National College of Ireland

Retail Inventory Management using Deep Learning Techniques

MSc Research Project
MSc.Data Analytics

Karthik Krishnan Iyer
Student ID: x21205485

School of Computing
National College of Ireland

Supervisor: Mr. Vikas Tomer

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Karthik Krishnan Iyer
Student ID	x21205485
Programme:	MSc. Data Analytics
Year:	2022-23
Module:	MSc Research Project
Supervisor:	Mr.Vikas Tomer
Submission Due Date:	14/08/2023
Project Title:	Retail Inventory Management using Deep Learning Techniques
Word Count:	9546
Page Count:	25

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	28th August 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Retail Inventory Management Using Deep Learning Techniques

Karthik Krishnan Iyer
x21205485

Abstract

The success of the Retail Industry is dependent on customer satisfaction and sales generated. It mainly depends on effective inventory management. If the store is fully stocked up with inventories it will lead to customer attraction and thereby result in the high volume of sales. Priorly the inventory management was done manually which was very time-consuming and also error-prone. Last two decades researchers have investigated the use of object detection using Faster R-CNN and YOLO for examining and keeping a count of stocks present on the shelf. But these researchers either examined the products present on the shelf or detected the empty spaces present on the shelf and didn't provide the count of empty spaces present in each shelf. The purpose of this research is to train the retail shelf data using YOLOv7 algorithm for detection of Products as well as the empty spaces together.To begin, the YOLO algorithm will be trained not just to detect but also count and display the count of Objects as well as empty spaces. Finally, the tesseract OCR will be employed for extracting the count of empty spaces and convert it to speech using gTTS(Text to speech algorithm).. This will allow the system to alert shop management of vacant areas, allowing them to refill inventory more effectively. Using these novel approaches has the potential to change inventory management in the retail business while also improving consumer satisfaction by decreasing out-of-stock situations.

***Keywords*— Object Detection and counting, YOLOv7, Tesseract OCR, gTTS(Text to speech conversion), Text extraction**

1 Introduction

Object Detection is the most popular and growing field in computer vision. Object detection involves detecting the object of interest in the image or video. Object detection is a technique generally used in the Medical industry, surveillance and security, robotics, augmented and virtual reality, and retail industry. The retail industry is growing industry day by day. The retail industry is mainly dependent on effective goods supply, supply chain management, and inventory management. Retail inventory is one of the main factors of the retail industry. The retail industry profits are mainly dependent on customer satisfaction and sales. Customer satisfaction is dependent on the fact that the retail store has all the products which customers are looking for and the store is all stocked up. If the store is not fully stocked it might lead to customers looking for other options thereby leading to a drop in sales and eventually generating a loss for the store. Priorly, the inventories used to be managed and examined manually, which is very time-consuming

and also error-prone. Sometimes it leads to products being overstocked or out-of-stock. In this digital era, researchers came up with the computerized way of examining inventories which can be time efficient and also less error prone. Xia and Roppel (2019) and Zhang et al. (2016) explained the use of RFID for managing inverntories.The use of RFID tags gives reduces human efforts and also provides somewhat accurate results. But, it is practically not possible to install RFID tags in each shelf. This hardware structure will be very expensive and it is not possible for every store to install the RFID tags. So, to avoid this the researchers made advancements and started using computer vision and AI for detecting the products in the shelf. Deng and Yang (2021) and Goldman et al. (2019) made use of computer vision technology for detecting the stock in the shelf. They used two-stage detectors faster R-CNN for detection. These detectors performed well in dense scenerios. But due to the two stages which includes the region proposed network(RPN) this leads to slower inference speed. So for this the researchers came up with one-stage detectors which have one layer and include feature pyramid network which leads faster detection speed.Wu and Lu (2022) introduces the detection and counting of stocks using YOLOv3. Saqlain et al. (2022) and Yilmazer and Birant (2021) propose a detection algorithm to detect empty spaces in the shelves.Detecting empty spaces is very essential as timely detection of empty and almost empty spaces will help the store management to refill the stock as soon as possible. Overall it summarizes that there have many advancements in retail inventory management through object detection algorithms. But there is no algorithm that detects both objects as well as empty spaces on the shelf. If the empty spaces are detected and moreover there is no algorithm that detects both empty spaces and objects and provides the count for both. In small stores it is possible for the store management to manually go through every image and look for an empty space that needs attention. But in large supermarkets it is very time-consuming. Therefore this proposed research involves two novel approaches which can be applied in further study:

- Firstly, the algorithm will be trained to detect as well as provide counts of Objects as well as Empty spaces.
- Secondly, the counts of empty spaces are extracted using an image to text converter, and convert the extracted text to speech so that the image can speak out to the store management that there are n number of empty spaces present on that particular shelf and it needs to be re-stocked as soon as possible.

Research Question

"How accurately can object detection and counting algorithms, evaluated by mean Average Precision (mAP), detect both products and empty spaces on retail shelves for inventory management in the retail industry? Furthermore, how can speech conversion be employed to notify store management of empty shelf spaces, and what are the potential benefits and limitations of this approach?"

This will be very essential in real-world scenarios. The empty shelf needs to be detected and restocked as feasible. So, this will assist the store management for efficient stock of inventories in the shelf and examine the products which are out-of-stock or in-stock to avoid the scenarios of over-ordering the products. In the small stores it is practically possible for store management to verify the camera and restock the products in the shelf. But, in large supermarkets it is very time-consuming to run through very detection image. So, in this situation the text extraction and text to speech technique will assist the

store management to inform the store management empty spaces are present in which shelf.

2 Related Work

The literature review consits of the researches from the past decade, their methodology, advantages and limitations. This section is divided into 6 subsections. subsection 2.1 introduces methods for counting the stocks in the shelf, subsection 2.2 about void space detection, 2.3 about Shelf Monitoring and Planogram Compliance, 2.4 about the Lighting conditions for object counting and 2.5 about Text extraction and Text to speech conversion.

2.1 Counting of Stocks in the shelf

Counting operations is essential in many businesses. The retail industry and many businesses run on profits and customer satisfaction. So, for good sales and customer satisfaction, the inventories should be managed effectively. There are effective counting methods which have been proposed in the last decade. The Wu and Lu (2022) and Jenkins et al. (2023) have described the use of counting algorithms using the retail store dataset.The Wu and Lu (2022) used the Yolov3 algorithm for the detection and counting of different coloured cups in the shelf. They state that the traditional methods were used priorly to count the stocks manually on the shelf which was very time-consuming also not that much accurate.This author suggested collecting real-time images of shelves using a camera and transferring the images to the Yolov3 network. There are three feature maps in the yolov3 network in which the author used two feature maps and removed the structure of the third feature map by employing the k-means algorithm which produces better anchor size. Although the changes made to the network made a significant increase in the performance of detection FPS from 48.15 to 54.88 which thereby contributed to a significant increase in the mean average precision(mAP) from 95.65 percent to 96.65 percent. Most probably the author suggests using this modified network for tiny object detection tasks or where the objects to be detected are of the same size. The second algorithm the author used was yolov3 with redefining the anchor size. By performing a clustering algorithm the author reset the anchor size from 96.65 percent to 96.82 percent. Finally, the author performed the Yolov4 algorithm in the data which lead to 96,06 percent accuracy. So, the overall author concludes that using Yolov3 with two feature maps and a k-means clustering with resetting the anchor size saw a significant rise in performance as compared to other algorithms used by the author. In the research Jenkins et al. (2023) stated that there has been a significant increase of the use of 3D scenes in recent years. The author states that the performance of 3d detectors can vary in the densely packed scenes. The author designed a novel regression-based approach for inventory management. Which includes two stages the first stage includes the use of 2D object detectors for fine-grained detection and classification of objects on the shelf and in the second stage the pointnet backbone is used. In the method, it provides the end-to-end count of objects in one row of the shelf. Overall, the 3D regression-based counting methods outperformed any other detection methods thereby giving the exact count of the objects in one row. The Countnet 3D is employed on the large real-world dataset where there is an extremely dense scenario is present so this method is used and achieves an error rate of 11.01 percent. First method shows the improved version of Yolov3 algorithm which performs well for

Author	Technique	Evaluation Used
Babila et al. (2022)	YOLOv4	mAP, F1, Recall
Babila et al. (2022)	Yolov5	Precision, F1
Deng and Yang (2021)	Multi-sampling faster R-CNN	AP, AP.75, FPS
Goldman et al. (2019)	Faster R-CNN layer with jaccard index	MAE, RMSE, AP, AP.75. MAE
Jenkins et al. (2023)	CountNet3D	MAE, MAPE, MSE
Chen et al. (2019)	EfficientDet and YOLOv5	mAP, f1, AP, AR
Wu and Lu (2022)	YOLOv3	mAP, Precision, Recall
Saqlain et al. (2022)	YOLOv3, YOLOv4, YOLOR	mAP,f1,recall, Precision

Table 1: Summary of Techniques and their Evaluation

the shelf of cups that are densely kept on the shelf. This is a 2D object detector. While the second detector shows the object counting in 3D scenerio using pointbeams where the object in a row can be detected.This is the advanced version of counting. Overall it shows that the 3D algorithms gives the exact count of the objects in the shelf but the hardware requirement can be expensive for it.

2.2 Void Spaces Detection

The main problem in the retail industry is that the items getting out-of-stock. To prevent the products from being out-of-stock it is important to keep an eye on the shelf and look for empty spaces. If there are any then restock it. The Jha et al. (2022) and Chen et al. (2019) suggest ways to detect and manage empty space scenarios. In the paper Jha et al. (2022) states that the main reason for a drop in sales is that the products are out-of-stock. So, the store management used to examine the stock manually which is very time-consuming and also not as accurate. So for this, the author came up with an automated way to detect the Out Of Stock on the shelf. The author used the Faster R-CNN algorithm which is a two-stage algorithm used to detect the object location in the shelf. The author combined the faster R-CNN with the gray level occurrence matrix and color features for out-of-stock detection. Overall this method produced good results as compared to other methods. Overall three methods were employed to detect the Out Of Stock in the shelf. The canny operator, gray level co-occurrence matrix, and color-based feature. The results of the canny operator stated that it outperformed other methods. It achieved a high recall, precision, and F1 score. Overall the out-of-stock also depends on the shelf type, lighting, and camera angle. So in the future author is advised to keep the edge information with color information for effective out-of-stock detection.Chen et al. (2019) described that on-shelf availability has to maintained in the retail store specifically in Fast Moving consumer goods. The main profit of the retail industry depends on the sales and customer satisfaction. If the product becomes out-of-stock the customer finds other options, eventually which causes loss in the long run. Due to this the On-shelf -availability should be maintained. For this, the retailers have come up with different techniques some of which are manual which is very time consuming or which uses hardware components which are very expensive. Recently the machine learning solutions have become popular. So, the author came up with an end-to end machine learning for empty spaces detection on the shelf. The author addresses that the images quality should be considered and the images should be annotated before deploying the model. So overall, the processes dealt in this machine learning pipeline building is

data collection which include collecting shelf images from store, then data cleaning by improving image quality by removing noise then annotating every image with labels. As, in this the empty spaces are to be detected so, in each image the empty spaces should be labelled using some tool. Then the annotated dataset is then split to training and testing before passing for model training. Then the model is tested on the inference images by tuning the hyperparameters if it works well for the other shelf images as well and then the model is deployed. The author used EfficientDet and yolov5 algorithm for this research. The yolov5 algorithm outperformed efficientdet algorithm in terms of all the factors.

2.3 Shelf Monitoring and Product Recognition:-

It is very important to ensure on-shelf availability for effective operation of business. For this the store should keep the track of products. The Sinha et al. (2022) descibes the product recognition in the shelf using deep learning. Yilmazer and Birant (2021) and Saqlain et al. (2022) describe the shelf monitoring and planogram compliance. Sinha et al. (2022) suggest using the automated inventory management system which is used to predict the class of the products in the shelves in the retail shop through images of the shelf and shelf class information. In this they have used two stage object detection and recognition pipeline which comprises of faster-R-CNN model which is used for object detection in the retail shelves and combination of ResNet-18-based image encoder which classifies the detected regions into appropriate classes. So, overall the author compared the results of the proposed method with the prior methods of full pipelines. They summarize that the combination of faster r-cnn , resnet50 and res-net 18 gives the most accurate results in terms of mAP score, F1 score and precision recall. Due to the combination of loacalization i.e. Faster r-cnn model and recognition framework which include ResNet based embedder, which is used for classification of products in the shelf it gives the appropriate recognition and class of the object present in the shelf. The method used by the author is light weight neural network so it requires a small memory and it is suitable to run on any device and this is also time efficient.Saqlain et al. (2022) came up with the hybrid approach known as Hyb-SMPC. The Hyb-SMPC algorithm consists of two parts in which the first part includes the detection of retail shelves using one-stage detectors. In this the author has fed the images to the three one stage detectors namely You Only Look Once (Yolov4), Yolov5 and YOLOR i.e. You Only Learn One Representation. They selected the one which gave the best results. The second part is planogram compliance for this the retail shelves company provided layout is used as a reference. The layout is first converted to Javascript Object Notation (JSON) then these json notations are matched with the postprocessed image from the output of one stage detectors. At the end the overall compliance report will be generated which indicates the level of compliance. So, overall according to the results of the three one stage detectors the yolov4 algorithm was chosen to train the model further. As the mean Average Precision(mAP), f1 score and recall is comparatively greater for the Yolov4 as compared to other two algorithms. Also priorly the planogram compliance used to take place through matching the color detection but the author has used generating json so using generating json method takes less processing time for planogram matching.Yilmazer and Birant (2021) states that most of the real world datasets doesn't have annotations with the data so, it is practically very difficult to annotate every image in the dataset. So to avoid this the author suggests using the new method which combines semi supervised machine learning approach with the on-shelf availability(SOSA) for the first time. The author

employed yolov4 algorithm for detecting On shelf Availability. It also includes as new software app called SOSA XAI which used SOSA and artificial intelligence. The dataset were divided into labelled and unlabelled data. The labelled data included annotations of "Object", "Almost Empty Shelf" and "Empty Shelf". Firstly the labelled data was fed into the yolov4 model to train the deep learning model. Then the trained model is used to predict the labels of unlabelled images there by increasing the accuracy. To evaluate the performance of this model this was tested on real world data of shelf images. So, overall this method outperformed other prior methods like retinanet and Yolov3.

2.4 Lighting Conditions for Object Counting:-

Lighting Conditions play an important role for counting the stocks present in the shelf. As in store, some places are exposed to low light areas which makes it difficult to detect and count the objects using computer vision algorithms. Likewise, some places are overexposed with light this also makes it difficult for object detection algorithm. Babila et al. (2022) and Sharma et al. (2019) explains the object counting scenerios in different lighting conditions. Babila et al. (2022) have reported a study of detection and counting of cellphone boxes namely, Cherry Aqua S9, Cherry Flare S8 in any orientation. Primarily they build the dataset using the images of the cellphone boxes which were taken in different angles. They gathered a dataset of 1623 images which included 32 target images in different angles and different lighting conditions. The main aim of this research was to apply an object detection algorithm to identify cellphone models placed on the shelf. The research will cover following parts: 1) Gathering and compiling photos of mobile models from various perspectives, including a) normal (0° from the z-axis), 180° (upside down), -90° and 90° (sideview), and vertical orientation. 2) Detecting objects with various light sources, such as 5W, 7W, and 9W, and assessing the detection accuracy with each light source. 3) Following detection, the YOLOv5 algorithm was used to count the mobile boxes and show the count on the screen.The author claims that the biggest benefit is that the YOLO algorithm identifies only learned items. This algorithm's output demonstrates good accuracy, recall, and precision.Sharma et al. (2019) introduces the object counting under different lighting conditions. Firstly the image is fed into the model and it captures the important features of the image using KAZE features. Then the detected features are then subjected to density based scanning algorithm. Then the clusters obtained are then processed by homography transform. Then the homography transform is used to predict the locations of the objects in the image.The homography predicts the results in the rectangular polygon box. Since there will be multiple prediction for a single object instances the predicted object instances are then combined with the density based scanning clustering algorithm so after this the count is obtained.The authors have also designed the user interface which is user friendly. Overall author compared this alogorithm with previously developed algorithms and summarizes that the algorithm outperforms the previous algorithm by performing well on different lighting conditions like low light and dim light.

2.5 Image to Text and Text to Speech Conversion:-

Pawar et al. (2019) explains the image to text conversion and how to extract important texts from the images. Liu et al. (2023) describes the extracted text to speech conversion. There some of the important textural data present in the data which needs to be stored

digitally. Pawar et al. (2019) used Optical Character Recognition algorithm with the help of Tesseract OCR Engine. Tesseract OCR Engine is used to extract important texts from the documents or any images which contain some important information.The author introduced different methods associated with OCR which include Connected Components method which detects the pixels differences in the text and the background of the image , Sliding window based method which locates and recognizes text, Hybrid approach which recognizes text in captcha images.Edge based method us used for text extraction in image segmentation, processing and computer vision algorithms, Texture based method and cofner based method is used to detect text and extract textural properties and recognizing text line from the bounding box respectively. Final method is stroke based method which is used to detect and recognize text from the video. So, then the extracted texts can be converted to audio format so that it can help the visually impaired people to hear the information they need and is important to know.Liu et al. (2023) specifies that the text to speech conversion and voice conversion are two different approaches and both of the tasks goal for the high quality speaking voice in different pitch voice. This paper proposes a novel approach of Unify speech which is a combination of Text to speech and voice conversion. The author defines that the text to speech conversion is done from text extraction method but the voice conversion is done from the source speech. So overall author specifies the bridge gap between the both the Text to speech conversion and voice conversion.By the evaluation on the basis of objective and subjective terms the author tells that the text to speech conversion is a better speaker ability as compared to Voice conversion

3 Methodology

This section involves the Research process flow of this research. The process flow follows the Knowledge Discovery Data(KDD) as it turns out to be a very good methodology for image datasets. This approach has 6 main stages i.e. Data Collection, Data Preprocessing, Data Transformation, Data Mining/ Model Building, and Evaluation. Each stage has different substages. Data collection is the primary step in this research. The next stage data preprocessing includes substages like image resizing and noise removal, dataset annotation, and feature extraction which is explained further. The next stage is data transformation is the step before the model building which is an essential step. This includes two substages that are data augmentation and dataset splitting. The next stage is the most important stage which is data mining/ model building. This research has 3 stages of model building, which include object detection, object counting, text extraction, and text-to-speech conversion. Then the final step is a model evaluation which includes evaluating the model's performance. The model's performance in the Object detection algorithm like Yolo is done using Mean Average Precision(mAP), Precision, F1 score, recall, and Intersection Over Union(IOU).The stages of the process flow are as follows:

3.1 Data Gathering

This is the initial stage of the methodology. This stage includes the collection of the data for the model building. In this research, the dataset of images is required for the object detection algorithm to run. The dataset should be chosen according to the requirement of which object to be detected. There are many things to be considered before collection of data:

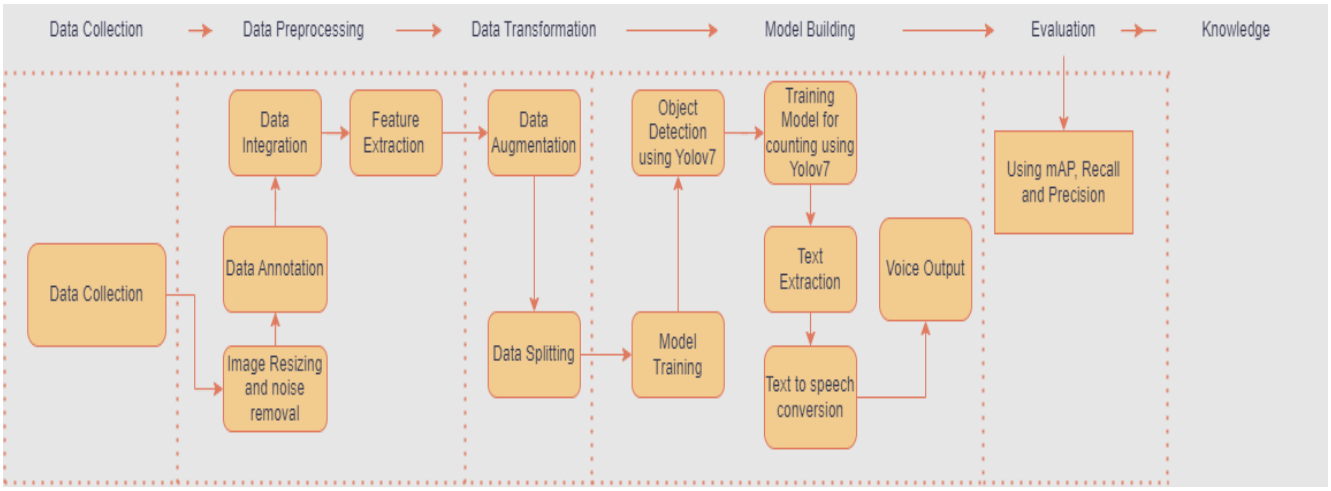

Figure 1: Process flow

- The dataset should have the elements which are essential for detection purposes. The author should make sure that there are enough objects in the images to perform the detection algorithm. In this research, the object detection algorithm is performed for getting the presence of objects as well as empty spaces in the retail shelves.

- The dataset can be any open dataset or any custom dataset which can be obtained from any website or Kaggle. The data can also be obtained from the retail store's real-time data but there should be an agreement between the author and the store management that the data will only be used for research purposes. In this research, the data collected is an open dataset which is obtained from the roboflow website which includes all the open-source computer vision datasets. So this dataset is the SKU-110k dataset. This includes the retail shelves data which has an annotation of objects with it. Each image has an annotation of an object with it.

- The second step is annotating the images so that they can be used for further processing. The dataset used has images of shelves of a retail store. This dataset includes images in different lighting conditions and different angles of retail shelves. These images can be used for training so that model can be trained for such scenarios.

- Finally, the research includes detecting objects as well as the empty spaces in the retail shelf so there should be the presence of empty spaces and objects in somewhat equal instances on the shelf as it will be trained for both cases in the same image.

- The downloaded dataset has 10000 images with annotation of objects. Only a instance of 2688 images with annotations are used for operation.

3.2 Data Preprocessing

Data preprocessing is an important stage in any deep learning task to make the dataset ready for the model to be applied effectively. This involves a series of sub-stages like image

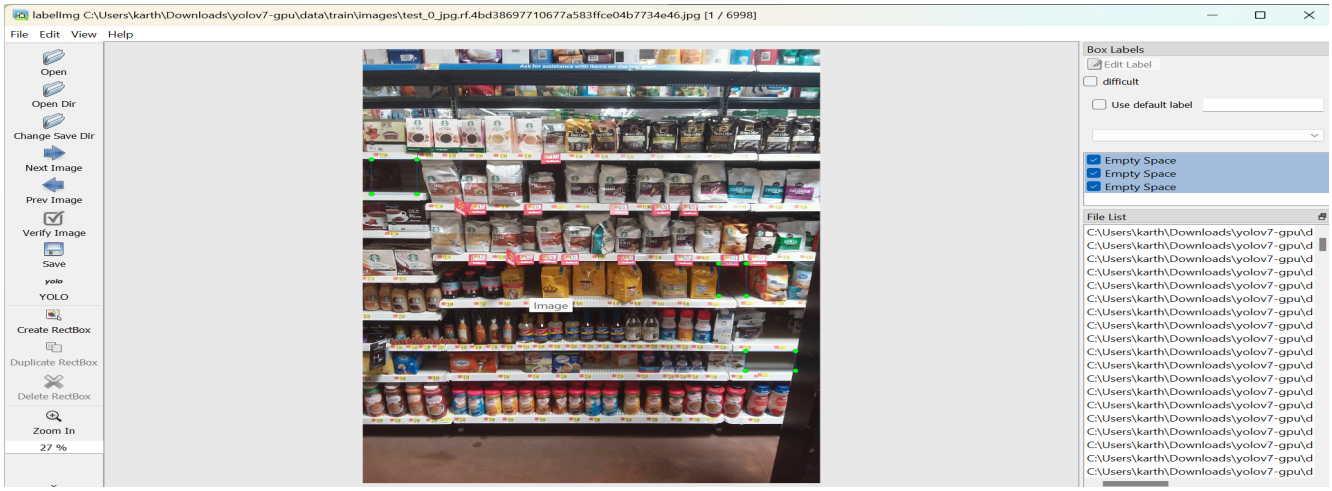

Figure 2: Data Annotation using labelimg

resizing and noise removal, image normalization, background removal, data annotation, and feature extraction.

3.2.1 Image scaling and noise reduction:

For object detection models like Yolo, R-CNN, and SSD to be applied the input images needs to be a fixed aspect ratio. The distorted image ratio will lead to less model accuracy. The main task is to maintain the aspect ratio. Mostly the width is kept as 416 pixels and the height is decided based on aspect ratio. The image needs to be made sure that the features are visible enough for the model to be applied. So, for removing the disturbance in the image and make it clear enough for detection noise removal is required. Different types of noise removal techniques can be applied like Gaussian filtering, median filtering, bilateral filtering, and specialized deep learning denoising networks.

3.2.2 Data Annotation:-

Data annotation is an important step in any object detection algorithm. Data annotation involves labeling the objects which need to be detected in the model. It is very essential for training the object detection model. This includes assigning specific classes to the objects by creating the bounding box around them.There are several steps involved in data annotation:

- To choose an appropriate tool for annotating objects in an image. Some of the popular tools include RectLabel, VGG Image Annotator(VIA), LabelImg, and Labelbox. So, in this research, the labelimg tool is used for annotation. For using the labelimg tool it is set up and installed in the Anaconda environment.

- Then open the images folder that needs to be annotated and create a text file that includes the class labels. In this case, there are 3 class labels which include 'Almost Empty', 'Empty', and 'Object' in order respectively.Yilmazer and Birant (2021) have segregated the classes as 'Empty', 'AlmostEmpty' and 'Object' respectively.Then load the images folder that needs to be annotated.Then by loading the image create a bounding box around the object and assign a particular label. In case

of empty spaces name it as 'Empty' and likewise for 'Almost Empty' and 'Object'. It should be made sure that the bounding boxes do not overlap with each other and the occlusions should be handled wisely. If there are groups of the same objects present in a single line they should be annotated separately instead of annotating them together so that the model is well-trained.

- Then the annotation should be saved in the same folder with the same name as the image so that the algorithm can identify the same file. The annotation includes (class index, center x, center y, width, height). The class index is different for each class label. In this research, the 'Almost Empty' is assigned 0, 'Empty' as 1, and 'Object' as 2.

3.2.3 Data Integration:-

Data integration is combining datasets from different sources to obtain a new dataset that can be trained or combining the created labels of the image with the preexisting image annotations so that the image can have more classes and the model can be trained for each of the classes. This involves collecting images that have the objects that need to be detected in the model. Make sure that the labels of the classes are in the format of the (class index, x, y, width, height) format. Each class has a different index. So in this research almost in every image, the objects were already annotated. So, it was required to manually annotate 'Almost Empty' and 'Empty'. So, the text file was created in the images folder. So this time, it was created with 'AlmostEmpty', 'Empty', and 'Object'. in the same order. So, that it can be annotated accordingly to the preexisting annotations. As the 'Object' had an index of 0, it had to be changed to 2 in each image. As the object was changed to 2 index. The images were annotated for 'AlmostEmpty' as 0 and 'Empty' as 1. So, it was done for each image. After completing annotations for each image it was saved in the .xml format which is compatible with the yolo. The annotated file of each image was saved under the same name as the image file. Then each file manual annotation was combined with the preexisting file of object annotations. As this will combine all three classes annotation and the model now can be trained for all three classes simultaneously.

3.3 Data Transformation:-

It is the most important phase before moIt is the crucial step before the model building as it makes the data ready for model building. This helps the model train in different scenarios to get better accuracy and get perfect results while testing the image. Data Transformation involves two stages data augmentation and data splitting:

3.3.1 Data Augmentation:-

Data augmentation is the process used to increase the size of the training set by application of various transformations to the images. It increases the robustness and performance of the model to train in those situations. It makes the model more efficient. There are a series of steps involved in data augmentation.

- Random Flipping: This is the horizontal flippling of the image and adjust the bounding boxes accordingly. This will assist the model to learn to detect objects in such scenarios.

- Cropping the image at random: This includes cropping the image for only getting the objects of interest that need to be detected. This helps the model to only train the images and detect only the objects which are required.
- Colour jittering : This includes applying random color to some of the images, such as increasing or reducing brightness, contrast, hue, and saturation. This will assist the model to learn in all the lighting conditions.
- Blur and Sharpen: This includes applying blur or sharpening filters to some of the images according to the requirement. This helps the model to learn effectively in such conditions.

3.3.2 Data Splitting

Data Splitting is the vital step for the preparation of the dataset for training and testing the model. The aim of this is to divide the dataset into 3 sets i.e. training, testing, and validation. This will ensure good evaluation and avoid overfitting. In this object detection algorithm to be performed in the data, it is divided into training, validation, and testing. All the annotated images and the annotation of each image are present in the same folder. And it will perform a split code to divide the data into training, validation, and testing.

- Training:- The training set of the data consists of the maximum part of the data. It is used to train the object detection model. In this research, the training set constitutes 80 percent of the data.
- Validation:- The validation set of data is used to examine the model's performance by tuning the hyperparameters. The validation set in this research constitutes 10 percent of the total data.
- Test Set: The test set is not used for training the data. The test set is used for testing the model's performance after its training. The test set in this research constitutes 10 percent of the total data.

3.4 Model Building:-

This is the most crucial stage in the object detection algorithm. After the data splitting stage, the split data is taken as input in the model-building stage. The model building stage has different sub-stages like model selection, model training, hyperparameter tuning, and model testing. In this research, Object Detection needs to be performed for detecting the objects as well as the empty spaces in the shelves. So, for this, the model needs to be trained for the detection of objects as well as empty spaces together. And apart from this, it should give the count of the objects and empty spaces in the image of the shelf. This is the first novel approach employed in this study. The second novel approach includes the count of the empty spaces extracted from the image and converting it into speech so that the machine will speak out to the store management about how many empty spaces are present on the shelf. So, in this research, the model-building stage has several stages which include Model Training, Object Detection with the model, Applying a counting algorithm for the detected image then extracting the text of empty spaces from the image and converting it to speech. So, these steps are given below:

3.4.1 Model Training:-

The split data is then passed to the model training stage for the training of the data. The first important step for model training is to choose the appropriate algorithm for training the data. In this research, the You Only Look Once Version 7 (Yolov7) is employed for training. YOLO is known for its real-time object detection technique as because of its speed and accuracy. YOLO is a one-stage detection algorithm. It is employed for both the detection and counting of objects in an image. As mentioned by Deng and Yang (2021) and Goldman et al. (2019) the other algorithms like faster R-CNN and SSD do not perform well for the dense scenarios where occlusions of objects are present as it is a two-stage detector that detects the presence of objects using multilayer feature maps which results in the algorithm to perform poorly and also the speed is also less. As discussed in the previous research theWu and Lu (2022) stated that the YOLO algorithms contain three feature maps as discussed in the study. Two smaller feature maps are kept and the structure behind them is removed and employed with clustering analysis which makes the YOLO algorithms perform better accurately as compared to other algorithms. So, in this research, the model training is performed using yolov7 algorithms.

3.4.2 Object Detection using Yolov7:-

The yolov7 algorithm is employed for the object detection model. The detection of an image includes creating the bounding boxes around the objects which are trained. Object detection can aid in determining which goods are present on the shelf and which need to be stocked up or ordered. Object detection was done using the pre-trained weights of the yolov7. The dataset contains the images of the shelf with their respective annotations. Overall in previous research either the objects were detected on the retail shelf or the empty spaces were detected in the shelves. This research includes a novel approach combining both detections of empty spaces as well as the objects present on the shelf. Space detection is important as it will aid the store management to know where the empty spaces are and what needs attention. In the previous research, the author developed a machine-learning pipeline to detect the void spaces in the shelves. Jha et al. (2022) performed two algorithms and tested their accuracy. They performed EfficientDet and Yolov5. So, overall after analyzing the yolov5 surpassed the efficientDet in terms of accuracy and performance. So in this research, both the empty spaces as well as the objects will be detected. The result of this stage will be the bounding boxes on the detected objects as well as the empty spaces naming them as objects and empty spaces.

3.4.3 Object Counting using Yolov7:-

This is the second novel approach as there is no prior method that detects the empty spaces as well as the objects on the shelf and also provides the count of both. Yolo performs well in dense store shelf or occlusions scenarios. In this stage, the trained images are fed into this stage so this gives the appropriate count of the detected objects and empty spaces in the shelf. In this stage, the counter is created which records the count of objects as well as empty spaces. This method will employ the use of object tracking to track and count the objects and empty spaces on the shelf accurately. Secondly, it will involve the use of a Non-Maximum Suppression approach which will reduce the overlapping of bounding boxes and increase the count accuracy.Wu and Lu (2022) used the Yolo algorithm for counting the cups on the shelf which used two smaller feature

maps and combined with the clustering analysis which gave the accurate counts of the cups in the shelf.

3.4.4 Image to Text Conversion/ Text Extraction:-

In this digital age, most of the important information is taken through photographs. So, to retrieve this important information, it is necessary to extract those texts from the images for further application. Pawar et al. (2019) explains the use of Tesseract OCR for Image to Text conversion. In this research, the images are shelf images it is used to detect the presence of objects and empty spaces. It employs the use of yolov7 for detection and displaying the count of objects as well as empty spaces. This is the final novel approach applied in this research as the count of the objects and empty spaces displayed in the image is extracted and converted to speech. So, extracting the text is performed by Optical Character Recognition(OCR) method. The Tesseract OCR library of Python is used for this purpose. The previous research specifies the use of Tesseract OCR for text extraction from the image.

3.4.5 Text to Speech Conversion:-

The extracted text is passed to the text-to-speech converter library in Python so that it can be converted to speech.Liu et al. (2023) and Siby et al. (2020) have performed the text to speech conversion using Tesseract OCR. The Tesseract OCR engine is used to convert the extracted text from the images to speech. In this research, text-to-speech conversion is employed so that the count of the empty spaces can be extracted and converted to speech so that the store management can be informed through speech that there are empty spaces on the shelf that need attention.

3.5 Evaluation:-

The main parameters for evaluating the one-stage object detection and counting model includes Mean Average Precision(mAP), Precision, Recall, Intersection over Union, F1 score and Frames per second.

IoU :-The Intersection over Union is the defined as how much the predicted bounding box is overlapped with ground truth box. After the model training in YOLO algorithm each object is detected using a bounding box and the class is predicted. The predicted bounding box of the algorithm is compared with the ground truth box. Higher the IoU value indicates the object detection algorithm is detecting the objects well.

Precision :- The Precision score indicates the ratio of True positives to the sum of false positives and true positives. The true positive indicates that the measure of correctly predicted objects in the bounding box . False positives indicates the number of predicted bounding boxes which doesn't have the object present in it. Sum of true positives and false positives indicates the total number of predicted bounding boxes. The good Precision score indicates that the object is present in predicted bounding boxes.

Recall :- Recall is the measure of the correctly predicted bounding boxes to all the present ground truth boxes. F1 score:- The f1 score is the harmonic mean of precision and recall. This measure is often used of class imbalances scenarios. The F1 score ranges

from 0 to 1. Close to 1 indicates that the model performs well.

Mean Average Precision(mAP):- mean Average Precision is the most important metric which is used to evaluate the performance of the object detection model. The mAP takes into account of all the other evaluation metrics like recall, precision and f1 score. The mean average precision is the measure of the model's performance on different confidence levels. The Average precision of each classes are taken into consideration. The mean of each classes average precision is taken to obtain the mAP for the trained model.

4 Design Specification

This research employs the use of Yolov7 which is the newer version of the YOLO family that is more accurate and can perform multiple detections simultaneously and is also very quick in detection. Yolov7 is the recent and up-to-date object detection model which has set a bar for its high-performance rate. The Yolo network is a fully connected neural network. The Yolo-based object detectors mainly contain three components i.e. Backbone Head and Neck. Firstly the input image of 640x640 is fed to the input of the yolov7 framework. Then following it passes through the backbone network. The backbone network consists of a combination of CBS, MPConv, and ELAN blocks. The CBS block in the backbone is the combination of the convolution network, batch normalization function, and Activation function SiLU. The batch normalization technique is commonly used in deep neural networks for making sure that the training stability and convergence are improved. It includes the newly added block of Extended Efficient Layer Aggregation Network. The E-ELAN is the computational block in the YOLOv7 backbone. This block helps to increase the speed and accuracy of the model. The E-ELAN uses expand, shuffle, and merge cardinality to achieve to enhance the learning ability of the model. Overall the E-ELAN block is only the added network in this yolov7 model, which increases the accuracy as well as thereby training the model to detect multiple object detection scenarios.

5 Implementation

5.1 Environment for Development

This object detection algorithm programming was carried out using Python. This entire research is performed on the Google Colab Pro Notebook as it offers the cloud and GPU for processing faster and more complex datasets. As Python has several libraries required for object detection it makes it easier. Some packages were used for object detection. Some of the libraries used include TensorFlow, Keras Pytorch, Matplotlib, sklearn, scikitpy, OpenCV, cv2, gTTS, Text to speech, pyteserract.

5.2 Model Training:-

Firstly the yolov7 model is used for training the object detection algorithm. The yolov7 algorithm is downloaded from the git repository and cloned it with the google collab pro. This research makes use of the pre-trained YOLOv7 model. It is retrained by changing

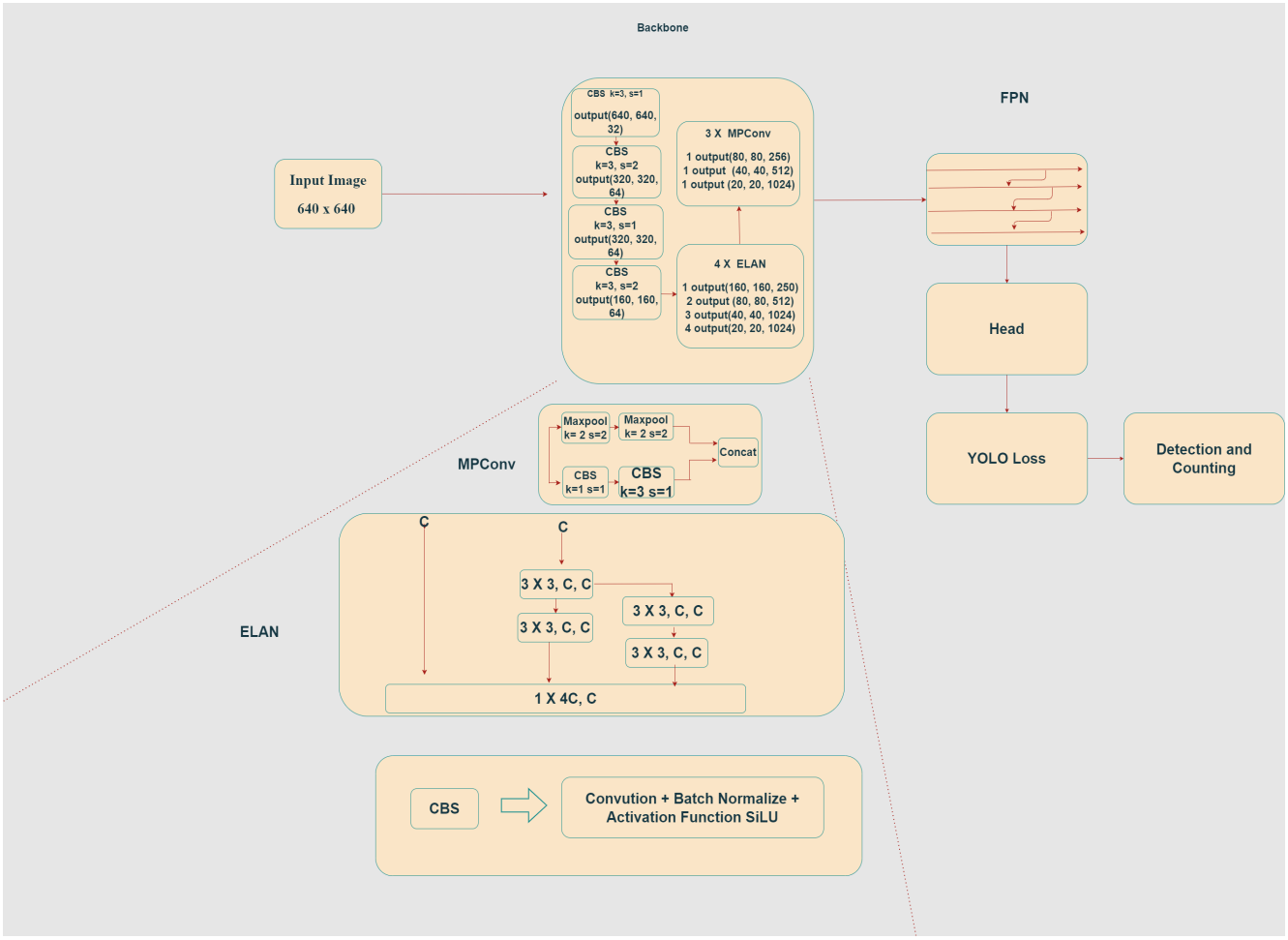

Figure 3: Yolov7 Framework

Algorithm 1 Object Detection and Annotation Algorithm

Step 1: Input the labeled dataset
$D = \{(x_i, y_i), (x_2, y_2), \ldots, (x_n, y_n)\}$
x_i contains images, y_i denotes labeled images with class labels.
Step 2: Manually annotate almost empty spaces and empty spaces
$A = \{(x_1, a_1), (x_2, a_2), \ldots, (x_n, a_n)\}$
a_i contains annotations of empty spaces and almost empty spaces.
Step 3: Unlabeled data
$U = \{x_{n+1}, x_{n+2}, \ldots, x_{n+s}\}$
s instances of shelves with products without annotations.
Step 4: Use YOLOv7 Model
Pretrained and initialized YOLOv7 Model.
Step 5: Combine annotations
Combine manual annotations y_i with a_i.
Step 6: Split dataset
$D_{\text{Train}} = \text{Split}(D \times 0.8)$
$D_{\text{Val}} = \text{Split}(D \times 0.1)$
$D_{\text{Test}} = \text{Split}(D \times 0.1)$
Step 7: Train YOLOv7 model
Train YOLOv7 model on labeled dataset D_{Train} using SGD.
for $c = 1$ to k **do**
  for each epoch **do**
    for each $(x_i, y_i) \in D_{\text{Train}}$ **do**
      $z_c = \text{Train}(x_i, y_i)$
    end for
  end for
  $Z = Z \cup \{z_c\}$
end for
Step 8: Test trained detectors
Test trained detectors on labeled dataset D_{Test}.
for $c = 1$ to k **do**
  for each $(x_i, y_i) \in D_{\text{Test}}$ **do**
    $Prediction = z_c(x_i)$
    $PredictionResult_c = PredictionResult_c \cup Prediction$
  end for
end for
$SD = \max(PredictionResult_c)$ {Selected detector}
Step 9: Perform testing on new images
Perform testing on new images to predict bounding boxes and class labels.
Step 10: Check and optimize predictions
Check for the correct predictions of objects, empty spaces, and almost empty spaces.
Optimize the confidence to improve bounding box accuracy.
Step 11: Initialize counters
Initialize counters:
$Object\_count = 0$
$empty\_spaces\_count = 0$
$almost\_empty\_space = 0$
Step 12: Loop through bounding boxes
Loop through bounding boxes, increment counters based on detected classes.
Step 13: Output results
Output the count of objects, empty spaces, and almost empty spaces in the sample image. $y' = TM(SI)$

Figure 4:

the weights and some hyperparameter tweaking. The YOLOv7 weights are downloaded from the GitHub which include 'yolov7ttraining.pt'. Firstly for training the model, the split dataset is imported into google drive by creating the .zip file. So the next step is extracting the dataset in the colab notebook. After extraction of the dataset in the Google Colab the data is trained using the YOLOv7 algorithm. For training the model, there is a command which is followed. There is one train.py configuration file that has pre-trained model training so this assists in performing training. The training command includes the train.py file, batch size, device, and weights of yolov7 i.e. yolov7training.pt are used for training the data. The batch size is set to 8 it includes that the batch of 8 images will be trained in each go, the device is set to '0 'as it indicates performing the training in the GPU memory, and '1' indicates training in the CPU memory. The file is created retailshelf.yaml which includes the dataset location and the class names. As three classes are annotated in the images namely 'AlmostEmpty', 'Empty', and 'Object'. So this is fed with the training command as this will segregate the data into classes. The training is performed for a total of 300 epochs. Then the results of the training are stored in runs/train which includes the batches of images that are trained and this involves the bounding boxes in each object on the shelf and labeling it as 'Object', and the same for 'AlmostEmpty' and 'Empty'.

5.3 Object Detection and Training:-

After training the model, the model is tested with some random images of the shelf and fed into inference/images. It is being tested for almost 6 images of the shelves. It follows a command for testing the new image. It includes file detect.py which is a

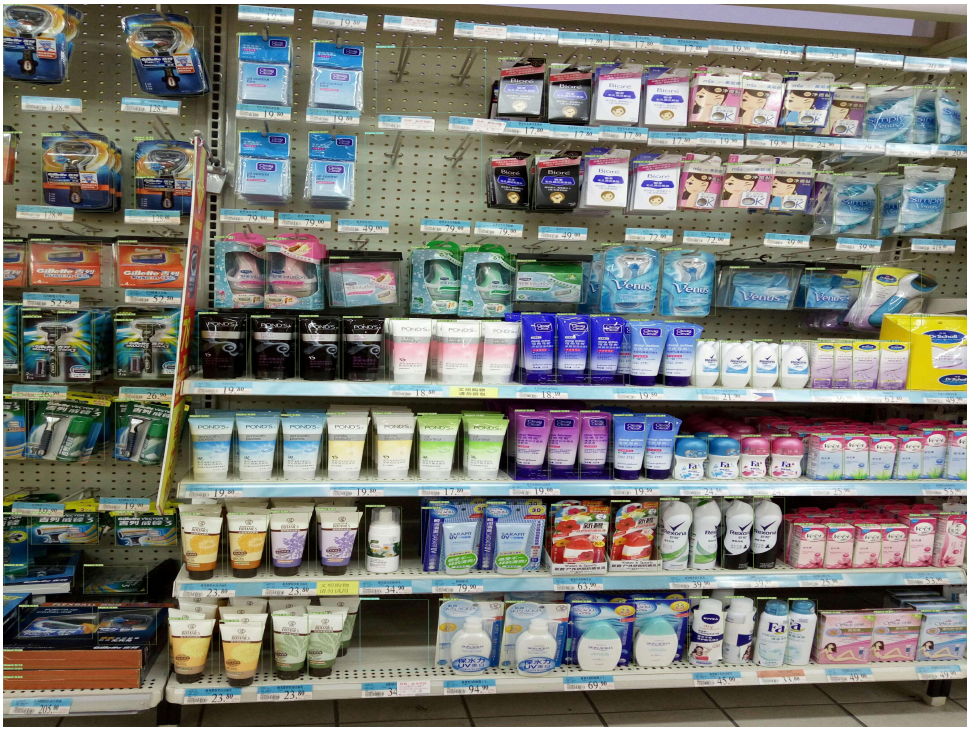

Figure 5: Object Detection Results

configuration file, and confidence threshold and the location where the image of the shelf is located in inference/images. The confidence threshold is used to determine the model's probability that the bounding box contains the desired object of interest. The confidence value ranges from 0 to 1, where higher values indicate higher confidence. The confidence threshold is chosen according to the trade-off between false positives and false negatives. The higher value of confidence provides better detections but it misses the objects with lower confidence threshold. So, for this, the confidence threshold is taken as 0.1. So, the output of the detection stage produces an image with the detection of 'Objects' and 'Empty' spaces with the bounding box and labeling it as object and empty spaces respectively.

After the detection result the object counting is employed using YOLOv7. The command for the counting of objects includes the detectandcount.py file which is a configuration file that contains all the code for initiating the counter for storing the count of the objects present. The command also includes the confidence threshold which is set to '0.1' and the location where the image is present which are inference/images. So, after executing the command the count of the 'Object' and 'Empty' spaces is obtained in the image as the counter.

5.4 Text Extraction and converting it to speech:-

After the object detection and counting stage the image is passed through the text extraction and conversion of speech stage. The counting of the objects gives the count of the objects as well as the empty spaces in the counter. The final novel approach proposed in this research included to extract the count of empty spaces from the detected image so that it can be converted to speech as the store management can be notified faster.

Figure 6: Object Detection and Counting Result

Figure 7: Audio Conversion

As in large supermarkets it is not practically possible to have a look at each image but if the audio output is given to each detected image the store management can get notified faster and can restock the empty shelf as soon as possible. For this objective to be achieved various libraries were used like gTTS(text to speech convertor), IPython.display, cv2, matplotlib and pytessearct. The pytessearct module is used to extract the text from the image. Using this module the Empty spaces count were extracted from the image. I.Python.display is used to display the image with the audio and cv2 is used to read the image file. gTTS library was used to convert the extracted text to speech.

6 Evaluation

As shown in the table 2, the training results of the model are given. The model training is performed for 300 epochs in batches. The image training of 267 batches were performed which included a total of 30629 labels. This was segreggated as 30404 for 'Object' class, 202 for 'Empty' class and 22 for 'AlmostEmpty' class respectively. So, overall it is seen that the objects are present in maximum amount. As compared to the object class, the empty and almost empty classes are very low.

So, this leads to class imbalances. Therefore, it thereby affects the overall mAP score, Precision, and Recall. As there are the highest number of objects present the object mAP score is 0.9 i.e. 90 percent. But due to the negligible amount of Empty class and Almost Empty class, the mAP score includes 0.63 and 0.0929 respectively. So overall the average of all the mAP scores is calculated as 0.541 which is 54 percentage. So, the overall Precision value is 0.674 i.e. 67 percentage for all classes and 0.363 for AlmostEmpty class, 0.745 for Empty class and 0.915 for Object class. Overall average of Recall is 0.538 which

Class	Images	Labels	P	R	mAP 0:5	mAP 5:95
all	267	30,629	0.674	0.538	0.541	0.308
AlmostEmpty	267	22	0.363	0.182	0.0929	0.0477
Empty	267	203	0.745	0.576	0.63	0.331
Object	267	30,404	0.915	0.857	0.9	0.546

Table 2: Summary of Evaluation of all the classes

is the average of AlmostEmpty class with 0.182 recall, Empty class with 0.576 recall and Object class with 0.857 recall.

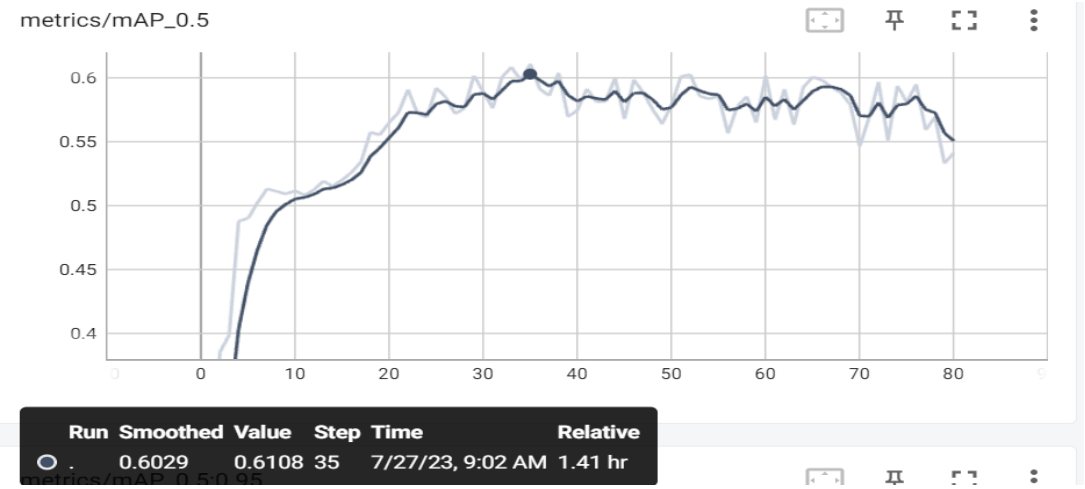

Figure 8: mAP 0:5 :- 0.6029

The highest mAP is recorded as 0.6029 in 35 th epoch.

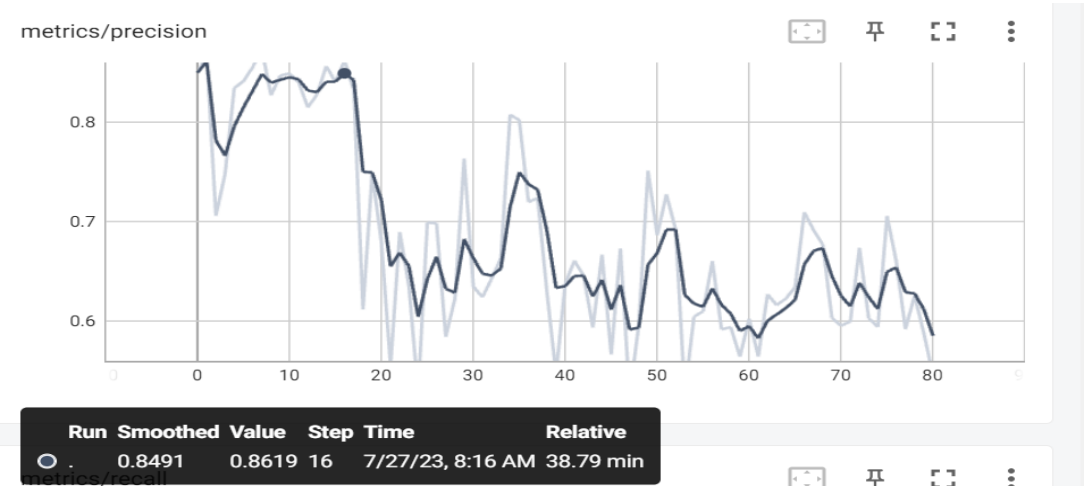

Figure 9: Precision:- 0.8491

The highest Precision is recorded as 0.8491 in 16 th epoch.

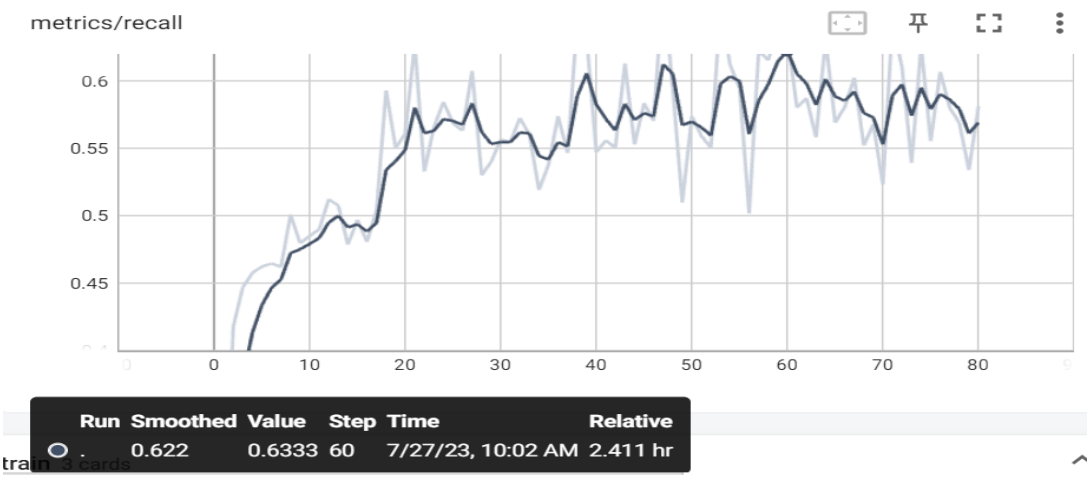

Figure 10: Recall:- 0.622

Likewise the highest recall is recorded as 0.622 in 60 th epoch.

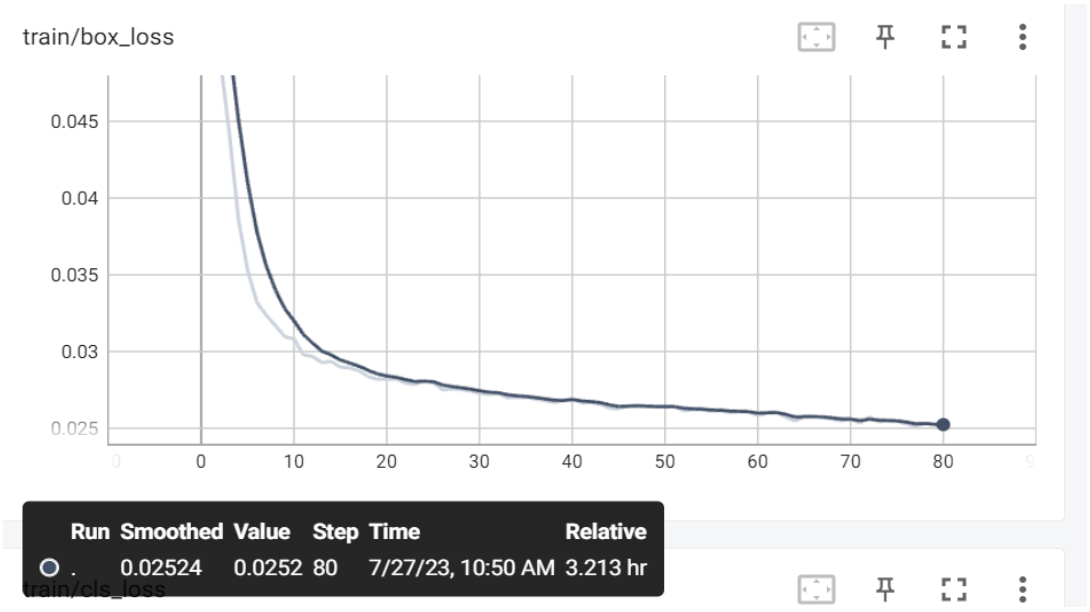

Figure 11: Box Loss: - 0.02524

The box loss is recorded as 0.02524 which is the lowest in 80 th epoch.

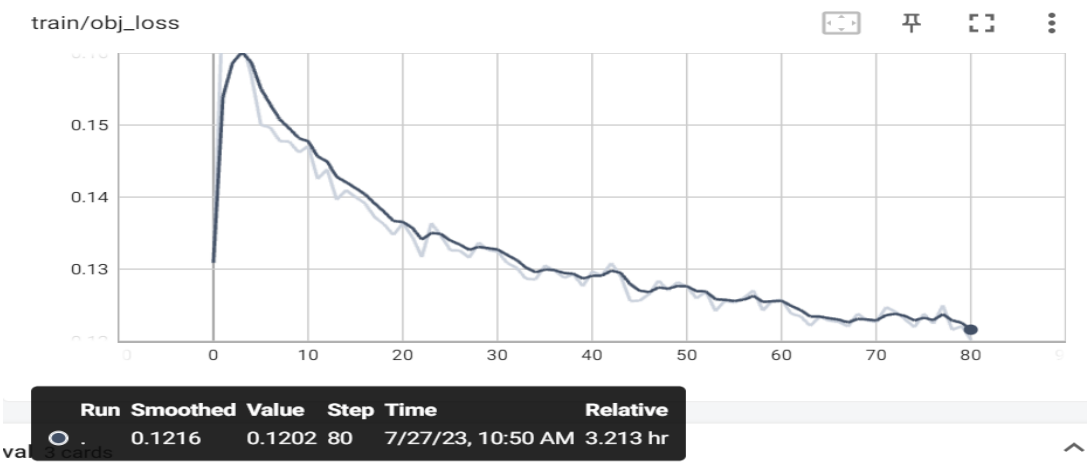

Figure 12: Object Loss:- 0.1216

The Object loss is recorded as 0.1216 which is the lowest.

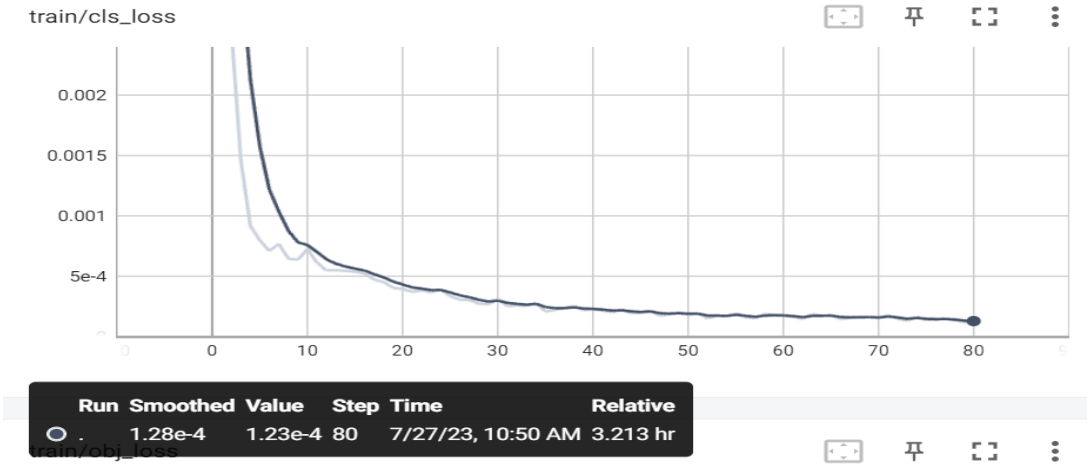

Figure 13: Class Loss:- 1.28e-4

The Class loss is recorded as 1.28e-4 which is very negligible.

There are several losses. The Object loss indicates that if the object is present in each predicted bounding box. The object loss should be as minimum as possible. In this the object loss is 0.1202 which is very low. This indicates that almost every predicted bounding box has object present in it. The class loss indicates that if the predicted bounding box object is recognized and predicts the correct class of the object. If the class loss is very low then it indicates that the images are annotated very well. The class loss in this include 1.28-4e which very negligible.The box loss indicates the accuracy of predicting the bounding box co-ordinates. This should match the ground truth box. The box loss in this is 0.025 which is very negligible.

7 Conclusion and Future Work

This research examines efficient Inventory Management in the retail sector. This proposed approach consists of two novel approaches. This includes using the Object detection algorithm YOLOv7 for detecting and counting the products and empty spaces simultaneously in the shelf.Second, Tesseract OCR was used to extract the count of empty spaces, which was then converted into voice using the gTTS (text to speech) module. The system implemented succussfully detected and counted the products as well as empty spaces in the shelf. This was integrated with the tessaract OCR for text extraction of empty spaces count and thereby using Text to speech converter for speech conversion. This research involves integration of object detection and counting with text extraction and text to speech conversion. These combined techniques can be used to revolutionize the retail industry operations.

In future the YOLO algorithm can be integrated with 3D Pointbeams so that the image can be converted to fine-grained format and the exact count of the objects in the shelf can be obtained as 3D Pointbeams are capable to detect the objects in a single row. The proposed research can be implemented in real-world scenerio by building an app which has integration of detection algorithm, text extraction and speech conversion.

References

Babila, I. F. E., Villasor, S. A. E. and Dela Cruz, J. C. (2022). Object detection for inventory stock counting using yolov5, pp. 304–309.

Chen, J., Wang, S.-L. and Lin, H.-L. (2019). Out-of-stock detection based on deep learning, pp. 228–237.

Deng, Z. and Yang, C. (2021). Multiple-step sampling for dense object detection and counting, pp. 1036–1042.

Goldman, E., Herzig, R., Eisenschtat, A., Ratzon, O., Levi, I., Goldberger, J. and Hassner, T. (2019). Precise detection in densely packed scenes.

Jenkins, P., Armstrong, K., Nelson, S., Gotad, S., Jenkins, J. S., Wilkey, W. and Watts, T. (2023). Countnet3d: A 3d computer vision approach to infer counts of occluded objects, pp. 3008–3017.

Jha, D., Mahjoubfar, A. and Joshi, A. (2022). Designing an efficient end-to-end machine learning pipeline for real-time empty-shelf detection.

Liu, H., Wang, T., Fu, R., Yi, J., Wen, Z. and Tao, J. (2023). Unifyspeech: A unified framework for zero-shot text-to-speech and voice conversion.

Pawar, N., Shaikh, Z., Shinde, P. and Warke, Y. (2019). Image to text conversion using tesseract, *IRJET International Research Journal and Technology* **6**(02).

Saqlain, M., Rubab, S., Khan, M. M., Ali, N. and Ali, S. (2022). Hybrid approach for shelf monitoring and planogram compliance (hyb-smpc) in retails using deep learning and computer vision, *Mathematical Problems in Engineering* **2022**: 4916818.
URL: *https://doi.org/10.1155/2022/4916818*

Sharma, T., Jain, A., Verma, N. K. and Vasikarla, S. (2019). Object counting using kaze features under different lighting conditions for inventory management, pp. 1–7.

Siby, A., Emmanuel, A. P., Lawrence, C. and Jayan, J. M. (2020). Text to speech conversion for visually impaired people, *2020 International Journal of Innovative Science and Research Technology* pp. 1253–1256.

Sinha, A., Banerjee, S. and Chattopadhyay, P. (2022). An improved deep learning approach for product recognition on racks in retail stores.

Wu, W.-S. and Lu, Z.-M. (2022). A real-time cup-detection method based on yolov3 for inventory management, *Sensors* **22**(18).
URL: *https://www.mdpi.com/1424-8220/22/18/6956*

Xia, X. and Roppel, T. (2019). Enabling a mobile robot for autonomous rfid-based inventory by multilayer mapping and aco-enhanced path planning, *International journal of robotics and automation technology* **1**(1).

Yilmazer, R. and Birant, D. (2021). Shelf auditing based on image classification using semi-supervised deep learning to increase on-shelf availability in grocery stores, *Sensors* **21**(2).
URL: *https://www.mdpi.com/1424-8220/21/2/327*

Zhang, J., Lyu, Y., Roppel, T., Patton, J. and Senthilkumar, C. (2016). Mobile robot for retail inventory using rfid, pp. 101–106.