ISSN 0511-5728
The West Indian Journal of Engineering
Vol.45, No.1, July 2022, pp.50-61

A Convolutional Neural Network Based Robust Automated Real-Time Image Detection System for Personal Protective Equipment

Jordon Hayles[a], Kolapo Sulaimon Alli [b,Ψ], and Latchman A. Haninph[c]

Faculty of Engineering, The University of the West Indies, Mona Campus, Kingston 7, Jamaica, West Indies;
[a]Email: jordonhayles@yahoo.com
[b]Email: kolapo.alli@uwimona.edu.jm; allikolapo@gmail.com
[c] Email: haniph.latchman@uwimona.edu.jm

[Ψ] *Corresponding Author*

(Received 08 March 2022; Revised 13 June 2022; Accepted 28 June 2022)

Abstract: *Statistically, casualties in engineering workplaces often result from one of the following accidents: when people get stuck in the rotating machines, electric shocks or collision with heavy equipment. Most of these accidents can be prevented if the workers make proper use of personal protection equipment (PPE). This paper presents the design and implementation of a functional image detection system that takes a picture of an employee, analyses it, and determines the employee he is appropriately attired to enter a potentially hazardous workplace. This system can help to reduce the liability of company owners, by extension their costs, and can provide level of accident prevention. In this study, a convolutional neural network (CNN) was used to develop three sets of models, namely hard hat model, boot model, and vest model. These were used to detect the appearance of workers and determine if the PPE being worn was in compliance with the stipulated requirements for entry to a particularly hazardous workplace. To determine the performance of the system, each model was validated with two classes of image datasets: normal colour RGB (Red, Green and Blue) and grayscale image. The overall average accuracy of the system, in real-time implementation, then was calculated and determined to be 83.33%.*

Keywords: *Convolutional Neural Networks, Image Processing, Deep Learning, Personal Protective Equipment, Safety, Tensorflow, Training, Accuracy*

1. Introduction

Accidents often occur on construction sites when there is limited availability of personal protective equipment (PPE) or when workers fail to use PPE provided by employers. The essence of PPE is to limit the level of hazard exposure for the employees when engineering and administrative controls are not able to minimize the risk to an approved level. According to the International Covenant on Economic, Social and Cultural Rights (ICESCR) - Article Seven, everyone should enjoy the right to favourable conditions of work (UN, 2016). Workplaces are regulated and are obligated to ensure the safety of the work environment. In addition, employers are required to provide a safe environment for their workers. This requires three lines of defences to protect employees against hazards.

Engineering controls are used to reduce and/or minimise the hazard in question and involve making changes to the environment. This is the most effective form of control. The second line of defence is administrative and work controls. This defence aims to reduce the severity, duration, and frequency of the exposure to the risk, for example, limiting worker's exposure time. The third line of defence is personal protective equipment (Nilfisk, 207). This is considered the last line of defence and should not be disregarded. PPE is an important part of worker safety. In recent times, emphasis has been placed on PPE used by medical and frontline workers fighting the novel coronavirus, SARS COVID-19.

Many engineering firms require employees to wear PPE to ensure the health and well-being of their workers, and to protect the company from liabilities, such as medical and legal fees. Workplaces have established rules and regulations. However, employees sometimes forget to bring their PPE to work, and may try to avoid fully conforming to the rules established by the employer. To ensure that employees comply with the established workplace standard, and help protect industrial companies from liabilities, an Automated PPE Image Detection System is proposed.

The high prevalence of causalities on construction sites has reached an alarming rate. Personal protective equipment helps prevent injury and promote workplace safety. From published statistics in the United States, 84 % of workers who sustained head injuries were not wearing hard hats and 99 % of workers who suffered facial injuries were not using face protection (New York

https://doi.org/10.47412/LXQC9076

University, 2005). This paper aims to use deep learning to create an effective and efficient oversight layer that will promote proper PPE usage. The system designed is intended for employers who require workers to wear special gear while carrying out their duties. This system is low cost, easy to implement, and is easily upgradable. Consequently, the system presents a proven solution for the problem of high causalities on construction sites and can be used as a permanent or interim solution.

Many of the existing automatic techniques used for detecting workers without PPE focus narrowly on field surveillance videos. The present research is to develop a robust and automated PPE detection, with the ability to verify and grant access to authorized workers before allowing access onto construction site or any other work environment, where protection and security are essentials. In addition, this research has adopted the use of a deep learning method, called a convolutional neural network (CNN) in developing an automated PPE Image detection system that incorporates a fingerprint identification system.

The rest of the paper is structured as follows: Section 2 discusses the existing related work and their limitations, and Section 3 describes the software and the hardware implementations of the system, as well as the CNN training and its architecture. Section 4 discusses the results and discussions, while Section 5 provides a summary and conclusions.

2. Literature Review

In 2013, a deep learning method called MobileNet-SSD structure was implemented in a construction equipment image detection system that was deployed on an embedded selection system (Arabi et al., 2013). The network class of the method involves depthwise separable convolution which factorize normal convolution into two different operations. The accuracy of the system in real-time learning is 90%. Barro-Torres et al. (2012) advocated the use of Zigbee and Radio Frequency Identification (RFID) technologies for the detection of PPE for monitoring how PPE are worn by the workers. This system was designed to inspect each worker's appearance in the use of a PPE where every worker must carry the system, and then reports that alerts a unit at a central location and provides information about each worker. The operation of the system relies solely on battery power which is a major limitation of this detection system. Any temporarily discharged battery, could cause failure in sensor node connection which would require reconfiguration of the entire sensor network. Another deep learning algorithm called YOLO has been reported in Hung et al. (2019) for a real-time personal protection equipment detection. This involves a road segment designed using an RFID, an infrared sensor (IR) and a camera which are installed to identify if workers accessed the construction site with the required protective equipment. This system achieved high precision but the detection of small objects in groups by this technique might give incorrect prediction and also the algorithm requires long computational time.

In another similar work, a CNN-based camera identification system for detecting workers with noncompliance PPE, capable of detecting twelve (12) classes of PPE has been presented in Wahyu Pradana et at. (2019). The accuracy of the system in real-time implementation was 85.83 % for respondents who were included in the dataset. However, this system encountered several errors in differentiating workers who used safety glasses and the respondents that did not use safety glasses in real time when using the CNN method.

A review of CNN applications for fruit image processing analysis has been presented in Naranjo-Torres et al. (2020). This study was limited to the classification, quality control and detection of agricultural fruit images. In addition, the use of an automated intelligent drone equipped camera and computer vision for safety inspections on construction sites was reported in Abbas et al. (2016) with the aim of detecting workers not wearing hard hats. However, the control of the drone requires a human expert to navigate in both indoor and outdoor construction sites.

A survey of on-site construction personnel hazard perception in the Middle East Country of Lebanon has been investigated in Abbas et al. (2018). The study centered on the awareness and perception of engineers, foremen and workers on various indoor hazardous activities and emphasized the significance of safety equipment on different construction sites in Lebanon. It also discussed the main constraints limiting proper enforcement of wearing PPE at construction sites. Similarly, a real time pattern recognition using digital video and applications for measuring safety in construction sites was presented in Bajracharya (2013). This system used a camera recording video as input and processed it to several image frames, and then used a classification method by developing a client interface and created an image box to detect the appearance of workers in compliance with the wearing of PPE on construction sites. However, there were no specific results reported that describe the level of performance of the system.

In 2014, a recognition algorithm system was developed that uses site images collections for an automated monitoring of construction progress (Azadani et al., 2014). This algorithm reported an accuracy of 70 % or less and does not perform well with images collected from construction sites. In another work, Du et al. (2011) considered face features, motion, and colour information in video processing technique used for detection of face and hard hat in construction sites. Every worker is expected to look at the camera before the system can detect 1) if hardhat is worn and 2) the colour of the hardhats is the same.

In image classification and object detection, CNN was

adopted due to its ability to self-learn from large training datasets in Roy (2020). Siddula et al. (2016) also described a CNN algorithm has used for detection of objects such as roofs, roof workers, and guardrails in construction area. This involves segmentation of each image using a Gaussian Mixture Model and later passed into the CNN model for detection. This study also investigated the effect of different networks architectures, to determine which topology is suitable for recognizing each object being tested and hence this affects the performance of the system. One challenge was in the use of the object images captured from a far distance which involved some other objects. These types of images would generate low contrast during segmentation of different objects which led to low contrast of images. However, these images are considered unsuitable for proper segmentation operation.

In another instance, Wu et al. (2019) developed a hardhat wearing detection system using a CNN method. In this study, the image feature extraction of different layers was achieved using reverse progressive attention which makes the final prediction of the detection results. This system was limited to detect workers with proper wearing of hardhat and the respective colour of the hardhat. Kyrkou et al. (2018) used a CNN architecture for detecting objects embedded in a lightweight processing system deployed in an unmanned aerial vehicle (UAV), with a reported overall system accuracy of 95%. However, this work was limited to a single object class and has fewer numbers of training set that could lower its applications and performances.

Another CNN-based architecture algorithm has been used in Audebert et al. (2017) for the identification of vehicles which involves segmentation, detection and classification of remote sensing images. Similarly, Rastegari et al. (2016) introduced a CNN network with binary weights and exclusive nor (XNOR) approximation networks for image classification. This technique involved the reduction of network size where a small amount of memory is applicable, to run the deep learning algorithm in real-time on portable devices.

Moreover, Chollet (2017) adopted a CNN architecture based on xception modules which involves depthwise separable convolution layers, for classification of a large image dataset. The time required to train xception modules is expensive but performs well in terms of transfer learning rate which makes it to be adaptable for any specific operation. Fang et al. (2018) used a Faster R-CNN method for detecting workers with non-hardhat use, carried out in twenty-five (25) different construction sites, taking into consideration the impacts of visual range, weather and illumination on far-field images captured at various ranges of working hours. This work was limited to detect workers' non-hardhart-use (NHU) but was not designed to identify the workers on the construction sites and is limited to one class of model. According to He et al. (2017), an extension of a Faster Recurrent Convolutional Neural Network R-CNN, called Mask RCNN, has been adopted for a high-quality object segmentation detection in an image. This method was limited to work with still images as it fails to detect dynamic objects which experienced motion blur at low colour resolution.

Ham et al. (2016) reported a review assessment of camera-equipped Unmanned Aerial Vehicles (UAVs) for visual monitoring the infrastructure in construction sites. This paper investigated most recent methods for collection, analysis, visualization, and communication of visual data obtained with or without a Building Information Model. This work also presented the potential of each of these methods that leads to automatic monitoring and civil infrastructure assessment.

Kim et al. (2016) proposed an object detection for autonomous vehicles and localization of objects on road areas using deep neural network. The detection accuracy was evaluated with some object classes and analyzed the identified results were analyzed fine-tuned single shot multibox detector on KITTI dataset. In another study, Murugan et al. (2019) confirmed applications with a vehicle logo recognition system in traffic monitoring, security systems and surveillance systems categorized with higher accuracy by RCNN. The limitation of the vehicle logo system for real-world applications using a CNN method was that the testing dataset kept changing the training set after detection which keeps the training set frequently updating. Hence, this system reported about 15 hours to train its network using 10,000 training images. Therefore, the proposed technique was very time-consuming. Fang et al. (2018) thus adopted a faster RCNN for an automatic non-hat use detection system. However, the main problems with the far-field surveillance videos are the continuous movement, equipment, objects and the environment which are all captured during the process of the non-hardhat workers' detection on construction sites.

Dimitrov and Golparvar-Fard (2014) developed an automatic monitoring construction progress and a 3D point cloud building information model for construction materials classification using a support vector machine (SVM) method. This algorithm achieves an average accuracy of 97.1 % for the classification of building materials using 200×200-pixel image patches. Du et al. (2011) also reported the use of video surveillance based on face features, motion and colour information for identifying a person without the hard hat in real time when entering the construction sites for ensuring safety of workers. Similarly, Gualdi et al (2011) reported the use of Covariance Descriptors with a LogitBoost classifier and a surveillance detector that could identify those workers who do not wear hardhat during construction work.

To enhance further, Memarzadeh et al. (2013) advocated an automated system for detection of construction equipment and workers from video streams using binary Support Vector Machine (SVM) classifier. With over 8,000 dataset containing the video frames of construction equipment and workers, the detection results

for standing workers, excavators and dump truck achieved average accuracy of 98.83 %, 82.10 %, and 84.88 % provided by the method, respectively. In addition, Radio Frequency Identification (RFID) technology was used in the construction site to estimate the distance between the worker and the heavy equipment to prevent the worker from entering a dangerous environment. Chae and Yoshida (2010) developed an RFID-warning system for preventing workers from collision accidents with heavy equipment like excavators and cranes.

Most fast deep learning methods are known for less accuracy in achieving a real-time object detection application, Zhou et al. (2016) developed a deep neural network (DNN) called YOLO for a vehicle detection and classification system using rear view images captured by a position road camera placed at a distance along a multi-lane highway. Two cases of vehicle classification experiments were conducted, namely normal images and dark images, with achieved accuracy of 93.3 % and 83.3 % respectively compared with 94.4 % and 86.2 % obtained from a Deformable Parts Model (DPM) method. Peng et al. (2016) also developed a safety video surveillance system that comprises a Gaussian Mixture Model and YOLO. This extension of convolution neural networks would detect pedestrians near the transformer substation with 20 % more accurate than the single method.

A similar work involved developed YOLO-v3 architecture models to detect if construction workers comply with the proper wearing of hardhat, vest or both, from video images in real-time (Nath et al., 2020). These models were limited to detect only hat and vest classes while there are some other types of PPE such as the gloves, safety goggles and boots which can be accommodated by modifying the last layer of the YOLO-v3 models. The work does not conduct verification and authorization of every worker. Therefore, it does not assure the confidentiality of each worker on the construction site.

Popescu et al. (2019) presented the use of optical methods to train the acquired data from existing concrete bridge inspection to reduce traffic disturbance, improve the efficiency and reliability of the bridges. These techniques were validated with three separate imaging datasets for the 3D-geometric modelling of existing structures: terrestrial laser scanning, close-range photogrammetry, and infrared scanning.

This focus of this present study is to examine the weaknesses in the existing works and propose a feasible solution that could train a CNN to detect the most essential safety equipment such as boots, hardhats and safety vests needed for the construction workers to be worn before they can be granted permission to enter a construction area. This would offer optimum protections especially where the handling of heavy equipment and machines is required. Additionally, CNN is becoming widely recognized as an efficient and powerful deep learning tool for solving different forms of classification problems due to its inherent advantages to perform automatic feature extraction as well as its better computational efficiency. This also allows CNN models to operate with any device and makes it a suitable selection method.

On the other hand, the majority of the existing work on computer vision have been able to address fewer classes of safety equipment to detect the compliance of workers with the proper use of personal protective equipment. Previous studies have not considered some PPEs model classes (such as the boots, the hardhats and the vests models) for a real-time image detection system using CNN. The specific objectives of this study are to design and build a low-cost Automated PPE Image Detection System using still frame images captured by a camera, and capable of detecting and determining if someone is entering an industrial workplace.

Figure 1 shows a typical framework of the Automated PPE Detection System. The system would be able to differentiate between registered and unregistered personnel by using a scanned fingerprint to look up the employee on the system database. Unregistered personnel could opt to enroll into the system through managerial input, while registered personnel would be allowed to the next stage of the system where a camera takes their picture for image processing. After processing and analyses, the employee would get feedback on the appropriateness of attire for the particular workplace. These results would be updated to a database that are accessible through a website.

3. Methodology

The first stage of the process involved in the development of the proposed system is to obtain the images used

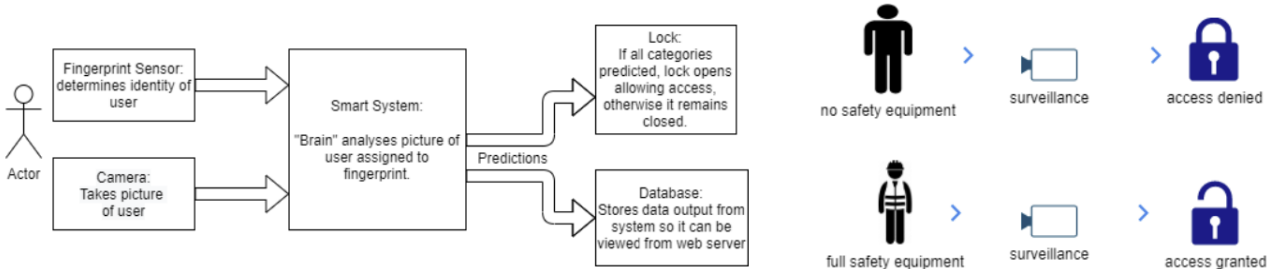

Figure 1. A Typical Framework of the Automated PPE Detection System

to train the CNN. To do this, a code is used that linked to Microsoft Bing's Image Search API. The code took a keyword, for example, "hard hat", and thoroughly searched the internet for related images and downloaded them to system memory. Once these categories were saved to system memory, dataset pruning began. Pruning took place in two ways: 1) by using a code that would process images saved to memory and remove duplicates leaving one original copy 2) having a human user scan every picture and deleting pictures from the dataset that were not related to any of the three categories (vest, hat, boot). Once the dataset was finalized, three different specialized CNN models were trained on same images. One model was specialized to predict hard hats, one model was specialized to predict high-vis vests, and one model was specialized to predict steel-toe boots. Once these models were developed, a system was implemented that accepted user data and took an image of user via a camera. The image taken was input and processed by each CNN model which made their independent predictions. The result was then tethered to the user data and a decision to provide or deny access to the facility was made depending on the final results.

The system is made up of a Raspberry Pi, an ATMega328P, a fingerprint sensor, a pressure mat, a solenoid lock, an input button, a buzzer and 2 LED indicators. Employees press the button, scan their fingerprint and stand on the mat. A camera takes a picture and the model and scans for PPE. If all is found to be good, the lock opens with a green light, else it stays closed with a red light and a buzzer sound. The system keeps monitoring the persons being scanned by using employees' fingerprints, and their performance records are stored on a web server which allow the administrator to view the activity on a website.

The circuitry hardware communication standard used in this work was the Universal Asynchronous Receiver/ Transmitter (UART) Serial standard of communication. This establishes a communication link between the microcontroller and the Raspberry Pi. Software serial communication was also used to establish a communication between the fingerprint sensor used and the microcontroller connected. Other standards involved in this work were related to the standard compression formats for the images used to train the models. The standard file formats that used were BMP, JPG, JPEG, and PNG. This work has been grouped into two categories – (i) the microcontroller circuit and (ii) the processing computer, and the functions of each group are described as follows:

3.1. Microcontroller Circuit

The microcontroller circuit deals with the linking of the physical environment to the system. The components connected to the microcontroller are a biometric sensor, a pressure sensitive security mat, a 12V DC solenoid style lock, an LCD, a momentary push button, a buzzer and two LEDs. The biometric sensor is a DFRobot fingerprint sensor module. It reads and collects unique user biometrics. The sensor returns data which is used to determine which database entry is associated with that specific user, and thus, which database attribute needs to be updated.

Next is the security mat. The voltage output, Vo, from the security mat is monitored by the microcontroller. When pressure in the form of human weight is applied to the mat, the Vo from the mat changes proportionately and the microcontroller is tasked to determine if Vo falls into a defined range. The lock is used to prevent and grant access to people based on the processing that occurs within the full system. The momentary push button is for user input, a 16 × 2 LCD is used to send output text message to the users, and a buzzer and two indicator LEDs provide audio-visual output to the users. The two main functions carried out by the microcontroller circuit are to verify if users are already on the database and enroll a new users into the system.

3.2. Processing Computer

The processing computer used for the system is the Raspberry Pi 4 Model B/4GB (RPi). The RPi stores most of the softwares used in the system. Stored on the RPi are three trained CNN models as well as the image processing and classification script. The RPi also hosts the database and the web server. The database used is MongoDB and the web server used is the Apache Web Server. The RPi and the microcontroller communicate with each other to coordinate the actions of the system. Both system components are wire-connected and communicate using the UART Serial communication protocol.

The RPi acts as the hub for the system. It plays the main role in defining the flow of data and system functions. The RPi houses the database and the classification script, which is connected to the camera and microcontroller peripherals. The script on the RPi polls both the serial port and the database.

When the RPi receives data via the serial port, it will trigger the camera to capture the image of the person. This data received from the serial port is sent once a user has been verified and is correctly positioned on the pressure mat. When the camera is triggered, the image is saved and loaded into the classification script where it is processed. The output from the classification is retrieved, along with the verification input from the microcontroller and then used to update the respective employee status on the database. The script will also send the processing results back to the microcontroller by way of the serial port and will then go back to polling both the database and the serial port. If the enroll button online is pressed, a python script is run that sends a command to the microcontroller. The RPi will then wait until it receives data from the microcontroller. If a negative code is received by the RPi,

it will delete the recently made entry and return back to polling.

3.3. CNN Training

The system is focused on finding instances of the defined PPE. The defined PPE categories outlined in the project objectives are hardhats, steel toe boots, and reflective vests. To achieve this, three specialized CNN models were trained, each one dedicated to detecting its own category. Each CNN was trained in similar fashion, with the only difference between them being the input dataset. The approach used to train the CNNs was based on the method used by Adrian Rosebrock on his PyImageSearch blog.

For the boot detection model, a total of 1,861 images were in the dataset. The breakdown of the image classes within the boot dataset are:

- Sandals (680 images)
- Heels (582 images)
- Boots (599 images)

For the vest detection model, a total of 2,814 images were in the dataset. The breakdown of the image classes within the vest dataset is:

- Dress (669 images)
- Tank Top (763 images)
- T-Shirts (582 images)
- Vest (800 images)

For the helmet detection model, a total of 1,682 images were in the dataset. The breakdown of the image classes within the hardhat dataset is:

- Cap (513 images)
- Hat (585 images)
- Random (584 images)

In this work, the performance of CNN approach was validated with two classes of image dataset for each of the models, namely, normal RGB coloured image dataset and grayscale image dataset. The same objects images used in training the CNN network for the grayscale images. Usually, the grayscale images format is obtained by training the network to accept RGB images as input dataset and conversion to grayscale images format is done by using cv2.cvtColor() command from the OpenCV library. Both RGB and grayscale are renamed and entered into different new folders. All the images in the specified folder are read and saved. Each of the converted grayscale images is resized into 128 × 128 pixels and stored in a new folder.

3.4. CNN Architecture

The CNN architecture used to train each model was the SmallerVGGNet model. It is a smaller version of the VGGNet network that was developed by Simonyan and Zisserman (2015). The network uses an architecture with very small 3×3 convolutional filters and increased depth. The volume sizes are reduced by max pooling and the fully connected layers at the end of the network use a softmax classifier.

The model is made of twenty-eight layers excluding the input layer. Each layer is stacked sequentially. All layers have a unique input and output, and each input and output are defined by an input and output shape. Each layer is broken down as follows:

1. The input to the first layer is defined by 96×96×3 (height, width, depth) or by 3×96×96 (depth, height, width) depending on whether the image is “channels first” or “channels last”. The first layer has 32 filters applied using 2D convolution, each is defined by 3×3 kernels. The first layer padding uses the same settings in order to maintain the spatial dimensions of the volume, which ensures that the output volume size matches the input volume size.
2. The next layer in the network is an activation layer that uses the rectified linear unit (ReLU) activation function which increases nonlinearity in images by returning the positive value it receives, or a zero otherwise. ReLU is used because images are naturally non-linear and upon going through the convolution, linearity can be imposed on an image.
3. The third layer in the network uses batch normalization, which normalizes the interlayer outputs of a neural network.
4. 2D Max pooling which reduces the size of the data, number of parameters, amount of computation needed, and controls overfitting.
5. Dropout layer. It randomly drops 25 % of the inputs to help prevent overfitting.
6. 2D convolution with 64 filters with 3×3 kernels and the padding being set to same.
7. ReLU activation layer.
8. Batch normalization layer.
9. 2D convolution with 64 filters with 3×3 kernels and the same padding being set to same.
10. ReLU activation layer.
11. Batch normalization layer.
12. 2D max pooling layer.
13. 25 % dropout layer.
14. 2D convolution with 128 filters with 3×3 kernels and the same padding being set to same.
15. ReLU activation layer.
16. Batch normalization layer.
17. 2D convolution with 128 filters with 3×3 kernels and the same padding being set to same.
18. ReLU activation layer.
19. Batch normalization layer.
20. 2D max pooling layer.
21. 25 % dropout layer.
22. This layer flattens the matrix from the previous layer into a single array of 8192 nodes.
23. The 23rd layer is a dense layer. Each neuron in this layer is fully connected to all neurons in the previous layer. The dense layer is connected to all 8,192 nodes from the layer above it.
24. ReLU activation layer.
25. Batch normalization layer.
26. 25 % dropout layer.
27. The 27th layer is a dense layer that outputs varying numbers of nodes to the final layer.

28. The final layer uses the softmax classifier that outputs probability ranges for the network's predictions.

The above information is represented in the simplified network in Figure 2. The above architecture was defined in a python script, and it was called in the python training script as a module. The training occurred over 100 epochs with 32 images per batch. The data for each model was divided into two categories: 80 % of the data was used for training the model, while the remaining 20 % of the data was used to test the model using the coloured image (RGB) dataset and grayscale image dataset, respectively. Before the images in the dataset are passed through the network for training, they are augmented using the ImageDataGenerator method. This helped to increase the performance of the network especially given the small size of the dataset used. When the training is completed, the output model with weights and biases is saved, as well as the label binarizer, while the training and loss accuracies of the training process is recorded.

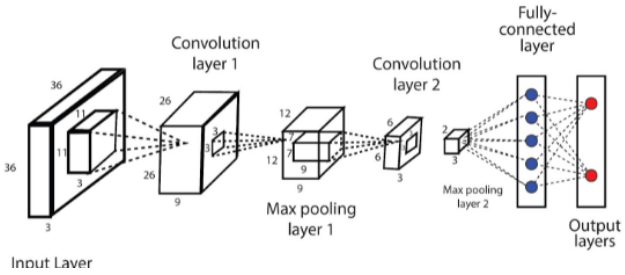

Figure 2. Simplified Representation of CNN Architecture Used to Train the Boot Detection Model

3.5. Dataset Augmentation

Dataset Augmentation also referred to as ImageGenerator is essential when the training model experiences overfitting which results in performance of the training model. Therefore, performing data augmentation allows the training model to have better performance when there is a limited training dataset, without overfitting the training set.

In most situations, when running a deep learning application such as in image classification problems, it is always challenging to obtain new training sets to perform classification tasks. In order to generate more data, data augmentation is performed to increase the size of the dataset. This generates new samples training dataset by performing random transformation of the existing set which decreases overfitting. The dataset augmentation performs the rotation, moving, resizing, position adjustment and contrast change of the original image dataset and generates a new training image set via these operations. These new training image sets do not alter the original image dataset. However, dataset augmentation has the ability to increase the size of the training set up to 50 times the original image dataset.

3.6. High-level Circuit Block Diagram

In the system, a majority of the connections are made to the microcontroller. This can be seen from Figure 3. The green and red LEDs, the solenoid lock, the LCD display, and the buzzer are all one-directional outputs for the microcontroller. There are two one directional 1D inputs connected to the microcontroller, these are the pressure mat and the input button. Two-way connected devices to the microcontroller are the fingerprint sensor and the RPi. The RPi houses only two connections - one to the camera it controls and the other to the microcontroller for communication. Both connections are bidirectional. The RPi receives and sends data between the microcontroller, and the camera receives commands from the RPi and sends pictures back to it.

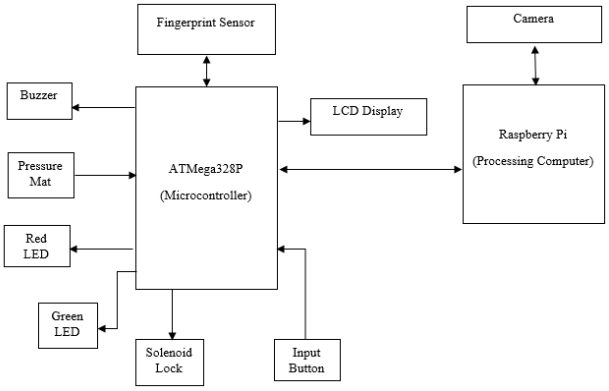

Figure 3. High-level Circuit Block Diagram of the Overall System

3.7 High-level Software Design

The overall processes of the system can be separated into two flow charts of verification and enrollment as described below.

3.7.1. Verification

The verification function is initiated by pressing a momentary push button which is processed by the microcontroller (see Figure 4). If a press is detected, the fingerprint sensor is activated and reads the fingerprint. The sensor will either return a success parameter, or a fail parameter. On fail, the user is notified by LCD that the verification process was unsuccessful, and he needs to press the button to re-initiate the process. On success, the user would be notified that he is required to stand on the pressure mat. The purpose of the pressure mat is to ensure that each user is within the capture area of the system camera. The pressure mat reads the voltage, and if the change in voltage that represents human weight is detected, the system will send a success code from the fingerprint sensor to the RPi via UART Serial Communication. Once this command is sent to the RPi, the microcontroller waits to receive the processed data from it. The RPi will therefore use the received command

to trigger the camera which will take the photo for image processing.

After processing the image, the database is updated, and the results will be returned to the waiting microcontroller. If the user passed inspection, a green LED would indicate success, and the solenoid lock will open, granting the user access to the work area. If the user failed inspection, i.e., if any PPE is missing, a red LED would indicate failure, a buzzer would sound, and the solenoid lock would remain closed.

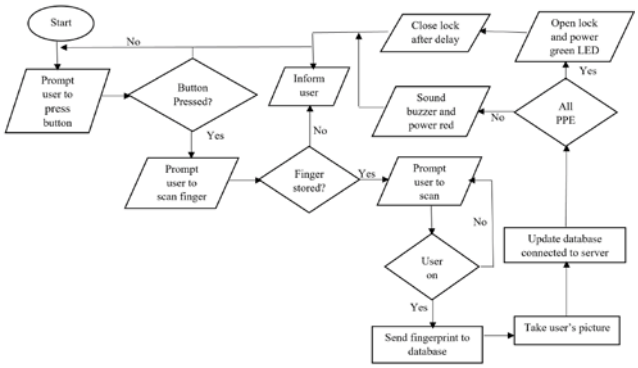

Figure 4. High-level Software Flow Chart for the Design of Validation Process

3.7.2. Enrollment

The enrolled function is an online-based operation. On the website, a manager has the option to add a new user. When the user parameters are entered and the administrator presses the submit button, then a user entry is created in the database. The website code is passed to a python script on addition of a new user that sends the user's fingerprint ID to the microcontroller circuit by way of UART Serial Communication. The microcontroller receives the ID and uses it to start the enrolled function. During the enrollment process, the fingerprint sensor reads the employee fingerprint and returns either a success or fail message which is then returned to the RPi.

On fail, the user is notified, and the code is sent to the RPi where the user entry in the database is deleted (see Figure 5).

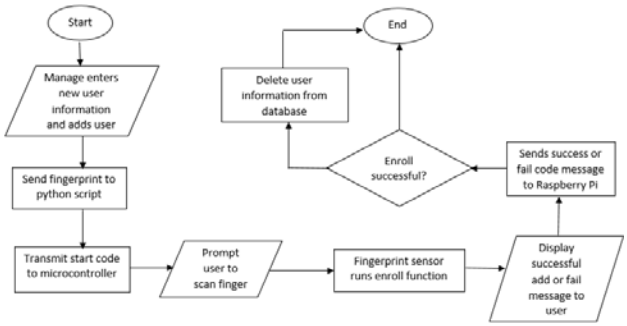

Figure 5. High-level Software Design Flow Chart for the Enrollment Process Combination of Loss and Accuracy Curves

4. Results and Discussion

The accuracies of the trained models are important factors when assessing the effectiveness of the system. To have a good estimate on the accuracies of the models, analysis is conducted on three areas:

- Model performance during training
- Model performance on training data after training
- Model performance on real world samples after training

4.1. Performance during Training

The performance of each model is tracked using the Training and Loss Accuracy plot produced after training each model. Figures 7, 8 and 9 show the loss and accuracy plots for the boot, hat, and vest models, respectively. The performance of the CNN algorithm for each of the models using normal RGB color image dataset and grayscale image dataset are presented in Tables 1 and 2. It is obvious that the accuracy results vary for each of the models and the best performances were obtained for the hat model with 97 % train accuracy, 91.69 % validation accuracy compared to the other two models results, presented in the Table 1.

Table 1. Model Performance during Training Using Normal Colour Image Dataset

Model Acc (%)	Train.	Val. Acc (%)	Train. Loss (%)	Val. Loss (%)
Hat	97	91.69	8.85	31.93
Boot	90.73	80.11	25.22	65.68
Vest	94.81	89.45	14.48	54.02

Table 2. Model Performance during Training Using Grayscale Image Dataset

Model Acc (%)	Train.	Val. Acc (%)	Train. Loss (%)	Val. Loss (%)
Hat	92.26	68.84	20.7	30.49
Boot	89.63	67.47	26.9	12.83
Vest	91.25	65.36	25.2	15.43

The high validation losses obtained in these models (as presented in Table 1) might be a case of overfitting of the dataset at the output layer of the CNN network. By adjusting the weights and biases of the network layer, the models get more accurately trained and the losses can be reduced.

Similarly, Table 2 presents the performance results for each of the models using the CNN algorithm tested on grayscale image dataset. The results obtained for the hat model achieved best performances with 92.26 % train accuracy, 68.84 % validation accuracy, 20.70 % train loss and 30.49 % validation loss compared to the other models results. It is observed from both Tables 1 and 2 that the CNN algorithm performs better with regards to training and validation accuracy results on the normal RGB image dataset for all the models compared to the results obtained

when the CNN algorithm is trained and tested with the grayscale image dataset.

4.2. Post-Training Performance on Training data

After the training process was completed, each model was trained on 10 random images from each image class in its respective dataset. In total there were 100 training images tested. It can be observed from the results (see Table 3) that the boot model achieves better predictions when compared with all other models. The average prediction accuracy of the system after testing was calculated to be 94.17 %. The models were evaluated for their ability to correctly identify the presence of hats, vests, and boots in each image. If none of these categories were present, the model was then evaluated to see if it would predict something else. The formula used to evaluate the Accuracy of each model is:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total Number of Images}} \times 100$$

The overall average performance of the system was calculated by averaging the performance of each individual model,

$$\text{Accuracy} = (\text{Acc. BOOT} + \text{Acc. VEST} + \text{Acc. HAT}) / 3$$

Table 3. Results Obtained after Using the Trained Models to Classify Images from the Normal Colour Training Datasets

Model	Image Category	Total	True Predict	False Predict.
	Sandals	10	10	0
Boot Model	Heels	10	9	1
	Boots	10	9	1
Total		30	28	2
Model Accuracy	**93.33 %**			
	Dress	10	10	0
Vest Model	Tank Top	10	9	1
	T-shirt	10	10	0
	Vest	10	8	2
Total		40	37	3
Model Accuracy	**92.50 %**			
	Hat	10	10	0
Hat Model	Cap	10	9	1
	Random	10	10	0
Total		30	29	1
Model Accuracy	**96.67 %**			

It can be observed that all these models begin with high losses (see Figures 6, 7 and 8). As the models got trained, they recognize patterns and make predictions. The losses are initially high because the models had no prior reference to what it was detected. A loss occurs when the model makes a prediction that is incorrect. It makes an incorrect prediction based on the label given to each image. During training, all incorrect predictions are accounted for, and internal weights and biases of the network are adjusted.

By adjusting the weights and biases, the model gets more trained, and losses can be reduced. The results show that the accuracy tests on training data were higher than the accuracy tests carried out on real world data. This was because the output predictions of the model were restricted to the data categories that it was learned with. Besides, the training accuracies are higher than the validation accuracies because during training, there are some data that are reserved from the network that it will not have seen before. It is on these unrecognized pictures that the model is tested to see how it learns. When validation accuracy increases, there is a sign of good learning. These models trained at a relatively fast rate and end with relatively good validation accuracies.

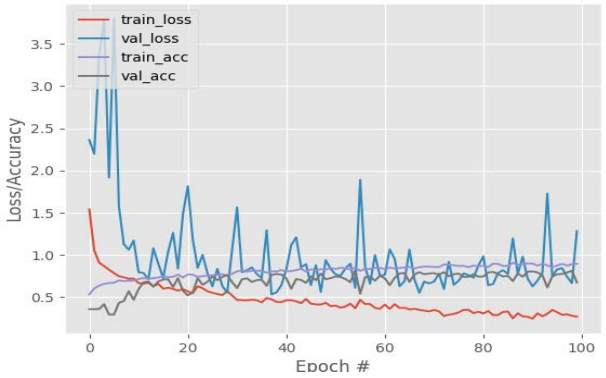

Figure 6. Training Loss and Accuracy Plot for Boot Model Using Normal Color Image Dataset

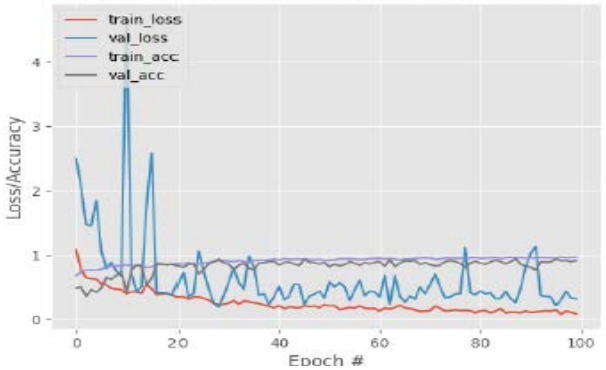

Figure 7. Training Loss and Accuracy Plot for Hat Model Using Normal Color Image Dataset

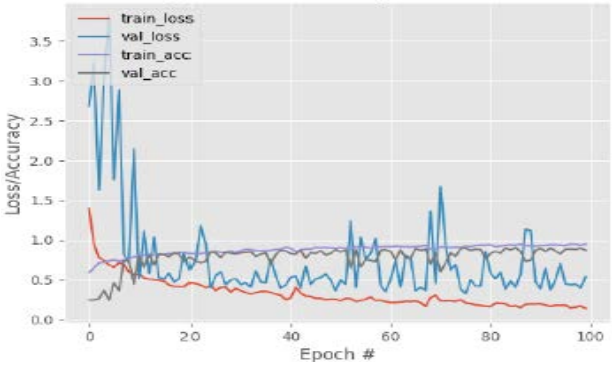

Figure 8. Training Loss and Accuracy Plot for Vest Model Using Normal Color Image Dataset

4.3. Model Performance after Training

The final test of the models occurred by testing these models with the real-world images captured by the camera which all the models have never processed before i.e., new unseen images.

In total, 10 images were tested, and the results are depicted in Table 4. Out of the 10 images analyzed, eight were hats present. The model correctly predicted when no-hat was present and wrongly predicted three of the images as no-hat while these images were actually present. Only seven images were correctly classified by the hat model. Under the average accuracy prediction for the hat, ARHAT was calculated to be 70 %. For the vest model, out of the 10 images, three were contained images of vests. The model was able to determine when no vest was present, however there was an instance where a vest was incorrectly classified. The average accuracy prediction for the vest, ARVEST was 90%. For the boot model, all 10 images had boots present. The model was able to predict nine of the 10 boot instances, setting the average accuracy prediction for the boot ARBOOT as 90 %. Figure 9 shows the correct prediction of the hat and boot, and the prediction without vest present. Figure 10 shows the correct prediction of boot and vest, with the incorrect prediction of hat. The overall average performance of the system was calculated by averaging the performance of each individual model, which gives ACC. TOTAL of 83.33 %.

The results obtained from the current study using CNN method are compared with the results obtained from the three approaches used in Nath et al. (2020) which are YOLO-V3 + Classifiers Neural Network + Decision Tree, YOLO-V3 and YOLO-V3 + Classifiers VGG-16 + ResNet-50 + Xception, to determine the efficiency of the approach. It can be observed from the Table 5 that the best models were obtained from the CNN approach with the better overall accuracy when compared with other models from the three methods reported in Nath et al. (2020).

Figure 9. Results of the Real-World Tests 1 and 2

Figure 10. Results of the Real-World Test

Table 4. Results Obtained after Using the Trained Models to Classify Images from the Training Datasets

Training Data	Image	Hat	Vest	Boot	Classified	Hat	Vest	Boot
Present	1	True	False	True	Predicted	True	False	True
Present	2	True	False	True	Predicted	True	False	True
Present	3	True	False	True	Predicted	False	False	False
Present	4	True	True	True	Predicted	True	True	True
Present	5	True	False	True	Predicted	False	False	True
Present	6	True	False	True	Predicted	False	False	True
Present	7	True	False	True	Predicted	False	False	True
Present	8	False	False	True	Predicted	False	False	True

Table 5. Comparison Results Obtained Using Real Time Images for Different Detection Models

Model	Hat (%)	No Hat (%)	Vest (%)	Boot (%)	Hat+Vest (%)	Overall Accuracy
Yolo-V3+ NN+DT (Nath et al., 2020)	74.29	63.84	74.32	-	63.1	63.1
YOLO-V3 (Nath et al., 2020)	79.81	68.12	84.96	-	72.3	72.3
YOLO-V3+ VGG-16 +ResNet-50+Xception (Nath et al., 2020)	79.81	63.13		-	80.49	67.93
Faster RCNN (Fang et al. (2018)	×	>90	-	×	-	>90
CNN	70	×	90	90	×	83.33

The performance of the proposed CNN approach is compared with Faster RCNN reported in Fang et al. (2018). It is observed that the Faster RCNN method achieves slightly better overall accuracy result than the CNN approach. However, the Faster RCNN approach was only applicable to detect when no hat was worn (Fang et al., 2018).

4. Summary and Conclusions

As safety is important in engineering, this work implemented a CNN method to carry out an image detection for detecting PPE equipment. This system incorporated a fingerprint system which authenticates every worker before granting access into a construction area. The system scans a person, and a convoluted neural network determines if they are wearing safety vests, boots, and hats. A picture of the user is taken and then analyzed by the model. If it sees all three PPE, access is given to a restricted zone by opening a lock, otherwise the lock remains closed. The system keeps track of who is being scanned by using employees' fingerprints, and a manager can keep a record of performance by viewing the activity on a website.

A learning curve is generated showing rate of the learning performance when each model is trained. All curves start off with very high loss. As the models go through more epochs, they are able to recognize more patterns and adjust weights and biases to make better future predictions. This is how the models learn. The losses (inaccurate predictions) decrease over time, and the accuracies increase.

After training, the models were used to classify some images they had never seen before. The models were tested on a total of 10 images with a mixed combination of PPE. The overall average accuracy recorded for this system in real world testing was 83.33 % when a relatively small dataset at a low cost was used. To improve the accuracy of the system, localized object detection can be carried out. This will focus model analysis on a specific area, increasing the confidence of each prediction.

References:

Abbas, M., Mneymneh, B.E., and Khoury, H., (2016), "Use of unmanned aerial vehicle and computer vision in construction safety inspection", *Integrated Solutions for Infrastructure Development*, ISEC Press, North Dakota, United States, ISBN: 978-0-9960437-3-1.

Abbas, M., Mneymneh, B.E., and Khoury, H. (2018), "Assessing on-site construction personnel hazard perception in a Middle Eastern developing country: An interactive graphical approach", *Safety Science*, Vol.103, pp.183-196, https://doi.org/10.1016/j.ssci.2017.10.026.

Arabi, S., Haghighat, A., and Sharma, A. (2013), "A deep learning based solution for construction equipment detection: from development to deployment," *Facta Universitatis, Series: Mathematics and Informatics*, Vol.27, No.3, pp.357-372.

Audebert, N., Le Saux, B., and Lefvre, S. (2017), "Segment-before-detect: Vehicle detection and classification through semantic segmentation of aerial images", *Remote Sensing*, Vol.9, No.4, Available: http://www.mdpi.com/2072-4292/9/4/368.

Azadani, E.N., Hosseinian, S.H., Moradzadel, B., and Hasanpor, P., (2014), "Vision-based material recognition for automated monitoring of construction progress and generating building information modeling from unordered site image collections", *Advanced Engineering informatics*, Vol.28, No.1, pp. 37-49.

Bajracharya, D. (2013), *Real Time Pattern Recognition in Digital Video with Applications to Safety in Construction Sites*, University of Nevada–Las Vegas Theses, Dissertations, Professional Papers, and Capstones.

Barro-Torres, S., Tiago, M., Fern´andez-Caram´es, J., H´ector, P´erez-Iglesias, and Carlos, J.T. (2012), "Real-time personal protective equipment monitoring system", *Computer Communications*, Vol.36, No. 1, pp.42-50. https://doi.org/10.1016/j.comcom.2012.01.005.

Chae, S. and Yoshida, T., (2010) "Application of RFID technology to prevention of collision accident with heavy equipment", *Automation in Construction*, Vol.19, No.3, pp. 368-374.

Chollet, F., (2017) "Xception: Deep learning with depthwise separable convolutions", *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, pp.1800-1807, doi: 10.1109/CVPR.2017.195.

Dimitrov, A. and Golparvar-Fard, M. (2014), "Vision-based material recognition for automated monitoring of construction progress and generating building information modeling from unordered site image collections", *Advanced Engineering Informatics Journal*, Vol.28, pp. 37-49.

Du, S., Shehata, M., and Badawy, W., (2011) "Hard hat detection in video sequences based on face features, motion and color information", *Proceedings of the IEEE 3rd International Conference on Computer Research and Development*, Shanghai, China, March, pp. 25-29, doi: 10.1109/ICCRD.2011.5763846.

Fang, Q., Li, H., Luo, X., Ding, L. Luo, H., Rose, T.M. and An, W., (2018) "Detecting non-hardhat-use by a deep learning method from far-field surveillance videos," *Automation in Construction*, Vol.85, pp.1-9, https://doi.org/10.1016/j.autcon.2017.09.018.

Gualdi, G., Prati, A., and Cucchiara1, R. (2011), "Contextual information and covariance descriptors for people surveillance: An application for safety of construction workers", *EURASIP Journal on Image and Video Processing*, Vol.1, pp.1-16 DOI:10.1155/2011/684819.

Ham, Y., Han, K.K., Lin, J.J., and Golparvar-Fard, M. (2016), "Visual monitoring of civil infrastructure systems via camera-equipped Unmanned Aerial Vehicles (UAVs): A review of related works", *Visualisation in Engineering*, Vol.4, No.1, pp.1-8.

He, K., Gkioxari, G., Dollar, P., and Girshick, R. (2017), "Mask R-CNN", *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, pp.2961–2969, https://doi.org/10.1109/iccv.2017.322

Hung, H.M., Lan, L.T., and Hong, H.S. (2019), "A deep learning-based method for real-time personal protective equipment detection", *Journal of Science and Technique* - Le Quy Don Technical University - No. 199 (6-2019)

Kim, H., Lee, Y., Yim, B., Park, E. and Kim, H. (2016), "On-road object detection using deep neural network", *Proceedings of the IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*, Seoul, Korea (South), pp. 1-4, doi: 10.1109/ICCE-Asia.2016.7804765.

Kyrkou, C., Plastiras, G., Theocharides, T.S., Venieris, I. and Bouganis, C.S. (2018), "DroNet efficient convolutional neural network detector for real-time UAV applications", *Proceedings of the 2018 Design, Automation and Test in Europe Conference and Exhibition (DATE)*, Dresden, Germany, 23 April, pp. 967-972, https://doi.org/10.23919/date. 2018.8342149.

Memarzadeh, M., Golparvar-Fard, M., and Niebles, J.C. (2013), "Automated 2D detection of construction equipment and workers from site video streams using histograms of oriented gradients and colours", *Automation in Construction*, Vol.32, pp.24-37.

Murugan, V., Vijaykumar, V.R., and Nidhila, A., (2019), "Vehicle logo recognition using RCNN for intelligent transportation systems", *Proceedings of the IEEE International Conference on Wireless Communications Signal Processing and Networking (WiSPNET)*, Chennai, India pp. 107-111, doi: 10.1109/WiSPNET45539.2019.9032733.

Naranjo-Torres, J., Mora, M., Hern´andez-Garc´ıa, R., Barrientos, R.J., Fredes C., and Valenzuela, A. (2020), "A review of convolutional neural network applied to fruit image processing", *Applied Sciences*, Vol.10, pp.3443; doi:10.3390/app10103443. www.mdpi.com/journal/applsci

Nath, N.D., Behzadan, A.H., and Paal, S.G. (2020), "Deep learning for site safety: Real-time detection of personal protective equipment", *Automation in Construction*, Vol.112, pp.103085. https://doi.org/10.1016/j.autcon.2020.103085.

New York University (2005), "OSHA personal protective equipment standard: Annual refresher training", Environmental Health & Safety, New York University, New York, N.Y. 10003, Revised August 2005.

Peng, Q., Luo, W., Hong, G., Feng, M., Xia, Y., Yu, L., Hao, X., Wang, X., and Li, M. (2016), "Pedestrian detection for transformer substation based on Gaussian mixture model and YOLO", *Proceedings of the 8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, Hangzhou, China, 27-28 August pp. 562-565, doi: 10.1109/IHMSC.2016.130

Popescu, C., Taljsten, B., Blanksvard, T., and Elfgren, L. (2019), "3D reconstruction of existing concrete bridges using optical methods", *Structure and Infrastructure Engineering*, Vol.15, No.7, pp.912-924, DOI: 10.1080/15732479.2019.1594315.

Rastegari, M., Ordonez, V., Redmon J., and Farhadi, A. (2016), "XNOR-Net: ImageNet classification using binary convolutional neural networks", *Proceedings of European Conference on Computer Vision*, Amsterdam, The Netherlands, October, Vol. 9908, pp.525-542, Springer, Cham. https://doi.org/10.1007/978-3-319-46493-0_32

Roy, R., (2020) "Using YOLO v3 for real-time detection of PPE and Fire", https://towardsdatascience.com/using-yolov3-for-real-timedetection-of-ppe-and-fire 1c671fcc0f0e.

Siddula, M., Dai, F., Ye, Y., and Fan, J. (2016), "Unsupervised feature learning for objects of interest detection in cluttered construction roof site images", *Procedia Engineering* Vol.145, pp. 428-435, https://doi.org/10.1016/j.proeng.2016.04.010.

Simonyan, K., and Zisserman, A. (2015) "Very deep convolutional networks for large-scale image recognition", *Proceedings of the 3rd International Conference on Learning Representations (ICLR2015)*, https://arxiv.org/abs/1409.1556.

UN (1976), *International Covenant on Economic, Social and Cultural Rights*, (Adopted by General Assembly resolution 2200A (XXI) of 16 December 1966 entry into force 3 in accordance with article 27), United Nations. https://www.ohchr.org/Documents/ProfessionalInterest/cescr.pdf

Wahyu Pradana, R.D., Sholahuddin, R.Y., Adhitya, M., Syai'in, R.M., Sudibyo, D.R., Ardhya Abiyoga, A., Jami'in, M.A., Subiyanto, L., Herijono, B., Wahidin, Ruddianto, A., Budianto, A. and Rochiem, N.H., (2019),"Identification system of personal protective equipment using convolutional neural network (CNN) method", *2019 International Symposium on Electronics and Smart Devices (ISESD)*, Badung, Indonesia, pp. 1-6, doi: 10.1109/ISESD.2019.8909629.

Wu, J., Cai, N., Chen, W., Wang, H. and Wang, G., (2019), "Automatic detection of hardhats worn by construction personnel: a deep learning approach and benchmark dataset," *Automation in Construction*, Vol.106, pp.102894, https://doi.org/10.1016/j.autcon.2019. 102894.

Zhou, Y., Nejati, H., Do, T., Cheung, N., and Cheah, L. (2016), "Image-based vehicle analysis using deep neural network: A systematic study", *Proceedings of the 2016 IEEE International Conference on Digital Signal Processing (DSP)*, Beijing, 16-18 October, pp.276-280, doi: 10.1109/ICDSP.2016.7868561.

Authors' Biographical Notes:

Jordon Hayles *is from Spanish Town, Jamaica and is currently a Junior Project Development Analyst at Soleco Energy. He graduated from the Faculty of Engineering at The University of the West Indies, Mona Campus in the Class of 2020 with a Bachelor's of Science Degree with Upper Second Class Honours. He has worked as a Supply-Side Energy System Engineer Associate at the Caribbean Centre for Renewable Energy and Energy Efficiency through the Regional Energy Apprenticeship Programme, an initiative under the Technical Assistance Programme for Sustainable Energy in the Caribbean. Jordon plans to further his professional career in the renewable energy space by working and advocating for sustainable energy implementation and true sustainable development regionally and globally.*

Kolapo Sulaimon Alli *did his B.Tech at the Dept. of Electronic and Electrical Engineering, LAUTECH, Ogbomoso, Nigeria. He obtained both his M.Sc and Ph.D from the Dept. of Electronic and Electrical Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria. He is currently with the Dept. of Electrical Power Engineering, The University of the West Indies, Kingston, Jamaica as a lecturer. His research interests are Instrumentation and Control Systems, Optimal Power Flow, Robust Control, Artificial Intelligence, Virtual and Online Experimentation, Computational Intelligence and its applications, Biomedical Engineering.*

Haniph A. Latchman *received the D.Phil. from Oxford University in 1986 as the Jamaica Rhodes Scholar (1983), and the B.Sc. degree (First Class Honors) from The University of The West Indies (St. Augustine), in 1981. Dr. Latchman's research and teaching focuses on analog and digital communication and control systems, and network engineering. He is also an innovator in the use of computer, communication and information technology to enhance the learning experience for traditional on-campus students as well as in distance education. Dr. Latchman is a Senior Member of the IEEE and has published some 200 technical journal articles and conference proceedings and four books in the general areas of Communication Networks and Control Systems. He served as Professor in Electrical and Computer Engineering at the University of Florida (1986-2016) and now serves as Professor in the Faculty of Engineering at The University of the West Indies (Mona), Jamaica.*

■