How to Cite:
Ghosh, S. R., Pemmaiah, A. B., Bharti, S., Gopika, D., & Joshy, N. (2022). An intelligent automated monitoring system using video surveillance based recognition. *International Journal of Health Sciences*, *6*(S6), 1978–1989.
https://doi.org/10.53730/ijhs.v6nS6.10098

An intelligent automated monitoring system using video surveillance based recognition

Shyamas Ree Ghosh
Dept of CSE CMRIT Bengaluru, India
*Corresponding author email: shyamasree.g@cmrit.ac.in

Ashika Pemmaiah B.
Dept of CSE CMRIT Bengaluru, India
Email: ashi17cs@cmrit.ac.in

Shruti Bharti
Dept of CSE CMRIT Bengaluru, India
Email: shru17is@cmrit.ac.in

Gopika D.
Dept of CSE CMRIT Bengaluru, India
Email: Gopikablue18@Gmail.Com

Neha Joshy
Dept of CSE CMRIT Bengaluru, India
Email: nejo17cs@cmrit.ac.in

***Abstract*---**The current pandemic situation makes it necessary to work in a contactless environment where human intervention is minimalized at most. Video surveillance is an important security asset for monitoring purposes at banks, department stores, highways, and crowded public places. With improved technology and a growing population, surveillance is becoming a key area in research. The best utilization of technology for surveillance is the focus area today. In recent times, object detection has come to the frontline as an important application in the field of Deep Learning. Unlike traditional methods, object detection in deep learning is characterized by its ability to learn features and depict the same. The proposed system aims at creating a platform that reduces/eliminates human intervention for monitoring purposes by using a CCTV camera assisted automated monitoring system. There is scope for automation, which would perform object detection and automatically open the door/gate when the same has been recognized. Here, CNN models are to be used for real-time object detection via the CCTV camera.

International Journal of Health Sciences ISSN 2550-6978 E-ISSN 2550-696X © 2022.
Manuscript submitted: 9 March 2022, Manuscript revised: 27 May 2022, Accepted for publication: 18 June 2022

***Keywords*---**object detection, recognition, deep learning, CCTV, real-time video surveillance.

Introduction

Rapid advancements in Artificial Intelligence and Machine Learning technologies have made lives easier by providing solutions to various complex problems across different domains. Beginning from text and image classification to video analytics, Computer Vision has made revolutionary accomplishments in modern technology. Human-level performance in visual perception tasks can be achieved by applying Modern Computer Vision algorithms. The current pandemic outbreak of COVID-19 makes it necessary to work in a contactless environment where human intervention is minimalized at most. At present, entry to various buildings/ organisations/ checkpoints is manually monitored by an individual. However recently, people across the world are keen on returning to normalcy and resume in-person work.

In order to start a normal workflow in various organizations like offices, schools, and shopping centers, it is important to make sure that no violations are made. Monitoring this via humans is not a feasible solution. Hence a video surveillance auto entry system is proposed, which aims at providing entry only if specified items are successfully recognized via a CCTV camera.
Example:

- In current situations, malls, shops, and banks provide entry only with masks. Hence, if a person is wearing a mask, the doors would automatically open without any human monitoring/assistance.
- Gates automatically open when a student with the right uniform appears.
- Nuclear power plants provide entry only in the presence of a complete bodysuit.

As per organizational needs, items such as mask/uniform/ID card etc can be fed as an input to the system, to monitor human entry to specific areas. Closed-circuit television systems (CCTV) have been installed across various offices, residences, societies, and public areas. CCTV aims at monitoring and controlling, detecting, observing, recognizing, and identifying situations and individuals to avoid potentially harmful activities [3]. Since viewing multiple camera views and identifying the objects is limited by the human factor. A solution to the problem is applying automated image processing algorithms, which reduces human intervention and generates an auto entry system.

Related Work

Various traditional machine learning algorithms focus on image analysis tasks such as hand-crafted features extraction, colour segmentation, normalization, etc. These methods are backed up by classification algorithms like Support Vector Machines (SVMs) and Regression [1]. As these methods are incapable of processing high-dimensional image feature sets, advanced methods like neural networks have come to the forefront and provided feasible solutions for the extraction process in automated high-dimensional image sets. Currently, neural networks are used extensively by implementing multi-layered networks.

CNN

Convolutional Neural Networks (CNNs/ ConvNets), shown in Fig. 1, are Deep Learning algorithms that take an image as input which is then assigned importance (biases and learnable weights) to various objects within the image, and hence, differentiates one from the other [1].

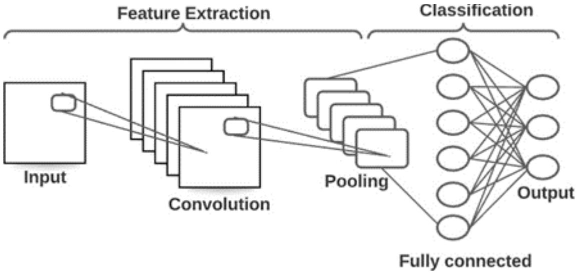


Fig. 1. Convolutional Neural Network

CNN requires much lower processing when compared to a variety of classification algorithms. While filters are hand-engineered and trained in traditional methods, CNN is capable of learning the same. It applies relevant filters to accurately capture spatial and temporal dependencies present in images [6] [5]. CNN reduces images into an easily processable form, by retaining features critical for achieving accurate predictions. This is crucial in designing architectures that are good for learning features and scaling to massive datasets.

Faster R-CNN

In the Region-Based Convolutional Neural Networks (R-CNN) family, the evolution between versions is with respect to computational efficiency, reduction in testing time, performance improvement and mean Average Precision (mAP) [6]. The network usually consists of:

- Region proposal algorithm that generates ‘bounding boxes’ or locations of specific objects in the image.
- Feature generation stage that obtains features of the objects, by applying CNN methods.
- Classification layer that predicts which class the object belongs to.
- Regression layer that makes the coordinates of objects bounding boxes more precise.

RCNN, as shown in Fig. 2, is a popular object detection architecture that uses CNNs like Single Shot Detector (SSD) and You Only Look Once (YOLO) [4]. Its architecture consists of two networks, namely,

- Region Proposal Network (RPN)
- Object Detection Network

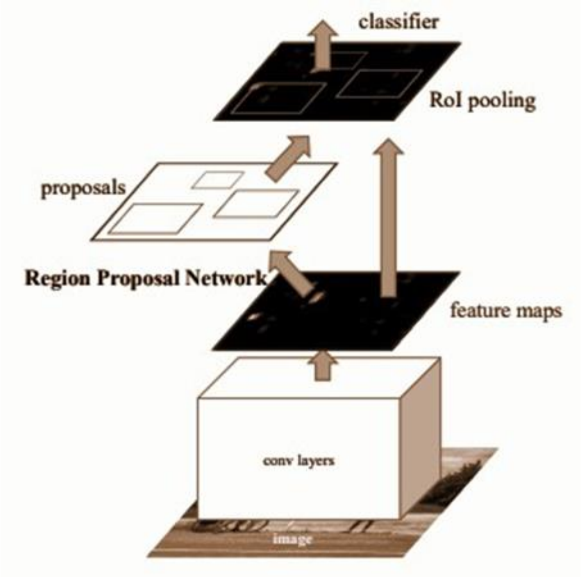


Fig. 2. Faster R-CNN

RPN is efficient in proposing multiple objects that can be identified within a particular image [4]. It consists of,

- Classifier : Responsible for determining odds of a proposal containing the target object
- Regressor : Responsible for regressing coordinates of the proposals

Faster R-CNN

SSD performs real-time object detection. Faster R-CNN utilizes RPN for creating boundary boxes and uses the same for object classification. Running at 7 frames/sec, the run time of this process is much lower than real-time processing requirements. In SSD, the elimination of RPN accelerates the process. Introducing default boxes and multi-scale features compensates the reduction in accuracy, making SSD's accuracy similar to that of Faster R-CNN, even though SSD uses lower resolution images, and in turn, increases the speed [2] [5]. Feature map extraction and convolution filter application are two stages of SSD object detection.

Faster R-CNN

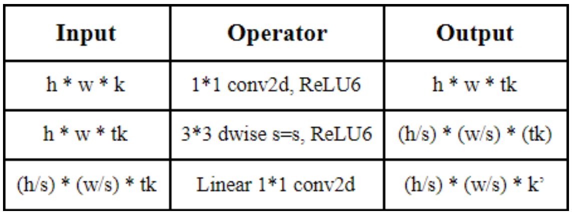
YOLO also performs real-time object detection [1]. Application of a solo neural network over the entire image enables this algorithm to split the image into several regions, and then predict bounding boxes along with predicted probabilities for each of them. Higher accuracy and ability to run in real-time accounts for YOLO's popularity. Here, 'only looks once' implies that only a single forward propagation is required via the network to make predictions [7]. Post non-max suppression it returns recognized objects along with their bounding boxes.

MobileNetV2

MobileNetV2 is a lightweight CNN model (shown in Fig 3) that accepts an image of dimension 224*224 as input and is composed of 53 layers. A pre-trained network model, trained over thousands of images using the ImageNet database can be loaded [2]. The same can then classify images into various categories of objects like mask, tree, mango, pen, cat, dog, etc. Thus, the convolutional network learns a high-quality representation of features over a variety of images. MobileNetV2 is predominantly built from the inverted residual structure. It acts as a backbone for extracting features and delivering state-of-the-art performance for detecting objects and semantic segmentation. MobileNetV2 has two types of blocks: A residual block with a stride of 1 and a downsizing block with a stride of 2 [6]. Both the blocks have 3 layers, as shown in Table I:

- Layer 1: A 1*1 convolution layer with Rectified Linear Unit-6
- (ReLU6)
- Layer 2: A depthwise convolution
- Layer 3: A 1*1 convolution layer with no non-linearity

Table I
Layers of residual & downsizing block



Input	Operator	Output
h * w * k	1*1 conv2d, ReLU6	h * w * tk
h * w * tk	3*3 dwise s=s, ReLU6	(h/s) * (w/s) * (tk)
(h/s) * (w/s) * tk	Linear 1*1 conv2d	(h/s) * (w/s) * k'

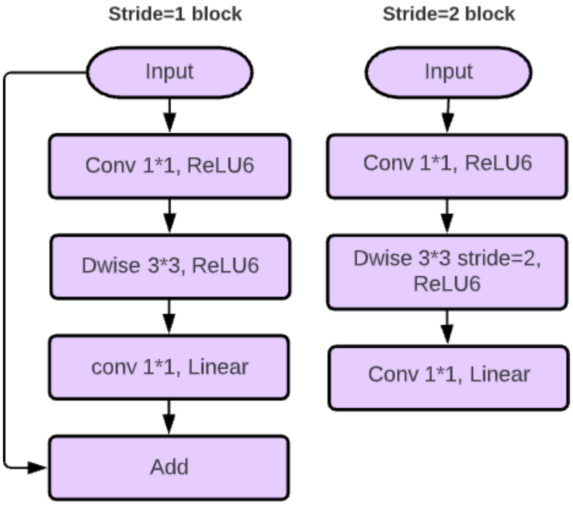


Fig. 3. MobileNetV2 Architecture

NASNet Mobile

Neural Architecture Search Network (NASNet), searches for an architectural building block over a trivial dataset and shifts it to a vast dataset [1]. Firstly, it searches for an ideal convolutional cell/layer on CIFAR-10 and then applies this over the ImageNet block by stacking multiple copies of this cell. ScheduledDropPath, an advanced regularization technique, proposed to significantly enhance the generalization in NASNet models. Finally, the NASNet model achieves state-of-the-art outputs with lower complexity (FLOPs) and smaller model size. In NASNet, even though the general architecture is pre-set as shown in Fig. 4, the cells (blocks) aren't predefined by authors. Instead, these cells are searched via the reinforcement learning search algorithm wherein the number of initial convolution filters and the number of motif repetitions(N) are free parameters and are used for scaling. Here, cells that return a feature map whose dimensions are reduced by a factor of two are called Reduction Cells, and those that return a feature map of exactly the same dimensions are called Normal Cells [7].

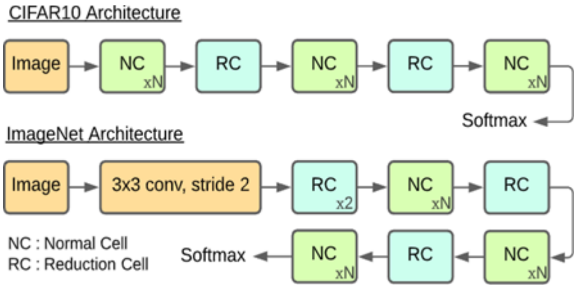


Fig. 4. Architecture of CIFAR10 and ImageNet

Hardware

To enable automatic entry based on video surveillance, provision of CCTV is a must for providing live footage. Alternatively, footage via webcam also can be used.

Methodology

The sudden outbreak of COVID-19 has made it necessary to work in a contactless environment where human intervention is minimalized at most. However recently, people across the world are keen on returning to normalcy and resume in-person work.

Goal

This system aims at creating a platform that reduces/eliminates human intervention for monitoring purposes by using a CCTV camera assisted automated monitoring system.

Gap

At present, entry to various buildings/ organisations/ checkpoints is manually monitored by an individual. Introducing this system, would eliminate the need for such human intervention, and instead automate it. This would serve as a precautionary measure highly required by the current pandemic situation.

Solution

A variation of three different modules is to be used as follows (Only for demonstration purpose. Based on the company/ user needs, the implementation can be customized) :

- Face mask detection
- ID card detection
- Hazmat suit detection

Datasets, such as the face mask dataset or ID card dataset are to be used for their respective models. These datasets consist of images in both categories. For example, within a face mask dataset, images of people wearing masks and people without masks are to be used. A similar mechanism will be followed in the detection of other objects such as IDs, hazmat suits, etc.

Proposed System

A multi-stage architecture for detecting specific objects as per user requirements is used [3]. The working of the same is demonstrated in Fig. 4.

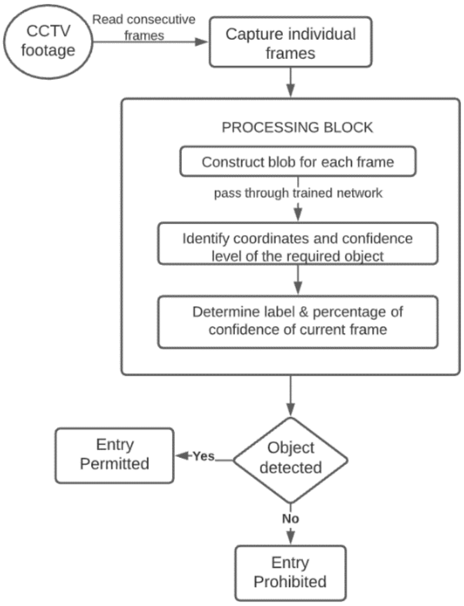

CCTV footage
Read consecutive frames
Capture individual frames
PROCESSING BLOCK
Construct blob for each frame
pass through trained network
Identify coordinates and confidence level of the required object
Determine label & percentage of confidence of current frame
Object detected
Yes
Entry Permitted
No
Entry Prohibited

Fig. 4. System Architecture

Firstly, real-time footage is fed via a CCTV camera [10]. It is then divided into frames/images and is then transferred to the processing block. Within this processing block, a Binary Large OBject (BLOB) is generated for each frame, which in turn is a group of connected pixels in a binary image represented in array format. These blobs are then passed through a trained neural network for:

- Identifying coordinates and confidence levels of the required objects.
- Determining accuracy percentage and corresponding label

Based on the label assigned entry is either permitted or prohibited. Implementation of the same is depicted in Fig. 5.

Image	Door/ Gates open
	✓
	✗

Fig. 5. Demonstration of mask detection

Implementation

As a first step, a face mask dataset consisting of 1915 with-mask images and 1918 without-mask images, is used for training MobileNetV2 and NASNetMobile models [8] [9]. The training of these 2 models is depicted in Fig. 6. Here, average pooling of the 7x7 kernel is used with flattening layers composed of ReLu and softmax as activation functions, along with a dropout rate of 0.5, to detect facemasks. Based on the result generated, the model with higher accuracy is to be chosen and used to implement other cases such as ID detection, hazmat suit detection etc.

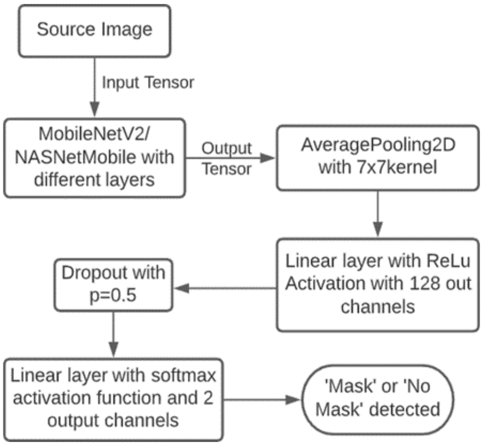

Source Image
Input Tensor
MobileNetV2/ NASNetMobile with different layers
Output Tensor
AveragePooling2D with 7x7kernel
Linear layer with ReLu Activation with 128 out channels
Dropout with p=0.5
Linear layer with softmax activation function and 2 output channels
'Mask' or 'No Mask' detected

Fig. 6. Training of face mask detection

Result

Pretrained Model Comparison

As shown in Table II, the model trained for face mask detection using MobileNetV2 generates a higher accuracy within a lesser training duration as compared to NASNetMobile. Based on this result, MobileNetV2 is chosen to be implemented for other object detection cases.

Table II
Comparison table

MODEL	ACCURACY (%)	TRAINING TIME (mins)
MobileNetV2	99.23	42
NASNetMobile	98.83	107

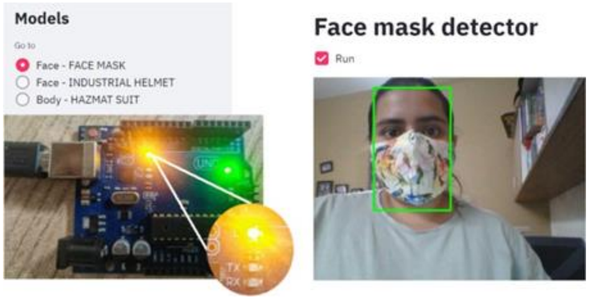
Models Implemented

Table III, demonstrates the accuracy generated on implementing and testing three different sample models.

Table III
Accuracy table of implemented models

MODEL	ACCURACY(%)
Face mask	99.23
Industrial helmet	98.78
Hazmat Suit	99.01

Using a base model that detects faces along with appropriate datasets containing images of the items that are to be detected on the face like mask, helmet etc., a multipurpose solution is demonstrated from Fig. 7 to Fig. 10.



Models

Go to

● Face - FACE MASK
○ Face - INDUSTRIAL HELMET
○ Body - HAZMAT SUIT

Face mask detector

☑ Run

Fig. 7. Face mask detection - Entry Granted

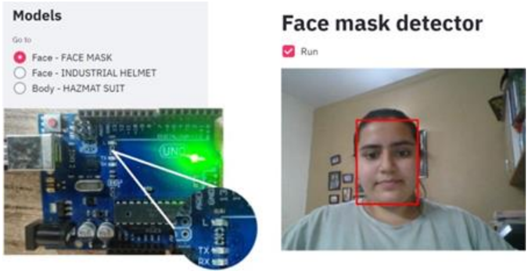


Fig. 8. Face mask detection - Entry Prohibited

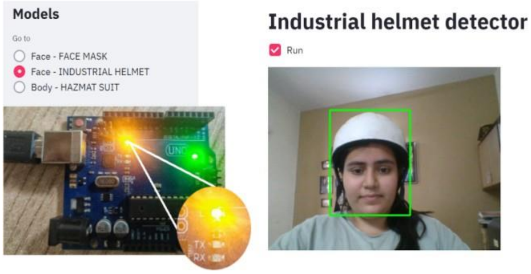


Fig. 9. Industrial helmet detection - Entry Granted

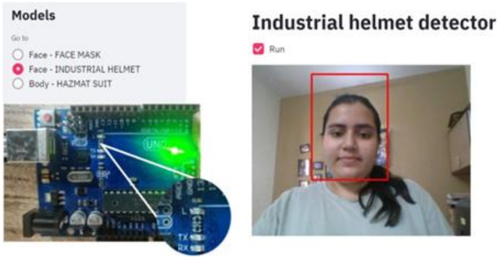


Fig. 10. Industrial helmet detection - Entry Prohibited

Similarly, using a base model that detects the human body, and utilizing a dataset of items that is to be detected on the body, a multipurpose solution is generated. A demonstration of hazmat suit detection using this approach is shown below, from Fig. 11 to Fig. 12.

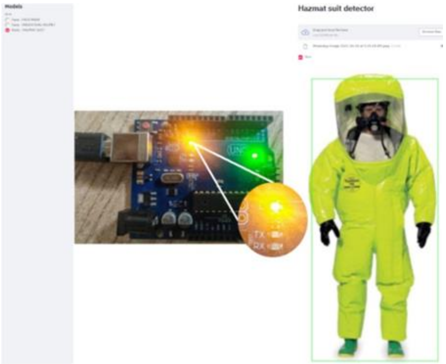


Fig. 11. Hazmat suit detection - Entry Granted

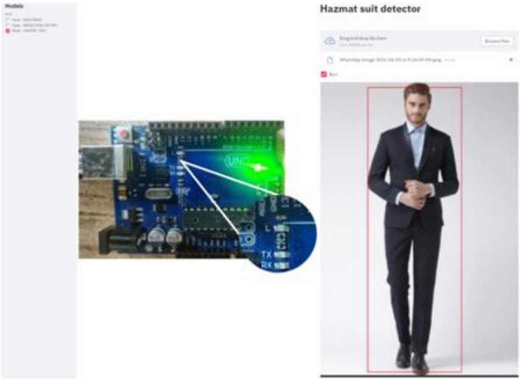


Fig. 12. Hazmat suit detection - Entry Prohibited

In all the above images, the glowing orange LED bulb on the arduino board indicates that entry is granted, while a turned off LED bulb indicates that entry is prohibited.

Conclusion

Employing this automated monitoring multipurpose system using video surveillance-based recognition, base models of broader categories can first be generated and utilized for the detection of specific items within them. In this manner, the entry process can be automated to minimize or eradicate human intervention.

Future Work

Currently, this system has been implemented only for face and body detection models. As a next step, base models for different cases like text or pattern detection can be generated to validate ID cards of different organizations just by changing the training dataset. Enhancement of datasets can be carried out to improve the accuracy of object detection. Additionally, suitable hardware can be used to simulate/demonstrate automated entry.

References

1. Chavda, A., Dsouza, J., Badgujar, S., & Damani, A. (2020). Multi-Stage CNN Architecture for Face Mask Detection. *arXiv preprint arXiv:2009.07627*.
2. Chandan, G., Jain, A., & Jain, H. (2018, July). Real-time object detection and tracking using Deep Learning and OpenCV. In *2018 International Conference on Inventive Research in Computing Applications (ICIRCA)* (pp. 1305-1308). IEEE.
3. Grega, M., Matiolański, A., Guzik, P., & Leszczuk, M. (2016). Automated detection of firearms and knives in a CCTV image. *Sensors*, *16*(1), 47.
4. Chowdhury, S., & Sinha, P. Real-Time Object Detection using Deep Learning: A Webcam Based Approach
5. Rohan, A., Rabah, M., & Kim, S. H. (2019). Convolutional neural network-based real-time object detection and tracking for parrot AR drone 2. *IEEE Access*, *7*, 69575-69584.

6. https://towardsdatascience.com
7. https://medium.com
8. https://keras.io/api/applications/mobilenet/
9. https://keras.io/api/applications/nasnet/#nasnetmobile-f unction
10. https://cloudxlab.com/blog/how-to-run-yolo-on-cctv-fee d/
11. Suryasa, W., Sudipa, I. N., Puspani, I. A. M., & Netra, I. (2019). Towards a Change of Emotion in Translation of Kṛṣṇa Text. *Journal of Advanced Research in Dynamical and Control Systems*, *11*(2), 1221-1231.
12. Suwija, N., Suarta, M., Suparsa, N., Alit Geria, A.A.G., Suryasa, W. (2019). Balinese speech system towards speaker social behavior. *Humanities & Social Sciences Reviews, 7*(5), 32-40. https://doi.org/10.18510/hssr.2019.754
13. Parmin, P., Suarayasa, K., & Wandira, B. A. (2020). Relationship between quality of service with patient loyality at general polyclinic of kamonji public health center. International Journal of Health & Medical Sciences, 3(1), 86-91. https://doi.org/10.31295/ijhms.v3n1.157