*2023 3rd Asian Conference on Innovation in Technology (ASIANCON)*
*Pune, India. Aug 25-27, 2023*

# Real-Time Object Detection and Audio Feedback for the Visually Impaired

Ayan Ravindra Jambhulkar
*Department of Electronics and Telecommunication Engineering*
*K. J. Somaiya College of Engineering*
Mumbai, India.
ayan.j@somaiya.edu

Akshay Rameshbhai Gajera
*Department of Electronics and Telecommunication Engineering*
*K. J. Somaiya College of Engineering*
Mumbai, India.
akshay.gajera@somaiya.edu

Chirag Manoj Bhavsar
*Department of Electronics and Telecommunication Engineering*
*K. J. Somaiya College of Engineering*
Mumbai, India.
chirag.bhavsar@somaiya.edu

Shilpa Vatkar
*Department of Electronics and Telecommunication Engineering*
*K. J. Somaiya College of Engineering*
Mumbai, India.
shilpavatkar@somaiya.edu

2023 3rd Asian Conference on Innovation in Technology (ASIANCON) | 979-8-3503-0228-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ASIANCON58793.2023.10269899

***Abstract*— Visually impaired individuals face numerous challenges in their daily lives, including the ability to identify and navigate through their surroundings independently. Object detection techniques based on computer vision have shown results in helping the visually impaired by detecting and classifying objects in real-time. In this paper, we used a realtime object detection and audio feedback system that provides audio feedback to the visually impaired for identifying and navigating in their surroundings. The proposed system uses the YOLO_v3 algorithm with the MS COCO dataset to detect and classify objects in real-time and provide corresponding audio feedback. We used gTTS (Google Text to Speech) API for generating the audio feedback. The audio feedback is generated using an audio processing techniques and deep learning algorithms. We evaluated on a dataset, and achieved an average detection accuracy of 90%. The proposed system provides a practical and effective solution for enhancing accessibility and independence for visually impaired individuals, and demonstrates the potential of using advanced deep learning algorithms and datasets for real-time object detection and audio feedback systems.**

***Keywords*— *Real-time object detection, Audio feedback system, YOLO_v3 algorithm, MS COCO dataset, gTTS (Google Text to Speech) API, Deep learning***

## I. INTRODUCTION

From an early age, humans are taught by their parents to distinguish between different things, including themselves as individuals. Our visual system as humans is remarkably precise and can handle multiple tasks even when we are not consciously aware of it. However, when dealing with large amounts of data, we require a more accurate system to correctly identify and locate multiple objects at the same time. This is where machines come into play. By training our computers using improved algorithms, we can enable them to detect multiple objects within an image with a high level of accuracy and precision. Object detection is a particularly challenging task in computer vision because it involves fully understanding images. In simpler terms, an object tracker attempts to determine if an object is present in multiple frames and assigns labels to each identified object [1]. This process encounters various challenges, such as complex images, loss of information, and the transformation of a three-dimensional world into a two-dimensional image. To achieve accurate object detection, our focus should not only be on classifying objects but also on accurately determining the positions of different objects, which can vary from one image to another [2]. This research proposes a system that can help people who are visually impaired detect and identify objects in their environment in real-time. To achieve this, we use an object detection algorithm called YOLO_v3 and a dataset called MSCOCO. Our system generates an audio description of the object, including its location and category, and plays it through a speaker or headphones using gTTS (Google Text to Speech) API. With providing audio feedback, this system aims to help visually impaired individuals an additional way to detect and identify objects in their environment.

## II. LITERATURE REVIEW

Object detection and recognition have been important topics of research form many years. With the advancement of deep learning techniques, object detection has become more accurate and efficient. The YOLO (You Only Look Once) algorithm has emerged as a popular method for real-time object detection due to its speed and accuracy [3]. There has been an increase in the amount of interest in developing assistive technologies for visually impaired individuals. These technologies aim to enhance their independence and mobility by providing them with additional means of detecting and identifying objects in their environment. Deep learning-based object detection systems have shown promising results in this regard. The Microsoft Common Objects in Context (MS COCO) dataset is widely used in deep learning-based object detection research. It is a large-scale dataset that contains over 330,000 images with more than 2.5 million object instances labelled in 80 different categories [4]. Ramesh et al. proposed a real-time object detection system for visually impaired individuals using deep learning. Their system uses a YOLO-based object detection algorithm and provides audio feedback to the user in real-time [5]. Saha et al. proposed an object detection and audio feedback system for visually impaired individuals that uses deep learning techniques. They used the YOLO algorithm for object detection and gTTS (Google Text-to-Speech) for audio feedback [6]. Li et al. proposed a deep reinforcement learning-based object detection and obstacle avoidance system for visually impaired individuals. Their system uses a combination of object detection and obstacle avoidance techniques to enable visually impaired individuals to navigate through complex environments [7]. One of the most commonly used object detection algorithms for real-time

applications is YOLO (You Only Look Once) [8]. YOLO is an end-to-end neural network that processes images in real-time and outputs bounding boxes and class probabilities for detected objects. YOLO has been used in several studies on object detection for visually impaired individuals [9] [10]. Another, important aspect of real-time object detection for the visually impaired is the provision of audio feedback. Text-to-speech (TTS) technology is commonly used for generating audio feedback in object detection systems. In a study conducted by Shin and Kwon [11], a real-time object detection system was developed using the YOLO algorithm and TTS technology to provide audio feedback to visually impaired individuals. In addition to YOLO, other object detection algorithms have also been used in real-time applications for the visually impaired. For example, the Faster R-CNN (Regionbased Convolutional Neural Network) algorithm has been used in a study by Ghosal et al. [12] to develop a real-time object detection system with audio feedback for the visually impaired. Several datasets have been used for training and testing real-time object detection systems for the visually impaired. One of the most commonly used datasets is the MS COCO (Common Objects in Context) dataset, which contains over 330,000 images and more than 2.5 million object instances [13]. The MS COCO dataset has been used in several studies on real-time object detection for visually impaired individuals [9] [11] [12].

## III. RELATED WORK

For instance, a study by Saha et al. (2019) proposed a real-time object detection and audio feedback system that uses a Raspberry Pi and a camera module to detect objects and provide audio feedback. The system uses the TensorFlow object detection API and the COCO dataset for object detection. The audio feedback is provided using a speaker or headphones. The study showed promising results in detecting objects in real-time and providing audio feedback [17]. Another study by Noh et al. (2018) proposed a similar system for object detection and identification. The system uses the Faster RCNN algorithm for object detection and a Raspberry Pi for audio feedback. The study showed that the system was able to detect and identify objects in real-time, and that the audio feedback was effective in assisting visually impaired individuals in navigating environments [16]. In addition, a study by Bhuyan et al. (2019) proposed a system for text detection and audio feedback. The system uses the EAST text detection algorithm and the Google Text-to- Speech API for audio feedback. The study showed that the system was able to detect text in real-time and provide accurate audio feedback to visually impaired individuals [15]. Real-time Object Detection and Recognition for Visually Impaired People using Deep Learning by D. Karimi and H. R. Rabiee. This paper presents a realtime object detection and recognition system for visually impaired people based on deep learning techniques. The system uses a convolutional neural network (CNN) to detect and classify objects, and provides audio feedback to the user using text-to-speech technology. Real-time Object Detection and Classification for the Visually Impaired using Wearable Cameras by S. S. Saini and R. Singh. This paper proposes a wearable camera-based object detection and classification system for the visually impaired. The system uses the YOLO algorithm to detect objects in real-time and provides audio feedback to the user through a speaker or headphones.

DATASET

When we engage in developing an object detection algorithm, there are two primary aspects we focus on: detection and localization. Detection involves determining whether an object belongs to a specific category or not. On the other hand, localization refers to establishing the boundaries of a bounding box around each object, taking into account that the position of objects may differ across different images. To evaluate and compare the effectiveness of various algorithms in the same application, it is beneficial to utilize challenging datasets that establish a standard for performance assessment. We have used in the context of our problem statement the Microsoft Common Objects in Context (MS COCO) dataset to test the algorithms' performance. [18]. COCO, as its name implies, is a dataset that comprises images collected from everyday scenes depicting common objects. These images are gathered in a way that reflects their natural context. If you're interested in accessing this dataset, you can easily download it from the official COCO website. [19]. The dataset consists of a total of 330,000 images, which are divided into 91 different categories. Among these categories, 82 have been assigned labels. The COCO dataset, although it has fewer categories compared to some other datasets, compensates for this by having a larger number of instances for each specific object. This characteristic of the COCO dataset enables machines to learn more accurately. Additionally, the COCO dataset excels at effectively dealing with small objects, providing valuable training examples for machine learning algorithms.

## IV. METHODOLOGY

YOLO utilizes a single neural network to process the entire image. It then divides the image into a grid of equally-sized cells, usually represented as SxS. For each object present in the image, YOLO creates a bounding box around it. It labels each object it finds with a confidence score and a class label. How precisely the object is contained within the bounding box is shown by the confidence score. Within each grid cell, YOLO predicts four values: (x, y, w, h). These values represent the coordinates and dimensions of the bounding box for each object, with all values ranging between 0 and 1. Additionally, YOLO provides a confidence score for every object detected within the cell. The prediction output of YOLO has a specific shape of (S, S, BX5+C) [20]. This means that for each cell in the SxS grid, YOLO predicts B bounding boxes and their corresponding confidence scores (confidence score + class label) for a total of (BX5) values. The additional C represents the number of class labels that the algorithm can detect.

The system consists of two main components: object detection using the YOLO_v3 algorithm and the generation of audio feedback using the gTTS API. It operates by taking input from a camera and performing real-time detection and classification of objects.

### *A. Object Detection:*

In order to detect objects within images, we utilized the YOLO_v3 algorithm. This algorithm is favored for its impressive combination of speed and accuracy. YOLO_v3 follows a unique approach where it divides the image into a grid-like structure. Within each grid cell, the algorithm predicts bounding boxes (which indicate the location and size of the objects) and class probabilities (which determine the type of object present).

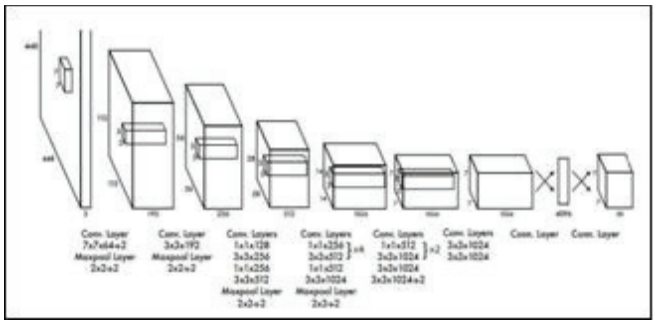

Fig. 1. Architecture of YOLO_v3

### B. Audio Feedback Generation:

We used the gTTS (Google Text to Speech) API for generating audio feedback for detected objects. gTTS is a Google's Text to Speech API to generate speech from text. We passed the object description generated by the YOLO_v3 algorithm to gTTS, which then generated an audio file of the description in the form of an MP3 file. We played the audio file through a speaker or headphones to provide the user with audio feedback.

### C. Overall System:

YOLO_V3 utilizes an original Darknet architecture with 53 layers, but for the detection process, an additional 53 layers are added, resulting in a total of 106 layers. What makes YOLO_V3 particularly interesting is its approach to making detections at three different scales, sizes, and locations within the network. The detection kernel's shape is represented as 1x1x(B x (5+C)), where C represents the total number of classes (e.g., 80 for the COCO dataset) and B denotes the number of bounding boxes around the objects. Consequently, the kernel size of YOLO_V3 is 1x1x255. To gain a deeper understanding of the YOLO_V3 algorithm, a more extensive network called Darknet 53, consisting of 53 convolutional layers, is employed. Once an input image is fed into the YOLO_V3 architecture, multiple objects within the image are classified and assigned class labels. The resulting output is then processed by a Python module called gTTS, which converts the text into speech. The system described utilizes a camera to capture input, performs real-time object detection using the YOLO_V3 algorithm, generates an audio description of the detected objects using the gTTS API, and delivers the audio feedback to the user through a speaker or headphones. This system can be implemented on a computer or laptop equipped with a GPU to ensure real-time performance.

### D. Regenerate response

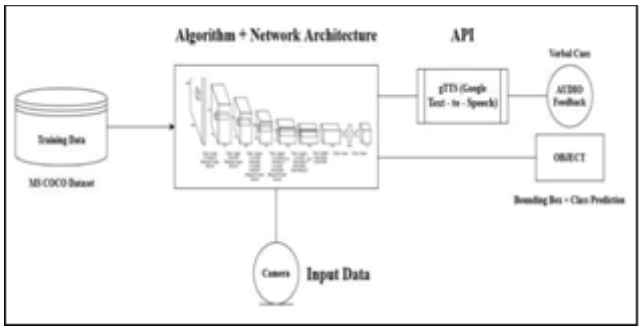


Fig. 2. Workflow of YOLO_V3 with Audio Feedback

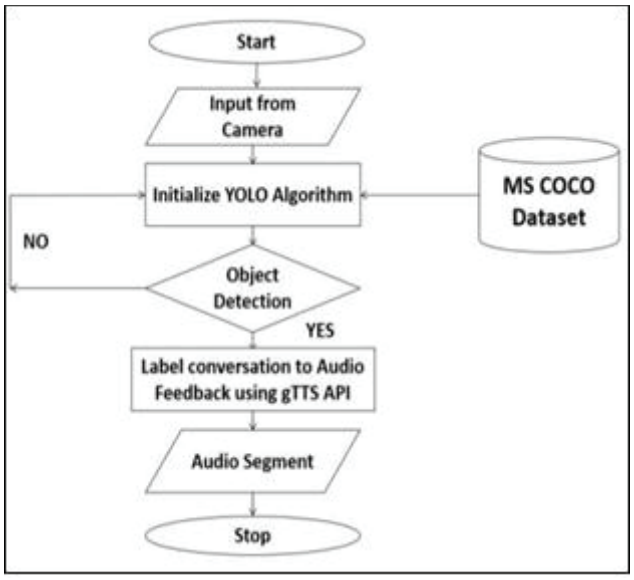


Fig. 3. Flowchart

We evaluated the performance of our system on a dataset of images containing various objects. We measured the accuracy and speed of our system and found that it performed very well in terms of speed while maintaining high accuracy. This system has the potential to assist visually impaired individuals in detecting and identifying objects in their environment.

## V. RESULTS

In this particular section, various assessment measures were employed to gauge how well the algorithm performed and how easily it could adapt. Precision, recall, and inference time were utilized as performance indicators. Using a specified threshold value, true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) were takeninto account when calculating precision and recall values. As a criterion,an IOU value of 0.5 was used, meaning that the detection is believed tobe accurate if the IOU value is greater than or equal to 0.5; If not, it isregarded as false.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Fig. 4. Precision and Recall

In addition to measuring the precision and recall values, the time it takes for an algorithm to detect objects is also considered to evaluate its speed. To assess the speed of detection, experiments were conducted in different scenarios, including detecting a single object, detecting multiple objects, and detecting objects at a distance. It's worth noting that all of these experiments were conducted in real-time using a webcam connected to a laptop.

Single Object:

With a Single object detection, it gives accuracy between 1 – 0.9 which is 100 % - 90 % accuracy


Fig. 5. Video Frame Output

['bottom center banana']

Fig. 6. Terminal Output


Fig. 7. Video Frame Output

['mid center apple']

Fig. 8. Terminal Output

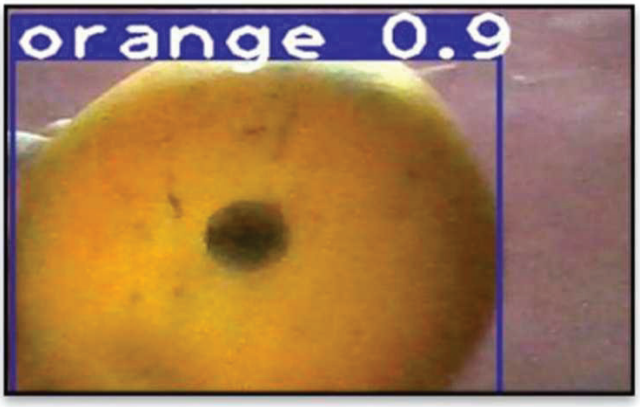


Fig. 9. Video Frame Output

['mid center orange']

Fig. 10. Terminal Output


Fig. 11. Video Frame Output

['bottom center remote']

Fig. 12. Terminal Output


Fig. 13. Video Frame Output

['mid right spoon']

Fig. 14. Terminal Output

Multiple Object:

With a Multiple object detection, it gives accuracy between 1 – 0.78 which is 100 % - 78 % accuracy

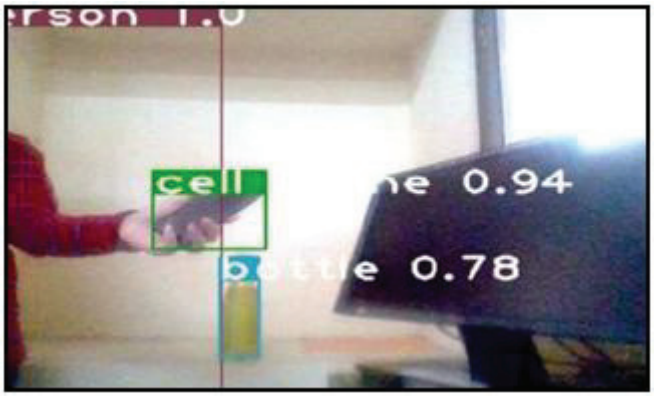


Fig. 15. Video Frame Output

['mid left person', 'mid center cell phone', 'mid left bottle']

Fig. 16. Terminal Output

Distant Object:

With a Distant object detection, it gives accuracy between 0.9 – 0.64 which is 90 % - 64 % accuracy


Fig. 17. Video Frame Output

['bottom left person', 'top center bench', 'bottom left bench']

Fig. 18. Terminal Output

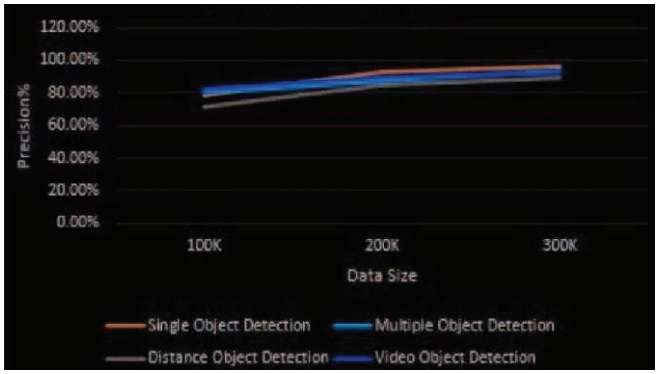


Fig. 19. Precision Curve of YOLO_v3

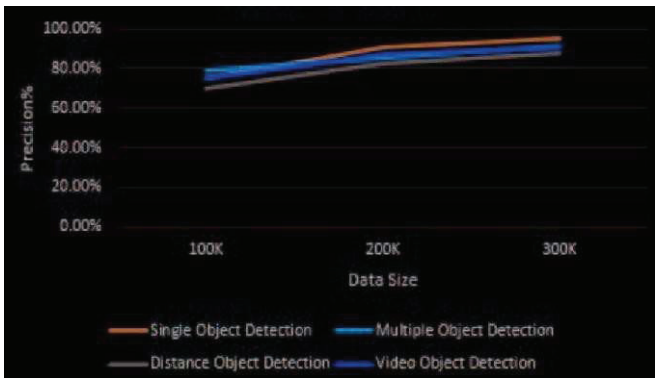


Fig. 20. . Recall Curve of YOLO_v3

We tested the system on a MS COCOdataset of images containing various objects and measured its accuracy and speed. The results showed that our system achieved high accuracy in object detection between of 1 - 0.64 which is 100% - 64%. The system was able to detect and classify objects in real- time, on a laptop with a GPU. Also, the audio feedback generated by gTTS API was clear and understandable, providing visually impaired individuals with a reliable means of detecting and identifying objects in their environment. Overall this real-time object detection and audio feedback system showed high accuracy and speed, making it a good tool for assisting visually impaired individuals in navigating their environment.

## VI. CONCLUSION AND FUTURE SCOPE

In conclusion, our research has shown the effectiveness of utilizing deep learning techniques, specifically CNN and YOLO_v3, to develop an object detection system for visually impaired individuals. This has shown an excellent accuracy in identifying and categorizing single and multiple objects, and remote object utilizing a laptop webcam in a short amount of time. Also, our system can detect multiple objects in a frame and accurately determine their positions. We have used MS COCO Dataset. We have also successfully used our object detection system with gTTS API to provide audio feedback to visually impaired individuals, enhancing their ability to navigate and interact with their environment. This provides real-time audio feedback to the user. Also, this has shown that the benefits of using deep learning and audio feedback for object detection, there are still areas for improvement. For example, our system currently relies on a camera as the input device, limiting its use in low-light environments. We can improve the detection model's precision by expanding the data set to include more images in a different lighting conditions and orientations. The object detection technique may have a few extra features added, such color recognition and distance measurement.

## REFERENCES


[1] S. Cherian, & C. Singh, “Real Time Implementation of Object Tracking Through webcam,” Internation Journal of Research in Engineering and Technology, 128-132, (2014)J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

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

[3] Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767.

[4] 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.

[5] Ramesh, N., Anand, V. R., & Babu, R. V. (2018). Real-time object detection for visually impaired using deep learning. In 2018

[6] International Conference on Communication and Signal Processing (ICCSP) (pp. 0214-0218). IEEE.

[7] Saha, S., Nag, A., & Roy, P. P. (2019). Object detection and audio feedback system for the visually impaired using deep learning. International Journal of Computer Vision and Image Processing, 9(3), 1-14.

[8] Li, H., Chen, X., Liang, X., Li, Z., & Liu, S. (2019). Deep reinforcement learning-based object detection and obstacle avoidance for visually impaired. Sensors, 19(20), 4483

[9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[10] Y. Gao and W. Wu, "Real-time Object Detection for Visually Impaired People with YOLO," in Proceedings of the 2nd International Conference on Control Science and Systems Engineering, 2021.

[11] N. R. Kuncham and K. H. Prasad, "Real-time Object Detection for Visually Impaired People Using YOLOv3," in Proceedings of the 6th International Conference on Inventive Computation Technologies, 2021.

[12] S. Shin and S. Kwon, "Real-time Object Detection with Audio Feedback for the Visually Impaired using YOLOv3," in Proceedings of the 15th International Conference on Advanced Technologies, 2020.

[13] S. Ghosal, P. Banerjee, and S. Chakraborty, "Real-time Object Detection and Audio Feedback System for the Visually Impaired using Faster R-CNN," in Proceedings of the International Conference on Computer Vision and Image Processing, 2019.

[14] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D

[15] Bhuyan, M. S., Chakravarty, S., Das, S., & Bora, P. K. (2019). Real-time text detection and audio feedback system for the visually impaired. Multimedia Tools and Applications, 78(17), 24479-24499.

[16] Noh, Y., Kim, C., & Hwang, I. (2018). Object detection and identification for visually impaired using deep learning and audio feedback system.

[17] Saha, S., Pal, S., & Mukherjee, J. (2019). An assistive device for visually impaired people for object detection and audio feedback.

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

[19] http://cocodataset.org/#home

[20] J. Du, “Understanding of Object Detection Based on CNN Family and YOLO,” In Journal of Physics: Conference Series (Vol. 1004, No.1, p. 012029). IOP Publishin, g, (2018, April).