A fully automated vision-based system for real-time personal protective detection and monitoring

Thi Lan Le
Hanoi University of Science and Technology
178 PUBLICATIONS **1,586** CITATIONS
SEE PROFILE

Hong Hoang Si
Hanoi University of Science and Technology
52 PUBLICATIONS **1,063** CITATIONS
SEE PROFILE

A fully automated vision-based system for real-time personal protective detection and monitoring

Quang-Huy Tran
School of Electrical Engineering
HUST
Hanoi, Vietnam
huy.tq151712@sis.hust.edu.vn

Thi-Lan Le
MICA Institute
HUST
Hanoi, Vietnam
thi-lan.le@mica.edu.vn

Si-Hong Hoang
School of Electrical Engineering
HUST
Hanoi, Vietnam

***Abstract*—Construction had the most fatal occupational injuries out of all industries due to the high number of annual accidents. There are many solutions to ensure workers' safety and limit these accidents, one of which is to ensure the appropriate use of appropriate personal protective equipment (PPE) specified in safety regulations. However, the monitoring of PPE use that is mainly based on manual inspection is time-consuming and ineffective. This paper proposes a fully automated vision-based system for real-time personal protective detection and monitoring. The proposed system consists of two main components: PPE detection and face detection and recognition. Several experiments have been conducted. The obtained detection accuracy for 6 main PPEs is up to 98% while that for face detection and recognition is 96%. The obtained results have demonstrated the ability to detect PPE and to recognize a face with high precision and recall in real-time.**

***Index Terms*—PPE detection, deep learning, object detection, automatic monitoring.**

I. Introduction

Construction has been identified as one of the most dangerous job sectors. In Vietnam, according to the Report of Ministry of labour-invalids, and social affairs in 2017, there were 8956 occupational accidents nationwide, causing 9173 victims, of which 5.4% of cases involved not wearing personal protection equipment [1].

Several onsite safety regulations have been established to ensure construction workers' safety. In the safety regulations, the appropriate use of appropriate personal protective equipment (PPE) is specified and the contractors must ensure that the regulations are enforced through the monitoring process. The monitoring of the use of PPE is normally conducted in two areas: the site entry and the onsite construction field.

Nowadays, most of the construction fields conduct the monitoring of PPE using manually by inspectors. This work is tedious, time-consuming and ineffective due to the high number of workers to monitor in the field.

Recently, several technologies have been proposed to enhance the construction safety. Among the proposed solutions, computer vision has been widely used [2], [3], [4]. However, most of recent works focus on detecting the use of hardhat on the onsite construction field. Besides hardhat, others equipment such as glove, shoes need to be detected in order to ensure the worker safety. Moreover, the monitoring of PPE using has to be conducted not only on the construction field but also at site entry.

Besides the PPE detection, the face and identity of the workers have to be defined. However, this is usually performed by manually or by others technologies that require extra hardware plate-form such as RFID (Radio Frequency Identification). In this paper, for the first time, we propose a fully automated vision-based PPE detection and monitoring. The proposed system consists of two main components: PPE detection and face detection and recognition. The main aim of PPE detection is to determine the presence of required PPE while face detection and recognition aims to determine the identity of the workers.

II. Related works

The enhancement of onsite construction safety has been increasingly received the attention of researchers and industrial practitioners. The proposed solutions can be divided into two groups: a noncomputer vision-based and computer vision-based technique.

The work of Kelm et al. [5] falls into the first group. The authors designed an RFID-based portal to check whether the workers' personal protective equipment (PPE) complied with the corresponding specifications. Dong et al. [6] use real-time location system (RTLS) and virtual construction are developed for worker's location tracking to decide whether the worker should wear a helmet and give a warning, while the silicone single-point sensor is designed to show whether the PPE is used properly for a further behavior assessment. However, these methods are limited in their respective ways. For example, the worker's identification card only indicates that the distance between the worker and PPE is close and the loss of sensors may be a consideration when applying.

Concerning computer vision-based techniques, taking into account the important role of hardhat, several works have been done for hardhat detection. Rubaiyat, et al. [7] incorporate a Histogram of Oriented Gradient (HOG) with Circle Hough Transform (CHT) to obtain the features of workers and hardhats. Du et al. [8] combine face detection and hardhat detection based on color information.

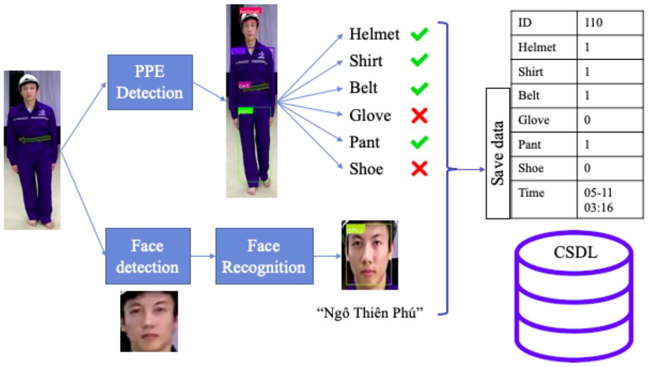
In recent years, deep learning developed extremely fast based on a huge amount of training data and improved computing capabilities of computers. The results for the problem of classification or detection of objects are increasingly improved. The object detection methods can be divided into two main approaches: one-stage (e.g., Faster Region-based Convolutional Neural Networks (Faster R-CNN) [9]) and two-stage detection (e.g., You Only Look Once (YOLO) [10], Single Shot Multibox Detector (SSD) [11]). In general, one-stage detectors is faster and simpler. Inspired by the performance of deep learning-based object detection method, in [12], [13], Fang et al. applied the Faster R-CNN algorithm to detect the absence of hardhats and discovering non-certified work. The proposed method has been evaluated in different situations and shown its high performance.

However, most of the current works focus on detecting the use of hardhat on the construction site. Besides hardhat, other equipment have to be detected. Moreover, to make sure that the workers use proper PPE, the present of PPE should be checked at the entry point of the construction field.

III. Proposed framework

A. Overview

Currently, monitoring the use of personal protective equipment is still done manually. It is costly and inaccurate because a large number of workers need to be checked over a period of time. In response to these restrictions, the overall objective of this paper is to propose a novel solution to address the unresolved problem of reliably identifying workers who comply with safety regulations at the entry point of the construction field. The proposed system illustrated in Fig. 1 consists of two main steps: PPE detection and face detection and recognition. The main aim of PPE detection is to determine the presence of required PPEs while face detection and recognition aim to determine the identity of the workers. Inspired from impressive results of deep convolutional neural network for different computer vision tasks, in this paper, PPEs and face are detected thanks to YOLO (You only look once) network while FaceNet is fine-tuned for face recognition.



ID	110
Helmet	1
Shirt	1
Belt	1
Glove	0
Pant	1
Shoe	0
Time	05-11 03:16

Fig. 1: The proposed system for real time PPE detection and monitoring at entry point of the construction field.

B. PPE and face detection

In our work, we employ the YOLO network for PPE and face detection. YOLO network has been introduced by Joseph Remon's team [10] for object detection.

The first version of YOLO is named YOLO v1. Different from two-stage methods, the core idea behind this fast detector is a single convolutional network consisting of convolutional layers followed by 2 fully connected layers. A single neural network predicts bounding boxes and class probabilities directly from full images in one evaluation. For this, YOLO network divides the image into an S × S grids. For each grid cell, it predicts B bounding boxes, their confidence score and C class probabilities as illustrated in Fig 2. The design of YOLO enables end-to-end training and real-time speeds while maintaining high average precision but still has limitations when detecting the small objects appeared in groups. Therefore, YOLO v2 is introduced [14] and after that is YOLO v3 [15]. YOLO v3 has significant improvements in both accuracy and speed, improved the capable of detecting small objects.

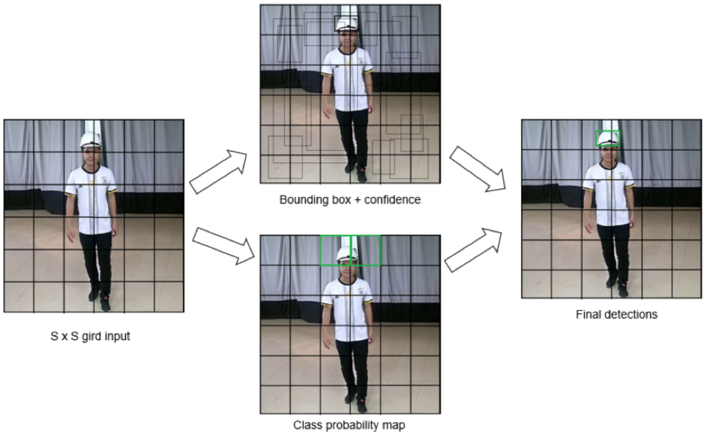


Fig. 2: Illustration of PPE or face detection based on YOLO network.

The YOLOv3 uses a variant of Darknet, which originally a 53 layer network trained on Imagenet called Darknet-53. Similar to the VGG models it uses mostly 3 × 3 filters and double the number of channels after every pooling step. In addition, it uses Network in Network (NIN) and batches normalization to stabilize the training process, speed up convergence, and regularize the model. Its structure is detailed on the layer type, input, output, filter number, filter size (including stride), as shown in Table I. We finetune YOLO from the pre-trained model on Imagenet for PPE and face detection with the weight decay of 0.0005, the momentum of 0.9, the learning rate of 10^{-3}. We also employ data augmentation including random crops, rotations, and hue, saturation, and exposure shifts. Besides, we use Multi-Scale Training for each of 10 batches with image dimension size to improve network accuracy with a variety of input dimensions [14].

TABLE I: Network architecture YOLOv3

	Type	Filters	Size	Output
	Convolutional	32	3x3	256x256
	Convolutional	64	3x3/2	128x128
	Convolutional	32	1x1	
1x	Convolutional	64	3x3	
	Residual			128x128
	Convolutional	128	3x3/2	64x64
	Convolutional	64	1x1	
2x	Convolutional	128	3x3	
	Residual			64x64
	Convolutional	256	3x3/2	32x32
	Convolutional	128	1x1	
8x	Convolutional	256	3x3	
	Residual			32x32
	Convolutional	512	3x3/2	16x16
	Convolutional	256	1x1	
8x	Convolutional	512	3x3	
	Residual			16x16
	Convolutional	1024	3x3/2	8x8
	Convolutional	512	1x1	
4x	Convolutional	1024	3x3	
	Residual			8x8
	Avgpool		Global	
	Connected		1000	
	Softmax			

The newer architecture boasts of residual skip connections and upsampling. The most salient feature of v3 is that it makes detections at three different scales. YOLO is a fully convolutional network and its eventual output is generated by applying a 1 x 1 kernel on a feature map. In YOLO v3, the detection is done by applying 1 x 1 detection kernels on feature maps of three different sizes at three different places in the network. The detection filters are calculated according to the number of classes and the attributes of the bounding box required as Equ. 1

$$Filters = bboxes * (5 + classes) \tag{1}$$

bboxs is the number of bounding box per cell, classes are the number of classes and "5" is for the 4 bounding box attributes and one object confidence as shown in Fig. 3. In this work, bboxes are 3 and classes are 7. In YOLO v3 trained on this dataset, 3 bounding boxes and 7 classes, so the kernel size is 1 x 1 x 36. YOLO v3 makes a prediction at three scales, which are precisely given by downsampling the dimensions of the input image by 32, 16 and 8 respectively. And at each scale gives us the corresponding output features map : 13 x 13 x 36, 26 x 26 x 36 and 52 x 52 x36. The 13 x 13 layer is responsible for detecting large objects, whereas the 52 x 52 layer detects the smaller objects, with the 26 x 26 layer detecting medium objects. Here is a comparative analysis of different objects picked in the same object by different layers.

The structure of an output cell as illustrated in Fig 3. From the output of the last layer, several bounding boxes can be generated for the same object. To filter out these bounding boxes, we use Non-maximum suppression and threshold. After that, we get the coordinates of bounding boxes with their highest probability class without duplicating the bounding boxes of the same object. These bounding boxes will be compared with those of ground-truth to compute the model's accuracy. In our work, we propose to use YOLO for detecting 6 main PPEs at the entry point of the construction field and workers face (see Fig. 4). The use of YOLO has two main advantages in comparison with state of the art methods proposed for PPE detecting. Firstly, YOLO is very fast. It can obtain the frame rate of 20fps at 416 × 416 resolution on a GPU-supported work station.

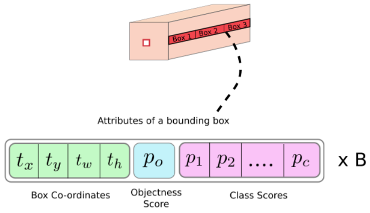


Fig. 3: Structure of one output cell in YOLO

Secondly, YOLO sees the entire image during training and test time. This is different from region proposal-based methods which only consider features within the bounding boxes. For that reason, YOLO has less than half the number of background errors than Fast R-CNN. [16].

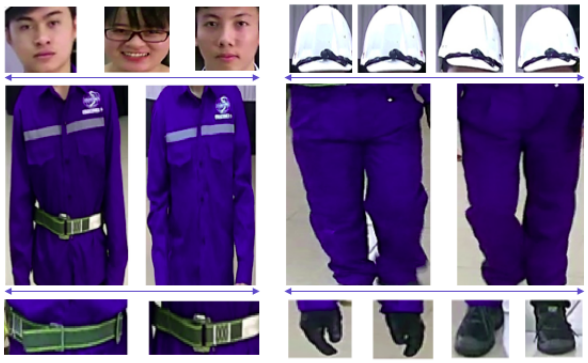


Fig. 4: Faces and six types of PPEs: hardhat, shirt, gloves, belt, pant, shoes.

It is important to note that the PPE and face detection can be done by using any image captured when the worker comes into the monitoring zone. However, to increase the reliability of the system, we determine the detection results from an image sequence based on the detection result of each image through voting technique. For this, a series of images during 3s are collected and processed. The number of frames is from 20 to 40 depending on the speed of the person. For each class, the ratio between the number of frames where the interested class is detected and the total of frames. If the ratio is equal or less than 0.5, the system will confirm the detection of the class. An example is illustrated in Fig 5. The green frames corresponds to the images that are detected correctly while the red frames are images with incorrect detection (e.g. miss detection of gloves and shoes).

Thanks to the voting technique, correct detection decisions can be made in some cases even with false alarm and miss detection at some images in the sequence.

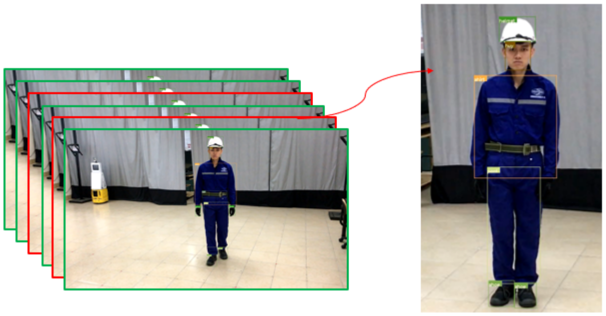


Fig. 5: Illustration of the voting technique used for PPE and face detection. The green frame corresponds to the images in which the PPE and face are detected correctly while the red frames are images with incorrect detection (miss detection of gloves).

C. *Face recognition*

FaceNet is introduced in 2015 by Google researchers, which is a start-of-art face recognition, verification, and clustering neural network. It has a 22-layers deep neural network that directly trains its output to be a 128-dimensional embedding. Once the FaceNet model having been trained with triplet loss for different classes of faces to capture the similarities and differences between them, the 128-dimensional embedding returned by the FaceNet model can be used to clusters faces effectively. Once such a vector space (embedding) is created, tasks such as face recognition, verification and clustering can be easily implemented using standard techniques with FaceNet embeddings as feature vectors. In a way, the distance would be closer for similar faces and further away for non-similar faces.

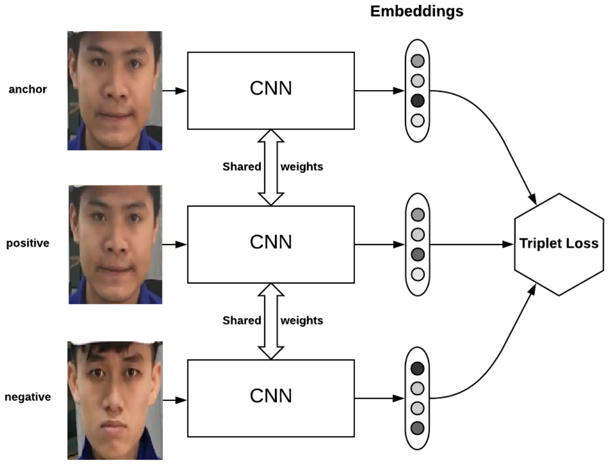


Fig. 6: Illustration of face recognition based on FaceNet [17]

The FaceNet model is based on the Inception which is the winning approach for ImageNet includes to extract features from images then uses an SVM for classification. The special in FaceNet is that it uses the Triplet loss (refer to Fig. 6) to minimize distance between similar faces and maximize the distance to faces that are not similar, so the FaceNet can distinguish very accurate person to person. FaceNet consists of a batch input layer and a deep CNN followed by L2 normalization, which results in the face embedding.

A training triplet contains three exemplars: anchor (an image of a person), positive (another image of the same person) and negative (an image of a different person). The triplet loss measures two Euclidean distances: one distance between the anchor and the positive image named A and another between the anchor and the negative image named B. The training process aims to reduce A while maximise B, such that similar images lie close to each other and distinct images lie far away in the embedding space as illustrated in Fig 7.

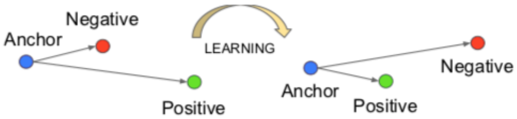


Fig. 7: The **triplet loss** minimizes the distance between an anchor and a positive, both of which have the same identity, and maximizes the distance between the anchor and a negative of a different identity.

The embedding is represented by f(x) $\in R^d$. It embeds an image x into a d-dimensional Euclidean space. The triplet loss learns to ensure that an image x_i^a (anchor) of a specific person is closer to all other images x_i^p (positive) of same person than it is to any image x_i^n (negative) of any other person. This is visualized in Fig 7. The loss fusion L that is being minimized is as follows:

$$\sum_{i}^{n} \left[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha \right] \qquad (2)$$

where α is a margin that is enforced between positive and negative pairs.

IV. Experimental results

A. *Dataset*

There is no off-the-shelf PPE dataset available and because the standards of personal protective equipment for construction site workers are clearly defined according to the nature of the works, we selected 6 types of typical protective equipment and face of workers in the construction site to create a data set for training and evaluation as illustrated in Fig. 4 . The dataset is collected outdoors by IP camera and collected over several days at different times so it has different lighting condition. The image acquisition camera is placed 2,5 meters high, the length of the road segment is from 3 to 4 meters. The images are captured at a rate of 10 FPS (frames per second). After collecting full images of models, we label each object on images from our dataset with the annotation tool Yolo-mark [18]. This work included marking bounding boxes of objects and generating annotation files that contain coordinates and box size parameters along with the label type. The dataset is collected with 10 subjects. The dataset is divided into 2 main parts: training set and testing set. The training set consists of 12,000 images of 6 types of PPE and 4,500 images of 10 different subjects. The number of testing images for PPE detection is 1200 images while those for face detection and recognition is 500. Detail information is given in Table II and Table III.

TABLE II: Number of images for PPE detetion training and testing

	Hardhat	Shirt	Glove	Belt	Pant	Shoe	**Total**
Training	3,000	2,000	2,000	2,000	1,000	2,000	**12,000**
Testing	350	250	200	150	100	150	**1,200**

TABLE III: Number of images used for face detection and recognition training and testing

	#subjects	#face images for each subject	**Total number**
Training	10	450	**4,500**
Testing	10	50	**500**

B. Evaluation measures

The metrics that we use to evaluate the model's accuracy are Precision, Recall and F1-score for object and face detection and accuracy metric for face classification. Firstly, we have to define the meaning of TP (true positive), FP (false positive), and FN (false negative). True Positive results when an object is correctly identified with IOU (Intersection over Union) between ground truth and bounding box predicted to be greater than a threshold (in our case, this threshold is set by 0.5). False Positive is the result of the wrong identity, which means that the wrong class can be identified, or $IOU < 0.5$. False Negative is the result of miss identification, meaning that the object appears but is not recognized.

Precision, Recall, and F1-score are defined as follows:

$$F1\ score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{3}$$

where $Precision = \frac{TP}{TP+FP}$, $Recall = \frac{TP}{TP+FN}$

C. Experimental results

1) Evaluation of PPE detection and face detection and recognition at image level: To determine the number of iteration for the network, we evaluate the network with different iterations. The obtained results are shown in Table IV with network resolution being 416 × 416 and threshold equal to 0.25. As is shown, training with a large number of iterations gives better results and allows us to avoid over-fitting. Based on these results, we use the weights after 4,000 iterations.

The performance of the method on the testing dataset with three network resolutions is shown in Tab. V.

TABLE IV: Obtained results for PPE and face detection with different numbers of iterations.

	Iteration number				
	1,000	2,000	3,000	4,000	5,000
Precision	0.85	0.92	0.94	0.98	0.95
Recall	0.86	0.92	0.99	0.96	0.99
F1 Score	0.85	0.92	0.96	**0.97**	0.97

Concerning face recognition, once a face detection module has detected the faces in images, the face recognition method based on FaceNet is executed. The face detection results are shown in the right-most column in Table V while the obtained face recognition accuracy is 96% for 500 testing images. The accuracy of face recognition is quite high. However, in some cases, the face is not correctly detected due to its small size in images (see Fig. 8) or the presence of emotions (see Fig. 9).

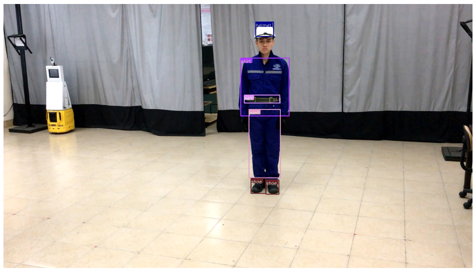


Fig. 8: Face is not correctly detected due to the small size.

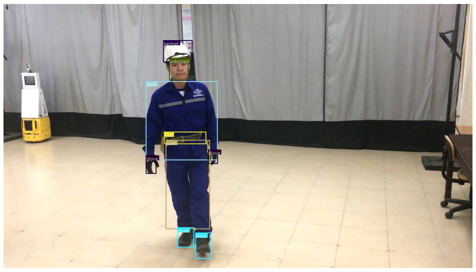


Fig. 9: Face is detected but incorrectly recognized due to the presence of emotions.

2) Evaluation of whole system: To evaluate the whole system, we prepare 4 scenarios. In all scenarios, the subjects do not wear all required PPE. In this evaluation, the chosen network resolution is 416 × 416. For each scenario, we perform 5 tests and evaluate the accuracy that is the ratio between the correct recognized test and the total number of tests. One test is considered as a correct recognized test if all interested PPE are correctly detected and the identity of the person is correctly classified. As shown in Tab. VI, thanks to the voting technique and the robustness of PPE detection and face recognition, the system produced the correct results even some PPEs are omitted or incorrectly detected and the identity of the person is wrong classified in some frames.

TABLE V: Obtained results for PPE and face detection with different network resolutions.

Network resolution	**Frame rate**	**Evaluation metric**	**Hardhat**	**Shirt**	**Glove**	**Belt**	**Pant**	**Shoes**	**Face**
320 × 320	10Hz	FN	8	5	0	5	5	0	0
		FP	10	10	14	12	13	15	0
		TP	332	245	210	155	105	150	500
		Precision	0.97	0.96	0.94	0.93	0.89	0.91	1
		Recall	0.98	0.98	1	0.97	0.95	1	1
		F1 Score	0.97	0.97	**0.97**	**0.95**	0.92	0.95	1
416 × 416	8Hz	FN	6	5	2	6	6	0	0
		FP	5	7	14	12	8	13	0
		TP	334	245	208	154	104	150	500
		Precision	0.99	0.97	0.94	0.93	0.93	0.92	1
		Recall	0.98	0.98	0.99	0.96	0.94	1	1
		F1 Score	**0.98**	**0.97**	0.96	0.94	**0.93**	**0.96**	**1**
608 × 608	6Hz	FN	10	5	2	6	6	0	0
		FP	6	7	15	11	14	15	0
		TP	330	245	208	154	104	150	500
		Precision	0.98	0.97	0.93	0.93	0.88	0.91	1
		Recall	0.97	0.98	0.99	0.96	0.94	0.99	1
		F1 Score	0.97	0.97	0.96	0.94	0.91	0.95	1

TABLE VI: Detection and recognition results of the proposed system with different scenarios.

Types of equipment	#tests	#false alarms	Accuracy (%)
Hardhat, shirt, belt, pant, shoe	5	0	100
Hardhat, shirt, pant, shoe	5	0	100
Shirt, glove,belt, pant, shoe	5	0	100
Hardhat, pant, shoe	5	0	100

V. Conclusions and Future works

Construction activities are dangerous and very risky. Work safety management in construction despite being concerned, still leaves many occupational accidents. From the practical needs and the urgent requirements laid down at the construction sites, we have developed a PPE detection and identification system to replace manual inspection and monitoring, and allow management and retrieval of data. Along with that is to build a database of images of 6 types of common protective equipment at construction sites. This database along with labeling information can be used for the research community. The results of detection and identification on the test dataset show that the model gives good results in both accuracy and miss rate. In the future, we will expand the data set with other equipment with different conditions. Furthermore, the code will be optimized to increase the speed of the system.

References

[1] M. of labour invalids and social affairs. (2018) Notification notice of the situation of labor accidents in 2017. [Online]. Available: http://vnniosh.vn/chitiet$_NCKH/id/7559/Thong-bao-tinh-hinh-tai-nan-lao-dong-nam-2017$

[2] B. E. Mneymneh, M. Abbas, and H. Khoury, "Automated hardhat detection for construction safety applications," *Procedia Engineering*, vol. 196, pp. 895 – 902, 2017, creative Construction Conference 2017, CCC 2017, 19-22 June 2017, Primosten, Croatia. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877705817331430

[3] ——, "Vision-based framework for intelligent monitoring of hardhat wearing on construction sites," *Journal of Computing in Civil Engineering*, vol. 32, 2019.

[4] Z. Zhenhua, P. Man-Woo, and E. Nehad, "Automated monitoring of hardhats wearing for onsite safety enhancement," in *International Construction Specialty Conference of the Canadian Society for Civil Engineering (ICSC)*, 2015.

[5] A. Kelm, L. Laußat, A. Meins-Becker, D. Platz, M. J. Khazaee, A. M. Costin, M. Helmus, and J. Teizer, "Mobile passive radio frequency identification (rfid) portal for automated and rapid control of personal protective equipment (ppe) on construction sites," *Automation in Construction*, vol. 36, pp. 38 – 52, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0926580513001234

[6] S. Dong, Q. He, H. Li, and Q. Yin, *Automated PPE Misuse Identification and Assessment for Safety Performance Enhancement*. [Online]. Available: https://ascelibrary.org/doi/abs/10.1061/9780784479377.024

[7] A. H. M. Rubaiyat, T. T. Toma, M. Kalantari-Khandani, S. A. Rahman, L. Chen, Y. Ye, and C. S. Pan, "Automatic detection of helmet uses for construction safety," in *2016 IEEE/WIC/ACM International Conference on Web Intelligence Workshops (WIW)*, Oct 2016, pp. 135–142.

[8] S. Du, M. Shehata, and W. Badawy, "Hard hat detection in video sequences based on face features, motion and color information," in *2011 3rd International Conference on Computer Research and Development*, vol. 4, March 2011, pp. 25–29.

[9] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *CoRR*, vol. abs/1506.01497, 2015. [Online]. Available: http://arxiv.org/abs/1506.01497

[10] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *CoRR*, vol. abs/1506.02640, 2015. [Online]. Available: http://arxiv.org/abs/1506.02640

[11] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg, "SSD: single shot multibox detector," *CoRR*, vol. abs/1512.02325, 2015. [Online]. Available: http://arxiv.org/abs/1512.02325

[12] Q. Fang, H. Li, X. Luo, L. Ding, H. Luo, T. Rose, and W. An, "Detecting non-hardhat-use by a deep learning method from far-field surveillance videos," *Automation in Construction*, vol. 85, pp. 1–9, 01 2018.

[13] Q. Fang, H. Li, X. Luo, L. Ding, T. M. Rose, W. An, and Y. Yu, "A deep learning-based method for detecting non-certified work on construction sites," *Advanced Engineering Informatics*, vol. 35, pp. 56 – 68, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1474034617303166

[14] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," *CoRR*, vol. abs/1612.08242, 2016. [Online]. Available: http://arxiv.org/abs/1612.08242

[15] ——, "Yolov3: An incremental improvement," *CoRR*, vol. abs/1804.02767, 2018. [Online]. Available: http://arxiv.org/abs/1804.02767

[16] R. B. Girshick, "Fast R-CNN," *CoRR*, vol. abs/1504.08083, 2015. [Online]. Available: http://arxiv.org/abs/1504.08083

[17] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," *CoRR*, vol. abs/1503.03832, 2015. [Online]. Available: http://arxiv.org/abs/1503.03832

[18] AlexeyAB. (2016) GitHub,Yolo-mark gui for marking bounded boxes of objects in images for training yolo. [Online]. Available: https://github.com/AlexeyAB/Yolo$_{mark}$