Project Synopsis Report

Navlika Singh (B20AI025)
Mitul Agrawal (B20AI021)
Vishnu Kumar (B19BB066)
Kshitij Singh (B19ME039)

Sub Project 1:

1) Title: **Cycle Repetition**

2) Abstract:

Cyclic processes ranging from industrial procedures, event patterns or day to day activities are of interest because of the variety of insights one can extract out of them. It may be that there is an underlying cause behind something that happens multiple times, or gradual changes in the scenario, or unambiguous action units, semantically meaningful segments that make up an action.

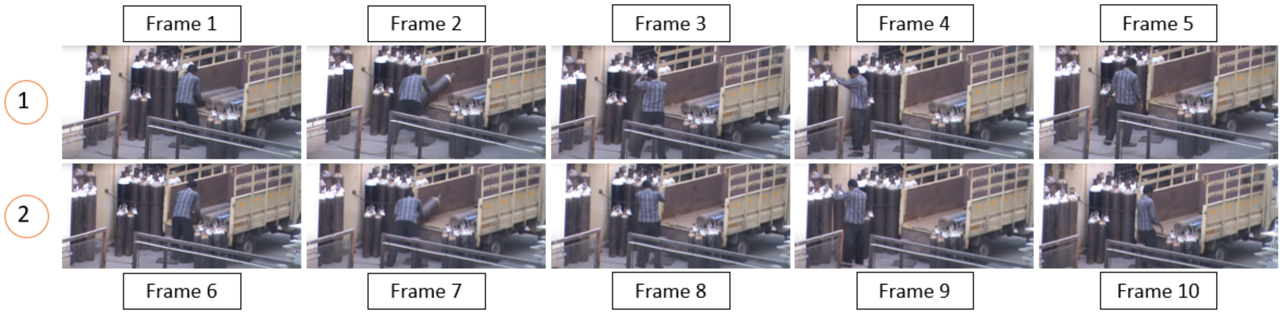
As a consequence the project focuses on classifying and estimating the period of repeating activities. The dataset gathered, consists of labelled videos with content focusing on varying repetitive activities. The team proposes two models to approach the problem, first based on centroid approach and the other on RepNet model developed by Google. We further discuss the innovations that can be implemented on the basic solution and the challenges we may face while implementing the proposed plan. The deliverables of the project are discussed towards the end of the report.

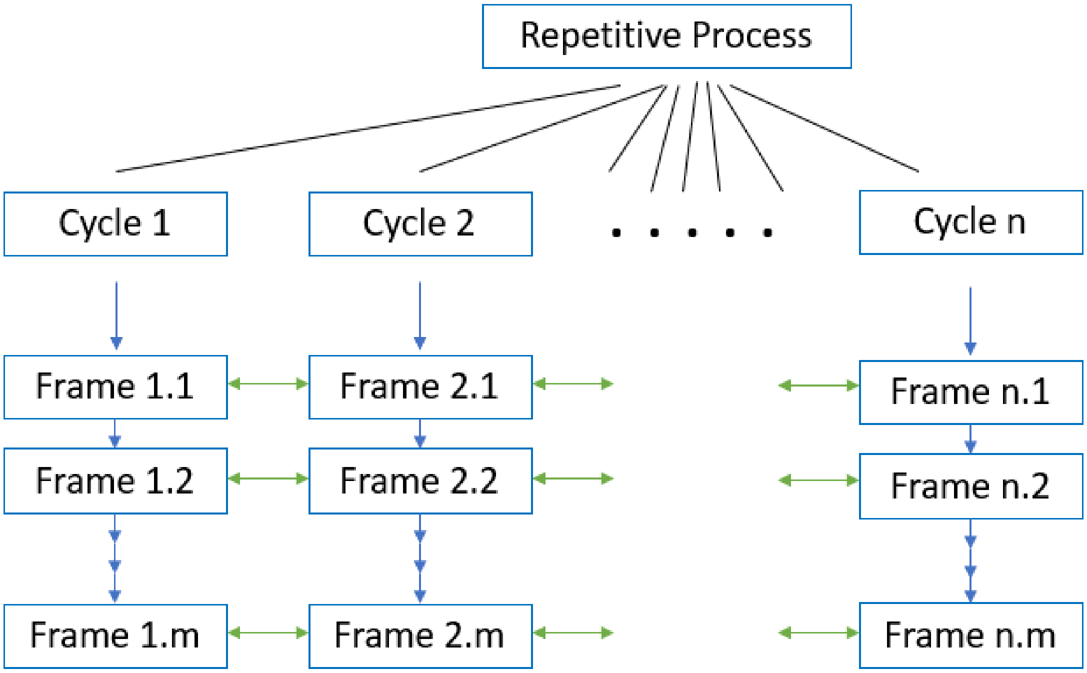
3) Project details:

   a) Description and Motivation:

The project aims to make a computer vision based model to count the number of times a repetitive activity was performed and to estimate the time period of a single event cycle.

We want to make a generalized model to count continuous repetitive tasks in the manufacturing line or any other industrial operational process.





The First Industrial Revolution began with the use of steam power. Industry 2.0 began with the use of electricity. Use of Computer Chips and Automated Systems marked the beginning of Industry 3.0.
Industry 4.0 is the era of smart machines and production facilities which can trigger actions without human intervention.

Till now, this process of counting has been done manually which takes human resources. By automating this, time is saved, the process becomes more organized and it is a step towards Industry 4.0.

b) Literature Review for related work:

Cyclic or periodic processes are an inevitable part of our lives. May it be industrial, involving manufacturing merchandise or processing it for shipping, or daily life activities like swinging on a swing or exercising, or may it be more universal like earth rotating on its axis or revolving around the sun. Hence, it is only natural to study such processes and try to draw meaningful insights from them.

‘Detection and Recognition of Periodic, Non-rigid Motion’ paper published by Ramprasad Polana and Randal C. Nelson in June, 1997. The highlights of the paper are discussed as follows:
Centroid Method -

- Centroid of frame : $(xt, yt) = \text{Summation}[(i, j)/N]$ where i,j are pixels in a frame.
- $(xt, yt) = (x0, y0)+(u,v)\ast t$, where (u,v) is the local velocity of the object.
- Just centroid wont work when multiple moving objects are present. Solution : Restrict centroid computation to the area that is most likely to have the object of interest.
- From the position estimates of the past K flow frames, we can get an estimate of the velocity of the object.
- $p(t + 1) = (xt + u, yt + v)$
- $S'(t + 1) = \{(x + u, y + v) : (x, y) \in S(t)\}$. [$S(t)$: Set of pixels to consider for centroid computation].
- $(x(t+1), y(t+1)) = w\ast p(t + 1) + (1-w)\ast c(t + 1)$ [w is between 0 & 1].

Google developed a model for ‘Video Understanding Using Temporal Cycle-Consistency Learning’ (article posted on August 8, 2019). This was focused on applying machine learning to understand each frame of a video, that is assessing the interdependency of frames. They proposed a potential solution using a self-supervised learning method called Temporal Cycle Consistency Learning (TCC). This uses correspondences between examples of similar sequential processes to learn representations for fine-grained temporal understanding of video.

Google further developed on this idea, and released an improved model called RepNet, focusing on Counting Repetitions in Videos (article posted on June 22, 2020). The architecture of the model consists of three parts: a frame encoder, an intermediate representation, called a temporal self-similarity matrix, and a period predictor. The frame encoder uses the ResNet architecture as a per-frame model to generate embeddings of each frame of the video, after which the TSM returns a matrix for subsequent modules to analyse for counting repetitions.

c) Technical plan:

i) Basic Solution Hypothesis:

Dataset
The dataset will consist of videos. The content of the video will consist of varying periodic activities. The video must also be labelled for classification purposes. To make the dataset robust the videos captures may follow the following criteria:

- Videos in a variety of lightning conditions, camera angles and of objects in different orientations.
- The proportion of the target item should be higher in the data set.

If the dataset is not labelled for classification purposes, label it using one of the many open source tools available for labelling images.

Data Augmentation
Performing data-augmentation techniques to increase the size of the data set. The techniques may involve:
Flipping horizontally
Zooming
Blurring

Transfer Learning
Model
The model deployed is RepNet. RepNet is developed by Google, for counting repetitions in Videos. The architecture of the model consists of three parts: a frame encoder, an intermediate representation, called a temporal self-similarity matrix, and a period predictor.

Frame Encoder: It uses the ResNet architecture as a per-frame model to generate embeddings of each frame of the video.

Temporal self-similarity matrix (TSM): This matrix is calculated by comparing the frame’s embedding with every other frame in the video, returning a matrix that is easy for subsequent modules to analyze for counting repetitions.

Period Predictor: For each frame, the Transformers are used to predict the periodicity, that is whether or not a frame is part of the periodic process and the respective period of repetition, directly from the sequence of similarities in the TSM. Then it obtains the per frame count by dividing the number of frames captured in a periodic segment by the period length. Summing this gives the number of repetitions in the video.

ii) Proposed Innovations over the Basic Solution:

Some of the proposed innovations on the basic model are as follows:

- Remove motionless background (set rgb=0,0,0)
- Think of rgb values as respective masses and find the centre of mass coordinates of r,g,b. Also the average mass can be found out for r,g,b. And this centre of mass and average mass should approximately perform repetition too.
- Further Extending the Idea : Centre,Average gives (2x3)+(3) = 9 values. Now the difference of these values between previous and next frame can be taken (local velocity) giving us 27 values. More such values can be taken.

iii) Benefit to the user agency:

There are multiple ways in which the user agency can benefit from this:

- Repeating processes are of interest to researchers for the variety of insights that can be obtained from them. It may be that there is an underlying cause behind something that happens multiple times, or there may be gradual changes in a scene that may be useful for understanding.
- Analyzing these repeating processes may provide us with unambiguous but semantically meaningful segments that make up an action.
- These units may be indicative of more complex activity and may allow us to analyze more such actions automatically at a finer time-scale without having a person annotate these units.
- Perceptual systems that aim to observe and understand our world for an extended period of time will benefit from a system that understands general repetitions.
- In the field of heavy machinery maintenance, it can be used to count the number of cycles that a particular machine part goes through and analyze thus if it is working perfectly or not.
- In manufacturing lines, it can be used to count the number of times all the substeps of a large manufacturing assembly occur and then use it to find if there is any wastage of parts/resources and whether any process can be optimized by

comparing multiple different repeating approaches to solving the same problem and finding which one leads to more repetitions in the given time frame.

d) Experimental Plan:

i) Examples and Use Cases:

Counting reparations of one’s exercise set.
Counting the No. of unloading or loading items in any factory.
Counting biological repetitive processes say heart beats.

ii) Benchmarking the Technical Plan with Industry Use Cases:

We will have to achieve a high accuracy.
We will need to observe if there is a change in speed of any repetitions. If some counts are longer or shorter.
We will need to also think about longer reparations, say if something takes more than a day or for very shorter reparations such as something happening in a second or maybe a fraction of seconds.

iii) Anticipated Challenges:

Different activities take different time to repeat themselves. And in some activities the time for one repetition may vary from another repetition. Different activities also makes the problem tougher as if in a person doing any exercise video vs a bouncing ball video. We will also have to maximize our accuracy with respect to real world deployment.

e) Deliverables - Milestones with Timeline:

Oct 1, 2021 - Oct 31, 2021 -

Trying out many different approaches, data preprocessing and initial ML Model development.

Nov 1, 2021 - Dec 15, 2021 -

Evaluating the different approaches and narrowing down on the best one, giving our inputs on the hardwares required for the problem (like computational power required, camera).

f) References:

- https://openaccess.thecvf.com/content_CVPR_2020/papers/Dwibedi_Counting_Out_Time_Class_Agnostic_Video_Repetition_Counting_in_the_CVPR_2020_paper.pdf

- https://www.youtube.com/watch?v=qSArFEIoSbo
- https://link.springer.com/article/10.1007/s11263-019-01194-0
- https://link.springer.com/article/10.1023/A:1007975200487
- https://ai.googleblog.com/2020/06/repnet-counting-repetitions-in-videos.html
- https://ai.googleblog.com/2019/08/video-understanding-using-temporal.html

Sub Project 2:

1) Title: **Inventory Tracking**

2) Abstract:

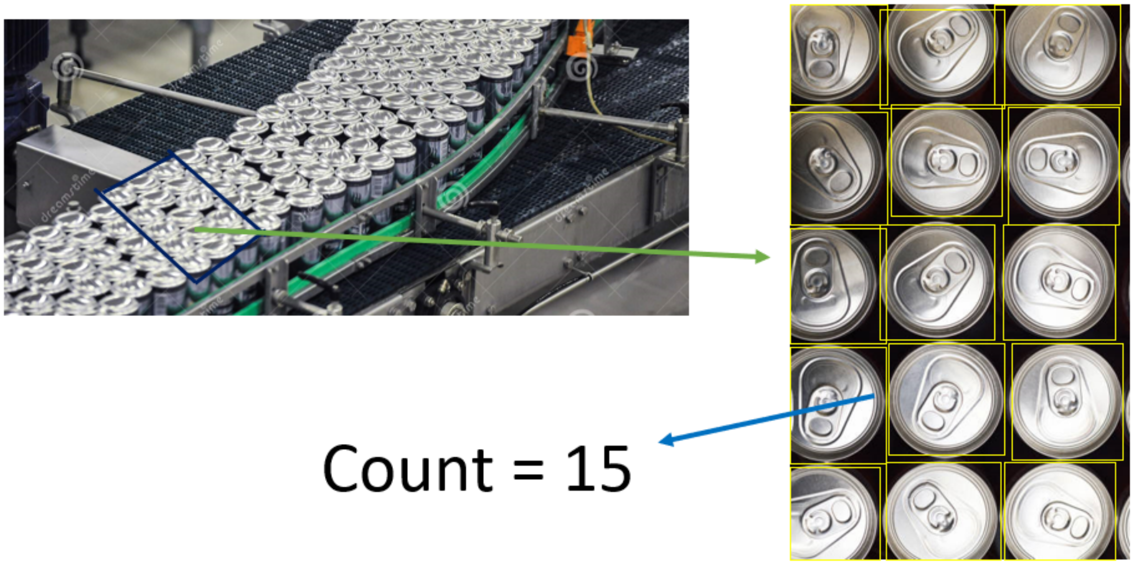
Most companies have to maintain inventory for various incoming, outgoing and internal flow of items such as raw materials to the plant, final product supply to customers, and intermediate product transfer between units, shops and warehouses. Majorly the tracking is done manually on paper which makes the process chaotic and difficult to devise meaningful insights from.

This project focuses on counting and keeping a track of orders, both, when it comes in the warehouse and when it leaves the warehouse. Also, the recorded data assists in predicting and sharing insights for optimal amounts of inventory buffers. The dataset gathered, consists of labelled videos with content focusing on various stocked workshops and warehouses. The team proposes two models to approach the problem, first based on yolo v-4 object detection and the other on ResNet object detection. After performing a comparative analysis, we would move forward with the model giving best results. We further discuss the improvements we can incorporate and the challenges we may face while implementing the proposed plan. The deliverables of the project are discussed towards the end of the report.

3) Project details:

   a) Description and Motivation:

We aim to make a model to count the number of items in an image. The system will keep a track and record of all the items moving in and out of the inventory to help automate the process and reduce mismanagement.



All types and sizes of firms need to maintain the inventory of various incoming,outgoing and internal flow of items. Most of this tracking is being done manually which makes the process chaotic and difficult to keep track of. Automating this process makes it more organized, saves human resources and is a step towards Industry 4.0.

b) Literature Review for related work:

Research on Inventory tracking problem till date can be roughly divided into three categories: counting by clustering, counting by regression, and counting by detection.

[1] describes a method to classify and sort objects based on their color and demonstrates the result by implementing it on the Phantom X Reactor robotic arm. [2] proposes a method to classify and sort objects according to their color and size (in pixels) from a camera mounted on top of the conveyor belt. The Method doesn't consider the shape of the object hence may confuse two objects with different shapes but the same sizes in pixels. [3] presents a method to classify cuboidal objects based on their dimensions and volume using a set of two cameras - one placed vertically and the other horizontally. It achieves an average accuracy of 87.5%. [4] presents an approach involving extracting Haar-like features and rectangular features from integral image representation and then training a classifier using SAMME (Stage-wise Additive Modeling using a Multiclass Exponential loss). The proposed technique has 30 frames per second speed with accuracy between 90% to 94%. [5] describes an object

classification method based on improved bag of features, which use SURF (Speeded Up Robust Features) and MSER (Maximally Stable External Regions) for local feature extraction capable of producing good results even on small datasets. [6] presents a method for object classification, localization, and segmentation using an RGB-D image from the Microsoft Kinect camera. The classification and localization are performed using Region-based Fully Convolutional Networks (R-FCN) and segmentation is performed using the GrowCut algorithm.

c) Technical plan:

i) Basic Solution Hypothesis:

Dataset

The dataset will consist of videos, from which static images need to be extracted for training the models for both the initial approaches. The content of the video will consist of varying items that we want to detect, stored in workshops or warehouses. To make the dataset robust the videos captures may follow the following criteria:

- Videos in a variety of lightning conditions, camera angles and of objects in different orientations.
- The proportion of the target item should be higher in the data set.

If the dataset is not labelled for classification purposes, label it using one of the many open source tools available for labelling images.

Data Augmentation

Performing data-augmentation techniques to increase the size of the data set. The techniques may involve:

- Flipping horizontally
- Zooming
- Blurring

We plan to use two basic approaches initially, both being object detection methods:

Method 1:

The model deployed is yolo v-4. YOLO is a real time object detection system which can recognise multiple objects in a single frame, developed by Joseph Redmon. It applies a single neural network to the full image. This network divides the image into regions and predicts bounding boxes and probabilities for each region. These bounding boxes are weighted by the predicted probabilities.

We will employ transfer learning, that is to train an already trained model, with our project specific dataset to obtain our customised model.

Method 2:
This approach will employ the use of Convolutional Neural Networks (CNNs). CNNs are state-of-the-art for many image classification problems/object detection problems. We plan to use two networks for CNN: ResNet-50 and Mask RCNN [6] will be employed to perform the task of classification, localization, and segmentation using a single network. Convolution Neural Networks (CNNs) like ResNet50 learn to extract features on their own and no handcrafted features are required. The success of any classification method depends on the selected features. Manually finding features robust enough to cope with variations in the images is a difficult task.

ii) Proposed Innovations over the Basic Solution:

Improvement for Method 1:

- Enhance the performance of yolo v4 by optimizing the hyperparameters using search libraries like optuna.
- Domain Adaptation is another technique which can be used, if there is a difference between training distribution and test distribution.
- Employ other versions of yolo, that is train multiple models and make inferences.

We will also try to work for depth in images by doing so we may achieve the counts which are tough as in camera 2D images all the edges of all the objects are tough.

We can make use of regression based models (rather than detection based) as their loss function is directly optimized for predicting the object count.
We can use novel loss function which outputs instance regions. Predicted count = number of predicted instance regions. So we can make use of point supervision rather than bounding box. This will make detection faster too. Point level annotations provide rough estimation of an object's location but not its size or shape. This makes the model flexible and our goal of counting the number of objects is still achieved. FCN (Fully Convolutional Network) is used for this approach.

iii) Benefit to the user agency:

- It will enable organizations to make well-informed business decisions.
- Operating expenses will get reduced.
- It will help in reducing wastage and loss of items.
- It will enable real time reporting (i.e. no delay time which is huge when the process is done manually)
- If the company has multiple warehouses across the country then doing this process manually can become very chaotic and it would be difficult to maintain a record. Automating this process makes it organized and efficient.

d) Experimental Plan:

i) Examples and Use Cases:

Keeping a track of available items in the inventory.
Counting your goods to ensure the final number matches up with your purchase orders helps you identify instances of loss or theft.
Tracking the whole inventory; what quantity comes in and what quantity comes out.
Predicting optimized buffers as well for different inventories.

ii) Benchmarking the Technical Plan with Industry Use Cases:

In industries counting items sometimes gets tedious and sometimes one may get less items than ordered. So, the automated counting model will just take a snap and give the count of whatever item is needed.

iii) Anticipated Challenges:

We have to count the number of any item so we can’t specifically train the model for one shape of product.
In some items it gets tough to define their outer boundary (say in case of plywoods) we will need to check if the algorithm is suitable for those cases as well.
We may also work on getting the DPIs of the images so that items which are not clearly visible to the camera can also be counted.
Since the assessment is continuous we will have to stop the model from counting items repeatedly.

e) Deliverables - Milestones with Timeline:

Dec 16, 2021 - Dec 31, 2021

Trying out many different approaches, data preprocessing and initial ML Model development.

Jan 1, 2022 - Jan 31, 2021

Evaluating the different approaches and narrowing down on the best one, giving our inputs on the hardwares required for the problems (like computational power required, camera).

f) References:

- https://ieeexplore.ieee.org/abstract/document/9268394/figures#figures
- https://www.seeedstudio.com/blog/2021/03/22/machine-learning-powered-inventory-tracking-with-raspberry-pi/
- https://drive.google.com/file/d/1elerd03mwoq8vAG3b6sU-2AiBulYUYXv/view
- https://www.youtube.com/watch?v=lhMu94uCzR0&t=1674s&ab_channel=MicrosoftTechCommunity
- https://www.youtube.com/watch?v=u7yABAolbts
- [1] Y. Jia, G. Yang and J. Sarnie, "Real-time color- based sorting robotic arm system," 2017 IEEE International Conference on Electro Information Technology (EIT), Lincoln, NE, 2017, pp. 354-358.
- [2] T. P. Tho, N. T. Thinh and N. H. Bich, "Design and Development of the Vision Sorting System," 2016 3rd International Conference on Green Technology and Sustainable Development (GTSD), Kaohsiung, 2016, pp. 217-223.
- [3] R. T. Yunardi, Winarno and Pujiyanto, "Contour-based object detection in Automatic Sorting System for a parcel boxes," 2015 International Conference on Advanced Mechatronics (ICAM), Surabaya, 2015, pp. 38-41.
- [4] H. Singh, D. Gupta and A. K. S. Kushwaha, "Multiclass Object Recognition and Classification Using Boosting Technique," 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Bangalore, 2018, pp. 1-6.
- [5] P. P. Ramya and J. Ajay, "Object Recognition and Classification Based on Improved Bag of Features using SURF AND MSER Local

Feature Extraction," 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT), CHENNAI, India, 2019, pp. 1-4.

- [6] K. Mouri, H. Lu, J. K. Tan and H. Kim, "Object Detection on Video Images Based on R-FCN and GrowCut Algorithm," 2018 International Conference on Information and Communication Technology Robotics (ICT-ROBOT), Busan, 2018, pp. 1-4.
K. He, G. Gkioxari, P. Dollar and R. Girshick, "Mask R-CNN," 2017 IEEE InternationalConference on Computer Vision (ICCV), Venice, 2017, pp. 2980-2988.