CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION TO THE PROJECT

Smart Video surveillance enables the monitoring of activity, behavioral patterns, or any other change in environmental conditions. It provides automatic perimeter monitoring and secure area protection. The systems requirements for video surveillance include; storage, encoders, interfaces, and management software. Machine learning and advanced image processing algorithms are playing a dominant role in smart surveillance and security systems. The main object of our project is to enhance the existing security system, like CCTV. The main drawback of CCTV is that it only monitors but won't detect and notify the irregularity in the environment. The real-time image that we get from the CCTV should be manually monitored. Continuous monitor from humans is needed in case of the restricted area if we are using the regular security system. Humans tend to do errors when doing continuous monitoring. Security systems are mainly used for preventing the disturbance occurring in the specified region. So, we can say continuous monitoring is unnecessary when there is no disturbance. We are giving a solution to this problem.

A smart surveillance system using the YOLO algorithm project does most of the work that is done by a person who is monitoring the real-time image captured by the security system. For example, if we consider a restricted area and there is a CCTV that monitors that area if a person enters the area, then the captured real-time will be processed and a warning message will be sent to the respected person. There is some confusion between object detection and image recognition but we can see the clear difference between object detection and image recognition. The main difference is the output of both. Image recognition output will be only the name of the object with how many objects are present. If we take an image where a dog is present then its output will be dog-1. The output of the object detection gives the name of the object(class), a number of the same objects present and the main point is that it will give the position of the object. We can say the advantage of it will be the output

Object detection is divided into two types Deep learning-based approach and a Machine learning-based approach. We are mainly focusing on the deep learning approach as we are focusing on the YOLO algorithm. YOLO stands for you only look once. All of this work is to enhance the security system and improve overall safety. The project can be used in the domestic

domain and also in the industrial domain for the prevention of trespassing and monitoring of the restricted area respectively.

1.2 LITERATURE SURVEY

Continued monitoring of restricted areas is a tedious job for any person. providing high security is challenging in every place. Therefore, several researchers have contributed to monitoring various activities and behaviors using object detection. There are a large number of authors, who have used different types of object detection and recognition methods; from which YOLO outperform other methods in terms of speed and accuracy.

[1] Sanam Narejo, Bishwajeet Pandey, Doris Esenarro vargas, Ciro Rodriguez and M. Rizwan Anjum has Published a research paper on Weapon Detection Using YOLO V3 for Smart Surveillance System. They developed a computer-based fully automated system to identify basic armaments, particularly handguns and rifles. Recent work in the field of deep learning and transfer learning has demonstrated significant progress in the areas of object detection and recognition.

Hu et al. [2] have contributed to detecting various objects in traffic scenes by presenting a method that detects the objects in three steps. Initially, it detects the objects, recognizes the objects, and finally tracks the objects in motion by mainly targeting three classes of different objects including cars, cyclists, and traffic signs. Therefore, all the objects are detected using a single learning-based detection framework consisting of a dense feature extractor and trimodal class detection. Additionally, dense features are extracted and shared with the rest of the detectors which head to be faster in speed and further need to be evaluated in the testing phase. Therefore, intraclass variation of objects is proposed for object subcategorization with competitive performance on several datasets.

[3] A new approach to vehicle license plate location based on the new model YOLO-L and plate pre-identification research paper presented by the author Weidong Min, Xiangpeng Li, Qi Wang, Qingpeng Zeng, Yanqiu Liao gives the content on how to detect the license plate under complex road environments such as severe weather conditions and viewpoint changes. The license plate location method may incorrectly detect similar objects such as billboards and road signs as license plates. To alleviate these problems, this article proposes a new approach to vehicle license plate location based on the new model YOLO-L and plate pre-identification.

The new model improves in two aspects to precisely locate the area of the license plate. First, it uses a k-means++ clustering algorithm to select the best number and size of plate candidate boxes. Second, it modifies the structure and depth of the YOLOv2 model.

[4] Joseph Redmon, Ali Farhadi, et al, published a research paper on "Real-Time object detection a state-of-the-art algorithm based on convolutional neural networks". They also made different changes to the original version of YOLO object detection. The latest enhanced YOLO algorithm is incredibly fast and precise. The YOLO machine learning algorithm proposed by the authors can recognize and detect over 9,000 object categories. They frame the problem of object detection as a regression problem in this design, straight from image pixels to bounding box coordinates and their class probabilities. While the training and testing period, YOLO examines the whole picture, which fully encodes contextual information about classes and their appearance. For training and testing techniques, they used the NVIDIA Titan X GPU. Convolutional layers of neural net select features from images in the designed process. There are 24 convolutional layers (CL) of their Neural Network followed by 2 fully connected (FC) layers. Up to 45 FPS can be processed in real-time with YOLO architecture images, which is a state-of-the-art speed. It defeats other well-known models of detection, such as the algorithm for Fast R-CNN and DPM. They trained the YOLO model on the COCO dataset and the Image-Net dataset using the optimization method for stochastic gradient descent (SGD).

[5] Aleksa Ćorović; Velibor Ilić; Siniša Đurić; Mališa Marijan; Bogdan Pavković presented a research paper on The Real-Time Detection of Traffic Participants Using YOLO Algorithm. Which provides the demonstration of the usage of the newest YOLOv3 algorithm for the detection of traffic participants. We have trained the network for 5 object classes (car, truck, pedestrian, traffic signs, and lights) and have demonstrated the effectiveness of the approach in a variety of driving conditions (bright and overcast sky, snow, fog, and night).

1.3 PROPOSED WORK

The flow chart shows the overall structure of the smart surveillance system using the YOLOv3 algorithm. This structure is entirely based on object detection using the YOLOv3 algorithm and consists of a Darknet-53 feature extraction network. The objective of the project is to enhance the performance of the present security system. In an unauthorized area when a person trespasses, he will be detected and the notification is sent to the landlord/ property owner.

CHAPTER 2

YOLO ALGORITHM

2.1 DEVELOPMENT OF OBJECT DETECTION TECHNOLOGY

Object detection is a challenging problem in the field of computer vision. The tasks of target detection include predicting the position information and category information of the target. It has been widely used in daily life. The most common application scenarios are pedestrian detection, defect detection, ship target detection, and obstacle detection in automatic driving. The traditional detection process is as follows: a specific region is selected as a potential region in the image, and the selection of a specific region will be framed by sliding windows of different sizes. Then, by analyzing the potential region, relevant image features are extracted. Finally, the appropriate classifier is selected to complete the classification based on the image features, but the time complexity is too high to meet the real-time requirements. In recent years, with the improvement of deep neural networks and hardware computing power, the detection method of deep learning has gradually replaced the outdated method, and the convolutional neural network (CNN) is mainly used in deep learning. By training the input images in the network, the convolution network can effectively extract and learn the features of the detected target. After repeated training, the performance of the training model is gradually improved, achieving an excellent target detection effect. [6] The methods based on deep learning can be divided into two categories: one-stage approaches and two-stage approaches. The two-stage models usually include two steps: potential region extraction and category prediction. Region-based Convolutional Neural Network (R-CNN) was put forward in 2014, this extracts image features through selective search instead of the traditional sliding window, and then uses classifiers to predict targets. But the cost of calculation is so large that a small dataset even takes several days to train. To solve this problem, Fast R-CNN was proposed to optimize the training process by simplifying the redundant calculation of overlapping regions. In addition, this method abandons the idea of using multi-classifiers and bounding box regression and achieves near real-time end-to-end training speed.

An excellent detection method needs both accuracy and computational efficiency. Although the two-stage methods have achieved high accuracy, the calculation speed of these methods is usually inferior to that of the one-stage methods. Therefore, the one-stage detection method represented by the You Only Look Once (YOLO) network is widely used in various

practical tasks, especially on lightweight platforms. The first generation of the YOLO model was developed in 2016. It transforms the detection task into a mathematical regression problem, which greatly inspires the development of the method in the subsequent one-stage methods. In the same year, the Single Shot Multibox Detector (SSD) provided a useful strategy to detect targets by combining features of different scales with default bounding boxes. After that, the YOLO series developed very rapidly. More methods are gradually added to the YOLO series, and the detection accuracy is improved. At present, the network based on YOLO has developed to YOLOv7.

2.2 THE NETWORK OF YOLOv3

	Type	Filters	Size	Output
	Convolutional	32	3 × 3	256 × 256
	Convolutional	64	3 × 3 / 2	128 × 128
1×	Convolutional	32	1 × 1	
	Convolutional	64	3 × 3	
	Residual			128 × 128
	Convolutional	128	3 × 3 / 2	64 × 64
2×	Convolutional	64	1 × 1	
	Convolutional	128	3 × 3	
	Residual			64 × 64
	Convolutional	256	3 × 3 / 2	32 × 32
8×	Convolutional	128	1 × 1	
	Convolutional	256	3 × 3	
	Residual			32 × 32
	Convolutional	512	3 × 3 / 2	16 × 16
8×	Convolutional	256	1 × 1	
	Convolutional	512	3 × 3	
	Residual			16 × 16
	Convolutional	1024	3 × 3 / 2	8 × 8
4×	Convolutional	512	1 × 1	
	Convolutional	1024	3 × 3	
	Residual			8 × 8
	Avgpool		Global	
	Connected		1000	
	Softmax			

Fig 2.1 Darknet53

YOLOv3 is a typical end-to-end target detection algorithm, so it runs very fast. The backbone network of YOLOv3 is Darknet53, which can effectively extract the features of the input image. There are no pooling layers in the network structure of YOLOv3, and the full convolutional network (FCN) is adopted to prevent the loss of feature information.[6] Darknet53 network is mainly composed of a series of 1 × 1 or 3 × 3 convolution layers, each of which contains a BN layer and a ReLU layer. It is called Darknet53 because it contains 53 convolution layers. The residual network is used to extract deeper features and avoid gradient fading. Five residual modules are added to the Darknet53 network, each of which consists of one or more residual units. YOLOv3 draws on the idea of FPN to detect targets of different sizes. FPN down-samples the input image five times and predicts the target through the last three down-sampling layers. The sizes of the output images corresponding to the last three down-sampling layers are 52 × 52, 26 × 26, and 13 × 13, respectively. The above three feature maps with different scales are used to detect small targets, medium targets, and large targets respectively. Small feature maps can provide deep semantic information, while large feature

maps contain a lot of fine-grained information. Therefore, YOLOv3 can not only make predictions at different scales but also fully learn the semantics of feature maps at different scales during the prediction process. In this work, the YOLO model trained on the COCO dataset is used. COCO consists of 80 labels namely bicycle, person, animals, airplane, cars, etc.

2.3 BOUNDING BOX AND CLASS PREDICTION

The yolov3 predicts bounding boxes using dimension clusters as anchor boxes.[7] The network predicts 4 coordinates for each bounding box, t_x, t_y, t_w, t_h. If the cell is offset from the top left corner of the image by (c_x, c_y) and the bounding box prior has width and height PW, p_h, then the predictions correspond to: $b_x = \sigma(t_x) + c_x$; $b_y = \sigma(t_y) + c_y$; $b_w = p_w e^{tw}$; $b_h = p_h e^{th}$**.** YOLOv3 predicts an object score for each bounding box using logistic regression. Each box predicts the classes the bounding box may contain using multilabel classification. Instead of softmax YOLO simply uses independent logistic classifiers. The softmax imposes the assumption that each box has exactly one class which is often not the case. A multilabel approach better models the data.

CHAPTER 3

METHODOLOGY

3.1 FLOWCHART

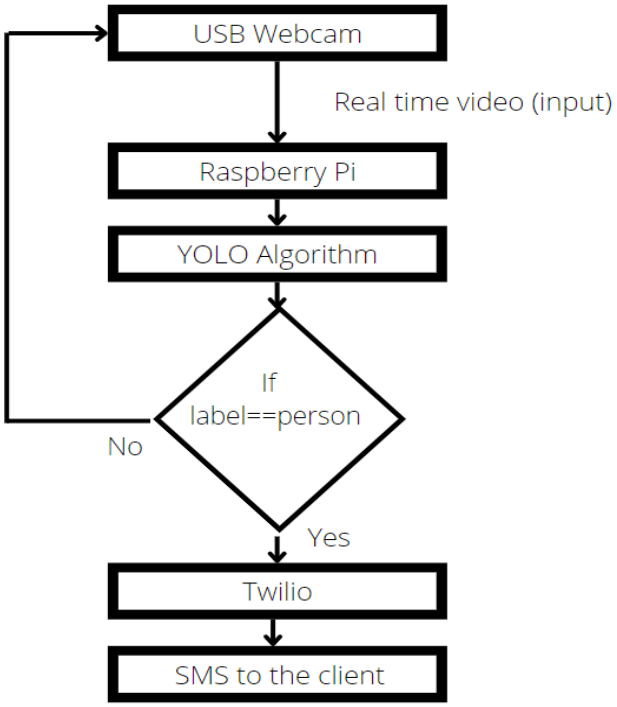

Fig 3.1: Flowchart

3.2 HARDWARE REQUIREMENTS

1. Raspberry Pi 3 model

The Raspberry pi is a single computer board with credit card size, that can be used for many tasks that a computer does, like games, word processing, spreadsheets, and also to play HD video. The Raspberry Pi is a very cheap computer that runs Linux, but it also provides a set of GPIO (general purpose input/output) pins, that allow to control of electronic components for physical computing and exploring the Internet of Things (IoT). The raspberry pi board comprises a program memory (RAM), processor and graphics chip, CPU, GPU, Ethernet port, GPIO pins, Xbee socket, UART, and power source connector. And various interfaces for other external devices. It also requires mass storage, and for that, we use an SD flash memory card. So that raspberry pi board will boot from this SD card similarly as a PC boots up into windows from its hard disk.[8]

2. USB webcam

A webcam is a video camera that feeds or streams an image or video in real time to or through a computer network, such as the Internet. Webcams are typically small cameras that sit on a desk, attach to a user's monitor, or are built into the hardware. All webcams work in broadly the same way, they use an image sensor chip to catch moving images and convert them into streams of digits that are uploaded over the Internet. The image sensor chip is the heart of a webcam.[9]

3.3 PROPOSED METHOD

Prerequisites: The Raspberry pi 3 model B+ is used. The peripherals devices such as keyboard, mouse, webcam, and desktop are interfaced with Raspberry pi. The power supply of 2A and 5V is given. Virtual environment sdjenv is activated in the command prompt. Then all the required packages such as OpenCV, NumPy, and argparse were installed, and some other files such as coco. names, yolov3.weights, yolov3.cfg were downloaded.

Working: To ensure the proper working of the YOLO algorithm, the inputs were taken in the form of jpg and mp4 formats. If the input passed is an image, then the YOLO algorithm takes the input and divides them into grids. The image classification and localization are applied to each grid.[7] Image classification aims at assigning an image to one number of different classes. Localization then allows locating the object in the given image and the output will be displayed in the console window with a bounding box around an object in this case i.e., a person with a certain confidence level. Suppose the input considered is an mp4 format the output is displayed as different frameworks, where each frame is processed.

A real-time video is considered as an input to the YOLO algorithm which will be captured using a USB webcam. If the object detected by the algorithm is of class ‘person’, the object will be displayed with the help of OpenCV. ‘Twilio’ is a modern communication API used by developers for establishing communication. Using a Twilio virtual phone number SMS will be received by the client. If the detected object is a person, then Twilio will send an alert SMS to the client or an administrator with the text “Person is detected”.

CHAPTER 4

EXPERIMENTAL RESULTS

4.1 RESULT

The YOLOv3 network used in this project work uses the lightweight convolutional neural network Darknet-53 as the feature extraction network i.e., the pre-trained model. The object detection code (YOLO) performance is verified with three different inputs namely image file, video file, and webcam feed. The trained yolov3.weight, yolov3.cfg, coco. names are used in the project. The forward () function of the cv2.dnn module returns a nested list containing information about all the detected objects which includes the x and y coordinates of the center of the object detected, height and width of the bounding box, confidence, and scores for all the classes of objects listed in coco. names. The class with the highest score is considered to be the predicted class, in this case, it is the person class. Here the predicted bounding boxes with more than 30% confidence are considered. Although the low confidence bounding boxes are removed, there is a possibility that multiple duplicate detections occur around the objects. For example, fig 4.1 Multiple bounding boxes around bicycle.jpg file. To fix this situation Non-Maximum Suppression (NMS) is applied**,** also called Non-Maxima Suppression**.** The confidence threshold value and NMS threshold value are used as parameters to select one bounding box. From the range of 0 to 1, an intermediate value like 0.4 or 0.5 is considered to make sure that the overlapping objects do not end up getting multiple bounding boxes for the same object. Fig 4.2 shows the final output after applying NMS.

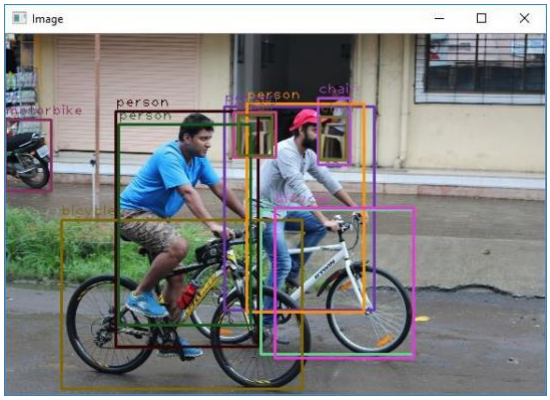

Fig 4.1: Without NMS

Fig 4.2: With NMS

The processing results of several typical scene examples for the image feed fig 4.3, and 4.4 respectively correspond to the original image and the detection results of the original image.

The YOLOv3 object detection algorithm is efficient in the detection of farther objects i.e., the person in this case as shown in fig 4.6. With the video file input pesdestrians.mp4, the processing results are shown in fig 4.7 (a), (b), and (c) with a good level of estimation. The real-time video is captured through the USB webcam using OpenCV. The results of the real-time video feed captured using a PC webcam are shown in fig 4.8.

Fig 4.3: Snapshot of CCTV footage

Fig 4.4: Output image after detection

Fig 4.5: People at the farthest distance (CCTV snapshot)

Fig 4.6: People detected at a long distance

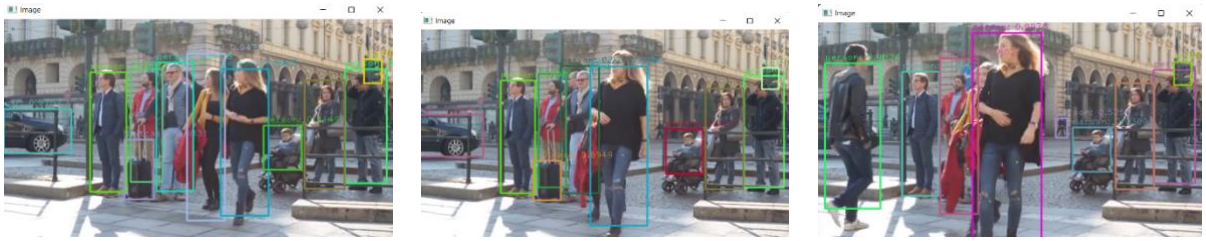

Fig 4.7(a), (b), (c): Output of the video feed

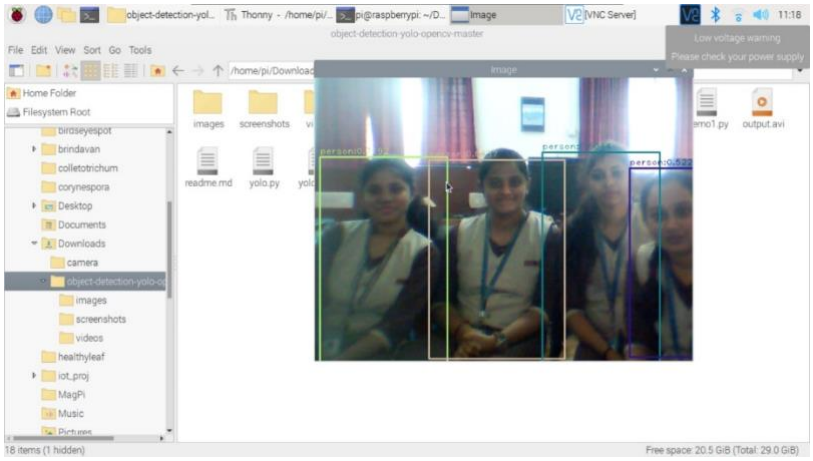

Fig 4.8: Real-time video feed output

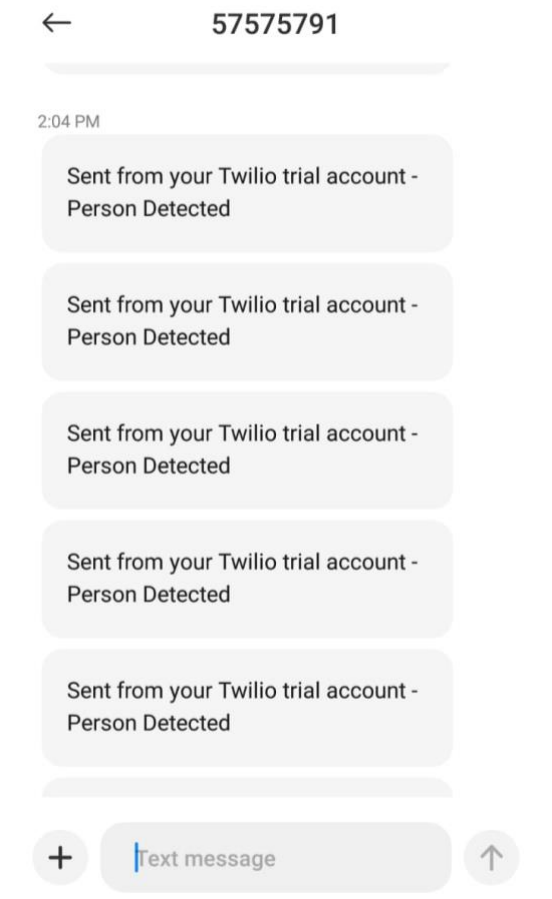

Fig 4.9: SMS received by the client

CHAPTER 5

CONCLUSION

We conclude that this project is very easy to implement on the current system. YOLO is incredibly fast. It is one of the best algorithms for object detection. YOLO has many advantages in practice. According to the experimental result, YOLOv3 has excellent detection performance for detecting a person or detecting a person in a complex group which can increase the efficiency of the smart surveillance system and reduces the burden of the client/administrator to look after the restricted area for a long period. Instead, the client can react at the instant when things go wrong. YOLO network understands generalized object representation; however, the spatial constraints limit the accuracy in the case of nearby and smaller objects. We have a newer version of this algorithm, YOLOv4 which addresses this problem and is more accurate and faster. Overall, YOLO's speed and accuracy make it a widely used algorithm for real-time object detection. YOLO is also better at generalizing object representation compared with object detection models and can be recommended for real-time object detection as the state-of-art algorithm in object detection. With these marks, it is acknowledgeable that the field of object detection has an expansion, comparing to other classifier algorithms this algorithm is much more efficient and the fastest algorithm to use in real-time.

REFERENCES

[**1**] Sanam Narejo, Bishwajeet Pandey, Doris Esenarro vargas, Ciro Rodriguez, M. Rizwan Anjum, "Weapon Detection Using YOLO V3 for Smart Surveillance System", Mathematical Problems in Engineering, vol. 2021, Article ID 9975700, 9 pages, 2021. https://doi.org/10.1155/2021/9975700

[2] Q. Hu, S. Paisitkriangkrai, C. Shen, A. van den Hengel, and F. Porikli, "Fast Detection of Multiple Objects in Traffic Scenes With a Common Detection Framework," in IEEE Transactions on Intelligent Transportation Systems, vol. 17, no. 4, pp. 1002-1014, April 2016, DOI: 10.1109/TITS.2015.2496795.

[3] Min, W., Li, X., Wang, Q., Zeng, Q. and Liao, Y. (2019), New approach to vehicle license plate location based on new model YOLO-L and plate pre-identification. IET Image Processing, 13: 1041-1049. https://doi.org/10.1049/iet-ipr.2018.6449

[4] J. Redmon and A. Angelova, "Real-time grasp detection using convolutional neural networks," 2015 IEEE International Conference on Robotics and Automation (ICRA), 2015, pp. 1316-1322, DOI: 10.1109/ICRA.2015.7139361.

[5] A. Ćorović, V. Ilić, S. Đurić, M. Marijan, and B. Pavković, "The Real-Time Detection of Traffic Participants Using YOLO Algorithm," 2018 26th Telecommunications Forum (TELFOR), 2018, pp. 1-4, DOI: 10.1109/TELFOR.2018.8611986.

[6] L. Kong, J. Wang, and P. Zhao, "YOLO-G: A Lightweight Network Model for Improving the Performance of Military Targets Detection," in IEEE Access, vol. 10, pp. 55546-55564, 2022, doi: 10.1109/ACCESS.2022.3177628.

[7] YOLOv3: An Incremental Improvement Joseph Redmon, Ali Farhadi University of Washington

[8] Opensoure.com, What is a Raspberry Pi?, https://opensource.com/resources/raspberry-pi#:~:text=The%20Raspberry%20Pi%20is%20a,Internet%20of%20Things%20(IoT) (accessed on 16 July 2022)

[9] Raspberry Pi Guide, working with USB webcams on Raspberry Pi, https://raspberrypi-guide.github.io/electronics/using-usb-webcams (accessed on 16 July 2022)