Real-time Scene Change Detection with Object Detection for Automated Stock Verification

A Project Report Submitted

in Partial Fulfilment of the Requirements

for the Degree of

B. Tech(Hons)

in

Computer Science and Engineering

by

SANDEEP KUMAR YEDLA

(Roll No. 2016BCS0031)

to

DEPARTMENT OF CSE

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY

KOTTAYAM - 686635, INDIA

June 2020

DECLARATION

I, **Sandeep Kumar Yedla** (**Roll No: 2016BCS0031**), hereby declare that, this report entitled **"Real-time Scene Change Detection with Object Detection for Automated Stock Verification"** submitted to Indian Institute of Information Technology Kottayam towards partial requirement of **Bachelor of Technology(Hon)** in **Computer Science Engineering** is an original work carried out by me under the supervision of **Dr. Panchami V** and has not formed the basis for the award of any degree or diploma, in this or any other institution or university. I have sincerely tried to uphold the academic ethics and honesty. Whenever an external information or statement or result is used then, that have been duly acknowledged and cited.

Kottayam-686635 **Sandeep Kumar Yedla**

June 2020

CERTIFICATE

This is to certify that the work contained in this project report entitled **'Real-time Scene Change Detection with Object Detection for Automated Stock Verification'** submitted by **Sandeep Kumar Yedla** (**Roll No: 2016BCS0031**) to Indian Institute of Information Technology Kottayam towards partial requirement of **Bachelor of Technology(Hon)** in **Computer Science Engineering** has been carried out by him under my supervision and that it has not been submitted elsewhere for the award of any degree.

Kottayam-686635
June 2020

(Dr. Panchami V)
Project Supervisor

ABSTRACT

Automation is a process of utilizing technology to reduce human efforts, a computer vision-based automated system for stock management in the supermarkets is proposed and designed for the system. The proposed system will help to reduce the manpower required in a supermarket by continuously monitoring the availability product in the supermarket and reporting the useful information to the concerned person automatically. The key idea behind the proposed scheme is that a few low-cost cameras will be placed in the supermarket which will help to capture the videos of the product racks in the supermarket. The presence of human beings are identified by using a structural similarity index (SSIM) based scene change detection technique, further, an object detection technique will be used to count the number of items present in the specific product rack. If the number of items present in a particular rack goes below a threshold limit, a short message service (SMS) and/or email will go the concerned authority. To make it more comfortable, a product identifier (printed) will be kept just below the product racks. An optical character recognition module in the proposed scheme will identify the product identifier and it will be mentioned in the SMS or email which will help the supervisor for scheduling the replacement of the items in the racks. The experimental study is carried out by placing sample items on a rack and the mobile camera is used as an IP camera with the help IP webcam android application for the monitoring purpose. The experimental study shows that the proposed scheme will work reliably in a supermarket environment.

Contents

List of Figures

Chapter 1

Introduction

Nowadays automation techniques are exploiting in its full potential to improve the productivity in manufacturing industries and service sectors. Computer vision techniques are utilized in many places to mimic the behaviour of human vision. The computer vision is a process extracting some useful information from images or videos. In general, the computer vision task involves two activities: object detection and object recognition. The object detection is a process identifying the regions in the image where some meaningful objects are present. The object recognition is the task of classifying the detected objects into certain classes.

In this project, a computer vision approach for monitoring the status of the products available in the supermarket. For this purpose, a few cameras need to be placed to capture the real-time videos of the product racks. From the real-time video, whenever the available number of products are going below a threshold limit, it will be identified and that information will be shared to the concerned person.

The novelty of the proposed scheme is that a structural similarity (SSIM) based scene change detection technique and it is used to optimize the number of frames required to process of automatic stock verification. The well-known object detection and recognition scheme called YOLO (You Only Look Once) is used in the scheme for identifying the products in the product racks.

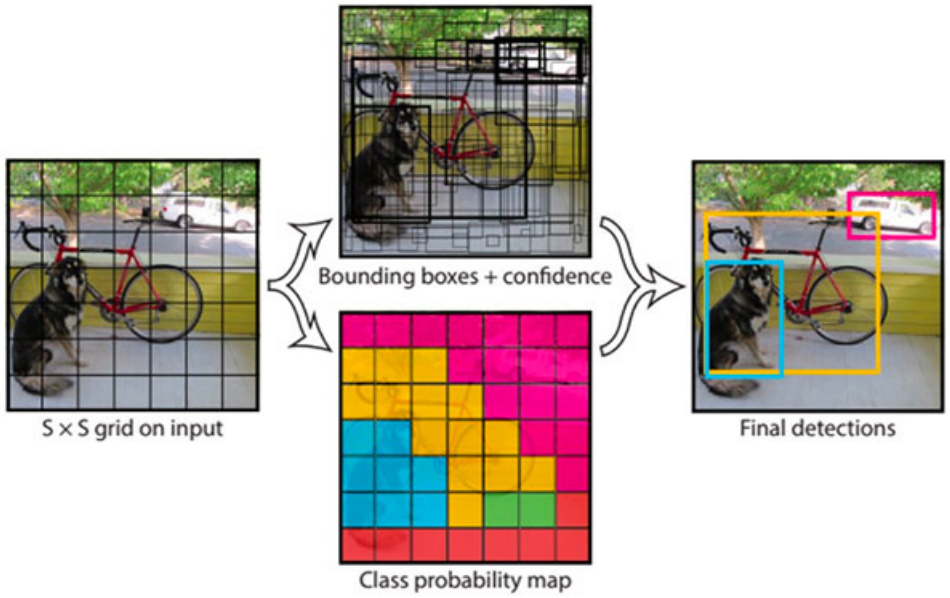

Figure 1.1: Detection approach

The major concepts explored are scene change detection, object detection, and optical character recognition and few communication modules are explored which might be an paid one, by private sectors but a trial version can be accessed and worked upon this project also an traditional mail approach is also explored based on the data from code or integers from the code are converted into string and attached to warning mail for using it to alerting purpose.

Chapter 2

Literature Review

Literature Survey that done for the implementation of this system is upon the image capturing and storing in the local storage(continuously replacing to save the storage place), two images conversion to matrices and gray scaling, comparing for frame difference,required object detection using the YOLO and its trained models and basic object detection in making boxes using. -ordinates formation of boundary.

- 'Fast R-CNN' 2015. [2]
  - This paper proposes a Fast Region-based Convolutional Network method (Fast R-CNN) for object detection. Fast R-CNN builds on previous work to efficiently classify object proposals using deep convolutional networks. Compared to previous work, Fast R-CNN employs several innovations to improve training and testing speed while also increasing detection accuracy. In Fast R-CNN, we feed the input image to the CNN, which in turn generates the convo-

lutional feature maps. Using these maps, the regions of proposals are extracted. We then use a RoI pooling layer to reshape all the proposed regions into a fixed size, so that it can be fed into a fully connected network.

    * As with the earlier two techniques, we take an image as an input.
    * This image is passed to a ConvNet which in turns generates the Regions of Interest.
    * A RoI pooling layer is applied on all of these regions to reshape them as per the input of the ConvNet. Then, each region is passed on to a fully connected network.
    * A softmax layer is used on top of the fully connected network to output classes. Along with the softmax layer, a linear regression layer is also used parallely to output bounding box coordinates for predicted classes.

– Problems with Fast R-CNN are that even Fast R-CNN has certain problem areas. It also uses selective search as a proposal method to find the Regions of Interest, which is a slow and time consuming process. It takes around 2 seconds per image to detect objects, which is much better compared to R-CNN. But when we consider large real-life datasets, then even a Fast R-CNN doesn't look so fast anymore.

– Conclusion:Compared to SPPnet, Fast R-CNN trains VGG16 3x faster, tests 10x faster, and is more accurate.

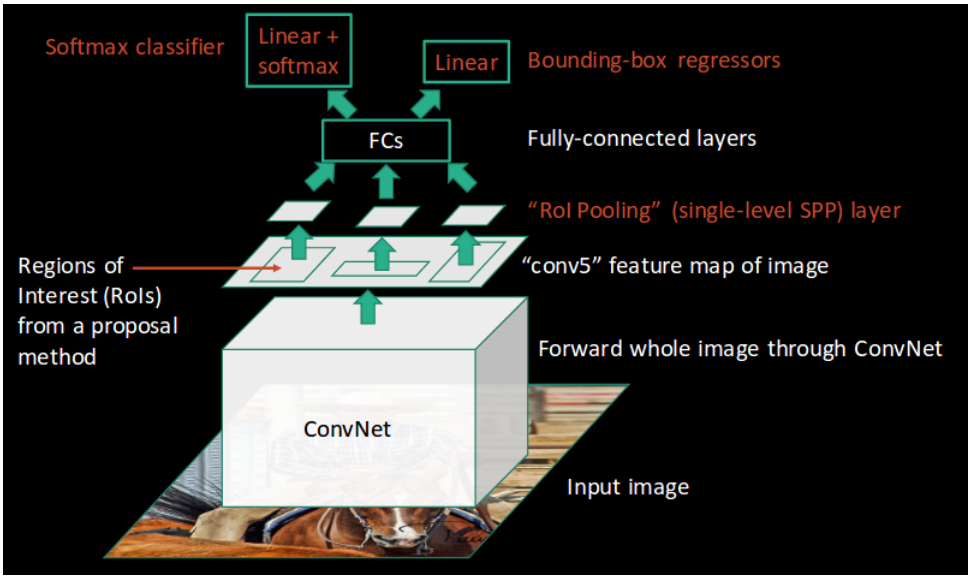

Figure 2.1: Working of Fast R-CNN

- 'Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks' 2015.[6]

  - State-of-the-art object detection networks depend on region proposal algorithms to hypothesize object locations. In this work, we introduce a Region Proposal Network (RPN) that shares full-image convolutional features with the detection network, thus enabling nearly cost-free region proposals.The algorithm requires many passes through a single image to extract all the objects. As there are different systems working one after the other, the performance of the systems further ahead depends on how the previous systems performed.

  - Problems with Faster R-CNN are all of the object detection algorithms we have discussed so far use regions to identify the objects.

The network does not look at the complete image in one go, but focuses on parts of the image sequentially. This creates two complications:

  - Conclusion:RPN and Fast R-CNN can be trained to share convolutional features

- 'You Only Look Once:Unified, Real-Time Object Detection', 2017 [5]

  - YOLO : you only look once is one of the good advancement for fast processing speed compared to other detection method,a fast yolo can process 155 frames per second, forming bounding boxes based on confidence and support calculation.YOLO makes more localization errors but is less likely to predict false positives on background. Since the whole detection pipeline is a single network, it can be optimized end-to-end directly on detection performance.
  - problems include, as It might leave behind the less probability and confidence images without boxing.
  - Conclusion: You only look once is a good method for detection as the supervisory system needs speed and better accuracy.

- 'Motion Detection Based on Frame Difference Method' 2014. [8]

  - Recent research in computer vision has increasingly focused on building systems for observing humans and understanding their look, activities, and behavior providing advanced interfaces for interacting with humans, and creating sensible models of humans for various purposes. The goal of motion detection is to recognize

motion of objects found in the two given images. Moreover, finding objects motion can contribute to objects recognition.

  - Conclusion: The obvious aim of the work is studying the principle of frame difference method and comparing different images and check for any motion in those images and to resolve the various problems.

- 'Weakly Supervised Object Localization on grocery shelves using simple FCN and Synthetic Dataset' 2018. [10]

  - It is a weakly supervised method using two algorithms to predict object bounding boxes given only an image classification dataset. First algorithm is a simple Fully Convolutional Network (FCN) trained to classify object instances. Here they enhance the FCN output mask into final output bounding boxes by a Convolutional Encoder-Decoder (ConvAE) viz. the second algorithm. ConvAE is trained to localize objects on an artificially generated dataset of output segmentation masks.

  - Conclusion:Pinhole Projection Formula can be used for distance estimation using a single camera input.

- 'EdgeFlow: a technique for boundary detection and image segmentation' 2011.[4]

  - The study is regarding the smart surveillance's using an edge flow algorithm and processing the images using a pin hole algorithm for better improved processing.

- Conclusion: Pinhole Projection Formula can be used for distance estimation using a single camera input.

- 'R-FCN: Object Detection via Region-based Fully Convolutional Networks' 2016.[1]
  - This is region-based, fully convolutional networks for accurate and efficient object detection. In contrast to previous region-based detectors such as Fast/Faster R-CNN that apply a costly per-region subnetwork hundreds of times, our region-based detector is fully convolutional with almost all computation shared on the entire image.
  - Conclusion: This is almost 2.5-20 times faster than the Faster R-CNN counterpart.

- 'A Case Study on smart Surveillance Application System using WSN and IP webcam' 2011. [7]
  - The study is regarding the smart surveillance's using a android application called IP webcam which is used for surveillance with many features such as pixel adjustment and running in background and many more. This is used as demonstration instead of real camera.
  - problems, As this application can be used in cell phones for surveillance this can make phone heat up and extra power usage with battery draining.

- Conclusion: Pinhole Projection Formula can be used for distance estimation using a single camera input.

- 'An Automated Computer Vision System for Extraction of Retail Food Product Metadata' 2018.[3]
  - Our study proposes an automation method to improve the extraction of unstructured product metadata from food product label images using computer vision (CV), machine learning (ML), optical character recognition (OCR), and natural language processing (NLP).Proposed an automatic image quality classification system to identify images that give a high degree of metadata extraction accuracy
  - Conclusion: Results show 95 % accuracy for attribute extraction from high-quality product images with machine-printed characters having contrasting backgrounds.

- 'An Overview of the Tesseract OCR Engine' May 2011. [9]
  - The Tesseract OCR engine, as was the HP Research Prototype in the UNLV Fourth Annual Test of OCR Accuracy.
  - The problem with OCR is there are not scalabe and not 100 percent accurate, they depend and rely on many factors such as font and distortion in the image.
  - Conclusion: Tesseract is now behind the leading commercial engines in terms of its accuracy. Its key strength is probably its unusual choice of features.

Chapter 3

Proposed Scheme

3.1 Block-Diagram

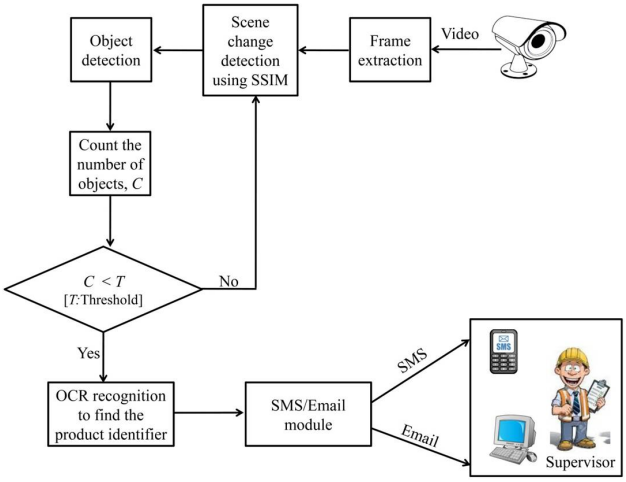

Figure 3.1: Proposed Scheme.

The overview of the proposed scheme is shown in above figure 3.1. In the proposed scheme, we need to place the video cameras to monitor product rakes in a store for surveillance. The captured video frames need to be

processed in a pre-defined way. The sequence of activities in a step by step procedure is described below as a algorithm and the working of the system.

Algorithm I :Computer-vision based supermarket management

1. Capture the real-time video, V, of the product rack. The video can be considered as sequence of N frames, $V = (f_1, f_2, f_3, ..., f_N)$
2. $R_f = f_1$ `//The first frame will be considered as the reference frame for scene change detection`
3. $p = 30 \times F_R$ F_R `is the frame rate of the camera; we are Considering frames in an interval of 30 seconds`
4. While $(f_p \neq NULL)$ `//Need to process all the frames from the given video stream`
5. Find the structural similarity index (SSIM) between f_p and R_f and keep it in F_D. The way to compute SSIM is given Equation (1).
6. If $F_D < T_1$ `//`T_1 `is a threshold value to determine a frame change (presence of human). We empirically identified,` $T_1 = 0.60$
7. Apply object detection and recognition on f_p and identify the number of objects, C, present in the image.
8. if $C < T_2$ `//`T_2 `is a threshold value and the concerned person should get a notification if the number of product available in the rack goes below` T_2
9. Apply optical character recognition to know the unique product identifier, P_{id}, that will be present in the bottom of the product rack.
10. Send an email and/or SMS to the responsible person to inform him that the product with product identifier, P_{id}, does not have enough stock.
11. $R_f = f_p$
12. $p = 30 \times F_R$ `Considering frames in an interval of 30 seconds`
13. EndWhile

3.2 Modules

The proposed computer vision based system will capture the video frames throughout the day. The whole proposed framework consists of the following modules:

- Frame extraction.
- Scene change identification.
- object detection.
- Optical character recognition (OCR).
- Communication module.

3.2.1 Frame Extraction

The real-time video, V , captured by the surveillance camera can be considered as a sequence of N frames and it can be denoted as V = (f 1 , f 2 , ..., f N). Each frame in the video can be treated as a colour image in RGB format, where RGB represents red, green and blue colour components. The frames in a frequent interval will be transferred to the next phase.

3.2.2 Scene Change Detection

The next phase in the proposed scheme is scene change detection. The key idea is that the further steps like object detection, counting the products in the rack, OCR and au-tomated emailing or SMS sending process can be

carried out only after scene change detection. Since the camera is working all the time and we are considering the frames in a frequent interval (every frame in an interval of 30 seconds), whenever the product is taking out by costumers then a scene change will be detected. No need to do all the phases in every frame. The scene change detection helps to optimize the overhead of object detection and recognition tasks. Since we have used SSIM based scene change detection technique, it can be carried out in real-time. Absolute frame difference is one of the most common approach to detect the scene change or frame change in a video sequence. The absolute frame difference between two images I and G having size of R C pixels can be computed as follows:

$$A_{FD} = \sum_{x=1}^{x=R} \sum_{y=1}^{y=C} \left| I_{x,y} - G_{x,y} \right|$$

Figure 3.2: Traditional Pixel difference.

Based on the situations, we need to fix a threshold value for A F D and if it goes below the threshold value, we can identify it as a scene change. The absolute frame difference approach is tried in the proposed scheme to detect the scene changes, but it was not giving a reliable result due to the lighting changes. This motivated us to use a new scheme for scene change detection which uses SSIM measure. For identifying the scene change, we have used the SSIM measure between the reference frame and the current frame. The SSIM is an image quality assessment technique to find the structural similarity between two images. We used the same measure to identify the frame changes in the proposed scheme. The structural similarity index (SSIM) is

a perception-based model that considers image degradation as a perceived change in the structural information. The SSIM between two images I and G can be computed as follows: The SSIM value may vary between 0 to 1. If the

$$SSIM = \frac{(2\mu_I\mu_G + C_1)(2\sigma_I\sigma_G + C_2)}{(\mu_I^2 + \mu_G^2 + C_1)(\sigma_I^2 + \sigma_G^2 + C_2)}$$

Figure 3.3: Theoretical SSIM Formulae

SSIM value between two images are 1, then it indicates that both the images are almost the same. The low SSIM values between two images declare that both the images do not have the same structural properties. In the proposed scheme, we are finding the SSIM value between reference frame, R f , and the current frame, f p . If the SSIM value between R f and f p is less than 0.6 then we are considering that in f p there is a scene change has happened. The scene change may happen due to various reasons: a person may walk between the product rack and the camera or a person is taking out a product from the rack. In all such cases, a scene change will be identified by the proposed scheme. The overview of the SSIM based scene change detection is shown as follows.

3.2.3 Object Detection

The object detection is carried out using YOLO. The YOLO is a real-time object detection scheme which helps to identify the objects in a frame very fastly. A fast YOLO can process 155 frames per second and the result will be an image which contains bounding boxes around the objects.

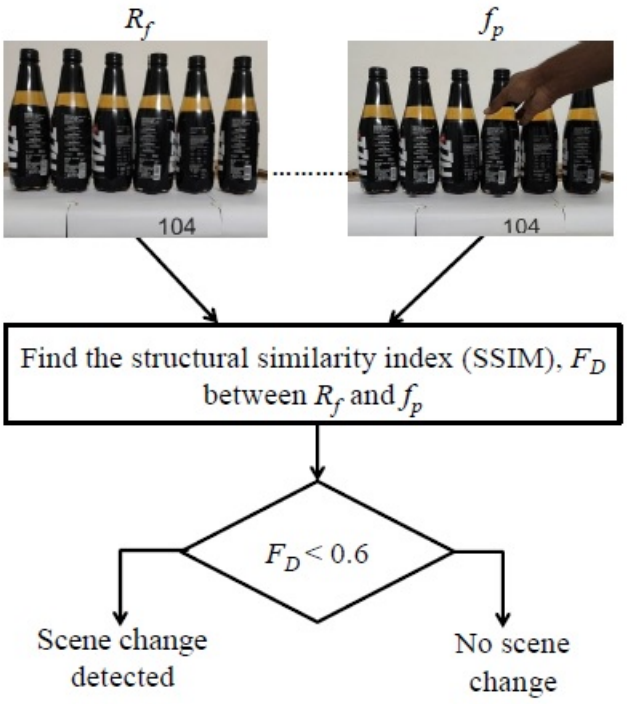

Figure 3.4: Theoretical SSIM Formulae

3.2.4 Optical Character Recognition

An OCR scheme is used in the proposed scheme to identify the product name or unique product identifier mention on the product rack. We assumed that the product identifier, probably based on characters such as 101,102 ... to uniquely identify an item and that will be kept below the product in the product rack. Region based or specific restricted range OCR is performed such that only the required identifier is captured.

3.2.5 Communication module

In this phase, the number of products in the selected rack will be counted and if it goes beyond the threshold limit an SMS or an email will be triggered to inform the supervisor. The product identifier or product name also will be included in the SMS or email that will help the supervisor to plan the refilling process.

3.2.6 Libraries

There are various libraries used in the project and worked to find out the optimal one. Various libraries used are

- numpy : NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array object
- urllib : Urllib module is the URL handling module for python. It is used to fetch URLs (Uniform Resource Locators)
- pytesseract : Python-tesseract is an optical character recognition (OCR) tool for python. That is, it will recognize and "read" the text embedded in images.
- sinchsms : sinchsms - a module to send sms using the Sinch REST apis
- skiimage : Used for importing SSIM index.
- matplotlib : Used for plotting the graphs
- smtplib : Used for Sending mails.

Chapter 4

Experimentation and Analysis

4.1 Input Images

Figure 4.1: Experimental Setup

Figure 4.2: Image Interference.

4.2 Experimental Results

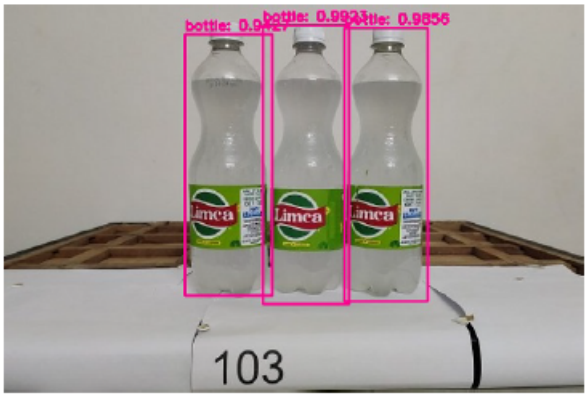

Figure 4.3: Detected Objects

The table in fig 4.4 below shows the results of the four products used and the accuracy of the algorithm which was recursively applied to test on the four products, as in the proposed scheme, the frames are captured and tested with SSIM support for real-time scene change detection where, when there are six to four items in the rack the the everything remains abridged and Processing layer is separated from other, but as per the constraint if less than or equal to 2 alert message triggered.

With processing the image and extracting the region of interest characters mentioned and the output of pytesseract ocr is converted into string attached to the warning mail and sent to the respective authority.

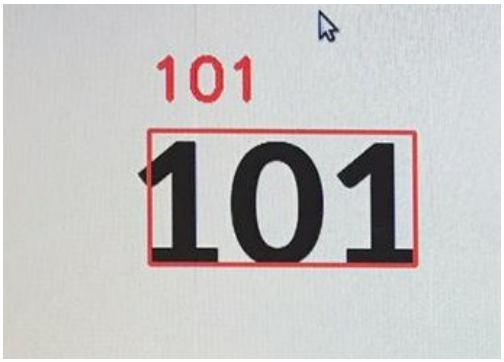

Figure 4.4: 101 by OCR

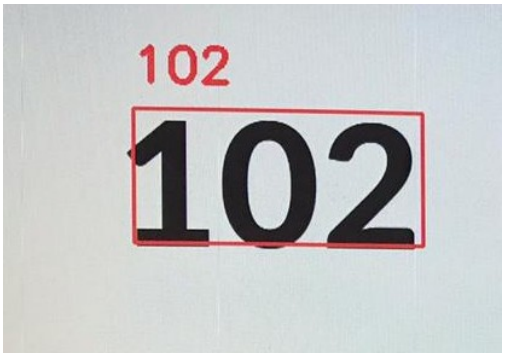

Figure 4.5: 102 by OCR

TABLE I
STATUS OF TRIGGERING THE MESSAGE BASED ON PRODUCT COUNT

Product	**Status of triggering warning message**						
Number of products ->	**6**	**5**	**4**	**3**	**2**	**1**	**0**
Product 1 (101)	No	No	No	Yes	Yes	Yes	Yes
Product 2 (102)	No	No	No	Yes	Yes	Yes	Yes
Product 3 (103)	No	No	No	Yes	Yes	Yes	Yes
Product 4 (104)	No	No	No	Yes	Yes	Yes	Yes

Figure 4.6: Experiment Result table.

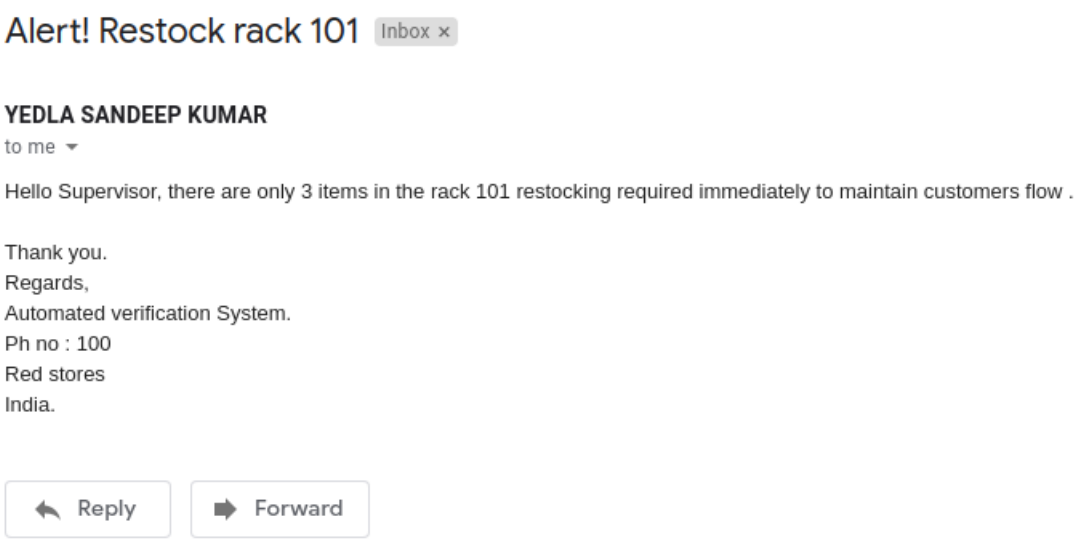

Figure 4.7: Alert mail

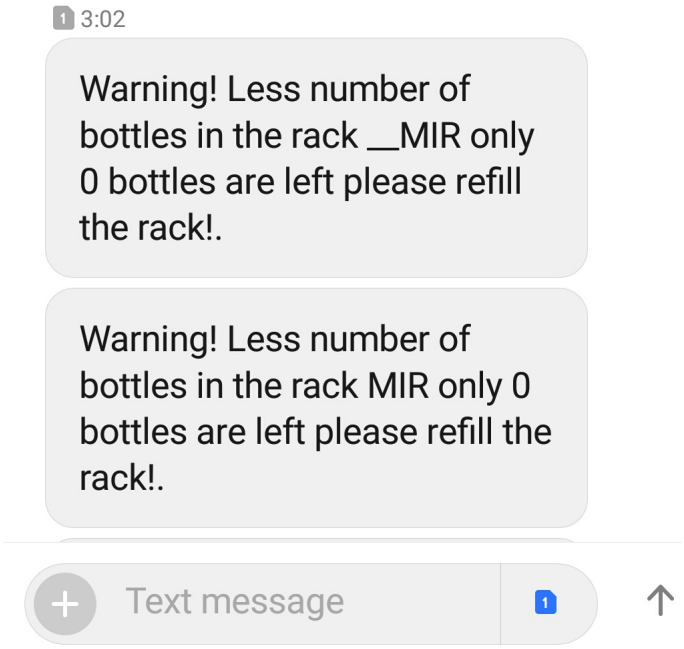

Figure 4.8: phase 1, mobile ping

4.2.1 CPU Performed Graph

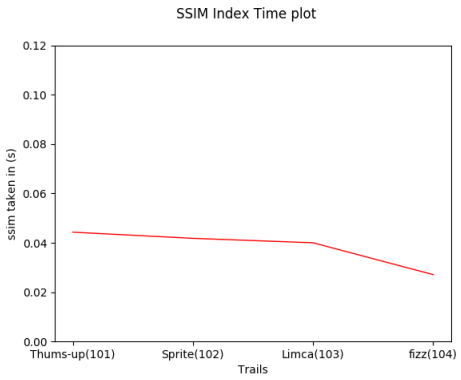

Figure 4.9: Graph 1

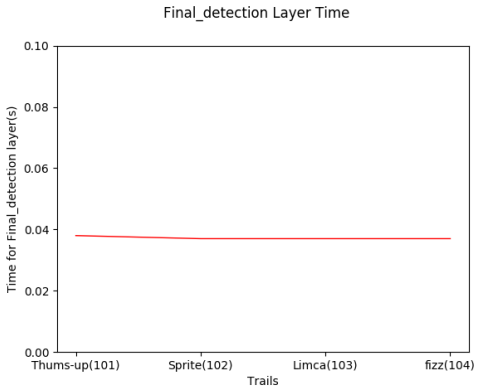

Figure 4.10: Graph 2

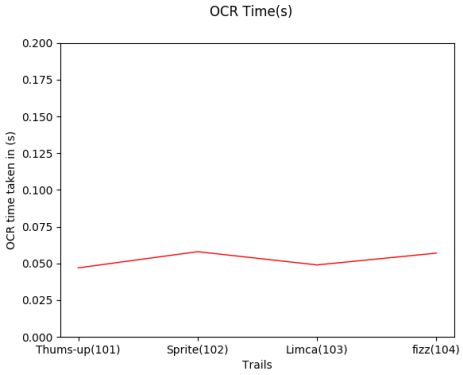

Figure 4.11: Graph 3

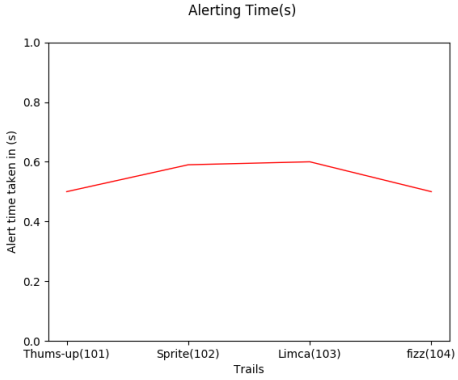

Figure 4.12: Graph 4

- The Four above graph figures(4.9,4.10,4.11,4.12) mentioned indicate the CPU performance(Time analysis) in individual modules like calculating similarity index, Object detection, OCR output and mail triggering in milliseconds and, compared among the four product trials. The Object detection layer in fig 4.10 which is almost linear around 0.039 seconds which is remarkable performance for real-time analysis. The rest of the graphs like fig 4.9, 4.11, 4.12 take around 0.06 seconds and below overall to perform. Thus the system running on CPU is reliable and is cost efficient, though we have compared with gpu performance which is provided below.

4.2.2 GPU Performed Graph

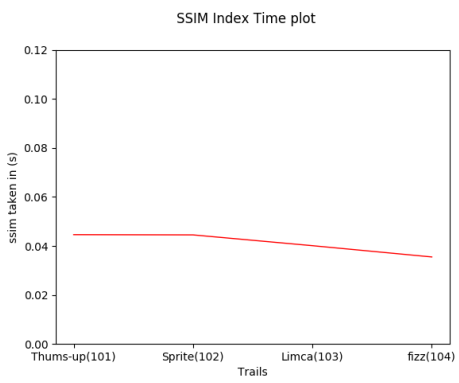

Figure 4.13: Graph 5

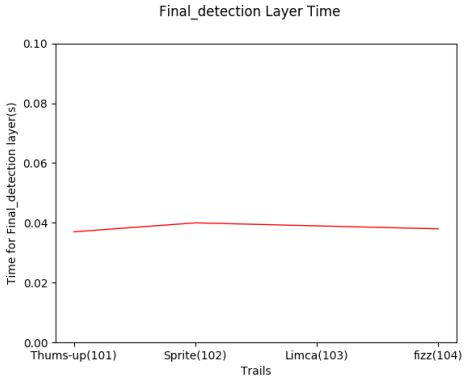

Figure 4.14: Graph 6

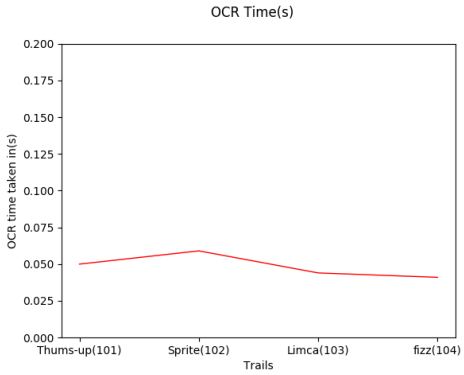

Figure 4.15: Graph 7

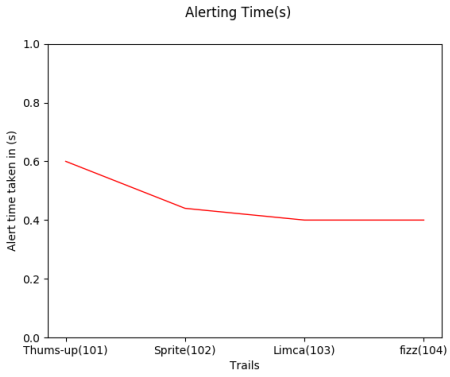

Figure 4.16: Graph 8

The above graphs also compare the same metrics but performed on an GPU(google colab), but though there are minute variations in the fig 4.13, 4.15, 4.16 which describe the ssim, ocr, and mail alert time plots, it doesn't matter as they are in mili seconds(ms) difference, almost comparable and equal performance when an cpu is used for the experimentation results.The yolo detection is also showing the linear and stable performance in detection. So these minute variation doesn't much effect the total overall performance with great variation.

4.2.3 CPU Vs GPU Performance

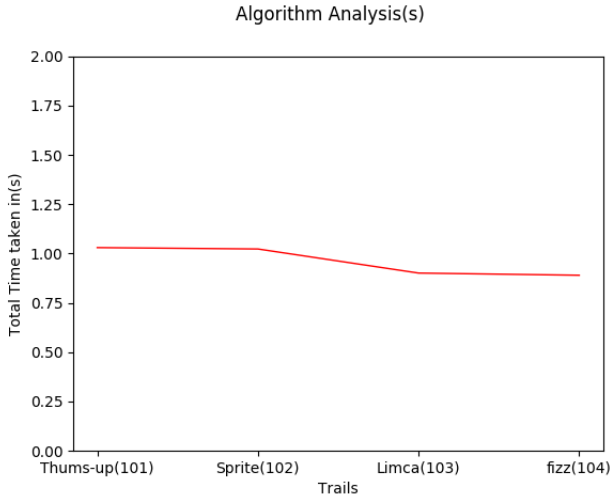

Figure 4.17: Processing time on CPU)

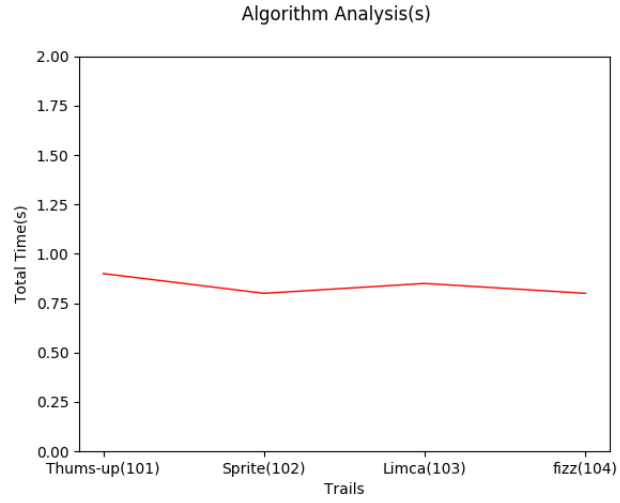

Figure 4.18: Processing time on GPU

The Graphs clearly describe that the system is performing at almost one second and less than one second while using an CPU and also almost same while using an GPU where its slightly less than one around 0.8 to 0.9 seconds. Hence through this data we can conclude the proposed system is scalable and reliable even while using an CPU.

Chapter 5

Conclusion

A computer vision-based approach for automated monitoring of the products in the supermarket is designed and introduced. The proposed framework can be adopted in busy supermarkets. The implementation of the proposed scheme in supermarkets will help the managing companies to reduce the required manpower, increase the profit and provides better customer satisfaction. In the proposed scheme, we used a new structural similarity index based scene change detection algorithm to avoid the processing of all the video frames, which supports the real-time optimisation. In this work, we have used one camera to keep track of the products in one single rack, but in a real case, a single camera may be used to cover a large area which contains racks for different products. This designed system can be developed to advanced and used to capture multiple item with unique identification.

Bibliography

[1] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. *Advances in neural information processing systems*, 2016.

[2] Ross Girshick. Fast r-cnn. *Proceedings of the IEEE international conference on computer vision*, 2015.

[3] Venugopal Gundimeda, Ratan S Murali, Rajkumar Joseph, and NT Naresh Babu. An automated computer vision system for extraction of retail food product metadata. *First International Conference on Artificial Intelligence and Cognitive Computing*, 2019.

[4] Wei-Ying Ma and Bangalore S Manjunath. Edgeflow: a technique for boundary detection and image segmentation. *IEEE transactions on image processing*, 2000.

[5] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *IEEE*, 2016.

[6] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 2015.

[7] Sayantani Saha and Sarmistha Neogy. A case study on smart surveillance application system using wsn and ip webcam. In *2014 Applications and Innovations in Mobile Computing (AIMoC)*, pages 36–41. IEEE, 2014.

[8] Nishu Singla. Motion detection based on frame difference method. *International Journal of Information & Computation Technology*, 2014.

[9] Ray Smith. An overview of the tesseract ocr engine. *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007).*

[10] Srikrishna Srivastava Varadarajan, Muktabh Mayank. Weakly supervised object localization on grocery shelves using simple fcn and synthetic dataset. *arXiv preprint arXiv:1803.06813*, 2018.