Advanced Engineering Informatics 43 (2020) 100980

Contents lists available at ScienceDirect

Advanced Engineering Informatics

journal homepage: www.elsevier.com/locate/aei

Review article

Computer vision for behaviour-based safety in construction: A review and future directions

Weili Fang[a,b,c,d], Peter E.D. Love[c], Hanbin Luo[a,b,*], Lieyun Ding[a,b]

[a] School of Civil Engineering and Mechanics, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China
[b] Hubei Engineering Research Center for Virtual, Safe and Automated Construction, Wuhan, Hubei 430074, China
[c] School of Civil and Mechanical Engineering, Curtin University, GPO Box U1987, Perth, Western Australia 6845, Australia
[d] Department of Civil Engineering and Engineering Mechanics, Columbia University, New York 10027, USA

ARTICLE INFO

Keywords:
Behaviour-based safety
Unsafe behaviour
Computer vision
Deep learning
Convolutional neural network

ABSTRACT

The process of identifying and bringing to the fore people's unsafe behaviour is a core function of implementing a behaviour-based safety (BBS) program in construction. This can be a labour-intensive and challenging process but is needed to enable people to reflect and learn about how their unsafe actions can jeopardise not only their safety but that of their co-workers. With advances being made in computer vision, the capability exists to automatically capture and identify unsafe behaviour and hazards in real-time from two-dimensional (2D) digital images/videos. The corollary developments in computer vision have stimulated a wealth of research in construction to examine its potential application to practice. Hindering the application of computer vision in construction has been its inability to accurately, and generalise the detection of objects. To address this shortcoming, developments in deep learning have provided computer vision with the ability to improve the accuracy, reliability and ability to generalise object detection and therefore its usage in construction. In this paper we review the developments of computer vision studies that have been used to identify unsafe behaviour from 2D images that arises on construction sites. Then, in light of advances made with deep learning, we examine and discuss its integration with computer vision to support BBS. We also suggest that future computer-vision research should aim to support BBS by being able to: (1) observe and record unsafe behaviour; (2) understand why people act unsafe behaviour; (3) learn from unsafe behaviour; and (4) predict unsafe behaviour.

1. Introduction

Within construction, research has repeatedly demonstrated that people's unsafe behaviour is a major contributor to accidents [97,54,56]. An array of theoretical models and metaphors have been propagated over the last century to explain people's unsafe acts and behaviours [83,32]. An notable theory is behaviour-based safety (BBS), which has been demonstrated to be an effective tool that can contribute to improving an organisation's safety performance [48,49,15,29,53,57].

A BBS approach can be used to observe and identify people's unsafe actions. Then, feedback can be provided directly to those who have committed an unsafe act with the aim of modifying their future behaviour [9,3,95,23]. It should be acknowledged, however, that BBS approaches have been widely criticised there is a proclivity for BBS to neglect the root cause of unsafe behaviour, ignore issues regarding values and attitudes, and can also hide management commitment and inadequacies [35–37,67,72]. In addition, BBS has been criticised for workers not reporting unsafe behaviours, near misses, and incidents, especially if it is related to penalties and punitive measures. Furthermore, BBS approaches may not be sustainable and in some cases fall back to the baseline when "reinforcers" are removed, explicitly indicating that the modified behaviour was controlled [13]. Regardless of the criticisms of BBS we consider it to be an invaluable approach that can be utilised to inspire people to be self-accountable and take responsibility for their unsafe actions through a process of reflection and learning [57].

To aid this process of reflection and learning, BBS comprises three phases: (1) observation; (2) feedback; and (3) training. Observation is central to monitoring and managing safety-behaviour on construction sites. Then, based on individual observable risky behaviour, feedback and training are implemented. The drawbacks of manual safety observation reporting (SOR) have been widely acknowledged [29,30]. For example, Oswald et al. [72] identified the problems with SOR that are

* Corresponding author at: School of Civil Engineering and Mechanics, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China.
E-mail address: luohbcem@hust.edu.cn (H. Luo).

https://doi.org/10.1016/j.aei.2019.100980
Received 11 February 2019; Received in revised form 16 August 2019; Accepted 26 August 2019
Available online 14 November 2019
1474-0346/

reliant on the use computer or handwritten cards, for subsequent action by the health and safety team to include (p.35):

- significantly increased administration to deliver predictable data;
- an unwelcome focus on the number rather than content of the reports;
- their use as a tool to ascribe individual or organizational blame; and
- the perception that reports are censored before they reach the health and safety team, which erodes trust between the workforce and management.

To overcome the shortcomings associated with SOR, technological developments aided by computer vision have been identified as a robust approach to automatically recognise and capture the unsafe acts committed by individuals during construction [29,30,73,4,93,14,16–19,21,80]. As a result, a rich collection of 2D images containing the actions that led to the unsafe event being performed can be obtained.

The advent of deep learning, in particular, Convolutional Neural Network (CNN), provide opportunities for computer vision-based data analysis and to overcome the problems associated with the manual observation and recording of unsafe acts. Despite the increasing attention being afforded to computer vision-based unsafe behaviour monitoring and identification in construction, there has been no comprehensive review that has been undertaken. Such a review is needed to track developments and provide an avenue for ensuring that future research not only has a robust theoretical underpinning, but is also relevant to improving the utilisation of BBS in practice. Notably, two-dimensional (2D) digital images/videos are widely used in construction. As a consequence, our research focuses on the use of 2D images that can be used to support the implementation of BBS. We commence our paper by reviewing existing CNN-based computer vision approaches. Then, a review of computer vision-based studies that have focused on identifying unsafe behaviour from 2D images are examined. Next research challenges and proposed solutions for implementing computer vision-based using deep learning within the context of BBS is presented.

2. Understanding computer vision and deep learning

Computer vision is an interdisciplinary scientific field that deals with how computational models can be made to gain high-level understanding from digital images or videos in order to automate tasks that the human visual system can do [11,39,41,51,77,79]. Developments in the field of machine learning have enabled computers to better understand what they see and as a result has bolstered developments in the area of computer vision. However, conventional machine learning approaches are limited in their ability to process natural data in their raw form [52]. This is due to there being a need to design a feature descriptor using engineering and expert experience [52]. To simplify the process of detection and pattern recognition, deep learning-based representation methods have been developed, which can automatically extract complex features end to end by learning from multiple data [52].

By combining deep learning methods (e.g. neural networks) with images that have been obtained from using computer vision, features that are not designed by human engineers can be automatically extracted and used to learn from training data. In comparison with Artificial Neural Networks (ANN), as denoted in Fig. 1, deep learning models are comprised of multiple processing layers based on neural networks that learn from representations of data with numerous levels of abstraction [52].

The most widely used deep learning method is the CNN, which consists of three main types of neural layers: (1) convolutional; (2) pooling; and (3) fully connected. The typical architecture of a CNN for detecting objects from images can be seen in Fig. 2. Within the domain of computer vision, CNN-based deep learning approaches have been widely adapted for an array of tasks such as image classification, object detection, object semantic segmentation, and pose estimation. We examine each of these tasks and deep learning-related technologies, as they form the underlying basis for applying computers to detect unsafe behaviour in construction. The structure of our review presented in this paper can be seen in Fig. 3.

2.1. State-of-the art applications in computer vision

2.1.1. Image classification

Image classification is an important task in computer vision, as it is used to identify an object that appears in an image. This task consists of labelling input images with a probability for the presence of a particular visual object class, as denoted in Fig. 4. Notably, each image in Fig. 4 has one ground truth label, followed by the top five estimates of their probability of occurrence.

A well-known classification network is the AlexNet CNN, which was able to achieve a top-5 error rate[1] of 15.3% outperforming the feature detection Scale-invariant feature transform (SIFT) algorithm with an accuracy of 26.2%. The upshot in this instance being that the CNN became a prominent classification model in computer vision with its detection accuracy being consistently improved over time with larger training sets from image classifications performed at ImageNet challenges[2] (Table 1). The Visual Geometry Group (VGG-16), for example, model proposed by Simonyan and Zisserman [81] had sixteen convolutional layers, multiple max-pooling layers and three fully-connected layers and achieved 7.3% top-5 error rate. Similarly, a deeper network called GoogLeNet (Inception V1) with 22 layers was developed by Szegedy et al. [84] and was able to achieve a 6.7% top-5 error rate.

2.1.2. Object detection

Object detection is fundamental to computer vision, as its aim is to identify an object's semantic features and locations contained within images. The work of Krizhevsky et al. [50] laid the foundation for the development of CNN-based object detection within the field of computer science. As a consequence of this pioneering work, developments with CNNs have been abounding. For example, Girshick [27] developed a Region-based convolutional neural network (R-CNN) model that was combined with a selective search, which was able to achieve a 31.4% mean Average Precision (mAP)[3] score using the 2013 ImageNet database. Likewise, Ren et al. [71] integrated a Faster R-CNN with a Region Proposal Network (RPN) to detect objects, which was identified as being the state-of-the-art for its accuracy on the PASCAL Visual Object Classes (VOC)[4] 2007, 2012 and MS COCO (Common Objects in Context) database[5] (Table 2).

Developments in object detection comprise two stages: (1) the generation of a set of candidate regions that contain objects such as a Selective Search [87], EdgeBoxes [96], DeepMask [76], and RPN [71]; and (2) the application of a CNN to classify obtained regions (i.e., the first stage) into different foregrounds or backgrounds. Object detection approaches that have been developed for their speed of detection at the expense of accuracy include You Only Look Once (Fig. 5) (YOLO), You Only Look Once 9000 (YOLO 9000) [70] and Single Shot Multibox

[1] Top-5 error rate is the fraction of test images for which the correct label is not among the five labels considered most probable by the mode.

[2] ImageNet Large Scale Visual Recognition Challenge (ILSVRC) evaluates algorithms for object detection and image classification at large scale.

[3] The mAP for a set of queries is the mean of the average precision scores for each query. The more details can be referred: https://en.wikipedia.org/wiki/Evaluation_measures_(information_retrieval)#Mean_average_precision.

[4] Pascal VOC data is a well-known set of standardized images for object class recognition. (https://github.com/shelhamer/fcn.berkeleyvision.org/tree/master/data/pascal).

[5] COCO is a large-scale object detection, segmentation, and captioning dataset (arXiv:1405.0312).

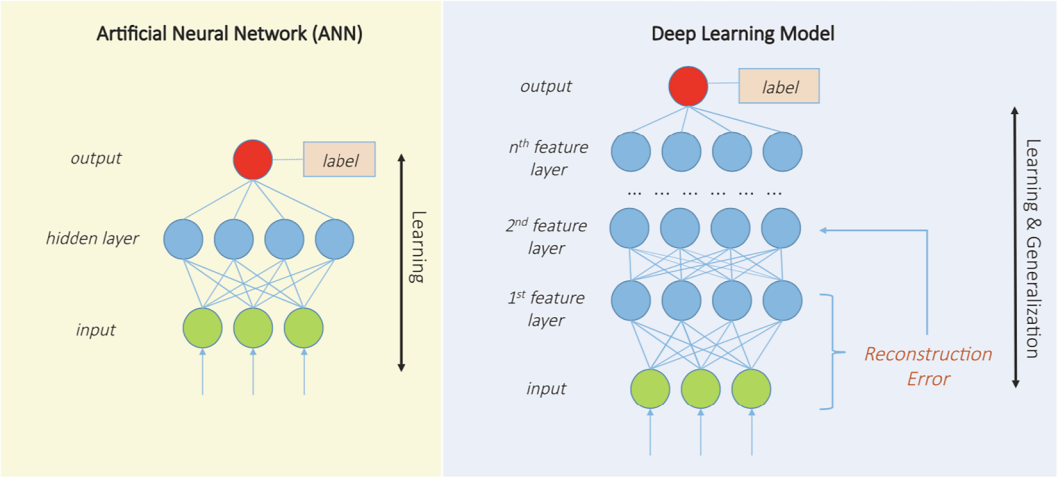


Fig. 1. Comparison between ANNs and deep learning architectures.
Source: Miotto et al. [63].

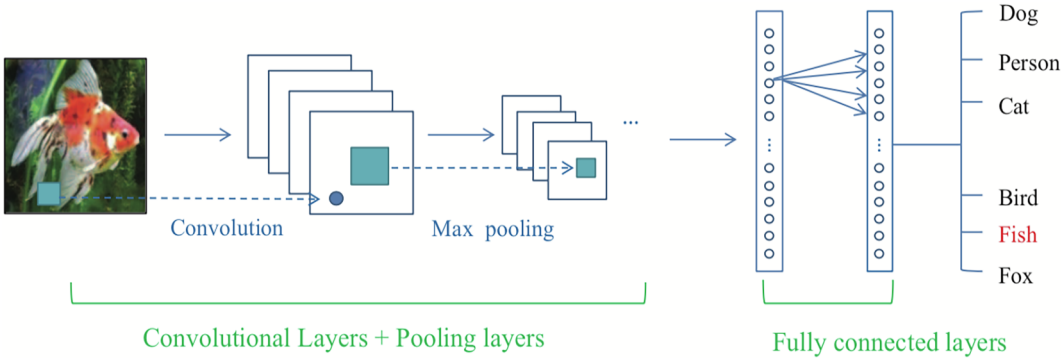


Fig. 2. The pipeline of the general CNN architecture.
Source: Guo [26].

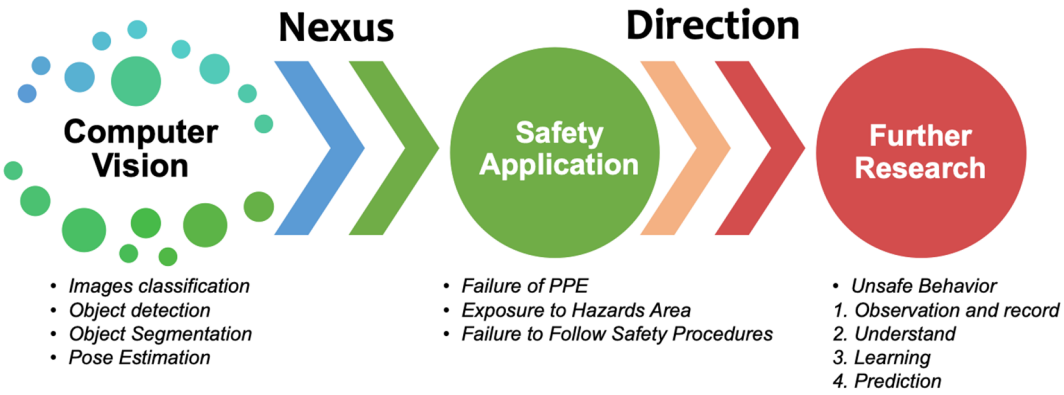


Fig. 3. Review structure.

Detector (SSD). An overview of CNN-based object detections on the PASCAL VOC and MS COCO databases are presented in Table 2.

2.1.3. Object segmentation

Semantic segmentation is typically used to locate objects and boundaries (i.e., lines, curves) from images to classify each pixel into a fixed set of categories without differentiating an object's instance [82]. The task semantic segmentation process has been simplified with the development of CNN models, which are capable of tacking the pixel-level predictions with the pre-trained network on large-scale datasets. In stark contrast to image classification and object detection, semantic segmentation requires output masks that have a 2D spatial distribution. For example, Fig. 6 presents an example where SegNet for object segmentation is used.

Three CNN-based semantic segmentation methods have been developed: [26]: (1) region-based semantic segmentation; (2) FCN-based semantic segmentation; (3) weakly-supervised segmentation. Table 3 presents prior works on the three different methods CNN-based object semantic segmentation for the COCO database.

2.1.4. Human pose estimation

The goal of human pose estimation is to determine the location of human joints from images (e.g. sequences and depth), or skeleton data as provided using motion capturing hardware [43] (Fig. 7).

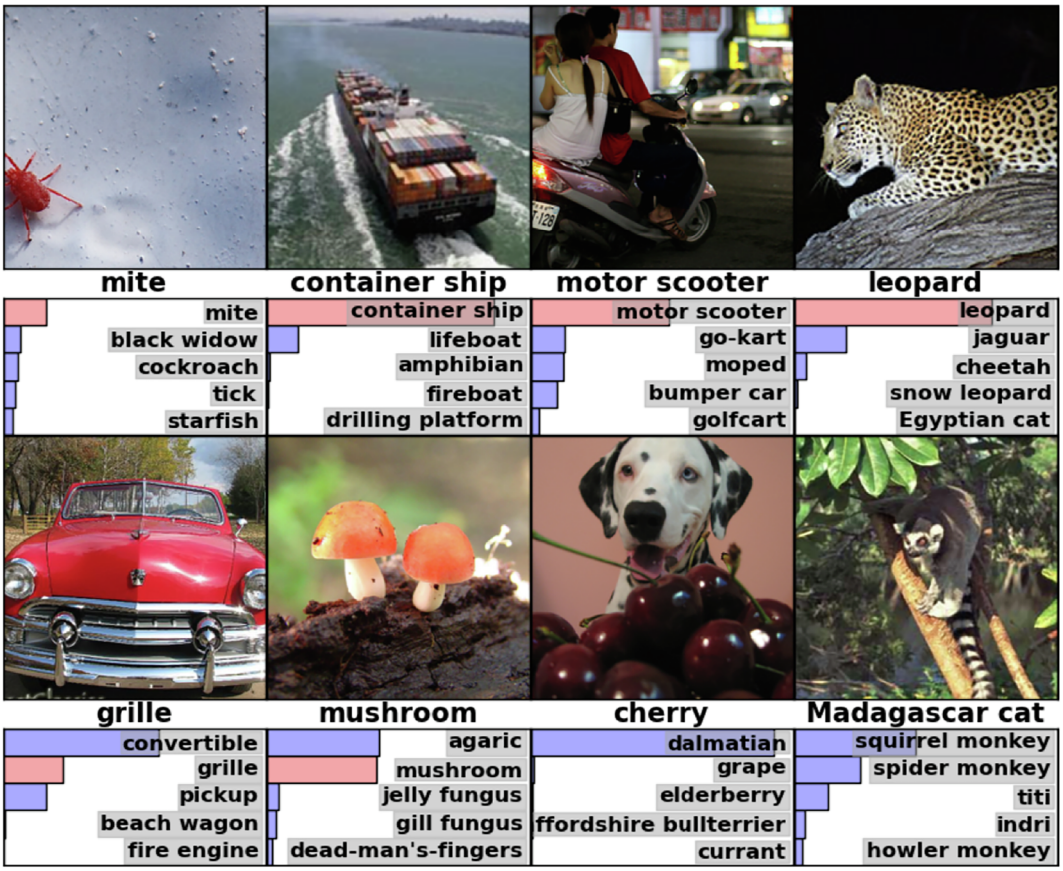


Fig. 4. Example of image classification from AlexNet.
Source: Krizhevsky et al. [50].

Table 1
Prior work on the top-5 error rates for image classification on ImageNet challenges.

Model	ImageNet 2012	ImageNet 2014	ImageNet 2015	ImageNet 2017
AlexNet	15.3%	–	–	–
VGG-16	–	7.3%	–	–
Inception V1	–	6.7%	–	–
Inception V3	5.6%	–	–	–
ResNet	3.58%	–	–	–
ResNet-152				5.5
ResNet-200				4.8
Inception-v3				5.6
Inception-v4				5.0
Inception-ResNet-v2				4.9
Inception-ResNet	4.49%	–	–	–
SE-ResNet-152	–	–	–	3.79
NASNet	3.8%	–	–	–

Determining the pose estimation of a person can be an arduous and challenging task, as consideration needs to be given to a number of aspects such as the viewpoint, illumination, and prevailing contextual backdrop, which can contain noise.

The emergence of CNNs has engendered considerable interest in human pose estimation, which can be essentially categorised as (Table 4): (1) Single-stage, which is based on a backbone network that has been well-tuned for performing image classification tasks; and (2) Multi-stage, which aims to refine the process of pose estimation.

2.2. Deep learning-related techniques

The development and application of a deep learning model to address real-world problems needs to consider the techniques of transfer learning and data augmentation to address issues associated with accuracy and reliability.

2.2.1. Transfer learning

A huge number of images are required to establish a database required for training a CNN model. In some instances, however, it is difficult to create such a database to examine specific tasks. When this situation arises, the machine learning method of transfer learning is adopted where a model developed for a task is reused as the starting point for a model on a second task. Here pre-trained models are used as the starting point on computer vision and natural language processing tasks provide the compute and time resources needed to develop neural network model. Several deep transfer learning methods have been developed, which include [40,85]: (1) instances-based; (2) mapping-based; (3) network-based; and (4) adversarial-based.

2.2.2. Data augmentation

To obtain good performance (i.e., accuracy) on computer vision with a small training database, data augmentation technique is required to produce new data by using different approaches to processing and/or combining (e.g., random rotation, crop or flips) together. The goal of data augmentation is to avoid overfitting, which can improve a model's ability to generalise. Data augmentation techniques can be divided into two types: (1) *position* (e.g., crop, resize, or horizontal flip); (2) *color*, (e.g. brightness, contrast or saturation). The AlexNet, identified above, employed two distinct forms of data augmentation: (1) generation of image translations and horizontal reflections; and (2) alternation of intensities of Red-Green-Blue (RGB) channels in training images [50]. Likewise, based on the AlexNet, Howard et al. [34] proposed an improved data augmentation strategy that extended image crops with

Table 2
CNN-based object detections on PASCAL VOC database and MS COCO database.

Model	PASCAL VOC 2007	PASCAL VOC 2010	PASCAL VOC 2012	COCO 2015	COCO 2017
R-CNN [27]	–	62.4%	–	–	–
Fast R-CNN [28]	70.0%	68.8%	68.4%	–	–
Faster R-CNN [71]	78.8%	–	75.9%	–	–
R-FCN [12]	82.0%	–	–	31.5%	–
YOLO [68]	63.7%	–	57.9%	–	–
SSD [55]	83.2%	–	82.2%	31.2(AP)	–
YOLOv2 [69]	78.6%	–	–	21.6(AP)	–
YOLOv3 [70]				33.0	
Mask R-CNN [33]	–	–	–	–	39.8%

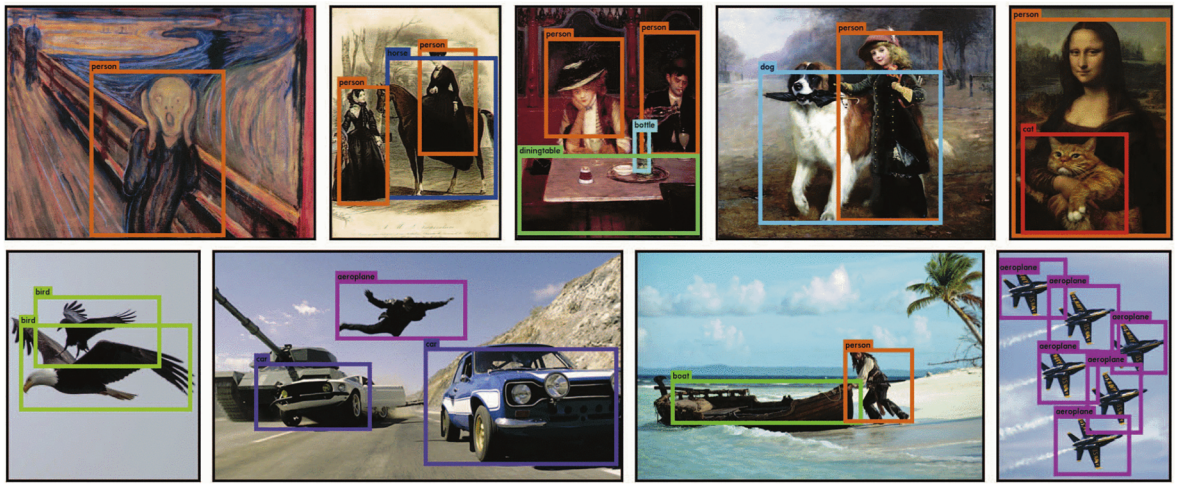


Fig. 5. Examples of YOLO-based object detection.
Source: Redmon et al. [68].

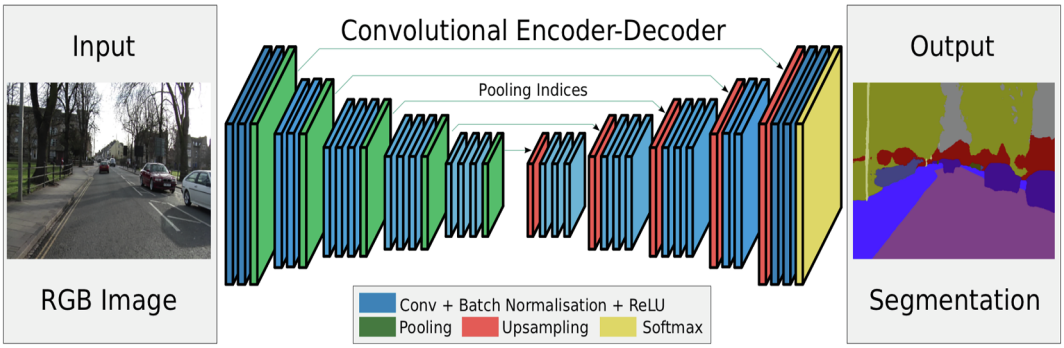


Fig. 6. An example of using SegNet architecture for object segmentation.
Source: Badrinaryanan et al. [2].

extra pixels and adding color manipulations to improve translation invariance and color invariance.

3. Research method

Our initial review of the literature relied on the use of Google Scholar's search engine to identify key works in the area of computer vision-base safety. We commenced our search by focusing on three of three keywords, 'computer vision', 'unsafe behaviour', and 'construction' that had been published between 2009 and 2019. Our search initially identified 107 papers. We manually examined each of these papers to determine their relevance to unsafe behaviour in construction and thus dismissed papers that focused on issues such as work activity recognition, and ergonomics monitoring. In sum, we found that there had been a limited number of computer vision-based papers that had sought to identify unsafe behaviour.

3.1. Definition of unsafe behaviour

Put simply, unsafe behaviour can be defined as dangerous acts that often result in injuries. When actions are likely to result in a negative outcome (i.e., injury) with high severity potential, we also view these as being unsafe. In construction, it has been demonstrated that 88% of accidents are caused by people's unsafe behaviour [97]. An analysis of 9358 accidents cases that occurred within the United States construction industry between 2002 and 2011, for example, revealed that the major three types of accidents were [8]: (1) falls from height (FFH) (43.9%); (2) being struck by a falling object (25.7%); and (3) caught in/between hazards (10.0%). Likewise, Williams et al. [90] analysis of accidents in Nigeria resulted in the following types: (1) contact with

Table 3
Comparison of CNN-based on semantic segmentation methods.
Resource: http://cocodataset.org/#panoptic-leaderboard.

	Panoptic Quality (PQ)	Segmentation Quality (SQ)	Recognition Quality (RQ)	PQ^{Th}	SQ^{Th}	RQ^{Th}	PQ^{St}	SQ^{St}	RQ^{St}
Megvii (Face + +)	0.532	0.832	0.629	0.622	0.855	0.725	0.395	0.797	0.485
Caribbean	0.468	0.805	0.571	0.543	0.818	0.659	0.355	0.785	0.438
PKU_360	0.463	0.796	0.561	0.586	0.837	0.696	0.276	0.736	0.356
ps	0.416	0.796	0.507	0.504	0.826	0.605	0.284	0.749	0.359
TeamPH	0.359	0.767	0.449	0.441	0.800	0.545	0.236	0.716	0.303
MMAP_seg	0.322	0.760	0.408	0.390	0.782	0.491	0.220	0.728	0.284
MPS-TU Eindhoven	0.272	0.719	0.359	0.296	0.716	0.394	0.234	0.723	0.306
LeChen	0.263	0.742	0.332	0.313	0.762	0.393	0.187	0.712	0.241
Artemis	0.168	0.716	0.220	0.168	0.724	0.220	0.167	0.704	0.219
grasshopyx	0.026	0.284	0.034	0.000	0.000	0.000	0.066	0.713	0.084
microljy	0.021	0.312	0.026	0.033	0.410	0.041	0.003	0.165	0.004

Note: $PQ(SQ/RQ)^{Th}$ is PQ (SQ/RQ) for things categories only, and PQ $(SQ/RQ)^{St}$ is PQ (SQ/RQ) for stuff categories only. The more details can be referred to Kirillov et al. [47].

Fig. 7. Example of human pose estimation.
Source: Lqbal and Gall [59].

objects; (2) vehicle/machine related; (3) slips and trips; and (4) falls.

According to a series of statistics/reports and the extant literature [61,62,8,29,54,24,25], the unsafe behaviour that have resulted in accidents is categorised: (1) failure of personal protect equipment (PPE); (2) exposure to hazardous area; and (3) failure to follow safety procedures.

4. Computer vision-based deep learning and unsafe behaviour

State-of-the-art CNNs (e.g., Faster R-CNN, SSD, Mask R-CNN, and YOLOv3) together with deep neural networks for object detection, such as the ZFNet, and ResNet, can be used to recognise people, plant and equipment on construction sites (Fig. 8).

Table 4
Prior works on CNN-based pose estimation.

Types of Approach	Descriptions	Authors
Single-based	Achievement of average precision of 0.649 on COCO database	Papandreou et al. [75]
	Achievement of AP^{50} at 0.859 on COCO database	He et al. [33]
	Achievement of average precision at 73.0 on COCO test-dev database	Chen et al. [7]
	Achievement of mAP of 73.7 on COCO test dev split	Xiao et al. [92]
Multiple-based	Achievement of AP^{50}[a] of 0.834 on COCO database	Wei et al. [89]
	Achievement of AP^{50} of 84.9 on COCO database	Cao et al. [5]
	Achievement of 92.0% PCKh score[b] at threshold of 0.5	Yang et al. [94]

[a] *Note:* AP^{50} (AP at OKS = 0.50). Object Keypoint Similarity (OKS) is the standard evaluation metric. The more details can be seen: http://cocodataset.org/#keypoints-eval.
[b] *Note:* PCKh (head-normalized probability of correct keypoint) score is a standard metric. The more details can be referred to Andriluka et al. [1].

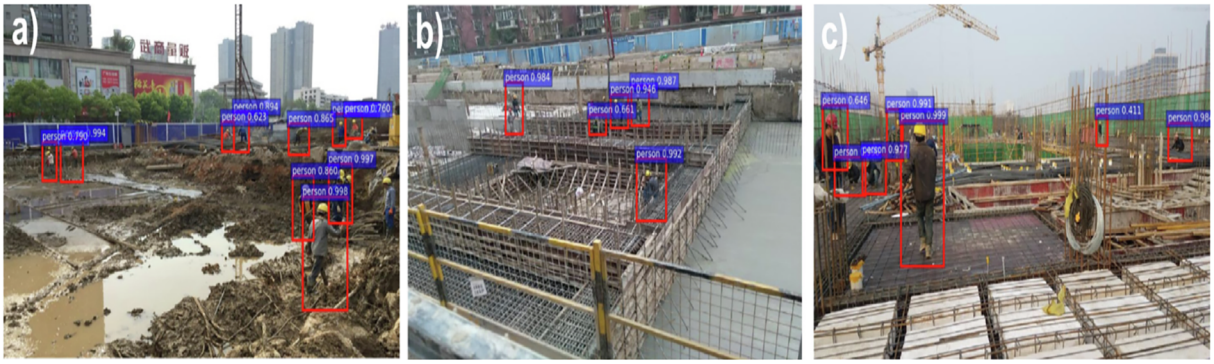


Fig. 8. Examples of deep learning-based object detection (i.e., people).
Source: Fang et al. (2018: p.148) [20].

Deep learning-based computer vision approaches therefore have the potential to accurately detect unsafe behaviour [73,53,16–19]. By reviewing studies that have utilised computer vision and deep learning to monitor safety behaviour, we identify the application areas and challenges of implementing these technologies so that future research directions can be propagated (Fig. 3).

4.1. Failure of PPE

Health and safety teams on site need to ensure that people are wearing their PPE such as: (1) hardhat; (2) high-visibility vest; (3) safety harness when working at height; and (4) appropriate footwear; (5) gloves; and (6) safety glasses. Yet, research has repeatedly shown that a significant amount of injuries that occur in construction materialise as a result of people simply not wearing their PPE [42,53]. In addressing this pervasive issue, a number of algorithms have been developed and used to recognise a person who is not wearing their PPE, which have been based on the following methods: (1) handcrafted features; and (2) deep leaning.

To extract features (e.g., shapes), from images or video descriptors such as Histogram of Oriented Gradients (HOG) [10], Histogram of Optical Flow (HoF) [88], and Bag-of-Features (BoF) [58] have all been employed. Hand-crafted feature-based methods usually employ a three-stage procedure, consisting of: (1) extraction; (2) representation; and (3) classification.

Research has been able to identify when people are not wearing their hard hat, high visibility vest and safety harness [73,64,16,18]. In the case of Park et al.'s [73] research, for example, people and hardhats are first detected by using a HOG descriptor. Then, their geometric and spatial relationships are matched (Fig. 9). Despite its success in being able to recognise when a person is not wearing their hard hat, this approach is dependent on manually designed features descriptors, which involves determining the right trade-off between detection accuracy and computational efficiency (i.e., speed). For example, one of the most powerful and robust feature detection algorithms is Scale Invariant features Transform (SIFT) [66,6].

Fang et al. [16], for example, developed a hybrid learning approach that integrated a Faster R-CNN and a deep CNN to detect people not wearing their safety harness while working at heights. Here the Faster R-CNN was used to detect people from images. Then, a new CNN model was used to classify those people who are and those that are not wearing their harness. This research, however, has limitations as it was based on a selected number of activities working at heights and the dataset was relatively modest in size. Similarly, Fang et al. [17] applied a Faster R-CNN to detect the people wearing their hard hats using a training database with approximately a 100 k images under varying conditions situations (e.g. different weather, different illumination) to validate its accuracy and reliability. While headway is being made to detect individuals that are not wearing their PPE, there has been to the authors knowledge no research that been able to identify when it is being incorrectly used.

4.2. Exposure to hazards area

Two types of hazards can generally found on construction sites [91,74,22]: (1) *Static hazards*, which form part of the design of a building, include temporary works, storage of the hazardous substances, site traffic control, and physical hazards such as an opening on a floor for services or stairwell; and (2) *Dynamic hazards,* which are the spatial-temporal movement or resources, such people and heavy equipment, and cranes with their being transported over working areas.

Research has tended to focus on how computer vision can be used prevent people from entering working areas while heavy equipment is being used [44–46]. For example, Kim et al. [44] integrated computer vision with a fuzzy inference method to monitor and assess a person's safety while working in the vicinity of heavy plant (Fig. 10). In this instance, crowdedness and proximity to the plant were used to assess safety levels. Despite the systems potential, Kim et al. [44] acknowledged that several improvements were required if their approach was to be applied during construction to enable real-time detection, which included: (1) consideration of the plant's operational status; and (2) achieving a greater level accuracy when dealing with high dimensional image data where there is a presence of clutter, numerous resources (e.g. people, and plant), varying poses and differing scales.

Similarly, Kim et al. [45] developed a hazard avoidance system by combining computer vision with augmented reality to proactively inform individuals of likely dangers (e.g., hazard orientation, distance, and safety level). However, occlusions had an adverse effect on the performance of the object detection.

Building on this earlier works, Fang et al. [21] utilised a Mask R-CNN and computer vision to determine when people entered a hazardous work area. In this instance Fang et al. [21] sought to recognise individuals that traversed structural supports from an array of images, and reported recall and precision rates with 90% and 75%, respectively. Akin to previous studies, occlusions hindered its accuracy, which stymie its use in practice.

While there have been several attempts to address static hazards, their detection remains a vexing problem (Table 4). For example, a recurring unsafe behaviour that people commit, despite consciously knowing that their actions are dangerous, is to enter excavations that are unsupported. Still, research, up until this point in time, has not been able to identify when dangerous work areas are unprotected. Furthermore, computer vision research has not been able to accommodate the changing nature of unsafe conditions. For example, when an individual becomes in close proximity to a crane's working. In this case, there is a need for digital technologies (e.g., sensors and Internet of Things) to be combined with computer vision to extract features and deep learning approach to improve the accuracy of detecting the individual's

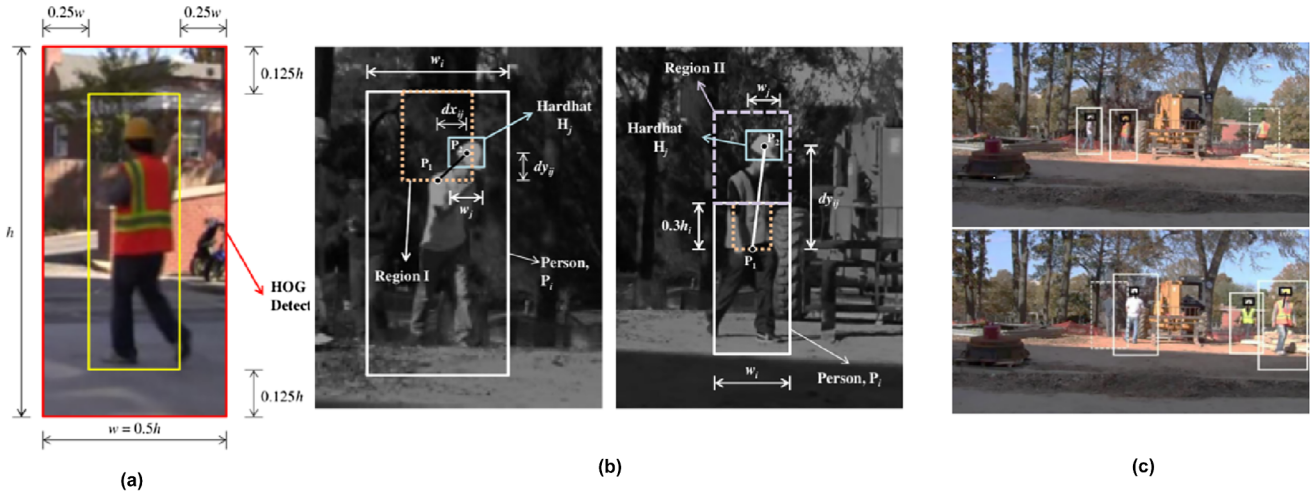
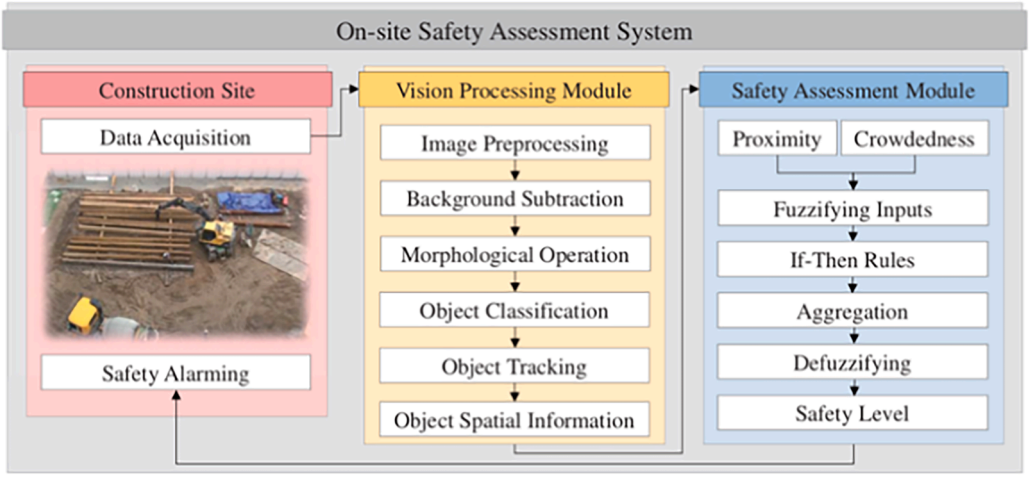


Fig. 9. HOG description approach to detect worker not wearing safety hardhat.
Source: Park et al. [73].

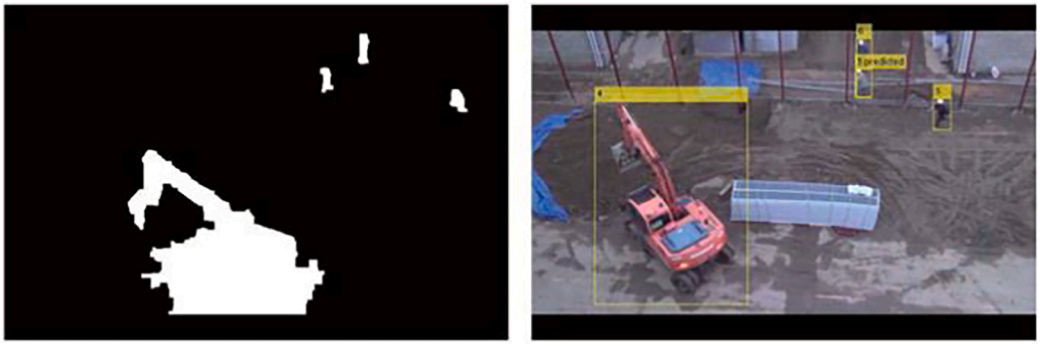
presence.

4.3. Failure to follow safety procedures

Previous studies identifying a person's abnormal behaviour have generally focused on using a skeletal and deep learning-based approaches [29–31,96,23,14,19]. Skeleton-based approaches have tended to rely on the use of depth sensors to extract three-dimensional (3D) models of a person to identify their unsafe actions. For example, Han and Lee [29] utilised a depth camera to collect motion data and develop a template to construct a 3D skeleton model. Then, Han and Lee [29] compared the predefined template with motion data to identify abnormal unsafe actions. This research, however, has limitations as it was only used in an outdoor environment where an individual was subjected to a limited range of movements. Furthermore, its accuracy was thwarted as the camera was sensitive to light.



(a) Overview of on-site safety monitoring system



(b) Detection Results

Fig. 10. Vision-based monitoring struck-by accidents with moving objects.
Source: Kim et al. [44].

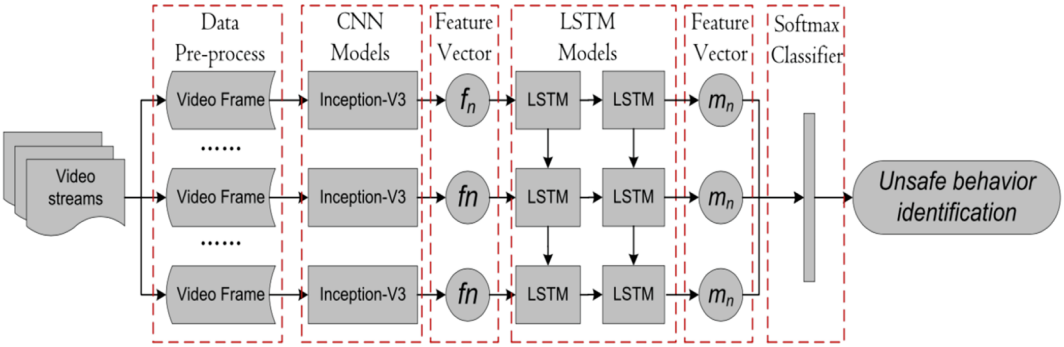


Fig. 11. Example of using CNN-LSTM model to identify worker unsafe behaviour.
Source: Ding et al. [14].

Using deep learning to automatically extract and learn features from videos Ding et al. [14] integrated a CNN and long short-term memory (LSTM) to extract spatial and temporal information of an individual's unsafe behaviour (e.g., abnormal climbing). Here Ding et al. [14] used the CNN to automatically extract visual features from videos and applied an LSTM to extract sequence features (Fig. 11). Likewise, Fang et al. [19] developed a deep learning-based framework to determine whether an individual was working in an area where they had been certified to do.

4.4. Challenges of deep learning

While strides have been made to identify unsafe behaviour, no fully automatic continuous vision-based system has been developed. Previous studies presented in this review relying on the use of CNNs have been limited to specific-tasks, where their accuracy of detection has been low and their ability to be generalised restricted. Thus, in developing a new approach there is a need for it to be (1) more accurate; (2) require less effort in hand-designing solution; and (3) generalisable well across related tasks and construction sites. To address the existing limitations, computer vision in conjunction with deep learning can enable headway to be made to automatically identify unsafe behaviour. There are, however, challenges that inhibit the development of a fully automatic continuous-based unsafe behaviour identification system that include:

- *Lack of training data*: Deep learning models for mapping and identifying unsafe behaviour require large datasets for training. To the best of our knowledge, there are no publicly available datasets of unsafe behaviours that are large enough for training in construction. Compared with public datasets in computer science (e.g., ImageNet and COCO), datasets need to have their own characteristics that can accommodate the nuances of construction (e.g., cluttered backgrounds, occlusions, varying poses and the scale of objects). In light of the lack of available datasets for training, researchers in construction are required to manually create their own by tagging images. This can be a time consuming, tedious, and expensive process.
- *Weak generalisation:* As a consequence of having to create databases, they tend to be small and thus are reliant on the use of supervised approaches, which can impede the ability to provide generalisations. The reasons are twofold: (1) previous studies have assumed that training and testing database are balanced. This is a theoretical assumption, but in reality, the dynamic nature and complexity of construction means that we need to assume they are unbalanced to better reflect practice; (2) underlaying machine learning models typically use a small database for training, which can limit inter and intra-class variability. As a result, this impedes their ability to accurately recognise unsafe behaviour and enable generalisations to different datasets [86].
- *Lack of metrics for performance evaluation*: For the analysis of experimental results, researchers tend to use different datasets. The varying size of the samples contained within each dataset and the reported evaluation metrics (e.g., precision, recall, and accuracy), renders it difficult to compare and contrast the performance of studies, particularly when different algorithms have been used. Thus, there is a need for common and objective criteria that can be used to evaluate the process of behaviour recognition.
- *Inability to identify unsafe behaviour due to changing safety requirements.* Prevailing computer vision approaches have been developed with low levels of information utilisation and therefore require higher levels accuracy to identify unsafe behaviour. As safety rules are modified due to changes in legislation, computer vision approaches will accordingly need to be adapted to accommodate such vagaries, otherwise it will become a tool that BBS approaches will not be able to utilize.
- *Inability to detect small or hidden objects.* Most of well-known CNN-based object detection approaches, such as SSD, YOLO, Faster R-CNN, are not able to effectively identify small objects [38,65]. This is a major problem when capturing and identifying people's behaviour from a distance (Fig. 12a). In addition, individuals can be difficult to detect due to occlusions and the limited number of cameras that may be made available for use on site (Fig. 12b).
- *Inability to extract multiple-attributes.* An unsafe act may involve breaking a series of safety rules and therefore multiple features may need to be extracted. For example, we may need to extract hoisting information (e.g., speed, status, and activities) to identify is the presence of an unsafe event such as "the speed is not smooth, uniform, with sudden braking during hoisting". However, it is not currently feasible to extract a wide range of features using computer vision.

5. Overcoming the challenges of deep learning

To address the above challenges and ensure computer vision can effectively and accurately identify unsafe behaviour, we provide suggestions for future research in the emergent area of deep learning, which include:

- *Solutions to address data problems.* To help improve the detection accuracy and generalisation of deep learning, we suggest that: (1) unsupervised learning or semi-supervised learning can be used to develop video streams as training data can be readily from extracted images. This process not only addresses the issues associated with limited training data, but also the problem of assuming that the distribution of training and testing database are identical; (2) data augmentation techniques can be utilised to increase training performance; and (3) the transfer of learning techniques can be used to pre-train models and fine-tune them with a small amount of manually created data.

(a) Varying size (b) Unseen object (partial)

Fig. 12. Examples of varying size's objects in construction sites.

- *Features extraction for aiding vision-based unsafe behaviour identification.* We suggest that computer vision can be aided by digital techniques (e.g., sensors, and Internet of Things). For example, gravity acceleration sensors can be used to monitor mechanical parameters during hoisting (e.g. its speed), and computer vision to identify the crane's activities. In this instance, for example, "the weight of suspended objects exceeds machine's rated load" can be monitored.
- *Content-aware-based unsafe behaviour understanding.* To address this problem, we suggest that computer vision can be combined with natural language processing (NLP) by using content-aware information to develop a reasoning model to assist with the identification unsafe behaviour. As a result, this will enable computer vision to accommodate changes in safety regulations that may materialise; and
- *Addressing small and hidden objects.* We suggest that: (1) an Unmanned Aerial Vehicle (UAV) or 360-degress cameras can be used to ensure that operations on site are constantly monitored and viewed in real-time; and (2) optimised deep learning approaches can be developed to address the problems of occlusions.

6. Directions of future research

The detection of unsafe behaviour is an innate feature of BBS. If we can be not able to identify the unsafe behaviour, then it is unable to be managed and changed. After accurately identifying unsafe behaviour from videos or images, we suggest that the obtained results can be used by the health and safety team to inform and educate individuals about the need to perform their work safely. We therefore suggest that when computer vision is combined with deep learning additional insights to support BBS can come to the fore. We therefore propose, in Fig. 13, at framework to illustrate how deep learning can computer vision can be utilised within a BBS program as can be used to: (1) observe and record; (2) understand; (3) learn; and (4) predict unsafe behaviour.

6.1. Observation and record

By identifying a wide range of unsafe behaviours, a health and safety team will need to be able to identify a culprit who performs such actions and then provide them with direct feedback about their unsafe actions. To achieve this goal, we suggest that an individual's identity needs to be recognised, which can be undertaken using computer vision. As a result, an individual's unsafe behaviour(s) can be recorded and analysed (e.g.., frequency, types, location and time). Thus, two solutions can be used to aid this process: (1) use of sensors to identity the person's identity and their location. Then, computer vision can be used to monitor the person's activities and derive their location from coordinates extracted from videos. Next, the information obtained from the sensors (e.g., identity and location) and computer vision (e.g., activities and location) are synchronised according to the person's coordinates; and (2) the development a deep learning approach to identify individuals from video streaming by integrating temporal and spatial information to extract features.

6.2. Understand

Understanding why people perform unsafe acts (i.e. violations) is an issue that health and safety team will seek to acquire knowledge about. Such acts may arise intentionally or unintentionally. Breaking rules has generally been associated with deviant behaviour, but there may be instances when committing a violation may have arisen out of taking initiative rather than negligence or malice. Furthermore, a violation may "even be a necessary way of testing rules and the truces around them" [4]. In making inroads to understanding a safety rule violation it is necessary for the health and safety team to realise the way people construct the intentions that lie behind it to ensure recidivism is mitigated. Computer vision can therefore be used to provide a context to better understand why unsafe acts have been performed.

6.3. Learning

Training, an active learning approach, has been regarded as an effective way to improve the awareness and competence of employees/subcontractors and to cultivate a positive safety culture [57]. Despite safety training being a beneficial mechanism for engendering learning, it has several limitations: (1) as it is unable provide people with realistic experiences of the conditions that will be experienced while performing their work [78]; and (2) training programs or models tend to be separated from a person's record of committing unsafe behaviour. This causes it to be difficult to design a personalized training system for individuals' who have performed unsafe acts [60].

We suggest that computer vision can be used to address these aforementioned limitations by introducing: (1) an interactive personalized safety training recommendation system that can be developed and designed to match an individual's training needs. In this system, direct feedback is reliant on the detection of unsafe behaviour using computer vision; and (2) a Virtual Reality system to provide a sense of being on-construction site where scenarios that have been captured from the computer vision can be experienced in a safe environment.

6.4. Predict unsafe behaviour

Previous studies have focused on the detection of unsafe acts after they have occurred. However, with increasing amounts of data generated from video and a greater understanding about the conditions that

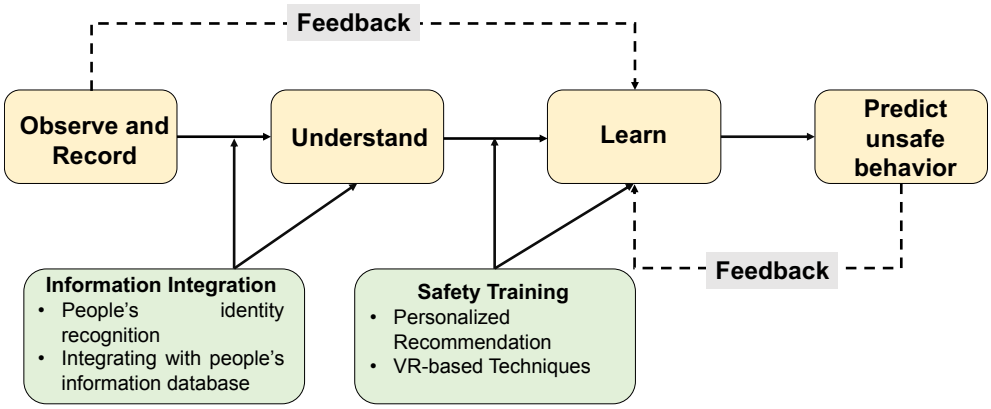


Fig. 13. Framework of deep learning-based computer vision approach for BBS.

lead to people committing unsafe acts, we will be better position to predict their occurrence. This will assist with hazard identification and enable site managers and their health safety teams to proactively manage safety performance. There remain challenges before we are able to predict unsafe behaviour, as computer vision is not yet able to model motion dynamics. Considering the developments with CNNs and LSTM for different sequence prediction tasks such as speech generation, we suggest that their combination can be extended to predicting unsafe behaviour.

7. Conclusions

In this paper we reviewed the developments of computer vision studies that have been used to identify unsafe behaviour from 2D images that arises on construction sites. Then, in light of advances made with deep learning we examined and discussed its integration with computer vision to support BBS. This leads us to propose the integration of computer vision and deep learning to aid the implementation of BBS in construction through a process of: (1) observing and recording; (2) understanding (3) learning; and (4) predicting unsafe behaviour. In light of deep learning and computer vision our proposed future research directions will not only improve BBS but also in multiple uses of these applications that could support other areas of project management in construction, such quality and real-time cost monitoring. The integration of deep learning and computer vision is an emerging area of research in construction, and thus the review we present will stimulate new lines of inquiry that will contribute improving the safety, performance and productivity of projects.

Declaration of Competing Interest

The authors declared that they have no conflicts of interest to this work. We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

Acknowledgement

The authors would like to acknowledge the financial support provided by the National Natural Science Foundation of China (Grant No. 71732001, No. 51678265, No. 51978302, No. 51878311, 71821001) and China Scholarship Council (CSC).

References

[1] M. Andriluka, L. Pishchulin, P.V. Geher, B. Schiele, 2d human pose estimation: New benchmark and state of the art analysis, 2014, pp. 3686–3693. http://openaccess.thecvf.com/content_cvpr_2014/papers/Andriluka_2D_Human_Pose_2014_CVPR_paper.pdf.

[2] V. Badrinaryanan, A. Kendall, R. Cipolla, SegNet: a deep convolutional encoder-decoder architecture for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (12) (2017) 2481–2495, https://doi.org/10.1109/tpami.2016.2644615.

[3] L. Ding, S. Guo, Y. Zhang, M.J. Skibniewski, K. Liang, Hybrid recommendation approach for behavior modification in the Chinese construction industry, Autom. Constr. (2019), https://doi.org/10.1061/(ASCE)CO.1943-7862.0001665.

[4] J. Busby, M. Iszatt-White, Rationalizing violation: Ordered accounts of intentionality in the breaking of safety rules, Organiz. Stud. 37 (1) (2016) 35–53, https://doi.org/10.1177/0170840615563590.

[5] Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh, Realtime multi-person 2d pose estimation using part affinity fields, arXiv preprint arXiv:1611.08050, 2016.

[6] H. Chen, L. Zhang, J. Ma, J. Zhang, Targer heat-map network: An end-to-end deep network for target detection in remote sensing images, Neurocomputing 331 (2019) 375–387, https://doi.org/10.1016/j.neucom.2018.11.044.

[7] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, J. Sun, Cascaded pyramid network for multi-person pose estimation, arXiv preprint, 2018.

[8] S. Chi, S. Han, Analyses of systems theory for construction accidents prevention with specific reference to OSHA accidents reports, Int. J. Project Manage. 31 (2013) 1027–1041, https://doi.org/10.1016/j.ijproman.2012.12.004.

[9] R.M. Choudhry, Behaviour-based safety on construction sites: a case study, Accid. Anal. Prev. 70 (2014) 14–23, https://doi.org/10.1016/j.aap.2014.03.007.

[10] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005, 2005, pp. 886–893, , https://doi.org/10.1109/cvpr.2005.177.

[11] H.B. Dana, M.B. Christopher, Computer Vision, Prentice Hall, 1982 ISBN:9780131653160.

[12] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, Y. Wei, Deformable convolutional networks, 2017.

[13] D.M. Dejoy, Behavior change versus culture change: Divergent approaches to managing workplace safety, Saf. Sci. 43 (2) (2005) 105–129.

[14] L. Ding, W. Fang, H. Luo, P.E.D. Love, B. Zhong, Q. Xi, A deep hybrid learning model to detect unsafe behaviour: integrating convolution neural networks and long short-term memory, Autom. Constr. 86 (2018) 118–124, https://doi.org/10.1016/j.autcon.2017.11.002.

[15] A.R. Duff, I.T. Robertson, R.A. Phillips, M.D. Cooper, Improving safety by the modification of behaviour, Constr. Manage. Econ. 12 (1994) 67–78, https://doi.org/10.1080/01446199400000008.

[16] W. Fang, L. Ding, H. Luo, P.E.D. Love, Falls from heights: a computer vision-based approach for safety harness detection, Autom. Constr. 91 (2018) 53–61, https://doi.org/10.1016/j.autcon.2018.02.018.

[17] Q. Fang, H. Li, X. Luo, L. Ding, H. Luo, C. Li, Computer vision aided inspection on falling prevention measures for steeplejacks in an aerial environment, Autom. Constr. 93 (2018) 148–164, https://doi.org/10.1016/j.autcon.2018.05.022.

[18] Q. Fang, H. Li, X. Luo, L. Ding, H. Luo, T.M. Rose, W. An, Detecting non-hardhat-use by a deep learning method from far-field surveillance videos, Autom. Constr. 85 (2018) 1–9, https://doi.org/10.1016/j.autcon.2017.09.018.

[19] 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. Inf. 35 (2018) 56–68, https://doi.org/10.1016/j.aei.2018.01.001.

[20] W. Fang, L. Ding, B. Zhong, P.E.D. Love, H. Luo, Automated detection of workers and heavy equipment on construction sites: a convolutional neural network approach, Adv. Eng. Inf. 37 (2018) 139–149, https://doi.org/10.1016/j.aei.2018.05.003.

[21] 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. Inf. 39 (2019) 179–1177, https://doi.org/10.1016/j.aei.2018.12.005.

[22] O. Golovina, J. Teizer, N. Pradhananga, Heat map generation for predictive safety planning: preventing struck-by and near miss interactions between workers-on-foot and construction equipment, Autom. Constr. 71 (2016) 99–115, https://doi.org/10.1016/j.autcon.2016.03.008.

[23] H. Guo, Y. Yu, Q. Ding, M. Skitmore, Image-based-Skeleton-based parameterized approach to real-time identification of construction worker's unsafe behaviours, J.

Constr. Eng. Manage. 144 (6) (2018) 04018042, https://doi.org/10.1061/(ASCE)CO.1943-7862.0001497.
[24] S. Guo, L.Y. Ding, H.B. Luo, X.Y. Jiang, A Big-Data-based platform of workers' behaviour: observations from the field, Accid. Anal. Prev. 93 (2016) 299–309, https://doi.org/10.1016/j.aap.2015.09.024.
[25] S. Guo, P. Zhang, L. Ding, Time-statistical laws of workers' unsafe behaviour in the construction industry: a case study, Phys. A: Stat. Mech. its Appl. 515 (2019) 419–429, https://doi.org/10.1016/j.physa.2018.09.091.
[26] Y. Guo, Y. Liu, T. Georgiou, M.S. Lew, Deep learning for visual understanding: a review, Neurocomputing 187 (2016) 27–48, https://doi.org/10.1016/j.neucom.2015.09.116.
[27] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, https://doi.org/10.1109/cvpr.2014.81.
[28] R. Girshick, Fast R-CNN, arXiv:1504.08083, 2015, 2014.
[29] S. Han, S. Lee, A vision-based motion capture and recognition framework for behaviour-based safety management, Autom. Constr. 35 (2013) 131–141, https://doi.org/10.1016/j.autcon.2013.05.001.
[30] S. Han, S. Lee, F. Peña-Mora, Vision-based detection of unsafe actions of a construction worker: case study of ladder climbing, J. Comput. Civil Eng. 27 (2013) 635–644, https://doi.org/10.1061/(asce)cp.1943-5487.0000279.
[31] S. Han, S. Lee, F. Pena-Mora, Comparative study of motion features for similarity-based modelling and classification of unsafe actions in construction, J. Comput. Civil Eng. 28 (5) (2013) 1, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000339.
[32] R.A. Haslam, S.A. Hide, A.G.F. Gibb, D.E. Gyi, T. Pavitt, S. Atkinson, A.R. Duff, Contributing factors in construction accidents, Appl. Ergon. 36 (2005) 401–415, https://doi.org/10.1016/j.apergo.2004.12.002.
[33] K. He, G. Gkioxari, P. Dolla'r, R. Girshick, Mask R-CNN, 2017 IEEE International Conference on Computer Vision (ICCV), 2017, https://doi.org/10.1109/iccv.2017.322.
[34] A. Howard, Some improvements on deep convolutional neural network based image classification, 2013. https://arxiv.org/pdf/1312.5402.pdf.
[35] J. Howe, A union critique of behaviour safety, Paper presented at the ASSE Behavioural Safety Symposium, Orlando, FL, (1998).
[36] A. Hopkins, What are we to make of safety behaviour programs? Saf. Sci. 44 (7) (2006) 583–597, https://doi.org/10.1016/j.ssci.2006.01.001.
[37] E. Hollnagel, P. Nemeth, S. Dekker, Resilience engineering perspectives, Remaining Sensitive to the Possibility of Failure vol. 1, (2008).
[38] J. Huang, V. Rathod, C. Sun, M.L. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, K. Murphy, Speed/accuracy trade-offs for modern convolutional object detectors, Available from:arXiv:1611.10012, 2017.
[39] T. Huang, Computer vision: evolution and promise, 19th CERN School of Computing, CERN, Geneva, 1996, pp. 21–25 https://doi.org/10.5170/CERN-1996-008.21 ; ISBN 978-9290830955.
[40] Z. Huang, S.M. Siniscalchi, C.H. Lee, A unified approach to transfer learning of deep neural networks with applications to speaker adaptation in automatic speech recognition, Neurocomputing 218 (2016) 448–459, https://doi.org/10.1016/j.neucom.2016.09.018.
[41] D.H. Hubel, t.N. Wiesel, Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex, J. Physiol. 160 (1962) 106–154, https://doi.org/10.1113/jphysiol.1962.sp006837.
[42] A. Hume, N. Mills, A. Gilchrist, Industrial head injuries and the performance of the helmets, Proceedings of the International IRCOBI Conference on biomechanics of impact, Switzerland, (1995).
[43] A. Kitsikidis, K. Dimitropoulos, S. Douka, N. Grammalidis, Dance analysis using multiple kinect sensors, prt, January, Proceedings of the 9th International Conference on Computer Vision Theory and Applications, VISAPP 2014, 2014, pp. 789–795.
[44] H. Kim, K. Kim, H. Kim, Vision-based object-centric safety assessment using fuzzy inference: monitoring struck-by accidents with moving objects, J. Comput. Civil Eng. 30 (4) (2016), https://doi.org/10.1061/(ASCE)CP.1943-5487.0000562, 04015075.
[45] K. Kim, H. Kim, H. Kim, Image-based construction hazard avoidance system using augmented reality in wearable device, Autom. Constr. 83 (2017) 390–403, https://doi.org/10.1016/j.autcon.2017.06.014.
[46] D. Kim, M. Liu, S. Lee, V. kamal, Remote proximity monitoring between mobile construction resources using camera-mounted UAVs, Autom. Constr. 99 (2019) 168–182, https://doi.org/10.1016/j.autcon.2018.12.014.
[47] A. Kirillov, K. He, R. Girshick, C. Rother, P. Dollar, Panoptic Segmentation, 2018. https://arxiv.org/pdf/1801.00868.pdf.
[48] J. Komaki, K.D. Barwick, L.R. Scott, A behavioural approach to occupational safety: pinpointing and reinforcing safe performance in a food manufacturing plant, J. Appl. Psychol. 63 (1978) 434–445, https://doi.org/10.1037//0021-9010.63.4.434.
[49] T.R. Krause, K.J. Seymour, K.C.M. Sloat, Long-term evaluation of a behaviour-based method for improving safety performance: a meta-analysis of 73 interrupted time-series replications, Saf. Sci. 32 (1999) 1–18, https://doi.org/10.1016/s0925-7535(99)00007-7.
[50] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Proceedings of the 25th International Conference on Neural Information Processing Systems, Lake Tahoe, 2012, pp. 1097–1105, , https://doi.org/10.1145/3065386.
[51] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (1998) 2278–2324, https://doi.org/10.1109/5.726791.
[52] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7533) (2015) 436–444, https://doi.org/10.1038/nature14539.
[53] H. Li, X. Li, X. Luo, J. Siebert, Investigation of the causality patterns of non-helmet use behaviour of construction workers, Autom. Constr. 80 (2017) 95–103, https://doi.org/10.1016/j.autcon.2017.02.006.
[54] H. Li, M. Lu, S.-C. Hsu, M. Gray, T. Huang, Proactive behaviour-based safety management for construction safety improvement, Saf. Sci. 75 (2015) 107–117, https://doi.org/10.1016/j.ssci.2015.01.013.
[55] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, SSD: Single Shot Multibox Detector, 2016. https://arxiv.org/pdf/1512.02325.pdf.
[56] 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) 270–279, https://doi.org/10.1016/j.ssci.2017.11.026.
[57] P.E.D. Love, S. Veli, P.R. Davis, P. Teo, J. Morrison, 'See the Difference' in a precast facility: changing mindsets with an experiential safety program, ASCE J. Constr. Eng. Manage. 143 (2) (2017), https://doi.org/10.1061/(ASCE)CO.1943-7862.0001224.
[58] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vision 60 (2004) 91–110, https://doi.org/10.1023/b:visi.0000029664.99615.94.
[59] U. Lqbal, J. Gall, Multi-person pose estimation with local joint-to-person associations, Lect. Notes Comput. Sci. (2016) 627–642, https://doi.org/10.1007/978-3-319-48881-3_44.
[60] S. Moon, B. Becerik-Gerber, L. Soibelman, Virtual learning for workers in robot deployed construction sites, Adv. Inform. Comput. Civil Constr. (2018) 889–895, https://doi.org/10.1007/978-3-030-00220-6_107.
[61] Ministry of Housing and Urban-Rural Development of the People's Republic of China, Quality and Safety Check Points of Urban Rail Transit Engineering, 2011. Retrieved from: http://www.zgjsjl.org.cn/uploadfile/201112/temp11121215128737.pdf.
[62] Ministry of Housing and Urban-Rural Development of the People's Republic of China, Standard for Construction Safety Assessment of Metro Engineering (GB 50715-2011), 2011. Retrieved from: http://www.spsp.gov.cn/page/CN/2011/GB%2050715-2011.shtml.
[63] R. Miotto, F. Wang, S. Wang, X.Q. Jiang, J.T. Dudley, Deep Learning for healthcare: review, opportunities and challenges, Brief. Bioinform. 19 (6) (2017) 1236–1246, https://doi.org/10.1093/bib/bbx044.
[64] B.E. Mneymneh, M. Abbas, H. Khoury, Evaluation of computer vision techniques for automated hardhat detection in indoor construction safety applications, Front. Eng. Manage. (2018), https://doi.org/10.15302/j-fem-2018071.
[65] V.N. Nguyen, R. Jessen, D. Roverso, Automatic autonomous vision-based power line inspection: a review of current status and the potential role of deep learning, Electr. Power Energy Syst. 99 (2018) 107–120.
[66] L. Nanni, S. Ghidoni, S. Brahnam, Handcrafted vs. non-handcrafted features for computer vision classification, Patt. Recogn. 71 (2017) 158–172, https://doi.org/10.1016/j.patcog.2017.05.025.
[67] M. Pillay, Taking stock of zero harm: a review of contemporary health and safety management in construction, Proc., CIB WO99 Achieving Sustainable Construction Health and Safety, Lund Univ., Sweden, 2014, pp. 75–85.
[68] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, Real-time object detection, The IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788 https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Redmon_You_Only_Look_CVPR_2016_paper.pdf.
[69] J. Redmon, A. Farhadi, YOLOv3: An Incremental Improvement, 2018. https://arxiv.org/pdf/1804.02767.pdf.
[70] J. Redmon, A. Farhadi, YOLO9000: better, faster, stronger, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, https://doi.org/10.1109/cvpr.2017.690.
[71] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, IEEE Trans. Patt. Anal. Mach. Intell. 39 (6) (2015) 1137–1149, https://doi.org/10.1109/tpami.2016.2577031.
[72] D. Oswald, F. Sherratt, S. Smith, Problems with safety observation reporting: a construction industry case study, Saf. Sci. (2018) 35–45, https://doi.org/10.1016/j.ssci.2018.04.004.
[73] M.-W. Park, N. Elsafty, Z. Zhu, Hardhat-wearing detection for enhancing on-site safety of construction workers, J. Constr. Eng. Manage. 141 (9) (2015), https://doi.org/10.1061/(asce)co.1943-7862.0000974, 04015024.
[74] J. Park, E. Marks, Y.K. Cho, W. Suryanto, Performance test of wireless technologies for personnel and equipment proximity sensing in work zones, J. Constr. Eng. Manage. 42 (1) (2016) 04015049, https://doi.org/10.1061/(ASCE)CO.1943-7862.0001031.
[75] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tomp-son, C. Bregler, K. Murphy, Towards accurate multi- person pose estimation in the wild, CVPR, vol. 3, 2017, https://doi.org/10.1109/cvpr.2017.395.
[76] P.O. Pinheiro, T.Y. Lin, R. Collobert, P. Dollár, Learning to refine object segments, Lect. Notes Comput. Sci. (2016) 75–91, https://doi.org/10.1007/978-3-319-46448-0_5.
[77] S. Ji, W. Xu, M. Yang, K. Yu, 3D convolutional neural networks for human action recognition, IEEE Trans. Patt. Anal. Mach. Intell. 35 (2013) 221–231, https://doi.org/10.1109/tpami.2012.59.
[78] R. Sacks, A. Perlman, R. Barak, Construction safety training using immersive virtual reality, Constr. Manage. Econ. 31 (9) (2013) 1005–1017, https://doi.org/10.1080/01446193.2013.828844.
[79] J. Schmidhuber, Deep learning in neural networks: an overview, Neural Netw. 61 (2015) 85–117, https://doi.org/10.1016/j.neunet.2014.09.003.
[80] J. Seo, S. Han, S. Lee, H. Kim, Computer vision techniques for construction safety and health monitoring, Adv. Eng. Inf. 29 (2) (2015) 239–251, https://doi.org/10.

1016/j.aei.2015.02.001.
[81] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, Comput. Sci. (2014) https://arxiv.org/pdf/1409.1556.pdf.
[82] L.J. Sun, Learning to segment object candidates, IEEE 13th International Conference on Signal Processing (ICSP), Nov. 6th–10th, Chengdu, China, 2016, https://doi.org/10.1109/icsp.2016.7877901.
[83] P. Swuste, C. van Gulijk, W. Zwaard, Safety metaphors and theories, a review of the occupational safety literature of the US, UK and The Netherlands, till the first part of the 20th century, Saf. Sci. 48 (8) (2010) 1000–1018, https://doi.org/10.1016/j.ssci.2010.01.020.
[84] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1–9, , https://doi.org/10.1109/cvpr.2015.7298594.
[85] C. Tin, F. Sun, T. Kong, W. Zhang, C. Yang, C. Liu, A survey on deep transfer learning, Lecture Notes in Computer Science, 2018. https://arxiv.org/pdf/1808.01974.pdf.
[86] J. Tompson, R. Goroshin, A. Jain, Y. Lecun, C. Bregler, Efficient object localization using convolutional networks, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 648–656, , https://doi.org/10.1109/cvpr.2015.7298664.
[87] L.R. Uijlings, K.E.A. van de Sande, T. Gevers, A.W.M. Smeulders, Selective search for object recognition, Int. J. Comput. Vis. 104 (2013) 154–171, https://doi.org/10.1007/s11263-013-0620-5.
[88] H. Wang, C. Schmid, Action recognition with improved trajectories, IEEE International Conference on Computer Vision, 2014, pp. 3551–3558, , https://doi.org/10.1109/iccv.2013.441.
[89] S.-E. Wei, V. Ramakrishna, T. Kanade, Y. Sheikh, Convolutional pose machines, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4724–4732.
[90] O.S. Williams, R.A. Hamid, M.S. Misnan, Causes of building construction related accident in the south-western states of Nigeria, Int. J. Build. Environ. Sustain. 6 (1) (2019) 14–22, https://doi.org/10.11113/ijbes.v6.n1.313.
[91] w. Wu, H. Yang, Q. Li, D. Chew, An integrated information management model for proactive prevention of struck-by-falling-object accidents on construction sites, Autom. Constr. 34 (2013) 67–74, https://doi.org/10.1016/j.autcon.2012.10.010.
[92] B. Xiao, H. Wu, Y. Wei, Simple Baselines for Human Pose Estimation and Tracking, arXiv preprint arXiv:1804.06208, 2018.
[93] X. Yan, Heng Li, C. Wang, J. Seo, H. Zhang, H. Wang, Development of ergonomic posture recognition technique based on 2D ordinary camera for construction hazard prevention through view-invariant features in 2D skeleton motion, Adv. Eng. Inf. 34 (2017) 152–163, https://doi.org/10.1016/j.aei.2017.11.001.
[94] W. Yang, S. Li, W. Ouyang, H. Li, X. Wang, Learning feature pyramids for human pose estimation, The IEEE International Conference on Computer Vision (ICCV) vol. 2, (2017).
[95] Y. Yu, H. Guo, Q. Ding, H. Li, M. Skitmore, An experimental study of real-time identification of construction workers' unsafe behaviours, Autom. Constr. 82 (2017) 193–206, https://doi.org/10.1016/j.autcon.2017.05.002.
[96] C.L. Zitnick, P. Dollár, Edge boxes: locating object proposals from edges, Lect. Notes Comput. Sci. (2014) 391–405, https://doi.org/10.1007/978-3-319-10602-1_26.
[97] H.W. Heinrich, Industrial Accident Prevention. A Scientific Approach, McGraw-Hill Book Company, Inc., New York & London, 1980.