Posture adaptive helmet wearing detection on construction sites

Kaipeng Zheng, Jie Shen*

University of Electronic Science And Technology of China, Chengdu, Sichuan, China
E-mail: kaipengm2@gmail.com sjie@uestc.edu.cn

Abstract. Automatic helmet wearing detection is a significant measure for the safety of construction sites. Recent studies based on face detection of workers have achieved higher accuracy than previous methods using pedestrian detection. However, such face detection based methods only use the coordinates of the detected face bounding box to predict the candidate region of the corresponding helmet, while the face posture is ignored. In our research, a posture adaptive transformation from faces to helmets is proposed. It uses a concise and efficient convolutional neural network to adaptively adjust the helmet candidate regions by extracting the features of the face image. Moreover, a new dataset is proposed to make up for the lack of small object samples in the current public dataset. Extensive experiment proves the effectiveness of the proposed method.

Keywords: Helmet detection, Face detection, Posture adaptive, Convolutional neural network

1. Introduction

Automatically detecting the objects in the image collected from the real world environment is a classic task in the image processing field[1,2], aiming to get the precise location of the target objects and make a prediction of its classification. It is widely used in various engineering applications, such as pedestrian detection[2,3,4] and traffic problems. Recent years, in civil engineering applications, image sensors have attracted growing attention, among which a large amount of object detectors based on computer vision technologies are developed. Detection models of defects and damages of the buildings including shield tunnels[5], pavements[6] and concrete bridges[7] are proposed to measure the safety status. [8] designed a detector to detect construction vehicles on the construction sites. Moreover, for the incremental accidents happened on the construction sites, individual safety on the construction sites has led unprecedentedly heated discussions these years. Personal protective equipment(PPE), especially helmet, is essential to protect people from danger on the construction sites. Automatically detecting personal protective equipment(PPE) such as helmets arouses much interest of researchers. By generating warning signals when detecting people with no such equipment automatically, it is interpreted as an efficient solution to advance safety quality on the construction site. Helmets play a significant role among PPE, on which the researches occupy the vast majority.

Most of the previous work uses pedestrian detection based method to detect helmets. It can be summarized into two major steps that detecting pedestrian from the original images and detecting helmets from the proposed region regarding to pedestrians. A majority of the researches conduct the conventional technologies of computer vision such as histogram of oriented gradient (HOG), scale-invariant feature transform (SIFT) and local binary patterns (LBP) or an integrated model using combined technologies to locate the pedestrians in the images. However, experiments carried out by these researches only focuses on very sparse scenes that only few people are contained in the images. For the crowded scenes, the detecting results will be poor as scale variances and occlusions appear in the images. Deep learning based pedestrian detection methods have received a high degree of attention in recent years. Several special models of neural network are designed to deal with scale variances and distortion, such as feature pyramid networks (FPN), achieving better performance than conventional detectors. A faster region-based convolutional neural networks (Faster R-CNN) based detector is proposed by [9] to directly detect people with no helmet. Though it gain some improvement compared with the conventional detection method, it gets a bad result suffering from occlusions.

Moreover, small object detection and occlusions are also severe problems which lead to poor performances, resulting from the vastness of the construction sites. Most surveillance cameras are located at the edge of the construction sites, which monitors the global status of the construction sites. In the images collected by these surveillance cameras, workers and helmets can be extremely small, which can be easily ignored by detectors with a high probability for the lack of features.

Instead of detecting pedestrian, a method using face detection to detect helmets is proposed by Shen[18], as the face detector is advanced in dealing with occlusions and small objects comparing to pedestrian detectors. After the face detection completed, the location of the helmet is obtained across a linear regression model, which is trained by the coordinates of the bounding boxes of faces. Finally, a binary classification is made to predict whether the bounding boxes contains a helmet. Though it achieves a higher performance in images with occlusion, distortion and scale variances, resulting from the state-of-the-art face detector, the following linear regression model remains several limitations. It only uses the coordinates of the bounding boxes of faces as independent variables of the regression model, which can not always get the correct bounding boxes of helmets resulting from the different poses of faces. For example, the helmet is located above a worker's face when he is standing. Nevertheless, when the worker is lying on the ground to complete some special work, the relative position of the helmet will change to the side of the face. Essentially, only using the coordinates of the bounding boxes of faces to train the regression model is unable to capture the pose change of faces, resulting in that only sizes of the helmet are correctly predicted while the relative orientations between helmets and faces are ignored.

Motivated by the method proposed by Shen[18], in our research, a novel integrated method to detect helmets are proposed using face detection. As different poses lead to different location of helmets, the proposed method aims to adapt to various poses. We adopt the dual shot face detector (DSFD)[22] as the tiny face detector, which is the same as Shen, for its strong robustness to occlusion and scale variances. After the face detection completed, a Convolutional Neural Network (CNN) model is built to regress the coordinates of the bounding boxes of the helmets. Instead of simply using the coordinates of the bounding boxes of the detected faces, we use the detected face images cropped from the original images as the input of the regression model. By extracting features of the face poses using the CNN models, the helmet can be precisely located with various poses such as looking up as a high place, bending down and lying on the ground.

Moreover, aiming to solve the small object detection, a new public dataset is proposed in our research, where small objects take a large proportion. In the field of helmet detection, the only public dataset so far is proposed in the research of Shen[18]. However, the amount of small objects contained in the dataset is small, which serves as an important factor for its poor performance in small object detection. In the proposed new dataset, images containing small objects captured in various scenarios are widely included. The total amount of the helmets contained in the dataset is 4512, among which small objects with less than 2500 (50 * 50) pixels make up 30%. The proposed dataset and method can advance the detection of helmets with small sizes.

To summarize, our research proposes an integrated helmet detection method based on face detection, which adapt well to various poses of people. Specifically, our main contributions are as follow: (a) We propose a CNN model which is posture adaptive to faces to predict the corresponding candidate regions of helmets. (b) We propose a new public dataset containing 4512 objects, where small objects make up more than 30%. (c) Extensive experiments show that the proposed method outperform the state-of-the-art method.

2. Related Work

Pedestrian detection based methods using conventional computer vision technologies take a majority of image sensors based helmet detection. Histogram of Oriented Gradient (HOG), a popular human detection method, is used in[11,12,13,14,15,16,17]. In these researches, HOG is used to detect the pedestrians in the image and serves as a feature descriptor of them. Following the HOG, some other feature extractors are involved to enhance the feature extraction of the detected human. The combined features are fitted by machine learning models to finally predict whether the detected people wear helmets.

K-nearest neighbours(KNN) is used to detect the motorcycle and classify whether the rider wears a helmet[10]. Following the human detection, special feature extractors are designed. [11,14]uses Circle Hough Transform (CHT) to extract features. [12] proposes a hybrid feature descriptor based on local binary patterns (LBP), hu moment invariants (HMI) and color histograms (CH) and a hierarchical support vector machine (H-SVM) to make classification. [16] uses Local Binary Pattern (LBP), Histograms of Oriented Gradients (HOG) and the Hough Transform (HT) to extract features. [13] directly uses a Support Vector Machine (SVM) to classify the detected humans described by the HOG features. [17] uses HOG, scale-invariant feature transform (SIFT), and LBP to extract features. However, only sparse and simple scenes are tested in the experiments of these researches. For the complicated environment with crowded people, the performances of these methods are always poor suffering from scale variances and occlusions.

Deep learning based methods have been studied recently. Motivated by the progress in general object detection achieved by deep learning, several researches involve deep learning technologies to conduct the helmet detection. FangQi's[9] method trains a Faster RCNN model to make a direct detection toward people without helmet. It is a pedestrian detection based method, and it has been verified that occlusions can severely affect its performance by Shen[18]. Moreover, such detection cannot identify who is not wearing a helmet. Shen[18] proposes a method using face detection. It first uses a deep learning based face detector to detect faces, and then regress to the coordinates of the candidate regions of helmets from the coordinates of the face bounding boxes. Using face detection, it can easily recognize the person who is not wearing a helmet. Since the face detection achieves high performance using deep learning technologies, the method also shows robustness in the scenes of occlusions and scale variances. However, it only uses the coordinates of the face bounding boxes to make the prediction while ignores the posture of faces which can lead to great changes of the corresponding helmets' position. In this work, a facial poses adaptive method is proposed.

3. Methodology

Our research proposes an integrated method to automatically conduct helmet detections. The whole architecture of our method is displayed in the Figure1. It can be summarized into 3 stages. Given an image collected from the real environment of the construction site, face detection is conducted first, aiming to precisely locate and make recognition of each individual. Since the different poses of the face will cause the relative position between the helmet and the face to change, the posture adaptive transformation (PAT) network is proposed in order to enable the model to adaptively locate the candidate area of helmets. it is formed into a regression task whose input is the image of a detected face and the output is the coordinates of the bounding box of the corresponding helmet. To better extract the features of the poses and sizes of the detected faces, we propose a simple but effective Convolutional Neural Network (CNN) model as the posture adaptive transformation. After obtaining the candidate regions of helmets, the next step is to determine whether the candidate area contains a helmet. We conduct a transfer learning of the Efficientnet[23] to adapt to this task and obtain the final results of the helmet wearing detection.

3.1. Face detector

A high-performance face detector, Dual Shot Face Detector (DSFD)[22], is made use in our method to detect faces for the first step. DSFD is a deep learning based face detector. In the architecture of the DSFD, an optimized feature pyramid network (FPN) involving the concatenation of multiple dilated convolutions and an advanced anchor matching mechanism are proposed to better adapt to face detections. DSFD achieves high performance under various situations. In our experiment, compared with other face detectors and pedestrian detectors, it shows stronger robustness to scenes with extreme scale variances, severe occlusions and tiny face detections respectively.

3.2. Posture Adaptive Transformation

After the face detection, the next step is to locate the candidate area of the corresponding helmet. We use the detection bounding box $H = (x_h, y_h, w_h, h_h)$ to describe the candidate position of the helmet. A detection bounding box is composed by two pairs of parameters, which are the center position coordinates (x_h, y_h), the length and width (w_h, h_h). The candidate area is de-

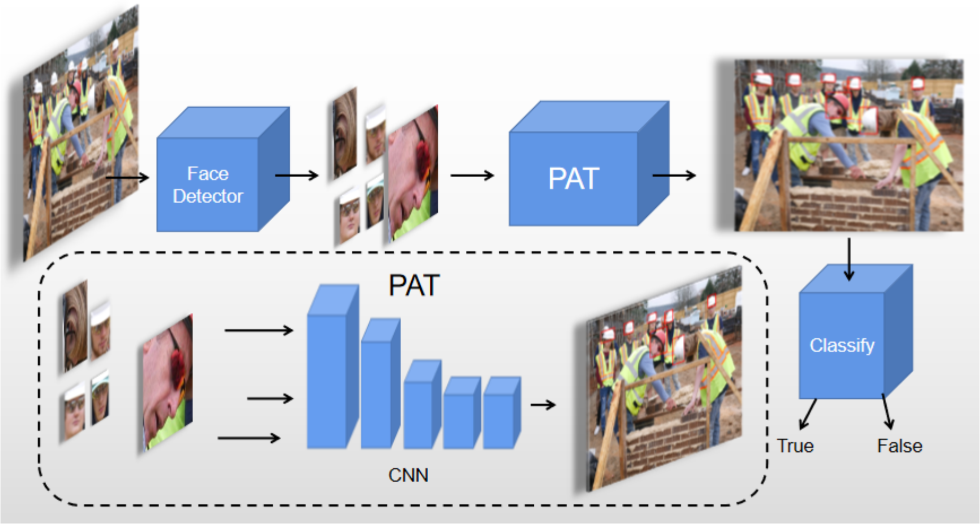


Fig. 1. The whole architecture of the proposed method. For an input image, it is first fed into the face detector to obtain faces in the image. After collocting the faces, the next step is to predict the correponding candidate region of the helmet of each face using the proposed PAT module. The proposed PAT module uses CNN to adaptively extract the posture features of faces to regress to the coordinates of the candidate regions. Finally, the candidate regions of helmet are made a classification that whether it contains a helmet or not.

termined by these two pairs of parameters. In addition, the center position coordinates of the helmet can be further converted to the relative position coordinates of the human face. It can eliminate the difference in the image position of the helmet and the difference in the size of the image itself, as follows:

Let the bounding box of the face be denoted as $F = (x_f, y_f, w_f, h_f)$. Among them, the center coordinates are (x_f, y_f), and the length and width are (w_f, h_f). The relative position coordinates are defined as the coordinates of the center position of the helmet with the coordinates of the center position of the face as the origin of the coordinates:

$$x_{hr} = x_h - x_f \tag{1}$$

$$y_{hr} = y_h - y_f \tag{2}$$

Therefore, the bounding box of the candidate position of the helmet represented by relative position coordinates is as follows

$$\begin{aligned} Hr &= (x_{hr}, y_{hr}, w_h, h_h) \\ &= (x_h - x_f, y_h - y_f, w_h, h_h) \end{aligned} \tag{3}$$

Different face poses will cause the relative position of the face and the helmet to change. Moltivated by this, we use a convolutional neural network to establish a regression model. The face image detected by the face detector is used as the input of the convolutional neural network. The expected output of the model is the four parameters of the detection bounding box of the helmet marked in the data set, which are the relative position coordinates and the length and width. Through training, the network can adaptively find the mapping relationship between the face pose and the position of the helmet. The parameters of the model are obtained by optimizing the following objective function:

$$\begin{aligned} \Gamma &= argmin \frac{1}{N} \sum_{i=1}^{N} (Hr_i^k - \hat{Hr}_i^k)^2 \\ &= argmin \frac{1}{N} \sum_{i=1}^{N} (Hr_i^k - f_\theta(I_i))^2 \end{aligned} \tag{4}$$

where N represents the number of samples, Hr_i^k represents the 4 parameters of the bounding box of the candidate position of the helmet to be predicted and $k = 1, 2, 3, 4$, f_θ represents the convolutional neural network model, and I_i represents the input face image. For the convolutional neural network model f_θ, we designed a simple and effective network structure to complete the mapping from the face pose to the candidate bounding box parameters of the helmet, as shown in Figure 2.

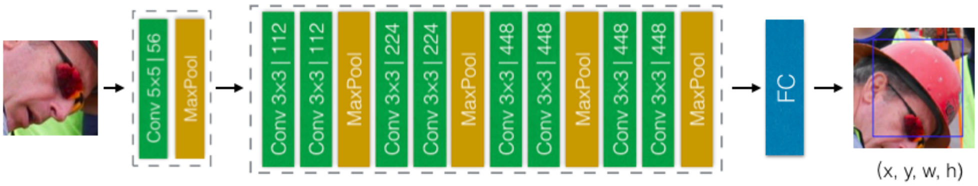
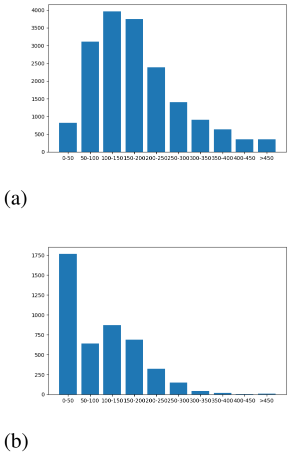


Fig. 2. The network architecture of the proposed PAT. The input is the cropped detected faces and the output is the 4 parameters to determine the bounding boxes o f the helmets' locaiton. $Conv$ represents the convolution operation defined in CNN. 5×5 and 3×3 represent the kernel size of the convolution. $56, 112, 224, 448$ represent the number of channels of each convolution, and FC represent the fully connected layer



(a)

(b)

Fig. 3. (a) indicates the distribution of the previous dataset's image sizes. (b) indicates the distribution of the proposed dataset's image sizes. It can be observed that the prposed dataset provide the implementary of the small objects.

After the face image is input to the network, it first passes through a convolutional layer with a convolution kernel size of (5, 5) to obtain the local receptive field of the image and capture the local features of the face pose. After that, through multiple convolutional layers and pooling layers with a convolution kernel size of (3, 3), a larger receptive field is gradually obtained, and at the same time, the number of feature channels is gradually increased according to 3, 56, 112, 224, and 448 to obtain Global characteristics of face pose. Finally, through the fully connected layer, the four parameters of the bounding box of the helmet detection are predicted.

Compared with some classic convolutional neural network models such as VGG[20] and ResNet[21], our proposed network is simpler and more effective, and is more suitable for solving this problem. The accuracy of the network can reach the same level as these classic network models, and the parameter scale is smaller, which makes it easier to deploy in practical applications and has a faster response time.

3.3. Classification of candidate regions

After the locations of helmets are determined, the following step is to predict whether the candidate location contains a helmet or not. Instead of using conventional computer vision technologies to directly recognizing the helmet as most previous work does, the task is formed as a binary classification problem solved by deep learning in our research. A deep learning network Efficientnet[23] is selected as the backbone of our proposed model. Efficient-net is very effective in the problem of image classification, and at the same time has the characteristics of small parameter. Efficient-net includes a feature extraction layer and a fully connected layer for classification. In order to migrate Efficient net to the problem of hard hat candidate area classification, we adjusted the output of the last layer of its fully connected layer from 1000 to 2, respectively. The candidate area contains safety helmets and does not contain safety helmets.

3.4. Dataset

To train the proposed CNN models, dataset of good quality is necessary. However, few public dataset has been proposed before. To our best knowledge, the DataFountain's dataset[18] proposed by Shen is the only public dataset in the feild of helmet detection. However, we find that small objects with less than 2500 (50 * 50) pixels included in the dataset is insufficient. Among the whole dataset, small targets account for less than 10. However, small target objects frequently appear in actual scenes, especially in vast construction sites. Motivated by this observation, a new dataset where small objects take large proportion is proposed in our research, aiming to make the model

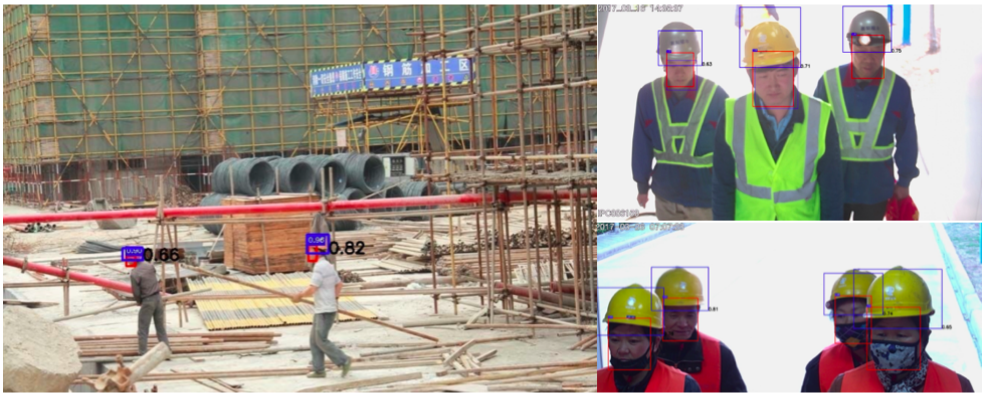


Fig. 4. candidate region prediction results of the real construction sites using the proposed PAT module

adapt better to small objects. The helmets in the dataset are all labeled with bounding boxes. Among the proposed dataset, 4512 helmets are collected from various scenes of construction sites, of which the objects with less than 2500 (50 * 50) pixels take 40%. The distributions of the objects' sizes of the proposed dataset and the DataFountain's dataset are showed in Figure3 (a) (b) respectively. It is evident that the proposed dataset provide a supplementary of the small objects.

4. Experiment

4.1. Metrics

To make the evaluation, the accuracy score and the recall rate are used, which can be described as follow. Let TP be the number of the true positive predictions, FP be the number of the false positive predictions, TN be the number of the true negative predictions, and FN be the number of the false negative predictions. The accuracy score is calculated as follow.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

The accuracy score is used to estimate how much rate of the samples are correctly predicted. In addition, the recall rate that estimating how much rate of the positive samples are successfully recalled as follow.

$$recall = \frac{TP}{TP + FP}$$

4.2. location prediction

In our experiment, the proposed PAT network which is used to make prediction of the candidate regions of helmets is evaluated. Comparative studies are conducted between the proposed method and a representative face detection method proposed by Shen[18], which uses linear regression to make such prediction. We first display several detection results of the real construction sites using the proposed PAT module. The results are shown in Figure4. It can be seen that the proposed PAT module can successfully locate the candidate area of the helmet with a high accuracy.

To further verify the robustness of the proposed method, we select images of challenging scenes including faces with abnormal poses, and add extra random rotations to some images as noise to make an extent verification. We compare the proposed method with Shen's[18] method. The results are shown in Figure5. As is shown in Figure5, the proposed method still keeps high accuracy while the results of Shen's[18] method method drop dramatically. The reason is that the proposed method make the prediction based on the posture features of the faces while Shen's[18] method only uses the coordinates. Thus, the proposed method can adapt to various changes in facial poses in the real construction sites.

Moreover, to make a specific evaluation, IOU (Intersection over Union) score between each predicted candidate region and the corresponding ground truth is calculated to measure the degree of overlap of the two regions, and 0.5 is set as the threshold to determine whether the object is successfully recalled. The results are shown in Figure 6.

(a) Shen's method

(b) The proposed method

Fig. 5. comparison on challenging scenes with strange postures between the proposed method and the state-of-the-arts method

It can be seen from the results that the proposed method achieves higher IOU scores than Shen's method. The average IOU score of Shen's method is 67.70%, and the proposed method achieves 71.48%. The proposed method gains 3.78% improvement compared with Shen's method. On the other hand, the recall rate of Shen's method is 88.12%, and the proposed method achieves 91.86%. The proposed method gains 3.74% improvement compared with Shen's method.

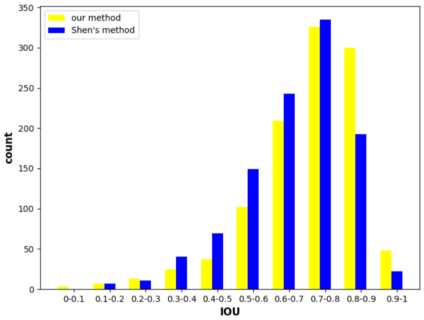


Fig. 6. Distribution of the IOU computation results among Shen's method and the proposed method

4.3. helmet classification

The predicted candidate area is then sent into the proposed Efficientnet based classifier to make the final decision that whether the candidate area contains a helmet or not. We also compare the results with Shen's[18] method. The accuracy of the proposed method is 97.87%, and the accuracy of the Shen's method is 95.26%, which proves the effectiveness of the proposed method.

5. conclusion

This paper proposes a novel helmet detection algorithm based on face detection. When workers perform work tasks on construction sites, their body postures often change drastically, causing the relative positions of their faces and helmets to be unstable. Previous helmet detection algorithms based on face detection cannot adapt to changes in human posture. Compared with the previous methods, the method proposed in this paper uses a simple and effective deep learning network to adaptively extract the pose features of the face, and use the extracted face pose features to adaptively determine the position of the helmet candidate region. So that when workers work on the construction site in various postures, such as bending over, raising their heads greatly, etc., the methods proposed in this paper can accurately find the candidate position of the helmet through the face. The experimental results show that, compared with the previous helmet detection method based on face detection, the helmet candidate frame obtained by the method proposed in this paper has a higher average IOU and a higher recall rate.

References

[1] L. Lamport, *LaTeX User's Guide & Reference Manual*, Addison Wesley Publishing Co, 1985.

[2] B. Newman and E.T. Liu, Perspective on BRCA1, *Breast Disease* **10** (1998), 3–10.

[3] Olmeda D, Premebida C, Nunes U, Armingol JM, de la Escalera A. *Pedestrian detection in far infrared images.* Integrated Computer-Aided Engineering. 2013;20(4):347-360. doi:10.3233/ICA-130441

[4] Li D, Xu L, Goodman ED, Xu Y, Wu Y. Integrating a statistical background- foreground extraction algorithm and SVM classifier for pedestrian detection and tracking. Integrated Computer-Aided Engineering. 2013;20(3):201-216. doi:10.3233/ICA-130428

[5] Xue Y, Li Y. A Fast Detection Method via Region-Based Fully Convolutional Neural Networks for Shield Tunnel Lining Defects. Computer-Aided Civil & Infrastructure Engineering. 2018;33(8):638-654. doi:10.1111/mice.12367

[6] Tong Z, Yuan D, Gao J, Wang Z. Pavement defect detection with fully convolutional network and an uncertainty framework. Computer-Aided Civil & Infrastructure Engineering. 2020;35(8):832-849. doi:10.1111/mice.12533

[7] Zhang C, Chang C, Jamshidi M. Concrete bridge surface damage detection using a single-stage detector. Computer-Aided Civil & Infrastructure Engineering. 2020;35(4):389-409. doi:10.1111/mice.12500

[8] Arabi S, Haghighat A, Sharma A. A deep-learning-based computer vision solution for construction vehicle detection. Computer-Aided Civil & Infrastructure Engineering. 2020;35(7):753-767. doi:10.1111/mice.12530

[9] Fang, Q., Li, H., Luo, X., Ding, L., Luo, H., & Rose, T. M. (2018). Detect- ing non-hardhat-use by a deep learning method from far-field surveil- lance videos. Automation in Construction, 85, 1–9

[10] Rubaiyat, A. H. M., Toma, T. T., Kalantari-Khandani, M., Rahman, S. A., Chen, L., Ye, Y., & Pan, C. S. (2016). Automatic detection of helmet uses for constructio

[11] Shine, L., C. V., J. Automated detection of helmet on motorcyclists from traffic surveillance videos: a comparative analysis using hand-crafted features and CNN. Multimed Tools Appl 79, 14179–14199 (2020). https://doi.org/10.1007/s11042-020-08627-w

[12] Wu, H., & Zhao, J. S. (2018). An intelligent vision-based approach for helmet identification for work safety. Computers in Industry, 100, 267–277.

[13] J. Li et al., "Safety helmet wearing detection based on image processing and machine learning," 2017 Ninth International Conference on Advanced Computational Intelligence (ICACI), Doha, 2017, pp. 201-205, doi: 10.1109/ICACI.2017.7974509.

[14] R. R. V. e. Silva, K. R. T. Aires and R. d. M. S. Veras, "Helmet Detection on Motorcyclists Using Image Descriptors and Classifiers," 2014 27th SIBGRAPI Conference on Graphics, Patterns and Images, Rio de Janeiro, 2014, pp. 141-148, doi: 10.1109/SIBGRAPI.2014.28.

[15] Li, K., Zhao, X. G., Bian, J., & Tan, M. (2018). Automatic safety helmet wearing detection. arxiv.org/pdf/1802.00264.pdf.

[16] R. Silva, K. Aires, T. Santos, K. Abdala, R. Veras and A. Soares, "Automatic detection of motorcyclists without helmet," 2013 XXXIX Latin American Computing Conference (CLEI), Naiguata, 2013, pp. 1-7, doi: 10.1109/CLEI.2013.6670613.

[17] K. Dahiya, D. Singh and C. K. Mohan, "Automatic detection of bike-riders without helmet using surveillance videos in real-time," 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, 2016, pp. 3046-3051, doi: 10.1109/IJCNN.2016.7727586.

[18] Shen J, Xiong X, Li Y, et al. Detecting safety helmet wearing on construction sites with bounding-box regression and deep transfer learning[J]. Computer-Aided Civil and Infrastructure Engineering, 2021, 36(2): 180-196.

[19] Shrestha K, Shrestha P P, Bajracharya D, et al. Hard-hat detection for construction safety visualization[J]. Journal of Construction Engineering, 2015, 2015(1): 1-8.

[20] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[21] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[22] Li J, Wang Y, Wang C, et al. DSFD: dual shot face detector[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 5060-5069.

[23] Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C]//International Conference on Machine Learning. PMLR, 2019: 6105-6114.