Advances in Computer, Signals and Systems (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2024.080316
ISSN 2371-8838 Vol. 8 Num. 3

A Study on Target Detection and Its Application and Development in the Identification of Unsafe Behaviour of Construction Workers

Tong Zhang[1,*]

[1]*School of Engineering, China University of Geosciences, Wuhan, China*
[*]*Corresponding author: zhangtong040504@163.com*

Keywords: Target detection, deep learning, evaluation metrics, construction worker safety

Abstract: The construction industry serves as the cornerstone of economic development, but the construction industry is also a very dangerous industry with workplace accidents occurring every year, so automatic identification and recognition of potentially unsafe behaviours and conditions is of great significance in safeguarding the safety of incoming recognised lives. In this paper, firstly, the three major classes of algorithms for target detection are elaborated in detail, the traditional target detection mainly relies on the method of machine learning, the two-phase target detection algorithm based on deep learning is mainly divided into two phases of candidate region production and target detection, while the one-phase target detection algorithm based on deep learning carries out end-to-end target detection without the need to produce a candidate region, and gives an assessment of the performance of the target detection indicators to evaluate the strengths and weaknesses, and summarises and analyses the current applications of target detection in the construction field and the new trends in the future.

1. Introduction

The construction industry, a cornerstone of economic development, has seen significant growth in recent years. However, globally, the construction industry is a very dangerous industry, with workplace accidents occurring every year, resulting in injuries and even deaths. Only 7% of the world's workforce is invested in the construction industry, yet as many as 40% of the accidents that occur each year result in fatalities [1]. Studies have repeatedly shown that more than 90% of safety accidents are due to unsafe behaviours and working conditions [2, 3]. Therefore, if we can control unsafe human behaviours and improve the working environment, the safety index will increase dramatically. In the past decades, a series of models have been studied to order human unsafe behaviours.

With the development of computer vision technology, the continuous evolution of computer vision assisted technology has been widely recognised by academics as a robust means to automatically identify and recognise potentially unsafe behaviours and situations [4-7]. Deep learning, as a key branch in the field of machine learning, has demonstrated its superiority in dealing with many target detection tasks. Target detection is one of the important research directions in the field of computer vision, and target detection methods based on convolutional neural networks have now become mainstream in this field.

2. Targeting methods and assessment systems

2.1 Traditional target detection methods

Traditional target detection algorithms are mainly based on the combination of feature extraction and classifiers. In the feature extraction phase, the algorithm extracts features such as shape, texture and colour of the target from the image or video. These feature extraction methods include statistical-based methods, edge-based methods, and gradient-based methods, among others. For example, Haar feature is a statistic-based method that describes the texture and shape features of a target by calculating the intensity difference between different regions of an image [8], while HOG feature is a gradient-based method that describes the shape features of a target by calculating the histogram of the gradient direction of each pixel point in an image [9].

After feature extraction is complete, traditional target detection algorithms use classifiers to classify the target. Commonly used classifiers are Support Vector Machine (SVM), AdaBoost, and Artificial Neural Network [10]. These classifiers classify the target into different categories based on the extracted features. Two traditional target detection algorithms commonly used in the construction field are HOG+SVM and Haar Cascade.

2.2 Deep learning based target detection algorithm

Prior to 2010, target detection mainly relied on traditional classical algorithms. However, traditional detectors had some inherent limitations that made it difficult to further improve their performance. Fortunately, the field of target detection has been revolutionised with the introduction of Convolutional Neural Network (CNN) architecture. Deep neural networks have been widely used for deep representation of image features, effectively improving the accuracy and efficiency of target detection and significantly reducing the error rate [11]. Therefore, the application of deep neural networks in the field of target detection has undoubtedly injected new vitality into the development of computer vision technology.

Common deep learning target detection algorithms include R-CNN, SPP-Net, Fast R-CNN, Faster R-CNN, YOLO, and SSD [12, 13], which have high accuracy and performance in target detection tasks and are widely used in various practical scenarios.

(1) R. Girshick, J. Donahue, and T. Darrell et al. pioneered a novel two-stage target detection technique, known as R-CNN, by utilizing convolutional neural networks (CNN) [14]. The proposal of this technique marks a major breakthrough in the field of target detection.The R-CNN algorithm consists of five key components [15]: region proposal, feature extraction, feature vector, feature classification and bounding box regression. The basic steps of R-CNN are shown in Figure 1.

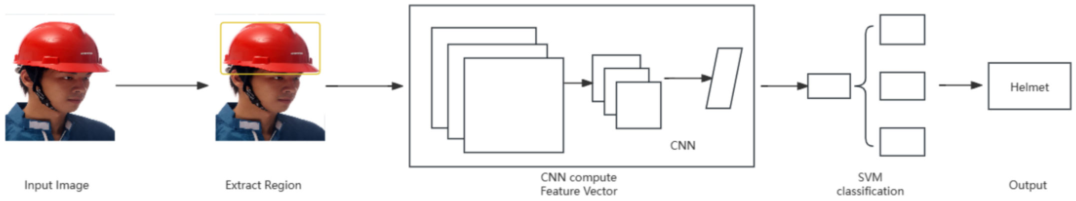

Figure 1: RNN basic steps

(2) In order to improve the shortcomings of R-CNN, K. He et al [16] introduced a new network structure SPP-Net in 2014.The SPP-Net network structure is an innovative deep convolutional neural

network structure, the core of which lies in the introduction of a Spatial Pyramid Pooling (SPP) layer.The SPP- Net network structure can solve the limitations of traditional convolutional neural networks (CNN) in processing images of arbitrary size or scale, and achieve more accurate recognition of images or sub-images.

(3) Girshic [17] further improved the RCNN and SPPNet by introducing a new architecture called Fast RCNN, which takes the complete image as input data and then uses deep convolutional layers to extract features from the image to generate a detailed feature map. Advantages of RCNN and SPP-Net. Although the speed is slightly lower than the previous two in proposal detection, the algorithm shows significant cost savings in terms of additional storage space. More importantly, Fast R-CNN achieves a significant breakthrough in improving the accuracy and efficiency of target detection, thus promoting the further development of the target detection field.

(4) Faster R-CNN proposed by S. Ren et al [18] in 2015 further improves the Fast R-CNN algorithm and achieves end-to-end detection in the true sense. The core of this algorithm is the introduction of a new region candidate network (RPN), which greatly improves the efficiency and accuracy of target detection.

(5) YOLO (You Only Look Once) is a real-time target detection algorithm which was introduced by Redmon J et al [19] in 2016.YOLO treats the task of target detection as a single regression problem to be solved, and its core idea is to divide the input image into SxS lattices, each of which is responsible for predicting a fixed number of bounding boxes as well as these bounding boxes' class probabilities.The various versions of YOLO include YOLOv1, YOLOv2, YOLO9000, YOLOv3, YOLOv4, YOLOv5, YOLOR, YOLOX, YOLOv6, YOLOv7, and YOLOv8.The development of YOLO algorithm is shown in Fig. 2.

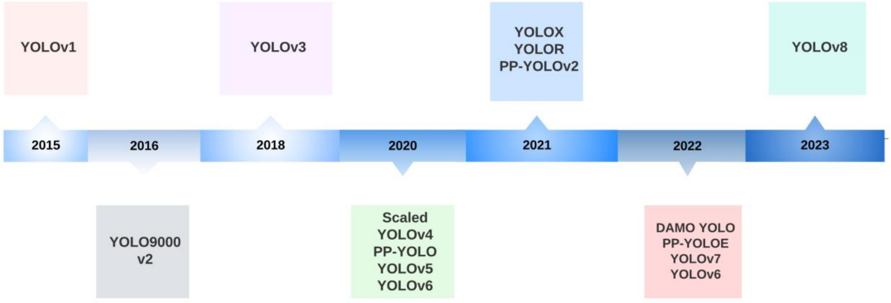

Figure 2: YOLO Development History

(6) SSD is divided into two parts, i.e., backbone models and SSD head. Backbone models are mainly responsible for feature extraction in SSD algorithms. It usually uses pre-trained deep convolutional neural networks such as VGG16 or ResNet. These networks have been trained on a large amount of image data and are able to learn a rich representation of visual features. The SSD head, on the other hand, is the part responsible for target detection on the feature map extracted by the backbone model. It contains multiple convolutional layers for further extraction and integration of feature information. At the core of the SSD head are multiple prediction layers that are connected to different layers of the backbone model for detection using multi-scale feature maps.

2.3 Evaluation indicators for target detection

Model evaluation metrics for target detection algorithms are a key basis for evaluating model performance, and they can comprehensively reflect the model's performance in target detection tasks. The following are some common target detection algorithm model evaluation metrics:

(1) Confusion Matrix

The Confusion Matrix has a statistical table of classification results that can show the model's prediction for each category. Through the confusion matrix, we can calculate the number of true cases (TP), false positive cases (FP), true negative cases (TN), and false negative cases (FN), and then calculate the other evaluation metrics [20].TP means that the model predicts a positive case and is actually a positive case as well, TN indicates that the model predicts a negative case and is actually a negative case as well, FP means that the model predicts a positive case but is actually a negative case, and FN indicates that the model predicts a negative case but is actually a positive case[21].

(2) Precision and Recall

The concepts of Precision and Recall are also mainly used to evaluate target detection methods [22]. Precision indicates the proportion of samples predicted by the model to be positive cases that are actually positive cases. It reflects the proportion of target frames predicted by the model that actually contain targets. The precision rate P is calculated using TP (true positive examples) and FP (false positive examples). Precision is calculated using the following formula:

$$P = TP / (TP + FP) \tag{1}$$

Recall indicates the proportion of samples that are actually positive examples that are predicted to be positive by the model. It reflects the proportion of targets that the model was able to detect out of the total number of actual targets. It is calculated using TP (true cases) and FN (false negative cases). Recall R is calculated using the following formula:

$$R = TP / (TP + FN) \tag{2}$$

(3) Average precision (AP) and mean average precision (mAP)

AP is a precision metric for a single category that takes into account the precision rate at different recall levels. In target detection, AP is usually obtained by plotting the precision-recall curve (P-R curve) and calculating its area. A higher value of AP implies a better performance of the method, and vice versa. mAP is the average value of AP for multiple categories, which is used to evaluate the average performance of the model over all categories. mAP is one of the most commonly used evaluation metrics in target detection tasks, and it provides a comprehensive picture of the model's detection performance on different categories [23]. mAP is calculated using the following equation:

$$mAP = \frac{1}{N} \sum_{i=0}^{N} AP_i \tag{3}$$

Here, N is the total number of classes and APi is the average precision of the ith class.

(4) Intersection and Union Ratio (IoU)

In a target detection task, the localisation of various classes of objects is achieved by predicting a bounding box around the objects. This step is crucial for accurately identifying and separating different objects in an image. IoU is mainly used to measure the degree of overlap between the predicted target bounding box and the true target bounding box. Specifically, IoU is calculated as the ratio of the intersection area to the concatenation area of the predicted border to the real border. If the value of IoU is higher, it means that the degree of overlap between the predicted and real borders is higher, and the accuracy of the model's prediction is also higher. The formula for calculating IoU is as follows:

$$IoU(BB_g, BB_p) = \frac{area(BB_g \cap BB_p)}{area(BB_g \cup BB_p)} \tag{4}$$

Where BBg denotes the true bounding box and BBp denotes the predicted bounding box.

(5) F1 Score (F1-Score)

The F1-Score takes into account both Precision and Recall to provide a comprehensive and accurate view of performance evaluation. F1-Score is the reconciled average of Precision and Recall, and is calculated by the following formula:

$$F1 = (2 \times P \times R) / (P + R) \tag{5}$$

By calculating the F1 score, we can obtain a single value to evaluate the performance of the classifier. The F1 score takes values between 0 and 1, where 1 indicates a perfect classifier and 0 indicates the worst classifier.

3. Recognition of unsafe behaviour of construction workers based on target detection

Target detection algorithms based on CNN networks (R-CNN, Faster R-CNN, YOLO, SSD, etc.) have been widely used to identify various factors in architectural scenes, as shown in Fig. 3.

Unsafe behaviours that lead to accidents are classified as (1) personal protective equipment (PPE) failures; (2) exposure to hazardous areas; (3) failure to follow safety procedures; and (4) unsafe devices. Below we provide an overview of the four areas.

3.1 PPE failure

In the specific environment of a construction site, there is a wide range of PPE, including but not limited to helmets, protective glasses, protective clothing, gloves, safety shoes, safety vests, and respirators. After extensive research, it has been shown that the main cause of accidents is not malfunctioning PPE but due to the fact that people simply do not wear PPE [24, 25]. In earlier studies, Du et al [26] proposed a machine learning combined with image processing helmet detection model, which incorporates Haar's facial features and detects the presence of a helmet based on the colour above the face. Park et al designed a new algorithm, which determines whether a worker is wearing a helmet or not by the spatial feature relationship between the helmet and the person's face.W. Fang et al used a Faster R-CNN to identify the wearing of seat belts and helmets, respectively, with detection accuracies as high as 80% and 95.7%.X. Long et al proposed an SSD-based helmet detection algorithm, which was supported by most of the people. The YOLO family is also widely used in the target detection of PPE faults, such as F. Wu et al to YOLOv3 to improve and design a helmet detection algorithm.

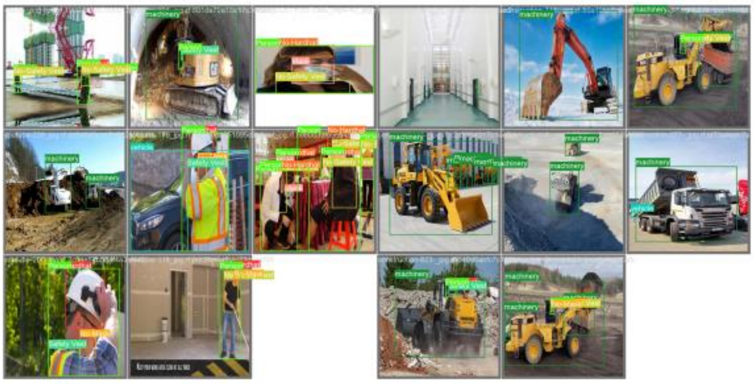

Figure 3: Example of target detection

3.2 Exposure to hazardous areas

In the construction site, which is an operating environment full of complexity and variability, unsafe behaviours exposed to hazardous areas undoubtedly pose a serious threat to the life safety and health of workers. It has been shown that such hazards are mainly classified into two categories, i.e., static hazards and dynamic hazards. For example, Fahad Lateef et al designed an algorithm based on the combination of semantic segmentation and deep learning to identify workers approaching or entering unprotected sides (floor edges, roofs, balconies, roof hatches, pits, etc.), and used semantic segmentation methods to identify the scene (i.e., pit, roof hatch) and the ontology in order to reason about the presence of safety barriers on a range of sides. Daeho Kim and others applied the Yolo-v3 method to locate objects in order to identify workers hitting or approaching bulldozers, excavators, etc. that were under construction, and used an image correction method to measure the distances between the objects, yielding accurate results. Nonetheless, so far the research has not been able to determine when a hazardous work area is not protected, and research in this area will continue.

3.3 Failure to comply with security procedures

For human unsafe behaviours and actions, etc., most of the studies are conducted in conjunction with 3D models of people, which tend to use depth sensors to extract 3D features of people. For example, Ding et al employed CNNs in order to achieve automatic extraction of visual features in videos, and then used LSTMs to capture the sequence information of these features, which led to a more comprehensive analysis of the spatio-temporal characteristics of unsafe behaviours.

3.4 Unsafe devices

The identification of unsafe devices is also a very important aspect in the identification of unsafe behaviour of construction workers. Some devices on construction sites may lead to serious work accidents and personnel injuries due to their design defects, improper maintenance, or improper use.M. Doherty et al made a review of applications in electrical safety based on the development of Artificial Intelligence, which provides readers with a reliable basis, and Xinyu Mei et al developed a computer vision-based personnel hazardous area intrusion detection method, mainly for static hazardous source scenarios.

4. Emerging trends in target detection research in construction

4.1 Transformer-based target detection

Traditional approaches to target detection often rely on hand-crafted features and complex processes, which are both time-consuming and labour-intensive. With the advent of deep learning, and in particular the introduction of the Transformer architecture, target detection has made significant progress in recent years. The Transformer architecture was initially proposed for natural language processing tasks, but it has recently been extended to the field of computer vision. The core component of the Transformer is the self-attention mechanism, which allows the model to assign different weights on the importance of different parts of the input data. This allows Transformer to capture long-range dependencies and global context, which is crucial for target detection tasks. The Transformer architecture shows great potential for capturing global dependencies and has been successfully applied to target detection tasks. Future research directions include improving the efficiency and accuracy of Transformer-based target detection methods, as well as exploring their application to other computer vision tasks.

4.2 Attentional mechanisms and multiscale detection

Attention Mechanism (AM) is an important technique in target detection that helps the model to focus on the critical regions more effectively by giving different weights to different regions in the image. This mechanism enables the model to identify the critical parts of the image that are relevant for target detection and ignore irrelevant background information. Multi-scale Detection, on the other hand, refers to target detection at different scales. This approach takes into account that the target object may appear in different parts of the image, and thus detection at different scales can improve the accuracy and robustness of the detection.

For example, in multi-scale detection, the YOLOv5s algorithm can be used to improve its feature extraction capability by introducing an attention mechanism to improve the accuracy of target detection. In addition, an attention-based approach can be used to generate anchor boxes (anchor boxes) to enhance the features of object regions and reduce the influence of background. Xiao et al proposed a new image target detection method capable of multi-feature selection on multi-scale feature maps.Ju et al proposed an adaptive feature fusion with an attention mechanism using the global attention and spatial location attention mechanisms to learn the correlation of channel features and the importance of spatial features at different scales, respectively; Zhang et al explored a method for small target detection, which improves the detection performance of small targets by combining a hierarchical attention mechanism and multi-scale separable detection.

4.3 Mabam-based target detection

Mabam is a novel sequence model developed based on the Selective State Space Model (SSM) or S4 model. This model exhibits performance that matches or even surpasses Transformer on multiple linguistic tasks with linear complexity and higher inference throughput.

LianghuiZhu et al proposed a new generic vision backbone called Vision Mamba (Vim), which uses bi-directional Mamba blocks. Vim compresses the visual representation by labelling the positional embeddings of the image sequences and using a bi-directional state space model. This research demonstrates the higher performance of Vim compared to established vision Transformers such as DeiT on ImageNet classification, COCO object detection, and ADE20k semantic segmentation tasks, as well as significantly improved computational and memory efficiency. Moreover, the current literature on Mamba is very limited to give more examples, but Mamba will eventually become a new research trend.

5. Conclusions

Target detection is a very popular topic in the field of computer vision and deep learning. In this paper, firstly, the three major classes of algorithms for target detection are elaborated in detail, traditional target detection mainly relies on the method of machine learning, the two-phase target detection algorithm based on deep learning is mainly divided into two phases of candidate region production and target detection, while the one-phase target detection algorithm based on deep learning carries out end-to-end target detection without the need to produce a candidate region, and it gives an evaluation of the performance of the target detection indicators to evaluate the advantages and disadvantages; then the current applications of target detection in the construction field and the new trends in the future are summarised and analysed.

References

[1] R. Y. Sunindijo, P. X. W. Zou.: Political skill for developing construction safety climate. ASCE Journal of Construction

Engineering and Management, 138 (5) (2012), pp. 605-612. https://doi.org/10.1061/(asce)co.1943-7862.0000482

[2] Heng Li, Miaojia Lu, Shu-Chien Hsu, Matthew Gray, Ting Huang:Proactive behavior-based safety management for construction safety improvement,Safety Science,Volume 75,2015,Pages 107-117,ISSN 0925-7535,https://doi.org/10.1016/j.ssci.2015.01.013.

[3] P.E.D. Love, P. Teo, J. Morrison:Unearthing the nature and interplay of quality and safety in construction projects: an empirical study Saf. Sci., 103 (2018), pp. 270-279

[4] Q. Fang, H. Li, X. Luo, L. Ding, T.M. Rose, W. An, Y. Yu :A deep learning-based method for detecting non-certified work on construction sites .Adv. Eng. Inform., 35 (2018), pp. 56-68.https://doi.org/10.1016/j.aei.2018.01.001

[5] Y. Lin, Z. Nie, H. Ma:Structural damage detection with automatic feature-extraction through deep learning. Computer-Aided Civil and Infrastructure Engineering, 32 (2017), pp. 1025-1046.https://doi.org/10.1111/mice.12313

[6] D. Kim, M. Liu, S. Lee, V.R. Kamat:Remote proximity monitoring between mobile construction resources using camera-mounted UAVs. Autom. Constr., 99 (2019), pp. 168-182, 10.1016/j.autcon.2018.12.014

[7] W. Fang, B. Zhong, N. Zhao, P.E.D. Love, H. Luo, J. Xue, S. Xu:A deep learning-based approach for mitigating falls from height with computer vision: convolutional neural network.Adv. Eng. Inform., 39 (2019), pp. 170-177.https://doi.org/10.1016/j.aei.2018.12.005

[8] Chengqi Li, Zhigang Ren, and Bo Yang "Binary image filtering for object detection based on Haar feature density map", Proc. SPIE 10613, 2017 International Conference on Robotics and Machine Vision, 1061303 (19 December 2017); https://doi.org/10.1117/12.2300505

[9] Jin Li, Hong Zhang, Lei Zhang, Yawei Li, Qiaochu Kang, and Yujie Wu "Multi-scale HOG feature used in object detection", Proc. SPIE 11069, Tenth International Conference on Graphics and Image Processing (ICGIP 2018), 110693U (6 May 2019); https://doi.org/10.1117/12.2524169

[10] Jin Liu, WenBing Tao, and Han Zheng "Confidence evaluation for cascade classifier in object detection", Proc. SPIE 8003, MIPPR 2011: Automatic Target Recognition and Image Analysis, 800306 (8 December 2011); https://doi.org/10.1117/12.901981

[11] Kaur, J., Singh, W. Tools, techniques, datasets and application areas for object detection in an image: a review. Multimed Tools Appl 81, 38297–38351 (2022). https://doi.org/10.1007/s11042-022-13153-y

[12] R. Girshick, "Fast R-CNN," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, pp. 1440-1448, doi: 10.1109/ICCV.2015.169.

[13] Mohamad Ali-Dib, Kristen Menou, Alan P. Jackson, Chenchong Zhu, Noah Hammond:Automated crater shape retrieval using weakly-supervised deep learning,Icarus,Volume345,2020,113749, ISSN00191035,https://doi.org/10.1016/j. i carus. 2020.113749.

[14] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014, pp. 580-587, doi: 10.1109/CVPR.2014.81.

[15] Liang, M. & Hu, X. Recurrent convolutional neural network for object recognition. In IEEE Conference on Computer Vision and Pattern Recognition (2015).

[16] K. He, X. Zhang, S. Ren and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 9, pp. 1904-1916, 1 Sept. 2015, doi: 10.1109/TPAMI.2015.2389824.

[17] R. Girshick, "Fast R-CNN," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, pp. 1440-1448, doi: 10.1109/ICCV.2015.169.

[18] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 1 June 2017, doi: 10.1109/TPAMI.2016.2577031.

[19] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 779-788, doi: 10.1109/CVPR.2016.91.

[20] Priya Ganapathy, Julie A. Skipper, "A novel ROC approach for performance evaluation of target detection algorithms," Proc. SPIE 6566, Automatic Target Recognition XVII, 656610 (7 May 2007); https://doi.org/10.1117/12.719063

[21] R. Padilla, S. L. Netto and E. A. B. da Silva, "A Survey on Performance Metrics for Object-Detection Algorithms," 2020 International Conference on Systems, Signals and Image Processing (IWSSIP), Niteroi, Brazil, 2020, pp. 237-242, doi: 10.1109/IWSSIP48289.2020.9145130.

[22] Hanczar, B., Nadif, M. (2019). Controlling and Visualizing the Precision-Recall Tradeoff for External Performance Indices. In: Berlingerio, M., Bonchi, F., G ärtner, T., Hurley, N., Ifrim, G. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2018. Lecture Notes in Computer Science (), vol 11051. Springer, Cham. https://doi.org/10.1007/978-3-030-10925-7_42

[23] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal

Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 1 June 2017, doi: 10.1109/TPAMI.2016.2577031.

[24] A. Hume, N. Mills, A. Gilchrist ,Industrial head injuries and the performance of the helmetsProceedings of the International IRCOBI Conference on biomechanics of impact, Switzerland (1995)

[25] H. Li, X. Li, X. Luo, J. Siebert, Investigation of the causality patterns of non-helmet use behaviour of construction workersAutom. Constr., 80 (2017), pp. 95-103, 10.1016/j.autcon.2017.02.006

[26] Shan Du, M. Shehata and W. Badawy, "Hard hat detection in video sequences based on face features, motion and color information," 2011 3rd International Conference on Computer Research and Development, Shanghai, 2011, pp. 25-29, doi: 10.1109/ICCRD.2011.5763846.