A Real-Time Automated Approach for Ensuring Proper Use of Personal Protective Equipment (PPE) in Construction Site

Shi Chen[1](✉), Kazuyuki Demachi[1], and Manabu Tsunokai[2]

[1] The University of Tokyo, Tokyo, Japan
shichen@g.ecc.u-tokyo.ac.jp
[2] IIU Corporation, Tokyo, Japan

Abstract. Construction sites are one of the most perilous environments where many potential hazards may occur. Even though workers are trained to stay away from potential dangers, there are still many types of risks that can occur within only a few minutes of carelessness. Personal Protective Equipment (PPE) is an important safety measure used to protect construction workers from accidents. However, PPE usage is not strictly enforced among workers due to all kinds of reasons. This paper proposes the combination of deep learning-based object detection and individual detection using geometry relationships analysis to automatically identify non-PPE-use (NPU); i.e., if a worker is wearing hardhat, eye protection visors, dust masks, or both, to help to facilitate the safety monitoring work of construction workers to ensure PPE are appropriately used. The experimental results demonstrate that the approach was capable of detecting NPU workers with high precision (84.13%) and recall rate (93.10%) while ensuring real-time performance (7.95 FPS on average).

Keywords: Construction safety · Personal Protective Equipment (PPE) · Deep learning · Object detection

1 Introduction

Construction work is much more dangerous than most other occupations, where many potential hazards may occur. According to the United States' Bureau of Labor Statistics, the number of construction fatalities in the US has gradually increased from 933 to 1013 between 2014 and 2017 [1]. Similarly, in Japan, there were 926 construction fatal accidents during 2016–2018 and the Ministry of Health, Labor and Welfare of Japan is aiming to reduce construction fatalities by at least 15% (relative to the 2017 level) by 2022 [2].

The construction fatalities are always caused by the combination of different factors that involve occupational risk exposures (e.g., electricity), tools (e.g., grinder), equipment (e.g., crane), and environment (e.g., dust and noise) in the construction sites. The consequences of head injuries caused by falling from height or being stuck by vehicles and other moving plants and equipment are one of the most serious of all construction accidents. A total of 2,210 construction fatalities occurred because of traumatic brain injury (TBI) which represented 25% of all construction fatalities during

E. Toledo Santos and S. Scheer (Eds.): ICCCBE 2020, LNCE 98, pp. 1115–1126, 2021.
https://doi.org/10.1007/978-3-030-51295-8_77

2003 and 2010 [3]. The Occupational Safety and Health Administration (OSHA) in the United States stipulated that workers working in areas where there is a possible danger of head injury from impact, or from falling or flying objects, or from electrical shock and burns shall be protected by hardhats [4]. Furthermore, construction-related occupational eye injuries are an important cause of vision loss. According to the National Institute for Occupational Safety and Health (NIOSH), an average of 2,000 U.S. workers require medical treatment for job-related eye injuries every day [5]. The majority of construction-related ocular injuries are preventable. The reasons cited for the majority of ocular injuries include the non-wearing of available eye protection or wearing of inappropriate eye protection for the current task [6]. On the other hand, fine dust and particles, gases and vapors can be produced when using machine tools and silica dust from bricks can cause lung and airway diseases such as emphysema, bronchitis and silicosis, and may increase cancer risks. Personal Protective Equipment (PPE) such as respirators or dust masks are used to controls these hazards [7]. OSHA indicates the workers shall be ensured to wear eye or face protection when exposed to eye or face hazards from flying particles, molten metal, liquid chemicals, acids or caustic liquids, chemical gases or vapors, or potentially injurious light radiation [8]. Nonetheless, the workers do not precisely follow the construction site's safety regulations due to all kinds of reasons, even if they have been previously educated and trained. Thus, an automated monitoring approach is necessary to be conducted to help to facilitate the safety monitoring work of construction workers to ensure PPE are appropriately used.

In this paper, we present a real-time approach to automatically identify non-PPE-use (NPU) in response to the limitations of monitoring systems in construction sites. The goal is to detect hardhats, eye protection visors and dust masks in each observed frame, and to identify whether individuals on construction sites are wearing PPE. Firstly, we detect entities in each observed frame using deep learning-based models: PPE(s) are recognized and localized using Yolov3 [9] and individual(s) are characterized by extracting their joint positions using OpenPose [10]. Subsequently, we associate detected PPE(s) with the detected individual(s). Finally, we identify whether PPE(s) are appropriately used by analyzing the geometric relationships of detected PPE(s) and individual(s). To summarize, this work contains the following contributions: (1) In contrast to commonly used object detection-based NPU identification approaches, this study provides a novel solution to automatically identify any failure to wear PPE by the combination of object detection and individual detection using geometry relationships analysis which is more effective and robust with viewpoint changes and different individual postures. (2) To the authors' knowledge, almost no research studies have been conducted concerning multi-class NPU identification other than NHU. This paper explores the possibility of multi-class NPU identification for non-hardhat-use (NHU), non-mask-use (NMU), and non-visor-use (NVU). (3) This paper demonstrates that the proposed approach can rapidly handle NPU identification to meet the industrial requirements of real-time processing.

2 Related Works

At present, a number of methods have been studied for automatic NPU detection [11–16], which can be divided into two categories: sensor-based detection and vision-based detection.

Sensor-based detection primarily relies on remote locating and tracking techniques, such as radio frequency identification (RFID) and wireless local area networks (WLANs). Kelm et al. [11] developed a mobile RFID portal for checking PPE compliance of personnel. The RFID readers were located at the construction site entrance, and therefore only the individuals who enter the construction site can be checked. Dong et al. [12] developed the real-time location system (RTLS) and virtual construction for worker location tracking to identify whether the worker should wear a hardhat by placing a pressure sensor in the hardhat to identify whether a hardhat was being worn, and if not, to transmit a warning. Generally, existing sensor-based methods relying on physical tags or sensors employed in PPE have difficulty in identifying whether any individuals on the sites are wearing PPE or not. Moreover, the practical use of the tags or sensors will lead to high costs due to the large volume of devices required.

Vision-based methods are nonintrusive and less device-intensive because of the wide application of on-site surveillance cameras. Shrestha, et al. [13] use edge detection algorithms to recognize the edge of objects inside the upper head region i.e. hardhats. However, this method relies on the recognition of facial features, therefore workers who turn their face away from the cameras cannot be recognized. Park et al. [14] proposed a vision-based NHU detection method that detects both a human body and a hardhat simultaneously in each obtained frame. The detected human body region and hardhat region are then matched for the detection of NHU. In general, these methods rely heavily on hand-crafted features to analyze the individuals. Consequently, they may fail in the case of complicated scenes with weather variability, different viewpoints and/or occlusions.

Recently, deep learning-based object detection methods have shown remarkable performance on most visual tasks in the construction industry. Fang et al. [15] proposed a Faster R-CNN based method to detect construction workers' NHU automatically. A total of 81,000 image frames were collected from various construction sites as a training dataset to train the Faster R-CNN model. In the training phase, the worker-of-interest (WOI) in the image was annotated as the ground truth for training. In the test phase, the NHU workers were detected and the rest were considered the background. Wu et al. [16] deployed a Single Shot Multibox Detector (SSD) based model combined with the presented reverse progressive attention (RPA) to propagate contextual information back to bottom layers discriminately. A benchmark dataset GDUT-HWD was generated by downloading Internet images retrieved by search engines to train the SSD-RPA model. However, existing deep learning-based detection methods are mainly focused on learning to localize only PPE(s) or NPU individual(s) in the obtained images, which may fail in cases of uncommon human gestures or appearance. Also, almost no research studies have been conducted concerning multi-class NPU identification other than NHU and are limited in practical application to real scenarios. In response to these limitations, the overall objective of this paper is to develop a new approach for monitoring workers and evaluate whether the proposed approach could be used to detect failure use of hardhats, visors and masks in the construction site.

3 Methodology

The associated generic pipeline is illustrated in Fig. 1, and it follows the subsequent stages:

(1) For each observed frame, individual(s) are detected, together with their keypoints coordinates using OpenPose [10]. PPE(s) are recognized and localized using YOLOv3 [9].
(2) Detected PPE(s) are associated with the detected individual(s).
(3) NPU identification is performed by analyzing the geometry relationships of the individual's keypoints and the detected PPE(s).

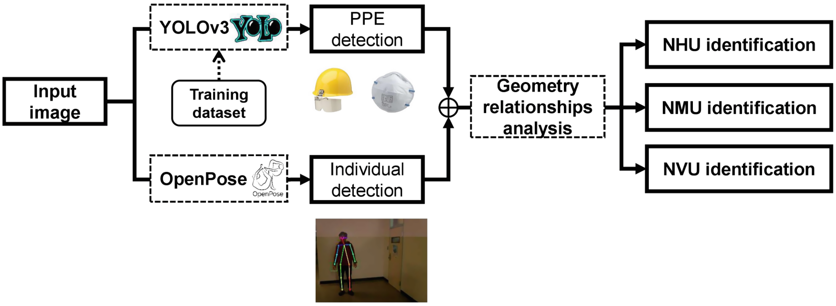


Fig. 1. Generic pipeline of the proposed approach.

3.1 Entity Detection

PPE Detection. We deploy YOLOv3 [9] to detect the PPE(s). YOLOv3 uses Darknet-53, a CNN with 53 layers, to extract image features. Subsequently, YOLOv3 makes predictions at three scales, which are precisely given by downsampling the dimensions of the input image by 32, 16 and 8 respectively. This method allows YOLOv3 to get more meaningful semantic information from the upsampled features and finer-grained information from the earlier feature map. The prediction result of the network is a 3-d tensor that encodes bounding box, objectness score and prediction over classes:

$$\mathrm{N} \times \mathrm{N} \times (3 * (4 + 1 + \mathrm{C})) \tag{1}$$

where $N \times N$ is the number of the grid cells of the system, and C is the number of the classes to train the network on. Besides, YOLOv3 predicts a confidence score for each bounding box using logistic regression, which is overwhelmingly beneficial especially considering that one image might enjoy multiple labels, and not all the labels are guaranteed to be mutually exclusive.

In a word, predictions of YOLOv3 are carried out from one single network, which can be trained end-to-end to improve accuracy. Higher efficiency and better performance on small object detection make YOLOv3 a reasonable option for real-time processing for industrial purposes.

Individual Detection. We characterize the workers' postures by extracting the joint positions of the individual in the images using OpenPose [10], which processes images through a two-branch multi-stage Convolutional Neural Network (CNN) and uses Part Affinity Fields (PAFs) to learn to associate body parts with individuals in the image to output the 2D keypoints for all people in the image. OpenPose provides the positions of 18 body joints (pre-trained using COCO 2016 keypoints challenge dataset [17], see Fig. 2). The choice of OpenPose is motivated for its functionality on RGB images or videos taken by on-site surveillance cameras in real-time. This provides a huge benefit in comparison to the skeletal tracking capability of RGB-D devices (e.g., Microsoft Kinect) which depend on depth information. To further improve the estimation speed, we deploy a light-weight architecture, Mobilenetv2 [18] as the feather extractor instead of VGG-19 [19] in the original paper.

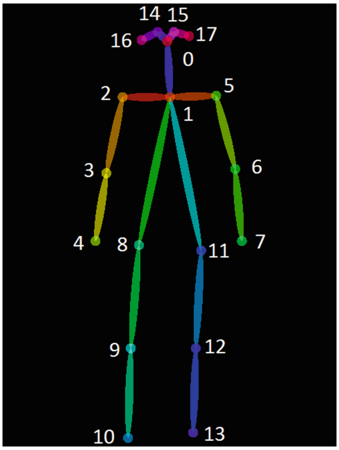


Fig. 2. OpenPose Output Format.

However, Openpose may fails in detecting individuals without their full body visible in the near field. In this case, we characterize individuals by localizing their faces using YOLOv3.

3.2 NPU Identification

We formulate the output of YOLOv3 as a set of objects bounding boxes $B = \{B_1, B_2, \ldots, B_I\}$, where I is the number of detected objects bounding boxes in the obtained frame. Each bounding box $B_i = (x_i, y_i, w_i, h_i, c_i)$, $I \in \{1, 2, \ldots, I\}$ contains five elements, where (x_i, y_i) and (w_i, h_i) are respectively the bounding boxes position and size and c_i represents the class of objects in the bounding box. Let $H = \{H_1, H_2, \ldots, H_J\}$ be the set of the detected individual(s) via OpenPose, where J is the number of detected individual(s) in the obtained frame and $H_j = \left\{\left(x_j^{(0)}, y_j^{(0)}\right), \left(x_j^{(1)}, y_j^{(1)}\right), \ldots, \left(x_j^{(17)}, y_j^{(17)}\right)\right\}$ represents the detected bodyparts of the j_{th} individual. We associate a detected PPE i^* to a specific individual j^* by searching the minimum Euclidean distance between bounding boxes B and detected neck keypoints $H^{(1)} = \left\{\left(x_0^{(1)}, y_0^{(1)}\right), \left(x_1^{(1)}, y_1^{(1)}\right), \ldots, \left(x_J^{(1)}, y_J^{(1)}\right)\right\}$ (bodypart 1 in Fig. 2) that satisfy the geometric constraints to make sure each PPE is in the upper area of the candidate associated individual:

$$i^*, j^* = \underset{i \in \{1,2,\ldots,I\}, j \in \{1,2,\ldots,J\}}{\operatorname{argmin}} \sqrt{\left(x_i - x_j^{(1)}\right)^2 + \left(y_i - y_j^{(1)}\right)^2} \quad s.t.\, y_{i^*} < y_{j^*}^{(1)} \tag{2}$$

Subsequently, for each associated PPE and individual, we identify whether the PPE is appropriately used by analyzing key lengths. We take advantage of Euclidean distance among detected neck keypoints and hip keypoints (bodyparts 8 and 11 in Fig. 2) as reference threshold when the distance between the individual and the camera changes, the reference threshold change synchronously:

$$\beta_{i^* \leftrightarrow j^*} = max\left(\sqrt{\left(x_{j^*}^{(1)} - x_{j^*}^{(8)}\right)^2 + \left(y_{j^*}^{(1)} - y_{j^*}^{(8)}\right)^2}, \sqrt{\left(x_{j^*}^{(1)} - x_{j^*}^{(11)}\right)^2 + \left(y_{j^*}^{(1)} - y_{j^*}^{(11)}\right)^2}\right) \cdot \gamma \tag{3}$$

where γ is the scaling coefficient to strike different NPU identification. If the Euclidean distance between the detected PPE and detected neck keypoint of the associated individual is smaller than the reference threshold $\beta_{i^* \leftrightarrow j^*}$, then the detected PPE is identified as being appropriately used by the individual (Fig. 3 (a)); otherwise the condition is identified as NPU (Fig. 3 (b)).

In the case of OpenPose failure, PPE-individual association is performed using detected face bounding boxes (Fig. 3 (c), (d)). Let $B' = \{B'_1, B'_2, \ldots, B'_K\}$ be the detected face bounding boxes, where K is the number of detected face bounding boxes in the obtained frame. Similarly, we associate a detected PPE i^* to a specific individual

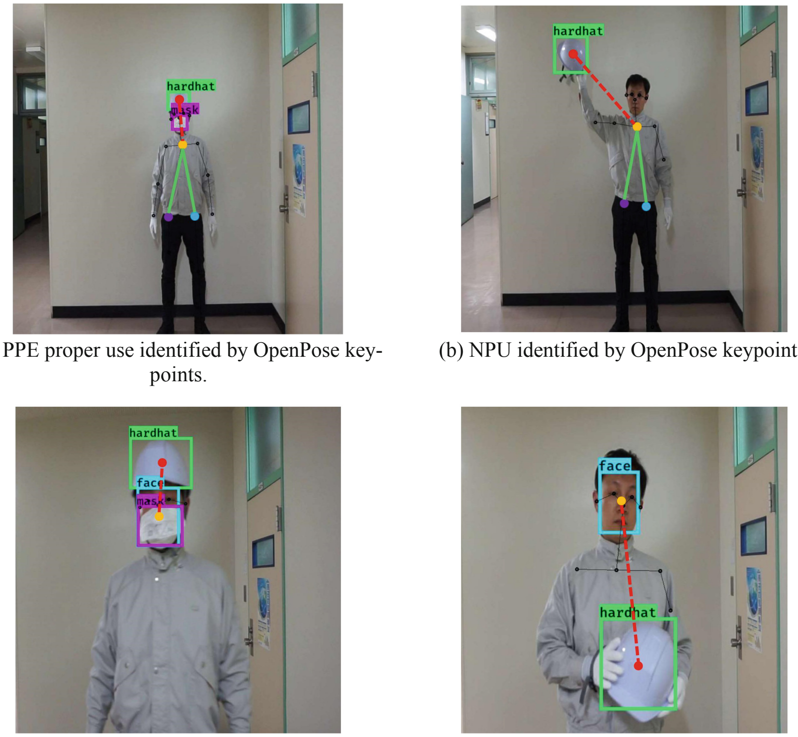
k^* by searching the minimum Euclidean distance between PPE and face bounding boxes:

$$i^*, k^* = \underset{i\in\{1,2,\ldots,I\},k\in\{1,2,\ldots,K\}}{\operatorname{argmin}} \sqrt{\left(x_i - x'_k\right)^2 + \left(y_i - y'_k\right)^2} \tag{4}$$

where (x_i, y_i) is the position of the i_{th} PPE bounding box in B and (x_k, y_k) is the position of the k_{th} face bounding box in B'. The reference threshold $\beta_{i^* \leftrightarrow k^*}$ is given as follow:

$$\beta_{i^* \leftrightarrow k^*} = max\left(w'_{k^*}, h'_{k^*}\right) \tag{5}$$

where $\left(w'_{k^*}, h'_{k^*}\right)$ is the size of the associated face bounding box in B'.



(a) PPE proper use identified by OpenPose keypoints.

(b) NPU identified by OpenPose keypoints.

(c) PPE proper use identified by detected face bounding boxes.

(d) NPU identified by detected face bounding boxes.

Fig. 3. NPU identification strategies.

4 Experiments and Results

4.1 Experimental Dataset

To create the training dataset of PPE detection, we collected hardhat, eye protection visor, and dust mask images from two sources: downloading Internet images retrieved by search engines using keywords, and capturing real-world images using the webcam as listed in Table 1. Also, we randomly selected 2,000 images from the WIDER FACE dataset [20] to learn the features of face. A total of 7,929 images were collected and annotated to train a YOLOv3 model.

Table 1. Information of collected training dataset.

	Internet images	Real-world images	Total
Hardhat	933	1,209	2,142
Mask	983	1,222	2,205
Visor	110	736	846
Face	2,000	736	2,736
Overall	4,026	3,903	7,929

Furthermore, to create the testing dataset to validate the performance of the trained model, six volunteers were instructed to perform normal working behaviors while wearing PPEs at different distances to the camera. As surveillance cameras are placed in different locations on construction sites and the trajectory of workers is stochastic, construction workers were captured in different resolutions in the surveillance videos. Thus, different distance conditions (1 m, 3 m, 5 m) were considered in our experiments to validate the robustness of our proposed approach. Finally, we randomly selected 1,500 images (500 images for each distance condition) from the collected image sequences and created the testing dataset. The details are provided in Table 2 where positive samples refer to the NPU (including NHU, NMU, and NVU) individuals and negative samples referred to the individuals who are wearing PPE properly.

Table 2. Information of collected testing dataset.

Distance (m)	NPU Categories	Images	Positive samples	Negative samples
1	NHU	500	318	182
	NMU		275	225
	NVU		334	166
3	NHU	500	263	479
	NMU		278	464
	NVU		452	290
5	NHU	500	274	457
	NMU		288	443
	NVU		444	287
Overall		1,500	2,926	2,993

4.2 Evaluation Metrics

We adopted precision and recall to evaluate the performance of our approach:

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

Where *TP* is defined as the number of correct detections of NPU individuals. *FP* is the number of wrong detections of NPU individuals, while *FN* is the number of the ground truth not detected as defined in Table 3.

Table 3. Definitions of TP, FP, and FN.

	Predicted	Ground truth
TP	NPU	NPU
FP	NPU	Proper use
FN	Proper use	NPU

4.3 Implementation Details

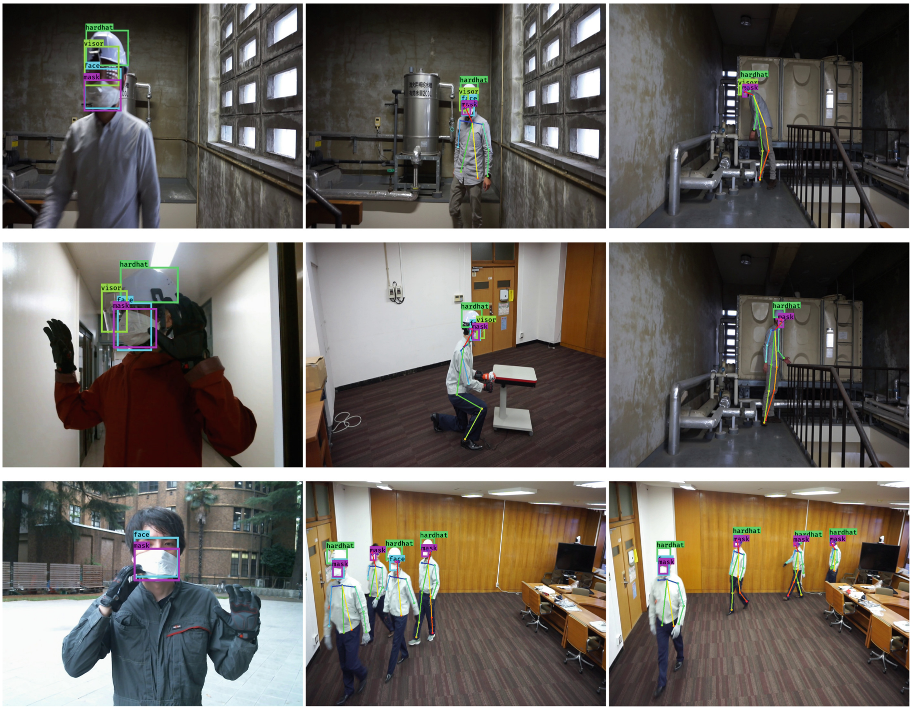
We initialized the YOLOv3 model based on pre-trained weights on the ImageNet dataset [21]. Training of YOLOv3 was performed in two stages. We first froze all convolutional layers up to the last convolutional block in Darknet-53 and train with frozen layers to get a stable loss. Subsequently, we proceed to unfreeze all convolutional layers of Darknet-53 to perform fine-tuning. The learning rates for the first and the second stage are 1e−3 and 1e−4 respectively. Adaptive Moment Estimation (Adam) optimizer was adopted to adjust the learning rate during optimization automatically. α (initial learning rate), β_1 (exponential decay rate for the first moment estimates), β_2 (exponential decay rate for the second-moment estimates) and ξ (a very small number to prevent any division by zero in the implementation) were set to 1e−3, 0.9, 0.99, 1e−8, respectively.

4.4 Results

We report the identification results on the testing dataset in Table 4. As the distance between the camera and individuals increases, the resolution of the individual's critical regions in the image gradually becomes smaller. Although the precision and recall rate of NHU identification gradually decreased, the precision and recall rate remained higher than 80% and 81%, respectively. For NMU identification, the precision and recall rate declined only slightly from 1 m to 3 m, and even the masks were quite small in far-field (5 m) images, our approach still achieved precision and recall rates of approximately 80%. For NVU identification, the precision and recall rate between 1 m to 3 m remained higher than 85% and 97%, respectively. However, the performance is decreased in the 5 m case since it is challenging to detect transparent visors with relatively low resolution. The overall precision and recall rates are 84.13% and 93.10%, respectively, which demonstrates the robustness of the proposed approach in PPE wearing identification at different distances. Figure 4 illustrates the identification examples on the testing dataset.

Table 4. Identification results under different distance.

NPU Categories	Distance (m)	TP	FP	FN	Precision (%)	Recall (%)
NHU	1	318	13	0	96.07	100
	3	241	31	22	88.60	91.64
	5	222	54	52	80.43	81.02
NMU	1	247	28	28	89.82	89.82
	3	250	40	28	86.21	89.93
	5	225	58	63	79.51	78.13
NVU	1	325	47	9	87.37	97.31
	3	452	75	0	85.77	100
	5	444	168	0	72.55	100
Overall		2724	514	202	84.13	93.10



(a) Distance: 1m (b) Distance: 3m (c) Distance: 5m

Fig. 4. Identification examples on the testing dataset.

4.5 Computational Efficiency Analysis

To meet the industrial requirements of real-time processing, we also conducted computational efficiency analysis experiments. Computational efficiency analysis results are presented in Table 5, the inference time of our approach outperforms other state-of-the-art methods while preserving high-quality results. It is able to run at about 7.95 FPS in a machine with a GeForce GTX 1080 Max-Q with 8 GB of GDDR5X memory and 2560 CUDA cores and it indicates that our approach is more effective compared to the Faster R-CNN approach adopted by Fang et al. [15] and SSD-RPA method proposed by Wu et al. [16].

Table 5. Computational efficiency analysis results.

Approach	Input size	FPS
Faster R-CNN [13]	300 × 500	4.88
SSD-RPA [14]	304 × 304	3.22
Ours	416 × 416	7.95

5 Conclusion

This paper has presented a new vision-based approach to address the difficulties of PPE proper use management in the construction sites. Firstly, we created a dataset using Internet images and real-world images to train the YOLOv3 model to recognize hardhats, eye protection visors, and dust masks. Subsequently, we conducted NPU identification using geometric relationships of the outputs of OpenPose and YOLOv3. The performance of the proposed approach was experimentally evaluated under various distance conditions. The test results indicate that the approach was capable of detecting NPU workers with high precision (84.13%) and recall rate (93.10%) while ensuring real-time performance (7.95 FPS on average). Further extensions of this work follow the consideration of on-site system implementation and performance improvement based on time-series analysis.

Acknowledgements. This work was supported by Japan Society for the Promotion of Science (JSPS) Grants-in-Aid for Scientific Research Grant No. 19K05324.

References

1. U.S. Bureau of Labor Statistics, Construction: NAICS 23. https://www.bls.gov/iag/tgs/iag23.htm. Accessed 13 Jan 2020
2. The Japanese Ministry of Health, Welfare, The 13th occupational safety health program. https://www.mhlw.go.jp/content/11200000/000341159.pdf. Accessed 13 Jan 2020
3. Konda, S., Tiesman, H.M., Reichard, A.A.: Fatal traumatic brain injuries in the construction industry, 2003–2010. Am. J. Ind. Med. **59**(3), 212–220 (2016)

4. U.S. Department of Labor, Occupational Safety and Health Administration, Head protection. https://www.osha.gov/laws-regs/regulations/standardnumber/1926/1926.100. Accessed 26 Jan 2020
5. The National Institute for Occupational Safety and Health, Eye safety. https://www.cdc.gov/niosh/topics/eye/. Accessed 13 Jan 2020
6. Dannenberg, A.L., Parver, L.M., Brechner, R.J., Khoo, L.: Penetrating eye injuries in the workplace: the national eye trauma system registry. Arch. Ophthalmol. **110**(6), 843–848 (1992)
7. Government of Western Australia, Department of Commerce, Guide to using dust masks in construction work (2019). https://www.commerce.wa.gov.au/sites/default/files/atoms/files/guide_to_using_dust_mask.pdf
8. U.S. Department of Labor, Occupational Safety and Health Administration, Eye and face protection. https://www.osha.gov/laws-regs/regulations/standardnumber/1910/1910.133. Accessed 13 Jan 2020
9. Redmon, J., Farhadi, A.: YOLOv3: an incremental improvement, arXiv preprint arXiv:1804.02767
10. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: Openpose: realtime multi-person 2D pose estimation using part affinity fields, arXiv preprint arXiv:1812.08008
11. Kelm, A., Laußat, L., Meins-Becker, A., Platz, D., Khazaee, M.J., Costin, A.M., Helmus, M., Teizer, J.: Mobile passive Radio Frequency Identification (RFID) portal for automated and rapid control of Personal Protective Equipment (PPE) on construction sites. Autom. Construct. **36**, 38–52 (2013)
12. Dong, S., He, Q., Li, H., Yin, Q.: Automated PPE misuse identification and assessment for safety performance enhancement. In ICCREM 2015, pp. 204–214 (2015)
13. Shrestha, K., Shrestha, P.P., Bajracharya, D., Yfantis, E.A.: Hard-hat detection for construction safety visualization. Journal of Construction Engineering **2015**, 1–8 (2015)
14. Park, M.W., Elsafty, N., Zhu, Z.: Hardhat-wearing detection for enhancing on-site safety of construction workers. J. Construct. Eng. Manage. **141**(9), 04015024 (2015)
15. Fang, Q., Li, H., Luo, X., Ding, L., Luo, H., Rose, T.M., An, W.: Detecting non-hardhat-use by a deep learning method from far-field surveillance videos. Autom. Construction **85**, 1–9 (2018)
16. Wu, J., Cai, N., Chen, W., Wang, H., Wang, G.: Automatic detection of hardhats worn by construction personnel: a deep learning approach and benchmark dataset. Autom. Construction **106**, 102894 (2019)
17. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: common objects in context. In: European Conference on Computer Vision, pp. 740–755. Springer, Cham (2014)
18. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: MobileNetV2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018)
19. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014). arXiv preprint arXiv:1409.1556
20. Yang, S., Luo, P., Loy, C.C., Tang, X.: Wider face: a face detection benchmark. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5525–5533 (2016)
21. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)