.2 seconds per image. The Mask R-CNN architecture is compared to YOLOv3 in [51], where Mask R-accuracy CNN's was shown to be much higher in performance than YOLO-v3's, though YOLO-v3 surpassed Mask R-CNN in terms of detection speed. YOLO-v4 is compared to SSD and Faster R-CNN in [52] with 2620 training and 568 test photos. It turned out that the accuracy of YOLO-v4 was found to be much greater than that of SSD and Faster R-CNN, while the detection speed of SSD was found to be significantly faster than that of YOLO-v3 and Faster R-CNN. Although some of the stated methods had a good inference time, the huge dataset of search and rescue application requires less time for analyzing the images and detecting the humans quickly. Run speed is a critical issue in such applications.

Chapter 3

Chapter 3: Models and Methods

The YOLO algorithm has gained many versions over time depending on applications, where it was applied for specific objectives. The improvements range from version 1 to 5. Below, we discuss each.

3.1 YOLO-v1

The YOLO findings concerning algorithm's strength is in its small model size and quick calculation speed. YOLO has an easy-to-understand structure. Through the neural network, it can immediately output the bounding box's position and categorization. Because YOLO just has to upload the image to the network in order to obtain the final results of detection. YOLO approach can also do video time-detection. Since YOLO detects objects directly using the global image, hence it is capable of encoding global information and minimizing the inaccuracy associated with detecting the backdrop as an object. For items that are very close to one another and ingroups, YOLO's test results are poor. This bad performance is caused by the fact that just two boxes in the grid are anticipated and they all belong to a new objects class in the same category, resulting in an unacceptable aspect ratio and the factors such as a lack of generalization capacity. The YOLO architecture in its initial form consisted of 24 convolutional layers followed by fully linked layers. YOLO predicts many bounding boxes per cell, but only with the highest Intersection Over Union (IOU) with the ground truth are chosen, a technique called non-maxima suppression [53]. YOLO has two flaws: the first is imprecise location; the second is a lower recall value than the method using area predictions. In YOLO-v1 [55] approach, the image is split into slots grids, with each grid predicting some B bits of bounding box information directly. While YOLO-v1 has a fast detection speed and a low false detection rate for background photos, its object identification accuracy is very low. It views target detection as a regression problem and suggests an approach to develop target detection algorithms. The Figure 1 depicts the network's structure.

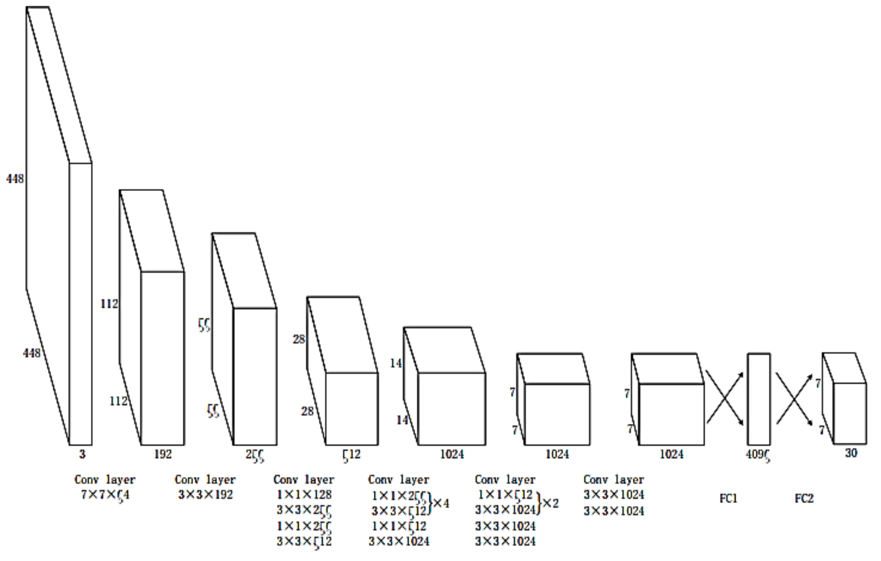

Figure 1: YOLO-v1 structure diagram [54]

3.2 YOLO-v2

YOLO-v2 [55] is also known as YOLO9000 which was released by Joseph Redmon and Ali Farhadi at the end of 2016. The key enhancements in this version are a quicker, and more powerful R-CNN that also includes an object identification method that utilizes a Region Proposal Network to recognize objects in an input image, and Single Shot Multibox Detector (SSD). This approach outperformed well-known detectors at that time, such as Faster R-CNN, on the VOC 2012 detection dataset. Owing to its unique structure, YOLO-v2 may run at approximately 40 frames per second on a GeForce GTX Titan X and between 20-25 frames per second on a GeForce GTX 970. The network is trained in a supervised manner, which means that for each object, the training method must be supplied with both true labels and bounding box coordinates. For each box, the YOLO-v2 approach predicts three properties [56] which are the Intersection over union (IoU) that estimates each anchor box's objectless score, the offsets of the anchor box which adjust the location of the anchor box, and finally the class probability that assigns each anchor box a class label based on the class likelihood. Additionally, because the original YOLO model (dubbed YOLO-v1) suffers from localization errors and low recall predictions, the research in [55] presents YOLO-v2, which incorporates novel and prior work-based improvements, namely SSD, to address the constraints and further improve the speed vs accuracy trade-off. The YOLO-v2 first trains the classifier on ImageNet at 224x224 resolution. Second, it finetunes the classifier for 10 epochs at 448x448 resolution. This forces the network's filters to adapt to higher resolution inputs. Furthermore, YOLO-v2 has a multi-scale training: On YOLO-v1, there is a problem in detecting items with varying sizes of input. This means that if YOLO is trained on smaller photographs of an object, it will have difficulty detecting the same object on larger images. This issue has been addressed to a large extent in YOLO-v2, which is trained on random pictures with dimensions from 320*320 to 608*608. This enables the network to learn and predict objects accurately using different input dimensions. The YOLO-v2 is built on Darknet-19 architecture, which includes nineteen (19) convolutional layers, five (5) max pooling layers, and a softmax layer. The Darknet-19 architecture is illustrated in Figure 2. Compared to YOLO-v1, YOLOv2 results in lower computing costs, higher speed, and increased mean average precision (mAP). Preprocessing the input data with batch normalization significantly enhances the performance of mAP [55].

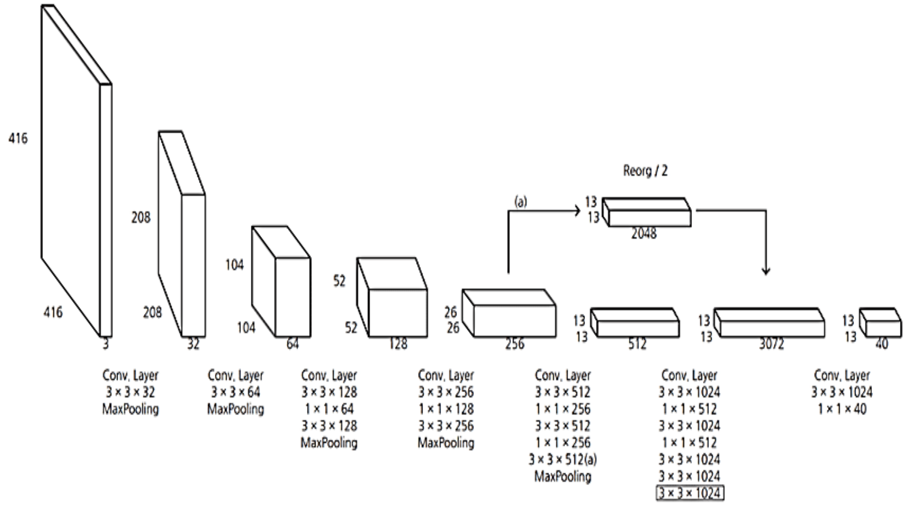

Figure 2: Darknet framework [57]

3.3 YOLO-v3

YOLO-v3 is a real-time object detection system that recognizes items in films, live feeds, or photos. YOLO-v3 is a significant upgrade over the previous two YOLO versions in terms of robustness. This model incorporates multi-scale detection, a more robust feature extraction network, and a few loss function tweaks. To gain a high-level knowledge, the network architecture is broken down into two primary components: the feature extractor and the feature detector (Multi-scale Detector).

YOLO-v3 [58], is an incremental form of YOLO-v2. The image is initially fed to the feature extractor, which extracts feature embeddings, and then to the network's feature detector, which outputs the processed image with bounding boxes surrounding the identified classes.

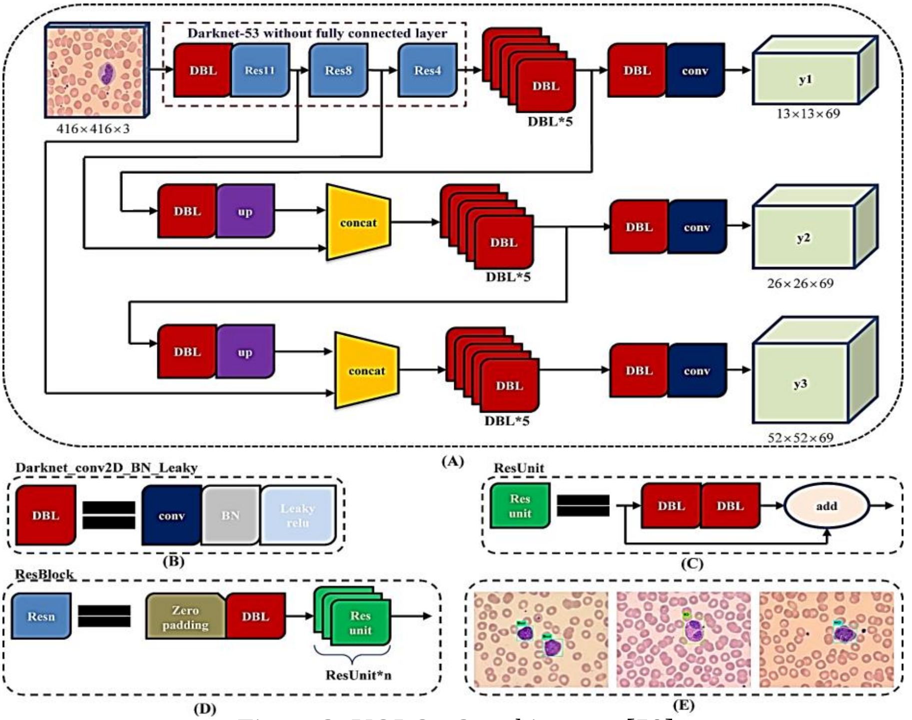

Figure 3: YOLO-v3 architecture [58]

Earlier YOLO versions employed Darknet-19 as a 19-layer feature extractor. YOLO-v2 introduced 11 additional levels to Darknet-19, bringing the total to 30. Nonetheless, the method encountered difficulties recognizing small objects due to the down sampling of the input image, which resulted in the loss of fine-grained features. YOLO-v3 proposed a more robust architecture in which the feature extractor was a combination of YOLO v2, Darknet-53 (an ImageNet-trained network), and Residual networks (ResNet). The network employs 53 convolutional layers (hence the name Darknet-53) and is constructed using a series of 3x3 and 1x1 convolutional layers followed by a skip connection. The darknet's 53 layers are layered with 53 additional layers for the detection head, giving YOLO-v3 a total of 106 layers of fully convolutional underlying architecture as indicated in Figure 3 [58]. Thus, a vast architecture is created, slowing it down slightly in comparison to YOLO-v2, but improving accuracy at the same time.

The multi-scale detector is utilized to guarantee that tiny items are identified as well, in contrast to YOLO-v2, which received continual criticism for this. Concatenating up-sampled layers with prior layers results in the preservation of fine-grained characteristics that aid in the detection of tiny objects.

In YOLO-v2, the last three terms are squared errors; while in YOLO-v3, they have been replaced with cross-entropy error terms. In other words, logistic regression is now used to estimate object confidence and class predictions in YOLO-v3. While training the detector, a bounding box is provided to each ground truth box, whose anchor has the greatest overlap with the ground truth box.

One of the improvements in the third version is that YOLO-v3 represents multilabel classification where no soft maxing the classes. The YOLO-v3 suggest a trade-off between speed and accuracy when YOLO is used instead of RetinaNet, since RetinaNet training time is longer. However, by utilizing a bigger dataset, the accuracy of recognizing objects using YOLO-v3 may be equivalent to that of RetinaNet, making it an attractive alternative for models that can be trained with huge datasets [55]. A frequent example of this is in common detection models such as traffic detection, where a large amount of data may be utilized to train the model due to the abundance of photos of various cars. On the other hand, YOLOv3 may be unsuitable for use with niche models for which huge datasets are difficult to get.

3.4 YOLO-v4

The authors [23] developed YOLO-v4. YOLO-v4 is a development of the YOLO-v3 architecture, which made use of the CSP darknet-53 classifier. Furthermore, it makes use of pyramid pooling and a path aggregation network (PAN) to link the YOLO-v3 head. It is the fastest and most accurate method for detecting several objects in a single frame. The significant improvement in YOLO-v4 is that it is twice as quick as EfficientDet, while maintaining comparable performance. Additionally, the AP metric rose by 10% and FPS metrics rose by 12%, when compared to YOLO-v3.

Typically, an object detector's backbone network is pre-trained on ImageNet categorization. Pre-training implies that the network's weights have been updated to recognize important aspects in an image. CSPDarknet53 is a unique backbone that can be used to enhance CNN's learning ability. The spatial pyramid pooling block is superimposed over CSPDarknet53 in order to expand the receptive field and isolate the most salient background information. Instead of the feature pyramid networks (FPNs) utilized in YOLOv3, the PANet is used to aggregate parameters for multiple detector levels.

The next phase in object detection is to prepare for detection by mixing and combining the features obtained in the ConvNet backbone [65]. YOLOv4 makes use of a "Bag of Freebies," so named because they boost network speed without increasing inference time during production. The majority of the items in the Bag of Freebies are related to data augmentation. YOLOv4 employs data augmentation to extend their training set and expose the model to semantic contexts it would not have encountered otherwise.

YOLO-v4 employs what are known as "Bag of Specials" strategies [65]. These methods add little to inference time but greatly boost performance, making them worthwhile. The authors conduct experiments using a variety of different activation functions. As features move across the network, activation functions change them. It might be challenging to get the network to drive feature creations toward their optimum point using typical activation functions such as ReLU.

A very recent application of YOLO-v4 can be diagnosed [60] for pedestrian detection, where they implement an application that detects people's social distance in public spaces during COVID-19, analyze the danger of infection in this region, and provide recommendations to those who are too near to one another.

As shown in Figure 4, YOLO-v4 [67], is a model that achieves optimal speed with great accuracy. YOLO-v4 is believed as an object detector for production systems and designed for parallel computing. YOLO-v4 enhances YOLO-v3's FPS (12%) and mAP (10%) [68] with resulting speed of double of EfficientDet.

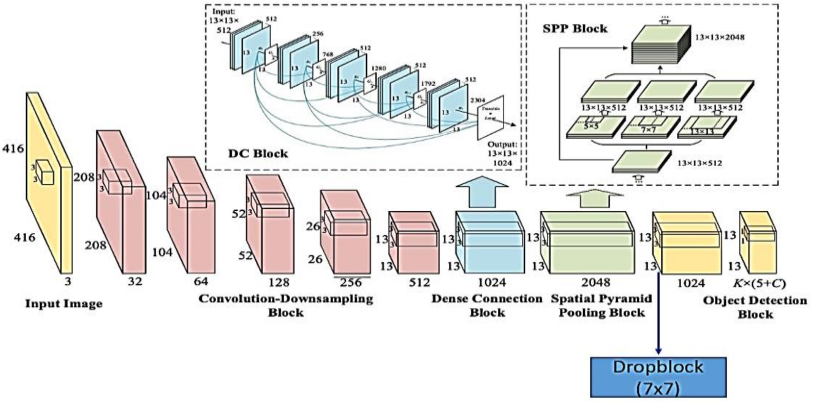

Figure 4: YOLO-V4 architecture [23]

3.5 YOLO-v5

YOLO-v5, the fifth generation of YOLO, is the most recent version that was not produced by the original inventor of YOLO. However, the YOLO-v5 [61] outperforms the YOLO-v4 in terms of accuracy and speed. Glenn Jocher introduced YOLO-v5 shortly after the release of YOLO-v4 utilizing the Pytorch framework.

The most cutting-edge developments in the field of computer vision have been included into YOLO-v5 and preceding versions. YOLO-v5 is a group of compound-scaled object detection models trained on the MSCOCO dataset. It contains straightforward capability for TTA, model assembly, hyperparameter evolution, and export to ONNX, CoreML, and TFLite. Moreover, YOLO-v5 inherits the advantages of YOLO-v4, including the addition of SPP-NET, modification of the SOTA technique, and introduction of new data improvement methods such as mosaic training, self-adversary training (SAT), and multi-channel feature replacing FPN fusion with PANet [62]. Along with network construction advancements in recent years, a group of researchers has been focusing on loss layer advancements. Wen Yandong pioneered the Center Loss monitoring technique [63]. It can significantly improve the capacity of a neural network to recognize deep learning characteristics, especially, lowering the intra-class variations while maintaining the characteristics of various separable classes. To do so, the center loss function was introduced [63]:

$$L_{c=\frac{1}{2}} \sum_{i=1}^{m} \| x_i - c_{y_i} \|_2^2 \qquad (1)$$

where $c_{y_i} \in R_d$ represents the y_{ith} class center with deep features. This formulation characterizes the intra-class differences effectively. Ideally, the c_{y_i} need to be updated when the deep features are modified. Mainly, the full training set should be considered with the average of the features of each class in every iteration, which is wasteful and unworkable. Thus, the center loss cannot be employed directly. This quite possibly be the reason that such a center loss has not been employed in convolutional networks till now.

To overcome this, two important adjustments were suggested. First, instead of updating the centers with regard to complete training set, the update is applied based on mini batch set. During each iteration, the centers are calculated by averaging the characteristics of respective classes. During this instance, some of the centers may not get updated. Second, to keep away huge perturbations due to few mis-labelled samples, a scalar α is used to adjust the learning rate. The gradients of L_C for x_i and update equation of c_{yi} are computed as: $\frac{\partial L_c}{\partial X_i} = x_{i-} c_{y_i}$

$$\Delta_{c_j} = \frac{\sum_{i=1}^{m} \delta(y_i = j).(c_j - x_i)}{1 + \sum_{i=1}^{m} \delta(y_i = j)} \quad (2)$$

where δ (condition) equals 1 when the condition is met and 0 when it is not, α is constrained to [0,1]. After including soft-max loss, the inter-class distance increases while the intra-class distance decreases.

The architecture of yolo-v5 as shown in Figure 5 consists of three parts: (1) Backbone: CSPDarknet, (2) Neck: PANet, and (3) Head: Yolo Layer. The data are first input to CSPDarknet for feature extraction, and then fed to PANet for feature fusion. Finally, Yolo layer outputs detection results (class, score, location, size).

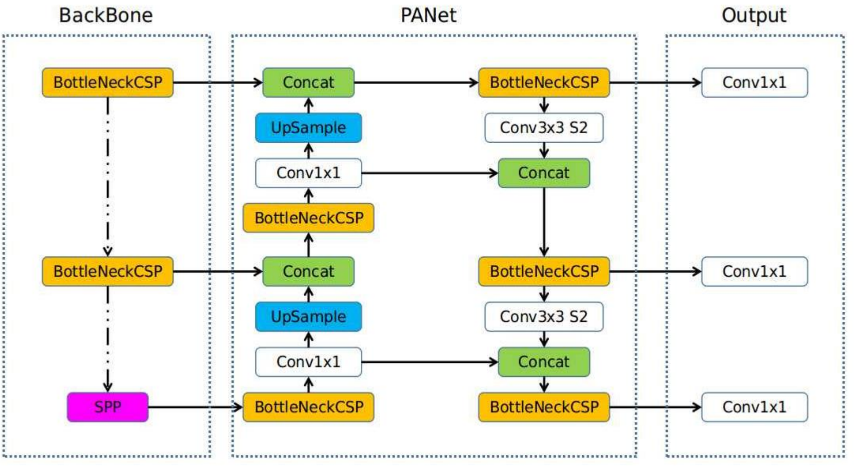

Figure 5: YOLO-v5 framework [63]

3.6 Dataset

Generally, the dataset includes scenes and positions of people that suits a particular goal and does not include search and rescue operation cases. In rescue operations, the main object under search is the human viewed from above. Typically, such records are not available in large datasets, which are used to train well-known detectors. To target high detection accuracy, the trained dataset must have similar conditions to those where the model is tested, so it becomes operative for the model to be trained with data viewed from above.

Recent research reports talk about deep neural networks trained on datasets such as MS COCO, VisDrone, Okutama, ImageNet and Pascal VOC. Each dataset includes scenes and positions of people that suits a particular goal and does not include search and rescue operation cases. Similar sights for SAR operations could be those of the human sitting, walking, setting in park, or lying on a beach. To support rescue cases, SAR dataset was used to train and test UAV images, in which diverse postures of wounded persons are present, in addition to normal positions like standing, sitting, lying, running, etc.

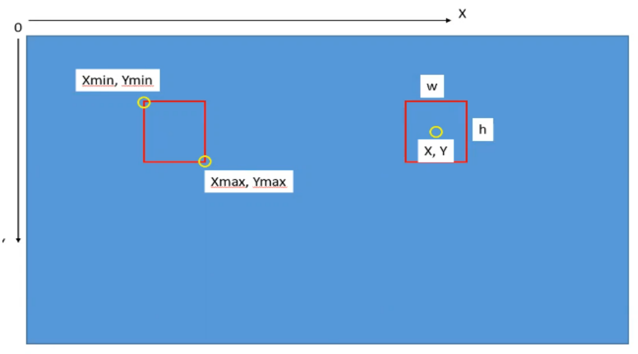

Figure 6: Bounding box(left) , Coordinates(right)

The picture annotation includes the bounding box position around each object of interest, its height and width, and person's class label (Walking, Running, Standing, Lying, Sitting, Not Defined)

For annotation purposes, the first two coordinates represent the center of the bounding box and the next two represent the width and height of the bounding box, as shown in Figure 6.

Using a high-definition camera attached to a drone equipped with a 3-axis solo gimbal platform, the video shooting, as an experiment, was done during daylight. At a frame rate of 50 frames per second, were captured at a resolution of 1920 x 1080 pixels. The UAV was operated at various heights (ranging from 5 to 50 meters) and camera angles (45 to 90 degrees). All the films were taken outside of the city limits of Moslavacka gora, Croatia. People's positions vary from standard -lying down, sitting, and standing- to positions of exhausted persons of different ages. Actors are also in a variety of locales, ranging from obviously visible to positions in the woods, shades, long grass, and other similar settings, making object detection much more difficult. Some sample images are shown in Figure 7.

Figure 7: Sample images from the dataset

3.7 Evaluation metrics

The authors in [64] provided measures based on the spatial intersection of ground-truth and system-generated bounding boxes and then produced multiple performance metrics, which were then averaged for all sampled frames. Various detector performance keys such as bounding-box coordinates of identified objects, the related class, and a parameter for reliability were measured on images that were unseen. The well-known evaluation measures employed were recall, mean average precision (mAP), and precision. In this research, just “person” class is considered, hence the mAP is equivalent to the average precision (AP). mAP is obtained by averaging the average precision of all classes, as indicated by equation (3), where q represents the number of queries and AveP (q) stands for the average precision for that query. mAP is also a metric for calculating the machine learning algorithms accuracy. True Positive in the emergency landing location recognition issue is the number of good (uncluttered and safe) landing spots found by the algorithm. The amount of non-good landing places mistakenly identified as excellent landing spots by the algorithm, and the number of good landing spots overlooked, are known as false positives and false negatives, respectively. Mathematically, mAP may be defined as:

$$mAP = \sum_{q=1}^{Q} \frac{AveP(q)}{Q} \quad (3)$$

Locating a person as quickly as possible is critical to a successful search and rescue operations, hence it is important to discover missing persons if present on the spot. Another term intersection over union (IOU) is used to determine if a prediction is false positive or true positive. It is the ratio of the ground truth and prediction labels' area of overlap and the area of union. Precision evaluates the accuracy of detection findings. It is calculated as the percentage of true positive (TP) detections to the total number of detections as indicated in Equation (4). In contrast to this, Equation (5) illustrates recall, which evaluates the true positive detections in relation to the total potential detections.

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

where FP denotes false positive, and FN represents false negative.

3.8 Proposed method

The research reported in this work addresses the use of real time object detection in human detection of SAR operations as shown in Figure 8. In contrast to prior developments that employed the DarkNet framework, the YOLOv5 implementation is investigated in Pytorch framework. This makes the model easier to comprehend, train with, and deploy. There have been no prior publications reported on using YOLO-v5 in SAR operations to identify persons. The YOLO models detect objects and localize them directly in one-shot, unlike ROI detection-based networks. YOLO-v5 model has been selected due to its run speed, which is a clear advantage since it is a single-stage object detection model. The YOLO-v5 architecture is Focus structure with CSPdarknet53 backbone. The benefit of employing a Focus layer is that it requires less CUDA memory, has a smaller layer, and allows for more forward and backpropagation. This structure aids in the efficient prediction of tiny to large items.

A single learning network is applied to the whole image. The image is broken into regions by this network to predict bounding boxes and probabilities for each region. The bounding box weights are based on projected probabilities. If box center is in the cell, it will compute the predicted class and bounding box coordinates. If there is no object, the score is zero. Otherwise, the score equals intersection over union (IOU) between predicted and ground truth.

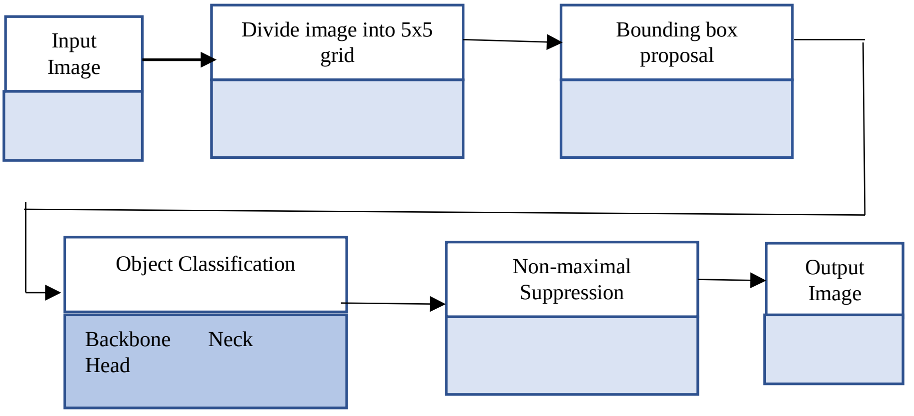

Figure 8: Block diagram of the Proposed Method involving YOLO-v5

For detection, the data are first input to CSPDarknet53 (backbone) for feature extraction, and then fed to PANet (Neck) for feature fusion. Finally, Yolo layer (head) outputs detection results (class, score, location, size). Non-max suppression appears to ensure the algorithm detects each object only once. Each cell outputs with probability Pc, where Pc is the probability if there is an object. The job of non-max suppression is that it takes the box with the highest value of Pc, throws away boxes with P$c \leq 0.6$ and "silences" the boxes with IoU ≥ 0.5.

SARD dataset resolution is 1920x1080, which indicates a full high-definition resolution. This cannot be fed directly as an input due to the tiny size of the targeted object in compared to size of the full image. To achieve a robust model involving YOLO-v5, it is required to resize and preprocess the input images before running the input through the network. The default training size needed for the proposed network is 640x640, as also recommended in [65]. Different resolutions have been tried for testing the images while maintaining the input images a1920x1080 size. Though, increasing the network resolution can produce better detection results, but as is to be shown in chapter four, the best resolution that produces the highest accuracy results is 832x832.

YOLO-v5 family has a total of 5 models [66]. YOLO-v5 nano (the smallest and quickest) to YOLO-v5 extra-large (the largest model). The YOLO-v5l is a large model, and is useful for datasets that involve detection of tiny items. This model suits the current research target of detecting humans in SAR applications.

For comparative purposes, all models need to be tested to investigate human detection in search and rescue scenarios. The models are to be tested YOLO v5l on SARD data set and compared it with the YOLO-v3, YOLO v4 and state of the art Faster R CNN. The detailed results are reported in the next section.

Chapter 4

Chapter 4: Results and Comparisons

In this section, experimental results from a sequence of photos acquired by a UAV for human detection purposes using the proposed method are presented. Selected models such as state of-the-art convolutional neural network-based object detectors (Faster R CNN, YOLO-v3, YOLO-v4, YOLO-v5) were pre-trained on MS COCO dataset. In order to allocate the missing people in search and rescue scenarios effectively with the highest recall and precision, a custom-made SARD test set was used to train the stated detectors. A comparison of different CNN networks was accomplished.

4.1 Preprocessing images

The SARD set images were obtained from eight videos with a resolution of 1920x 1080. The images are split 60:40 between training and testing. The training set has 1189 photos with 3921 people marked on them, whereas the testing set has 792 images with 2611 people marked on them. The network design as well as the input data format must be carefully considered while building an effective network model. The number of pictures, image height, image width, number of channels, and number of levels per pixel are the most often used image data input parameters. We usually have three data channels corresponding to the colors: Red, Green, and Blue (RGB) [0,255] are the most common pixel level.

There were 1,981 single frames with individuals on them picked from a total of 35 minutes of recordings. People were manually labeled in the selected photographs so that dataset could be employed to train the supervised model. The labeling tool was used to tag individuals. The picture annotation includes the bounding box position around each object of interest, its height and width, and the person's label (Walking, Standing, Running, Sitting, Lying, Not Defined). Labels are saved as XML files in YOLO format.

4.2 Multi scale training

The YOLO framework adjusts the size the image while maintaining the aspect ratio to resolution set by width and height parameters in the .cfg weights file. Network resolution refers to these factors. Transforming the image resolution in Yolo architecture may be defined as stated in equation (6):

$$Img_{train\_width} = Net_{\_width} \tag{6}$$

$$Img_{train\_height} = \frac{Net\_width}{Img\_width} \cdot Img_{height}$$

For example, if an image's input resolution is 1920x1080 and network resolution is marked as width (Net_width = 832), and height (Net_height = 832), YOLO will change the input image's resolution to the set Net_width, while changing the height of the input image to gain the ratio of Net_width to Img_width multiplied by the original image height. e.g., 1920×1080 will be transformed to 832×468. For illustration purposes, this comparison of input and network image's resolution is shown in Figure 9.

Figure 9: Input image resolution (a) and network image resolution (b)

Another approach for improving detection performance, primarily for small objects detection would be to utilize higher resolution of the images and then use these images to train the network. Since the YOLO network resizes the image down by 32, the width and height should be ensured to be a multiple of 32. YOLO obtains images of size 320×320, 352×352 ,…, 512×512,… 608×608,…and 832×832 during training with a step of 32 as stated in equation (7).

$$Net_{width} = Net_{width} + k \qquad k = 32n \tag{7}$$

In our simulations, SARD input image size is 1920x1080, where the standard resolution size to choose for YOLOv5 model is 640x640 [65]. Different network resolutions have been used for testing the images while maintaining the input images at the same size of 1920 x 1080. During comparing, it can be shown that increasing the network resolution during testing can produce better detection results. Table 1 below shows that a network resolution of 832x832 produces the best accuracy results.

Table 1: Network resolution (%) YOLOv5 detection performance

Network resolution$_{test}$	**AP**	AP_{50}
320 x 320	38	78
640 x 640	60	95
832 x832	64	96

4.3 Implementation platform and time

The model that is previously trained on MSCOCO dataset was imported to Kaggle in order to use the pretrained weights on our custom-made dataset (SARD), then the trained model was exported to google Collaboratory for validation and testing SAR human images, The hardware machine specification of google colab were as follows:

Model name : Intel(R) Xeon(R) CPU @ 2.30GHz

CPU MHz : 2299.998

Cache size	: 46080 KB
CPU cores	: 2
RAM	: 12GB
GPU	: Nvidia K80 / T4
GPU Memory Clock	0.82GHz / 1.59GHz
Performance	4.1 TFLOPS / 8.1 TFLOPS
Max execution time	12 hours
Max idle time	90 min

The preprocessed image took 0.9ms for execution, while the proposed method's execution time is 24.5ms per image with a maximum of 12 hours where the network size is 832px X 832px. Since Faster-R-CNN had a maximum execution time of 20 hours with an inference time of 1 second per image [52]. The objective of a SAR operation lies in searching the shortest amount of time to locate a lost individual. Therefore, in comparison with other neural network models, YOLOv5l is the fastest in terms of predicting the presence of a human.

4.4 Comparison on detection results

The results on the SARD dataset for object detection are given in Table 2. With YOLOv-5, the best results were obtained. Although it was slightly better than YOLO-v4, the Faster R-CNN detector's performance were significantly worse than YOLO-v5.

In all tested detectors, AP50 achieved the highest values. The best average precision has been conducted by YOLO-v5 with a value of 96.9% compared with YOLO-v4 model where AP50 reached 96%. On the contrary, Faster R CNN acted worse with lowest AP (91%). When detector precision was altered from AP50 to AP75, all detectors performed worse, but the highest mean accuracy of 74.3% noted again by YOLO-v5.

Table 2: Comparative Results on SARD dataset

Model	**Class**	**Images**	**labels**	**mAP@0.5**	**mAP @0.75**	**mAP**
YOLOv5	All	792	2605	0.969	0.743	0.643
YOLOv4	All	792	2605	0.96	0.71	0.61
YOLOv3	All	792	2605	0.925	0.63	0.902
Faster R-CNN	All	792	2605	0.91	0.51	0.50

The accuracy and recall ratios for all evaluated models are shown in Table 3. YOLO-v5 has the best precision-to-recall ratio, with 97 percent precision and a recall of better than 93 percent, indicating that it was the highest performer in detection results and has spotted the largest number of humans/objects in the given image. Faster R-CNN (SARD) has the highest recall, but with a lower precision of just 67% and with good number of false positive detections than YOLO-v5.

Table 3: Precision – recall ratios for different models

Model	Precision	Recall
YOLOv5	0.971	0.932
YOLOv4	0.96	0.91
YOLOv3	0.962	0.892
Faster R-CNN	0.67	0.936

Following the calculation of recall and precision for various IOU thresholds, a recall and precision plots for a single classifier at various IOU thresholds are constructed in Figures 10 and Figure 11 respectively. After then, the precision-recall curve is used to compute the average precision as shown in Figure 12. Equally crucial is having as few false detections as possible to avoid wasting human resources. It was decided that if the intersection of the associated ground truth bounding box and the detected bounding box to that of the union is 50% or greater, the detection is considered positive. This metric is known as intersection-over-union (IoU).

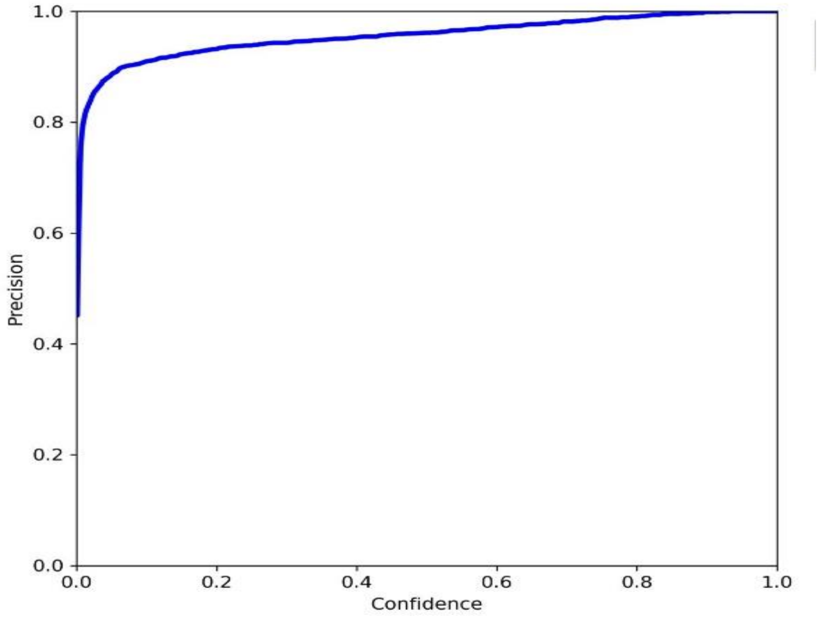

Figure 10: Precision curve

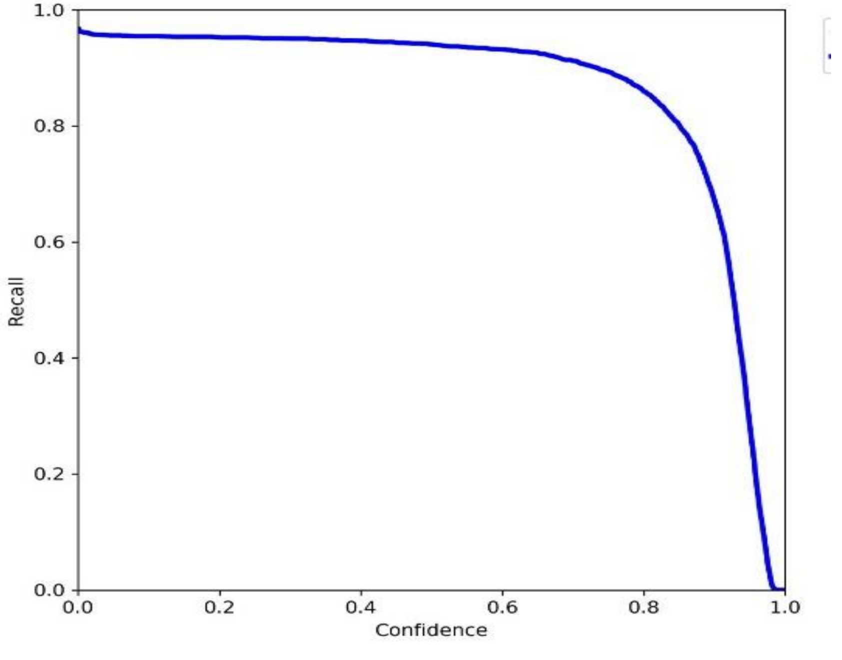

Figure 11: Recall curve

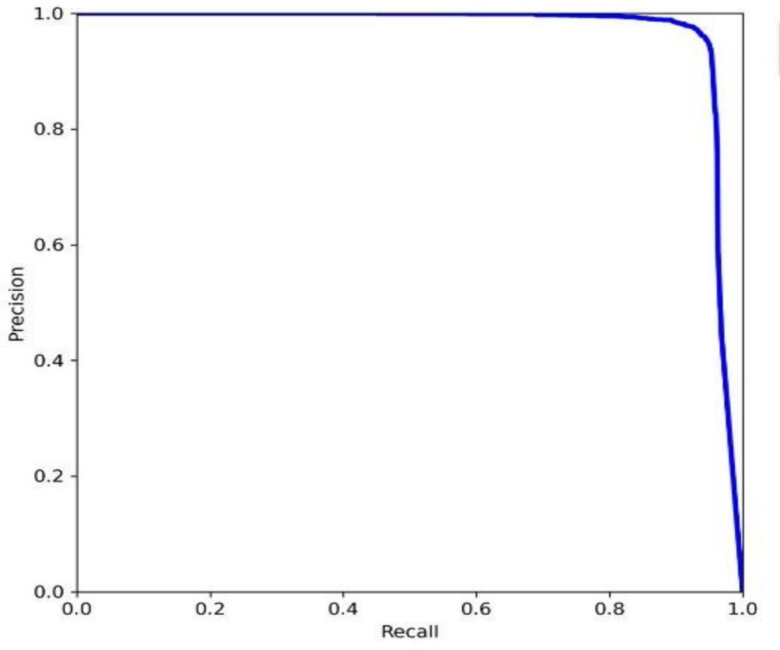

Figure 12: Precision – recall curve

The harmonic-mean of recall and precision, as indicated in Equation (8), is the F1 Score, which represents the test accuracy of the model. The maximum possible F1 score is 1, indicating excellent recall and precision, while the lowest possible score is 0, indicating neither recall nor precision has been registered. Figure 13 represents the F1 score that was obtained as a consequence of our experiment. This score was as high as 0.95, which proposes that there was a very good correlation between precision and recall.

$$F1\ score = 2 * \frac{(precision * recall)}{precision + recall} \tag{8}$$

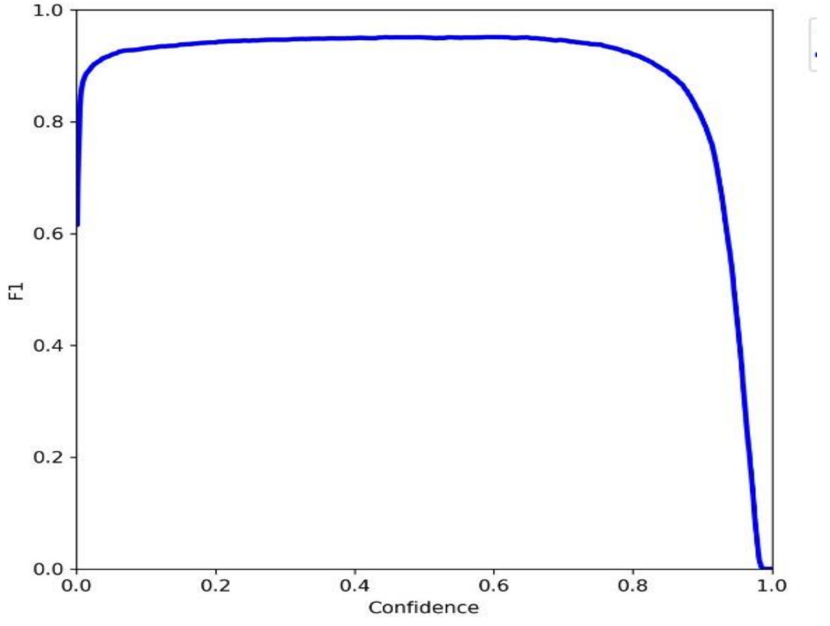

Figure 13: F1 score curve

The main aim of a search and rescue mission is to find all people who are on the spot. However, the detector's accuracy is also essential in order to avoid wasting resources on erroneous detections. As a consequence, the YOLO-v5 detector was chosen for comparative investigation using the acquired findings of accuracy-to-recall ratio and the average precision.

Examples of detection results on persons with YOLO-v5 model trained on SARD dataset are displayed in Figure 14. Figure 14 depicted all possible detection outcomes: a positive detection in which a person has been detected and the corresponding intersection over union of the bounding box and that of the person's ground truth is greater than fifty percent; a negative detection in which a person has not been detected and the corresponding intersection over union of the bounding box and that of the person's ground truth is less than fifty percent; and a false-positive detection in which a fraction of the image that does not include a person is understood as a person.

In Figure 14, YOLO-v5 successfully detects all subjects in the images where the confidence level is shown as well. Images in the first column (left side) are the labelled ones, while images shown in the second column (right side) are the predictions. YOLO-v5 was obviously the most successful at detecting people in SAR events, according to the qualitative analysis of the selected cases. However, there have been instances where the YOLO-v5 model has performed poorly, as shown in Figure 15. The situations of shadows (shown in first row) and when the detector perceives dark portions of the vegetation are the most common example of erroneous detection (second row). In search and rescue efforts, it is nearly expected for a person to fade into the surroundings.

Figure 14: A sample of positive detections with high confidence

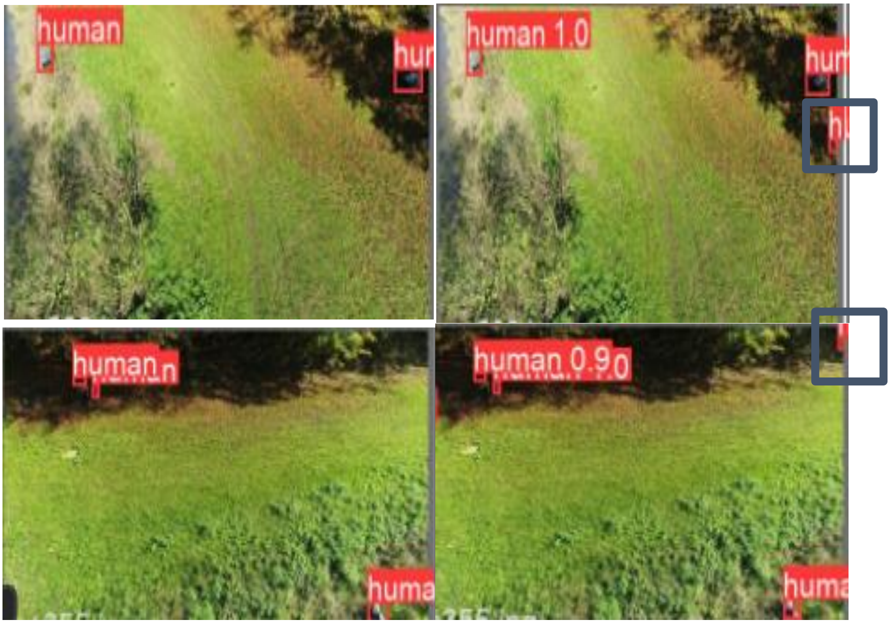

Figure 15: False detections of YOLO-v5 model (shadows, dark areas)

Chapter 5

Chapter 5: Conclusion and Future Work

In this section, conclusions based on results presented in chapter four are discussed. Furthermore, future directions that may help investigators to improve on the existing model to increase precision and reduce run time of the test images.

5.1 Conclusion

Analyzing a big number of high-resolution images for smaller objects and the details may be a time-consuming operation, and as a result, numerous mistakes are likely to occur. In human search, "searching but not seeing" is a well-known problem, while in an image processing system, this problem will be eliminated. This thesis presented a search and rescue system that combines UAV technology with real-time computer vision and deep learning algorithms. Since the datasets for SAR operations are limited due to lack of similar conditions typically found in real life scenarios, this work required search for rescue UAV images to find a suitable training dataset. To ensure the highest possible detection accuracy, the dataset that is likely to be used for training would have to include conditions that are identical to those when the model is tested in a similar environment, necessitating the use of recordings taken from above. Quick and in-time locating of the missing person(s) is a critical item in search and rescue operations because sometimes, it makes the difference between life and death. Hence, it was aimed to examine YOLO-v5 algorithm for increasing human detection accuracy.

The model utilized in this research was a YOLO-v5 architecture running on the Nvidia K80/T4 embedded autonomous computing platform. Processes for training, validation, and testing have also been discussed. The study of real flights proved the superiority of the deep learning algorithms used. The mAP of the proposed model was at 96.9 percent, which is a good performance for an on-board human detection system compared to well-known detectors such as Faster R-CNN, YOLO-v4, and YOLO-v3. Other metrics were examined as well such as precision and recall. The highest precision was achieved by the proposed approach, which evaluates the accuracy of detection findings. YOLO-v5 had the best precision-to-recall ratio, with 97 percent precision and a recall of better than 93 percent, indicating that YOLO-v5 was the most successful human detection and had spotted the greater number of persons in the image (ground truth). Despite a better value of precision for YOLO-v5 model, the recall of Faster R-CNN model was exceeded. Precision is more critical than recall when a fewer False Positives are required in favor for more False Negatives, which indicates getting a False Positive is costly, but getting a False Negative is not.

5.2 Future work

Future work will involve the use of a thermal camera [9] to enhance detection performance where detecting people with thermal cameras is reliable with weather conditions. For example, in winter or cold regions, the normal temperature of the human body is higher than the environment, so humans appear bright and clear using thermal imaging. While in summer and tropical areas the body temperature is much lower than the environment. In addition, dataset could be improved by adding photos that replicate additional weather conditions, such as fog, snow, and ice, that may occur in genuine search and rescue operations. Also blur images could be involved to represent camera movement and aerial photography in action. In particular, a sufficient training dataset for deep learning model training is essential for enhancing the accuracy of image classifications and object detection. To further enhance the performance and precision of deep learning models, it is necessary to construct a training database containing photos of various object classes for future research. To expand the training database, photos of additional occurrences should also be sorted into distinct object groups.

Furthermore, drone-based mountain surveillance systems [68] might be coupled to forest fire detection sensors to undertake wide-scale forest monitoring across a large area., mountains surveillance systems [68] based on drones may be linked with forest fire detection sensors to conduct wide forest monitoring over a vast region. This approach allows for a continuous and remote watch on a flame in woods and mountains, all while the UAV is flying and gathering elevated data, allowing clients to keep the number and area of flame foci the same. Observing programming expands capabilities, such as Fire: source identification, area, and LCD module.

Additionally, an effort is needed to investigate the benefits of alternative models of object saliency detection [69] and attempt to create a model that is tailored and specialized for this purpose by analyzing methods for detecting salient objects in natural air images for search and rescue missions. It's a challenging and complex task in which computer-assisted proposals for target object positions would be quite useful. Examining a huge number of high-resolution images for tiny items and details may be a time-consuming operation, and as a result, numerous errors may arise. In human search, "looking but not seeing" is a well-known problem. In an image processing system, this problem will be eliminated.

References

1. Koester, R. J. (2008). Lost Person Behavior: A search and rescue guide on where to look-for land, air and water. dbS Productions LLC. Charlottesville, VA, August, 3.

2. Bejiga, M. B., Zeggada, A., & Melgani, F. (2016). Convolutional neural networks for near real-time object detection from UAV imagery in avalanche search and rescue operations. In 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS) (pp. 693-696). IEEE.

3. Feed, E. (2022). The Alliance for Drone Innovation - UAV | Unmanned Aerial Vehicle. Retrieved 30 June 2022, from https://www.uav.org/the-alliance-for-drone-innovation/

4. Martinez-Alpiste, I., Golcarenarenji, G., Wang, Q., & Alcaraz-Calero, J. M. (2021). Search and rescue operation using UAVs: A case study. Expert Systems with Applications, 178, 114937.

5. Kampouraki, M., Wood, G. A., & Brewer, T. R. (2008). Opportunities and limitations of object based image analysis for detecting urban impervious and vegetated surfaces using true-colour aerial photography. In Object-Based Image Analysis (pp. 555-569). Springer, Berlin, Heidelberg.

6. HMICFRS, (2017). Planes, drones and helicopters: An independent study of police air support. Her Majesty's Inspectorate of Constabulary and Fire & Rescue Services (HMICFRS)

7. Zhang, P., Zhong, Y., & Li, X. (2019). SlimYOLOv3: Narrower, faster and better for real-time UAV applications. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (pp. 0-0).

8. Memon, Q. (2019). On assisted living of paralyzed persons through real-time eye features tracking and classification using Support Vector Machines. Medical Technologies Journal, 3(1), 316-333.

9. Memon, Q. (2013). Smarter Healthcare Collaborative Network. Building Next-Generation Converged Networks: Theory.

10. Rudol, P., & Doherty, P. (2008). Human body detection and geolocalization for UAV search and rescue missions using color and thermal imagery. In 2008 IEEE aerospace conference (pp. 1-8). IEEE.

11. Burke, C., McWhirter, P. R., Veitch-Michaelis, J., McAree, O., Pointon, H. A., Wich, S., & Longmore, S. (2019). Requirements and limitations of thermal drones for effective search and rescue in marine and coastal areas. Drones, 3(4), 78.

12. Sasa S., Marina Ivasic-Kos. (2021). SEARCH AND RESCUE IMAGE DATASET FOR PERSON DETECTION - SARD. IEEE Dataport. https://dx.doi.org/10.21227/ahxm-k331

13. Gaszczak, A., Breckon, T. P., & Han, J. (2011). Real-time people and vehicle detection from UAV imagery. In Intelligent robots and computer vision XXVIII: algorithms and techniques (Vol. 7878, pp. 71-83). SPIE.

14. Baker, C. A., Ramchurn, S., Teacy, W. L., & Jennings, N. R. (2016). Planning search and rescue missions for UAV teams. In Proceedings of the Twenty-second European Conference on Artificial Intelligence (pp. 1777-1778).

15. Yun, K., Nguyen, L., Nguyen, T., Kim, D., Eldin, S., Huyen, A., ... & Chow, E. (2019). Small target detection for search and rescue operations using distributed deep learning and synthetic data generation. In Pattern Recognition and Tracking XXX (Vol. 10995, pp. 38-43). SPIE.

16. Gallego, A. J., Pertusa, A., Gil, P., & Fisher, R. B. (2019). Detection of bodies in maritime rescue operations using unmanned aerial vehicles with multispectral cameras. Journal of Field Robotics, 36(4), 782-796.

17. Geraldes, R., Goncalves, A., Lai, T., Villerabel, M., Deng, W., Salta, A., ... & Prendinger, H. (2019). UAV-based situational awareness system using deep learning. IEEE Access, 7, 122583-122594.

18. Murphy, S. O., Sreenan, C., & Brown, K. N. (2019). Autonomous unmanned aerial vehicle for search and rescue using software defined radio. In 2019 IEEE 89th vehicular technology conference (VTC2019-Spring) (pp. 1-6). IEEE.

19. Lygouras, E., Santavas, N., Taitzoglou, A., Tarchanidis, K., Mitropoulos, A., & Gasteratos, A. (2019). Unsupervised human detection with an embedded vision system on a fully autonomous UAV for search and rescue operations. Sensors, 19(16), 3542.

20. Leira, F. S., Johansen, T. A., & Fossen, T. I. (2015). Automatic detection, classification and tracking of objects in the ocean surface from UAVs using a thermal camera. In 2015 IEEE aerospace conference (pp. 1-10). IEEE.

21. Sun, J., Li, B., Jiang, Y., & Wen, C. Y. (2016). A camera-based target detection and positioning UAV system for search and rescue (SAR) purposes. Sensors, 16(11), 1778.

22. Sambolek, S., & Ivasic-Kos, M. (2021). Automatic person detection in search and rescue operations using deep CNN detectors. IEEE Access, 9, 37905-37922.

23. Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.

24. Jiao, L., Zhang, F., Liu, F., Yang, S., Li, L., Feng, Z., & Qu, R. (2019). A survey of deep learning-based object detection. IEEE access, 7, 128837-128868.

25. Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28.

26. Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision (pp. 2980-2988).

27. Cai, Z., & Vasconcelos, N. (2018). Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6154-6162).

28. Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).

29. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). Ssd: Single shot multibox detector. In European conference on computer vision (pp. 21-37). Springer, Cham.

30. Ali, Z., & Memon, Q. A. (2013). Time delay tracking for multiuser synchronization in CDMA networks. Journal of Networks, 8(9), 1929.

31. Zhao, Z. Q., Zheng, P., Xu, S. T., & Wu, X. (2019). Object detection with deep learning: A review. IEEE transactions on neural networks and learning systems, 30(11), 3212-3232.

32. Valappil, N. K., & Memon, Q. A. (2021). CNN-SVM based vehicle detection for UAV platform. International Journal of Hybrid Intelligent Systems, (Preprint), 1-12.

33. Rivera, A. J. A., Villalobos, A. D. C., Monje, J. C. N., Mariñas, J. A. G., & Oppus, C. M. (2016). Post-disaster rescue facility: Human detection and geolocation using aerial drones. In 2016 IEEE Region 10 Conference (TENCON) (pp. 384-386). IEEE.

34. Yong, S. P., & Yeong, Y. C. (2018). Human object detection in forest with deep learning based on drone's vision. In 2018 4th International Conference on Computer and Information Sciences (ICCOINS) (pp. 1-5). IEEE.

35. Memon, Q., & Valappil, N. (2021). On multi-class aerial image classification using learning machines. 10.1049/PBPC042E_ch15.

36. Tan, M., Pang, R., & Le, Q. V. (2020). Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10781-10790).

37. Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning (pp. 6105-6114). PMLR.

38. Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). Ieee.

39. Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In European conference on computer vision (pp. 740-755). Springer, Cham.

40. Zhang, X., Zhu, K., Chen, G., Tan, X., Zhang, L., Dai, F., ... & Gong, Y. (2019). Geospatial object detection on high resolution remote sensing imagery based on double multi-scale feature pyramid network. Remote Sensing, 11(7), 755.

41. Pires de L., R., & Marfurt, K. (2019). Convolutional neural network for remote-sensing scene classification: Transfer learning analysis. Remote Sensing, 12(1), 86.

42. Rostami, M., Kolouri, S., Eaton, E., & Kim, K. (2019). Deep transfer learning for few-shot SAR image classification. Remote Sensing, 11(11), 1374.

43. Karimi, D., Dou, H., Warfield, S. K., & Gholipour, A. (2020). Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. Medical Image Analysis, 65, 101759.

44. Gąsienica-Józkowy, J., Knapik, M., & Cyganek, B. (2021). An ensemble deep learning method with optimized weights for drone-based water rescue and surveillance. Integrated Computer-Aided Engineering, 28(3), 221-235.

45. Gotovac, S., Zelenika, D., Marušić, Ž., & Božić-Štulić, D. (2020). Visual-based person detection for search-and-rescue with uas: Humans vs. machine learning algorithm. Remote Sensing, 12(20), 3295.

46. Dousai, N. M. K., & Loncaric, S. (2021). Detection of humans in drone images for search and rescue operations. In 2021 3rd Asia Pacific Information Technology Conference (pp. 69-75).

47. Kundid Vasić, M., & Papić, V. (2022). Improving the Model for Person Detection in Aerial Image Sequences Using the Displacement Vector: A Search and Rescue Scenario. Drones, 6(1), 19.

48. Papić, V., Šolić, P., Milan, A., Gotovac, S., & Polić, M. (2021). High-resolution image transmission from UAV to ground station for search and rescue missions planning. Applied Sciences, 11(5), 2105.

49. Mishra, B., Garg, D., Narang, P., & Mishra, V. (2020). Drone-surveillance for search and rescue in natural disaster. Computer Communications, 156, 1-10.

50. Božić-Štulić, D., Marušić, Ž., & Gotovac, S. (2019). Deep learning approach in aerial imagery for supporting land search and rescue missions. International Journal of Computer Vision, 127(9), 1256-1278.

51. Dorrer, M. G., & Tolmacheva, A. E. (2020), Comparison of the YOLOv3 and Mask R-CNN architectures' efficiency in the smart

refrigerator's computer vision. In Journal of Physics: Conference Series (Vol. 1679, No. 4, p. 042022). IOP Publishing.

52. Kim, J. A., Sung, J. Y., & Park, S. H. (2020). Comparison of Faster-RCNN, YOLO, and SSD for real-time vehicle type recognition. In 2020 IEEE international conference on consumer electronics-Asia (ICCE-Asia) (pp. 1-4). IEEE.

53. Jamtsho, Y., Riyamongkol, P., & Waranusast, R. (2020). Real-time Bhutanese license plate localization using YOLO. ICT Express, 6(2), 121-124.

54. Lu, J., Ma, C., Li, L., Xing, X., Zhang, Y., Wang, Z., & Xu, J. (2018). A vehicle detection method for aerial image based on YOLO. Journal of Computer and Communications, 6(11), 98-107.

55. Redmon, J., & Farhadi, A. (2017). YOLO9000: better, faster, stronger. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7263-7271).

56. Boudjit, K., & Ramzan, N. (2022). Human detection based on deep learning YOLO-v2 for real-time UAV applications. Journal of Experimental & Theoretical Artificial Intelligence, 34(3), 527-544.

57. Seong, S., Song, J., Yoon, D., Kim, J., & Choi, J. (2019). Determination of vehicle trajectory through optimization of vehicle bounding boxes using a convolutional neural network. Sensors, 19(19), 4263.

58. Digging deep into YOLO V3 — A hands-on guide Part 1. (2020). Retrieved 30 June 2022,, from https://towardsdatascience.com/digging-deep-into-yolo-v3-a-hands-on-guide-part-1-78681f2c7e29

59. Li, J., & Wu, Z. (2021). The application of Yolov4 and a new pedestrian clustering algorithm to implement social distance monitoring during the COVID-19 pandemic. In Journal of Physics: Conference Series (Vol. 1865, No. 4, p. 042019). IOP Publishing.

60. Huang, Z., Wang, J., Fu, X., Yu, T., Guo, Y., & Wang, R. (2020). DC-SPP-YOLO: Dense connection and spatial pyramid pooling based YOLO for object detection. Information Sciences, 522, 241-258.

61. Thuan, D. (2021). Evolution of Yolo algorithm and Yolov5: The State-of-the-Art object detention algorithm.

62. Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., NanoCode012, Kwon, Y., ... & Michael, K. (2022). ultralytics/yolov5: v6. 1-TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference. Zenodo, Feb, 22.

63. Wen, Y., Zhang, K., Li, Z., & Qiao, Y. (2016, October). A discriminative feature learning approach for deep face recognition. In European conference on computer vision (pp. 499-515). Springer, Cham.

64. Bashir, F., & Porikli, F. (2006). Performance evaluation of object detection and tracking systems. In Proceedings 9th IEEE International Workshop on PETS (pp. 7-14).

65. How to Train YOLO v5 on a Custom Dataset | Paperspace Blog. (2021). Retrieved 30 June 2022, from https://blog.paperspace.com/train-yolov5-custom-data/

66. Custom Object Detection Training using YOLOv5 | LearnOpenCV. (2022). Retrieved 30 June 2022, from https://learnopencv.com/custom-object-detection-training-using-yolov5/

List of Publications

Bachir, & Memon., “Investigating YOLO-v5 for Search and Rescue Operations involving UAVs International Conference on Control and Computer Vision (ICCCV), Xiamen University, China, 2022.

UAE UNIVERSITY MASTER THESIS NO. 2022:32

The objective of a SAR operation is to search the farthest area feasible in the shortest amount of time to locate a lost or wounded individual. A trained deep learning system can detect persons from a variety of perspectives. In particular, YOLO-v5 was trained on SARD dataset to overcome the issues of previous state-of -the art detectors. It outperforms other detectors with the highest speed and accuracy, as well as the smallest number of false detections.

Namat Ahmad Bachir received her Master of Science in Electrical Engineering from the Department of Electrical Engineering, College of Engineering at UAE University, UAE. She received her BSc from the Department of Electrical Engineering, College of Engineering at UAE University, UAE.