Advanced Engineering Informatics 42 (2019) 100981

Contents lists available at ScienceDirect

Advanced Engineering Informatics

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

Full length article

Recognizing people's identity in construction sites with computer vision: A spatial and temporal attention pooling network

Ran Wei[a,b], Peter E.D. Love[c], Weili Fang[a,b,c,*], Hanbin Luo[a,b], Shuangjie Xu[d]

[a] Dept. of Construction Management, School of Civil Engineering and Mechanics, Huazhong University of Science and Technology, Wuhan, Hubei, China
[b] Hubei Engineering Research Center for Virtual, Safe and Automated, Wuhan, Hubei, China
[c] Dept. of Civil Engineering, Curtin University, Perth, Western Australia 6023, Australia
[d] School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, Hubei, China

ARTICLE INFO

Keywords:
Recognition
Convolutional neural network
Recurrent neural network
Videos
Computer vision

ABSTRACT

Several prototype vision-based approaches have been developed to capture and recognize unsafe behavior in construction automatically. Vision-based approaches have been difficult to use due to their inability to identify individuals who commit unsafe acts when captured using digital images/video. To address this problem, we applied a novel deep learning approach that utilizes a Spatial and Temporal Attention Pooling Network to remove redundant information contained in a video to enable a person's identity to be automatically determined. The deep learning approach we have adopted focuses on: (1) extracting spatial feature maps using the spatial attention network; (2) extracting temporal information using the temporal attention networks; and (3) recognizing a person's identity by computing the distance between features. To validate the feasibility and effectiveness of the adopted deep learning approach, we created a database of videos that contained people performing their work on construction sites, conducted an experiment, and then performed *k*-fold cross-validation. The results demonstrated that the approach could accurately identify a person's identity from videos captured from construction sites. We suggest that our computer-vision approach can potentially be used by site managers to automatically recognize those individuals that engage in unsafe behavior and therefore be used to provide instantaneous feedback about their actions and possible consequences.

1. Introduction

Construction has a poor safety record when compared to other industrial sectors worldwide; workplace accidents are a common occurrence [15,12,48,50,28]. Previous research has revealed that approximately 88% of accidents in construction are related to people's unsafe behavior [38]. With this in mind, a considerable amount of attention has focused on mitigating unsafe behaviors in construction using a wide variety of institutional, regulatory, and legislator mechanisms. With the presence of cost and time pressures and the regular need to perform unplanned work (e.g., rework) people are prone to taking risks to make their work more efficient [47,49,50]. The upshot in this instance is that there is a proclivity for people to commit unsafe acts, especially when they know they are not being supervised.

Behavior-based safety (BBS) is an effective approach that can be used to observe and identify people's unsafe actions [73]. The direct feedback provided to people who have committed an unsafe act is used to modify their future behavior [9,4,51,31]. The process of observation is labor-intensive and time-consuming, and as a result, events may be inaccurately recorded and prone to selective bias. Developments in technology, aided by computer vision have been identified as an effective approach to automatically recognize people's unsafe behavior [34,37,58,11,16,18,77,30]. For example, Fang et al. [18] developed a computer vision approach that integrated a Mask R-Convolution Neural Network (CNN) to identify individuals who traversed structural supports. Despite being able to recognize when an individual is committing an unsafe act from videos/images, we have not been able to determine their identity. Once we can identify the person's identity, the site managers can provide specific feedback about their unsafe behavior. Computer vision, therefore, can play a pivotal role in implementing an effective BBS program.

Several non-visual sensor techniques, for example, have been developed to recognize a person's identity [67,42]. Non-visual sensor-based approaches require an individual to wear sensors, but these can

* Corresponding author at: Dept. of Construction Management, School of Civil Engineering and Mechanics, Huazhong University of Science and Technology, Wuhan, Hubei, China.
E-mail address: weili_f@hust.edu.cn (W. Fang).

https://doi.org/10.1016/j.aei.2019.100981
Received 21 February 2019; Received in revised form 19 August 2019; Accepted 26 August 2019
Available online 11 September 2019
1474-0346/

be expensive to install and maintain and intrusive [54]. We, therefore, develop a computer-vision approach to identify the person who performed an unsafe action automatically. Our approach provides a cost-effective solution that can be used to enable real-time safety management during construction. The contributions of our research are two-fold: (1) being able to identify the individuals who commit unsafe acts with computer vision; and (2) integrating a temporal attention mechanism to remove redundant video information to obtain higher levels of accuracy and therefore recognize a person's identity.

2. Literature review

2.1. Vision-based unsafe behavior management

Computer vision has been identified as an effective and robust approach to capture and understand information from images or videos in construction [29,72,60,5,17]. Areas where computer vision has been applied, include:

- progress monitoring [53,22,23,24,33],
- productivity measurement [25,26,75,6],
- proximity analysis and location tracking for safety monitoring [8,14,19,40,65]; and
- occupational health assessment ([37,69]).

Previous computer vision-based studies that have been undertaken to recognize unsafe behavior have tended to be based on: (1) three-dimensional (3D) skeletons; (2) features; and (3) deep learning. Skeleton-based approaches have focused on collecting motion data from depth sensors to reconstruct a 3D model of a person. Then, a pre-defined unsafe behaviour template is used to model and compare the as-built 3D skeleton model with established unsafe behaviours [63,34,35,37,36,64,46]. For example, Han and Lee [34] utilized a depth camera to collect motion data and developed a pre-defined template to construct a 3D skeleton model. Han and Lee [34] then compared the pre-defined template with motion data to determine unsafe actions. However, these approaches have been demonstrated in an indoor environment under experimental conditions (i.e., limited ranges, and sensitive to lighting) and therefore have not been able to accommodate the nuances of a construction site. Moreover, skeleton-based approaches are prone to suffering from the 'noisy' skeleton problem [59] when dealing with occlusions (e.g., self-occlusion or occluded by plant, equipment, and materials).

Feature-based unsafe behavior recognition approaches generally contain: (1) extraction and representation; and (2) classification. For example, Park et al. [61] applied the Histogram of Oriented Gradient (HOG) to extract features for a hardhat and people, and then used a support vector machine (SVM) classifier to detect their presence. The individual not wearing their hardhat is identified by matching the geometric and spatial relationship between them. While the approaches above have been able to detect unsafe acts accurately, they are prone to overfitting, which therefore weakens their ability to derive generalizations from a small dataset.

With developments in deep learning and the incorporation of CNN, there now exists an ability to automatically extract and learn features end-to-end by using stack multiple convolutional and pooling layers [43,80]. The ability to extract and learn features has resulted in 'CNN's being used in construction to recognize unsafe behavior on-site [11,16,18]. For example, Fang et al. [16] pioneered a hybrid learning approach that integrated a Faster R-CNN with a deep CNN to detect the use of harnesses while working at heights.

As a result of the numerous computer vision-based approaches that have been developed, we can accurately detect different types of unsafe behavior. We are, however, been unable to identify the culprit of such behavior. While face recognition techniques have been able to identify a person accurately, their use on construction sites can be thwarted as: (1) people are required to look into a camera to be identified directly. However, having to look directly into it does not resonate with the way work is being performed; and (2) they are sensitive to lighting and different scales, which can affect their accuracy [27]. In light of these limitations, there is a need to develop a practical approach to automatically identify a person from videos on a construction site when they have performed an unsafe act.

2.2. Vision-based HIR in computer science

Human identity recognition (HIR) has received considerable attention within the field of computer science, particularly in the areas of surveillance and human-computer interaction [3,74]. Two HIR approaches that dominate the literature are the image and video-based methods [55,76,78,57,70].

Image-based HIR methods focus on extracting reliable feature representations and learn to determine a distance metric. Most existing image-based identification methods focus on either addressing visual variances of pose and viewpoint [44], learn relative distances of triplet training samples [13], or learn similarity metrics of any pairs [1]. Likewise, a CNN with a triplet loss-based approach can acquire a higher level of accuracy when identifying an individual due to its end-to-end learning ability between its input image and desired embedding space. The use of a triplet loss-based approach may have a relatively sizeable intra-class variation and result in an inability to recognize a person's identity due to its weak generalization [7,45]. In comparison to video-based identification, however, datasets in image-based approaches are generally too small for the benefits of a deep model to be realized.

With increases in the size of deep learning databases, video-based person re-identification can occur [72,56,45]. Several video-based person re-identification approaches have been developed. For example, Wang et al. [72] developed a Discriminative Video Ranking (DVR) model for person re-identification by using discriminative video fragments to capture space-time feature information. Likewise, Mclaughlin et al. [56] developed a temporal deep neural network architecture that combined optical flow, recurrent layers, and mean-pooling achieving excellent performance on person re-identification. While these approaches have demonstrated that a person can be accurately re-identified the issue of mutual influence between frames has been overlooked. In this instance, the video can contain redundant information between frames as results of variations in a person movement between two frames.

To address this problem, we draw on the work of Xu et al. [74] who proposed a recurrent-convolutional network in conjunction with an Attentive Spatial-Temporal Pooling (ASTPN) to create a video-based approach to determine a person's identity on a construction site. With a well-designed temporal attention mechanism, the ASTPN model can select useful information which is valid for the final representation [74].

3. Research approach

In accordance with previous studies that have utilized computer vision and deep learning in construction, we have adopted a design science research approach to design and develop a CNN that can automatically determine a person's identity using videos (e.g., [11,16,18]). Design science focuses on describing, explaining and predicting the current natural or social world, and can be used to understand a problem, design solutions and improve human performance [2,71]. Design science has been widely used in many fields, such as engineering, computer science, or business [68,21]. The research process used to develop the ASTPN model to automatically recognize a person's identity is presented in Fig. 1.

Fig. 2 presents the workflow of the applied ASTPN approach. Our applied computer vision approach consists of: (1) Spatial Attention Networks, which is used to extract spatial features; (2) Temporal

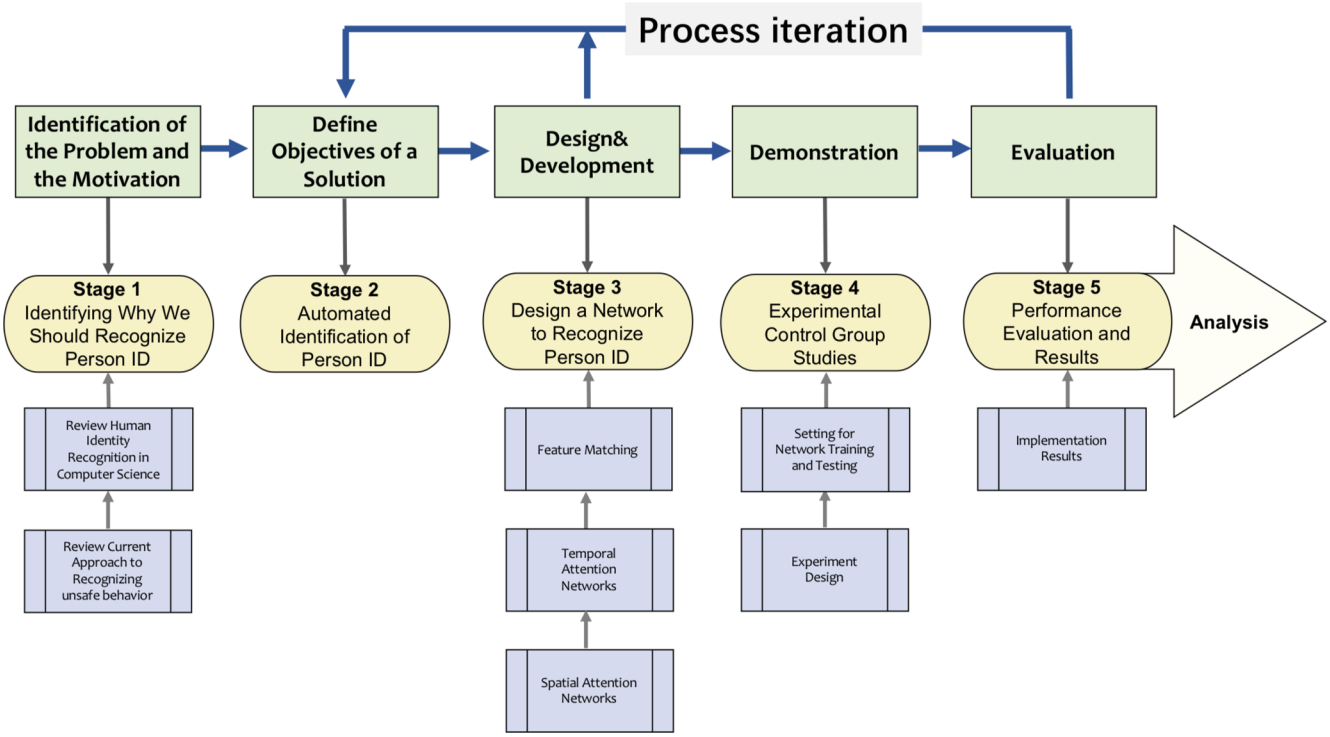

Fig. 1. Design science approach: Research process (Adapted from [10]).

Attention Networks, which is used to extract temporal features; and (3) Feature matching approach, which is used to identify a person by computing Euclidean distance between features.

A video consists of multiple frames, which were chosen randomly at each epoch during the training process. Each frame is an RGB image, which was converted to the YUV color space. Horizontal and vertical optical flow channels were calculated between each pair of frames using the Lucas-Kanade algorithm [52]. Each step takes one frame as input and consists of three color (YUV) and two optic flow channels (horizontal and vertical optical flow). Here, the color channels encode the details of a person's spatial information, for example, their clothing and appearance. The optical flow channels were used to encode a person's temporal information, such as their gait.

A Siamese network is introduced to map the gallery and probe videos containing people into two respective feature vectors [32]. Primarily, Siamese networks consist of two identical neural networks, where each takes one of the two input images. The last layers of the two networks are then fed to a contrastive loss function, which calculates the similarity between the two images and is expressed as:

$$L(W, Y, \vec{X_1}, \vec{X_2}) = (1 - Y)\frac{1}{2}(D_w)^2 + (Y)\frac{1}{2}\{\max(0, m - D_w)\}^2 \quad (1)$$

where D_w is defined as the Euclidean distance between the outputs of the sister Siamese networks. m denotes the margin to separate features of different persons Y is a prediction of the network. The procedure we adopted to implement our approach is comprised of the following three steps:

(1) *Spatial feature extraction*: Frames were inputted into the Siamese CNN to extract spatial features maps from the last convolutional layer. Then, these features were fed into the Spatial Pyramid Pooling (SPP) to obtain an image-level representation for one frame.
(2) *Temporal feature extraction*: The image-level representation for one frame obtained from step 2 were fed into the temporal attention network. A Recurrent Neural Network (RNN) model was used to generate the feature set of a video sequence by taking temporal information into consideration.
(3) *Feature matching:* All time steps obtained from Step 3 were combined by attentive temporal pooling to form the sequence-level representation. Then, a person's identity was determined by computing the Euclidean distance between the identity feature with previously saved features of other peoples' identities.

Our applied approach is further explained in greater detail below.

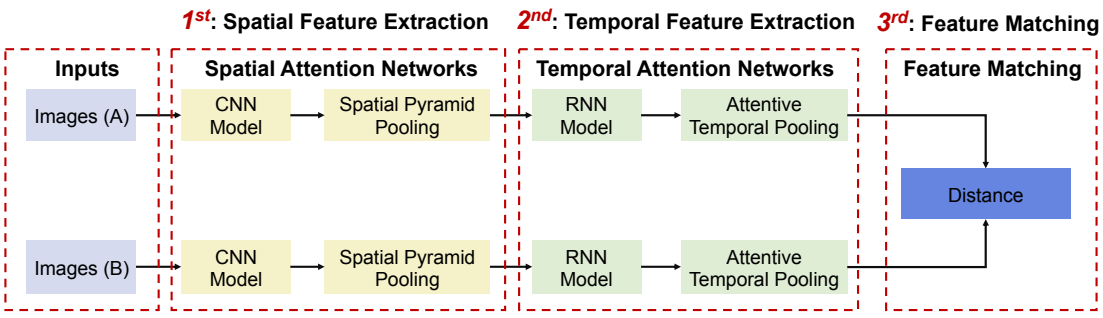

Fig. 2. The workflow of our proposed ASTPN approach.

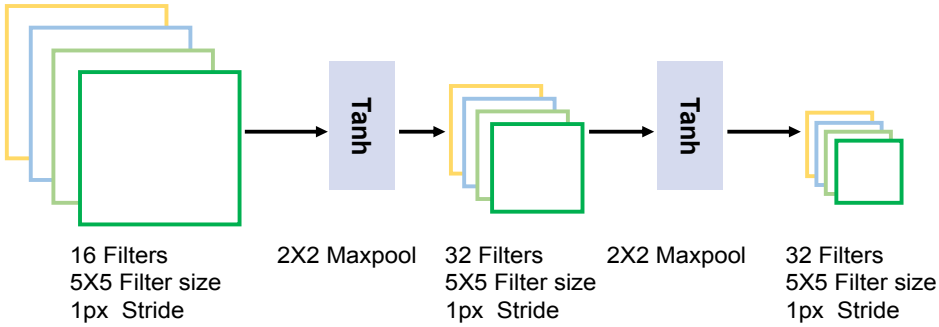

Fig. 3. Structure of CNN models.

3.1. Spatial attention networks

3.1.1. Convolution neural network

The structure of CNN models is adopted to automatically extract feature from each time-step image (Fig. 3). Firstly, each image is passed through the CNN models to generate a vector as an output. Then the output is considered as an input for the deeper layer. Finally, the output of the final layer $f^{(t)}$ is fed into the recurrent layer. In the CNN architecture, a dropout is used to reduce over-fitting; this is located between the CNN and RNN models. The details of the CNN architecture can be seen in Fig. 3.

3.1.2. Spatial pooling layer

When video surveillance is used during the construction process, a person's scale and pose can change. However, a person's size forms only a small proportion of an image. To enable deep learning models to learn sequence information from videos surveillance, local spatial attention is proposed. SPP layers are introduced to assist the deep learning models to focus their attention on the region of interest (RoI) in a spatial dimension. In the case of a person's pose and scale varying, the SPP layer has multi-level spatial bins to extract their spatial representations to enable the model to be insensitive to these variants. Next, these extracted multi-scale representations are combined into a fixed-length image-level representation. With the attentive spatial pooling mechanism mentioned above, the employed model is able to focus on regions with effective information in their spatial dimension.

Let $V = \{v^1, v^2, \cdots v^T\}$ be the inputs of the CNN models. Then the feature map set $C = \{C^1, C^2, \cdots C^T\}$ can be obtained from the CNN models (Fig. 3). Each $C^i \in R^{c\times w\times h}$ is then fed into a spatial pooling layer to obtain image-level representation r^i. Assuming that the size set of spatial bins is $\{(m_\omega^j, m_h^j)|j = 1, \cdots n\}$, and window set $\omega in^j = (\lfloor \frac{\omega}{m_\omega^j} \rfloor, \lceil \frac{h}{m_h^j} \rceil)$ and pooling stride $str^j = (\lfloor \frac{\omega}{m_\omega^j} \rfloor, \lfloor \frac{h}{m_\omega^j} \rfloor)$ for the j^{th} spatial bin are determined [74]. The r^i is formalized as:

$$\begin{cases} b^{ij} = f_R\{f_p(C^i; win^j, str^j)\} \\ r^i = b^{i,1} \oplus b^{i,2} \oplus \cdots b^{i,n} \end{cases} \tag{2}$$

where, f_R presents the reshape operation, which reshapes a matrix to a vector. f_p stands for the max poling function. $\lceil\cdot\rceil$and $\lfloor\cdot\rfloor$represent ceiling and floor operations, respectively. $\oplus$ denotes vector connection.

Let $r = \{r^i \in R^L | i = 1, \cdots T\}$ be the representation sequence, which is forwarded into the RNN models for extracting information. Where L is$L = \sum_j m_\omega^j m_h^j$.

3.2. Temporal attention networks

3.2.1. Recurrent neural network

To improve the ability to capture temporal information from videos, recurrent connections can be introduced to incorporate the CNN models and temporal pooling layers. When the output of the final layer $f^{(t)}$ is generated, it is fed into the RNN models to extract temporal information and is formalized as:

$$\begin{cases} o(t) = Uf^{(t)} + Ws^{t-1} \\ s^t = \tanh(o^t) \end{cases} \tag{3}$$

where $s^{t-1} \in R^N$ is the hidden state containing information from the previous time-step, and o^t is the output. U projects the recurrent layer input and W the hidden state.

3.2.2. Attentive temporal pooling layer

An RNN can successfully capture the temporal information with a hidden state, even though it may contain considerable redundancy [74]. For instance, a series of continuous frames within a surveillance video may be subjected to a minor temporal change. That is, while the background may not vary, the motion of a person may change between serial frames as cameras are positioned in a fixed location. To avoid the network's insensitivity to provide an effective representation between frames due to high levels of information redundancy, an attentive temporal pooling network is proposed to ensure the deep learning models focus its attention on attracting effective information. The use of attentive temporal pooling enables layers to perceive the probe and gallery in a time dimension. In this structure, the attentive temporal pooling layers are located between the RNN models and distance computation layers (Fig. 4).

Matrices $P \in R^{T\times N}$ and $G \in R^{T\times N}$ present the output of the recurrent layer in the *i-th* time step with probe and gallery data respectively, and then the attention matrix $A \in R^{T\times T}$ is formalized as:

$$A = \tanh(PUG^T) \tag{4}$$

where $U \in R^{N\times N}$ represents the information sharing matrix, which is updated using backpropagation. In addition, it enables the deep model to focus its attention on providing an effective representation by influencing the convolution parameters and in the hidden state.

Then, the temporal weight vectors $t_p = R^T$ and $t_g = R^T$ can be obtained by applying column-wise and row-wise max pooling on A, respectively. Meanwhile, the Softmax function is applied to the temporal weight vectors t_p and t_g to generate attention vectors $a_p = R^T$and $a_g = R^T$, correspondingly.

Finally, the sequence-level representations $v_p = R^N$ and $v_g = R^N$ are formalized as:

$$\begin{cases} v_p = P^T a_p \\ v_g = G^T a_g \end{cases} \tag{5}$$

3.3. Feature matching for recognizing a Person's identity

As part of the training process, we initiated match features to recognize a person's identity, a pair of sequences (I_P, I_g) for a person p and g is given, and the sequence-level representation (v_p, v_g), which is acquired from the Siamese network. The Euclidean distance Hinge loss is used to train the deep model, which is formalized as:

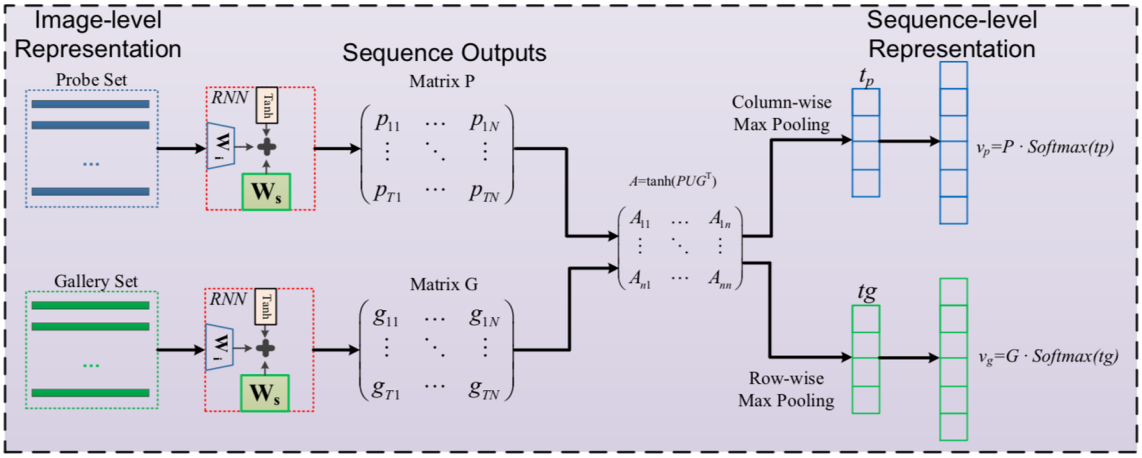

Fig. 4. Structure of attentive temporal pooling (Adapted from [74]).

$$E(v_p, v_g) = \begin{cases} \|v_p - v_g\| & p = g \\ \max(0, m - \|v_p - v_g\|^2) & p \neq g \end{cases} \quad (6)$$

where m stands for the margin to separate features of different people in Hinge loss.

In the training phase, positive and negative input pairs are alternatively fed into our model. While, in the testing phase, the input of a new sequence is copied to form a new pair, and it is passed through the Siamese network to enable an identity feature to be generated. By computing the distance between features that are extracted from two different sequences in an input pair, their identities are able to be re-recognized. In line with Mclaughlin (2015), an identity classification loss is introduced to recognize people within its architecture [79]. Then, the identity of a person in the sequence can be predicted using the softmax function, which is formalized as:

$$\begin{cases} I(v_p) = P(q = c|v_p) = \frac{\exp(W_c v_p)}{\sum_k \exp(W_k v_p)} \\ I(v_g) = P(q = c|v_g) = \frac{\exp(W_c v_g)}{\sum_k \exp(W_k v_g)} \end{cases} \quad (7)$$

where q is the identity of the person, W_c, and W_k represent the $c^{th}$and k^{th} column of W, the softmax weight matrix, respectively.

The ability to accurately identify people can significantly improve due to the joint learning that takes place between the Siamese and Hinge loss functions we have introduced. Therefore, the final training objective combines the Siamese and the identity loss is expressed as:

$$L(v_p, v_g) = E(v_p, v_g) + I(v_p) + I(v_g) \quad (8)$$

4. Experiment control group studies

4.1. Experimental design

To validate the effectiveness and feasibility of our developed computer vision approach, there is a need to create a video database of people for training and testing the applied ASTPN model. The video database was created from a construction project where video surveillance had been implemented to monitor works in Wuhan, China. The dataset was created using eight security Hikvision cameras that were installed by a contractor to monitor people's daily activities in real-time.

To create a database where a person's identification can be re-cognized we use two cameras as denoted in Fig. 5. In this case, if a person exists in camera A, they can also appear in camera B; the two videos are considered as one. For the purpose of this research, a total of 12 pairs videos were collected from the selected construction project.

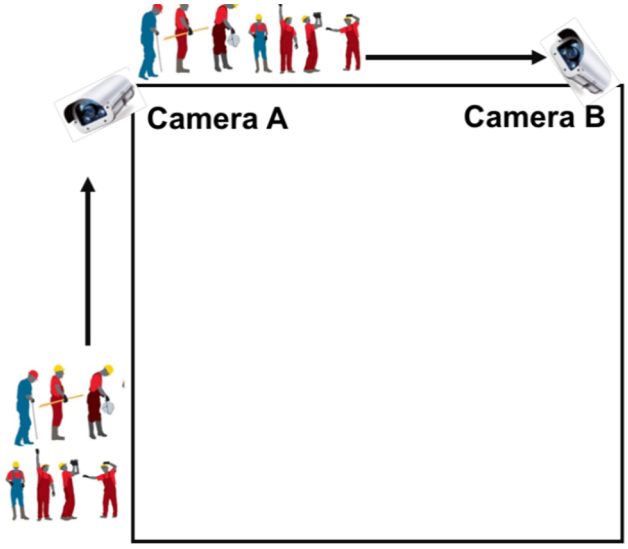

Fig. 5. Model of collecting data from video surveillance.

Each video pair was obtained from the same camera pair.

Fig. 6 presents an example of our created video database. The database was randomly divided into training and testing data according to a 1:2 ratio. Here, the training database is used to fine tune our ASTPN model, which we pre-trained using the iLIDS Video Re-Identification (iLIDS-VID) database that is comprised 300 different people that have been observed across two disjoint camera views in a public open space.

4.2. Settings for network training and testing

We developed our applied ASTPN approach using Python and used a computer equipped with a 2.5 GHz Intel® Xeon® E5-2680 CPU, an NVIDIA^R Tesla™ Titan XP and 64 RAM. We used PyTorch-0.4.0 [62] to build our network, Anaconda to manage our python environment and OpenCV to implement the Lucas-Kanade algorithm to compute the optic flow [66]. The following steps were followed:

(1) The ASTPN model was first pre-trained using an iLIDS-VID database that comprises 600 image sequences of 300 distinct individuals, with one pair of image sequences from two camera views for each person.
(2) Each video within the training database was clipped into 24 consecutive frames. Then an augmentation technique to randomly crop frames were used to ensure the model could randomly detect a

Fig. 6. Examples of created database for training and testing.

person's identity. Data augmentation technique was 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 was to avoid overfitting, which can improve a model's ability to generalize [41,39].

(3) The training database was used to finetune the pre-trained model to ensure its parameters could fit with the observations that are identified. Here, the key parameters of our model were set as follows: the stochastic gradient descent (SGD) with momentum was used as an optimizer with a learning rate of 0.001; the momentum was 0.9; and the embedding size of the RNN was set to 128.

(4) The tested consecutive frames were inputted into the ASTPN model to obtain our results.

5. Performance evaluation and results

Due to the limited size of the training sets, *k*-fold cross-validation was used to evaluate our approach. To reduce the computation (i.e., model training) period, this research assumed that *k* was set to 3. The database was randomly divided into three parts and labeled as: (1) Part 1(person 9 - person 12) for fine tuning pre-trained ASTPN model, and Part 2 (Person 1-Person 4) and Part 3 (Person 5-Person 8) are used for testing (Table 1); (2) Part 2 for fine tuning, and Part 1 and Part 3 are used for testing (Table 2); and (3) part 3 for fine tuning, and Part 1 and Part 2 are used for testing (Table 3).

To evaluate if our created training database has improved the ability to accurately identify a person's identification, we also make a comparison with our pre-trained ASTPN and non-fine-tuned models. The results are presented in Table 4. We, therefore, can conclude that our model's accuracy to recognize a person's identity was 79.2% (i.e., the average of 75%, 75%, and 87.5% of the three tests). In addition, Fig. 7 presents examples of our detection results for the pre-trained ASTPN fine-tune model, which has a 79.2% level of accuracy to recognize a person's identity and therefore outperform the model that has not been fine-tuned.

From Tables 1 and 2, we can conclude that person 5 is mis-detected as person 7. Here, we present an example of the incorrect detection of a person's identity (person 5 and person 7) in Fig. 8. In Figs. 7, 8 frames of each person 5 and 7 are selected and presented. We suggest that the main reasons for the error detection are twofold: (1) the features (i.e., color, background) of person 5 are similar to person 7; and (2) the low-resolution ratio of images.

Table 1
The Euclidean distance for the identification results for testing the database (average accuracy: 75%).

	1	2	3	4	5	6	7	8
1	0.583	1.294	0.762	1.368	1.145	1.637	1.284	1.512
2	1.294	0.627	1.628	1.371	1.537	0.813	1.688	1.597
3	0.762	1.628	0.493	1.647	1.472	1.729	1.764	1.697
4	1.368	1.371	1.647	0.356	1.714	1.389	1.522	1.683
5	1.145	1.537	1.472	1.714	0.984	1.728	0.936	1.263
6	1.637	0.813	1.729	1.389	1.728	0.524	1.371	1.416
7	1.284	1.688	1.764	1.522	0.936	1.371	0.634	0.937
8	1.512	1.597	1.697	1.683	1.263	1.416	0.937	1.014

Table 2
The Euclidean distance for the identification results for testing the database (average accuracy: 75%).

	5	6	7	8	9	10	11	12
5	1.109	1.202	0.934	1.409	1.234	1.445	1.320	1.673
6	1.202	0.589	1.476	1.350	1.537	1.165	1.371	1.832
7	0.934	1.476	0.561	0.944	1.318	1.366	1.303	1.508
8	1.409	1.350	0.944	1.644	1.366	1.318	1.428	1.353
9	1.234	1.537	1.318	1.000	0.584	1.384	1.428	1.353
10	1.445	1.165	1.366	1.318	1.384	0.419	1.583	1.743
11	1.320	1.371	1.303	1.428	1.428	1.583	0.905	1.118
12	1.673	1.832	1.508	1.353	1.353	1.742	1.118	0.596

Table 3
The Euclidean distance for the identification results for testing the database (average accuracy: 87.5%).

	1	2	3	4	9	10	11	12
1	0.391	1.229	0.713	1.592	1.249	1.845	1.722	1.782
2	1.229	0.745	1.829	1.445	1.586	1.635	1.362	1.556
3	0.713	1.829	1.064	1.874	1.743	1.460	1.435	1.969
4	1.592	1.445	1.874	0.372	1.380	1.598	1.714	1.648
9	1.249	1.586	1.743	1.00	0.356	1.709	1.614	1.231
10	1.845	1.635	1.460	1.598	1.709	0.712	1.404	1.610
11	1.722	1362	1.435	1.614	1.614	1.404	0.322	1.736
12	1.782	1.556	1.969	1.648	1.231	1.610	1.736	0.693

Table 4
A comparison of using our data fine-tuned and non-fine-tuned ASTPN models.

Type of testing database	Accuracy (Fine-tuned Model)	Accuracy (Non-fine-tuned Model)
Part 1 and part 2	87.5%	62.5%
Part 2 and part 3	75%	37.5%
Part 1 and part 3	75%	50%
Average Accuracy	79.2%	50%

Fig. 7. Examples of the accurate detection of a person's identity from videos.

6. Discussion

A pervasive challenge for site management is to ensure that people perform their work safely, but also to the desired quality, within a specified time and to budget. However, when confronted with time constraints and having to undertake unplanned work, then the likelihood for "violations (i.e. a conscious intention to break a rule or to be non-conforming to a standard" [20] occurring increases. Once a violation arises such as such not wearing hard-hat or harness it is important to detect and report the unsafe act as quickly as possible and then identify the person committing this action to enable an intervention to take place. However, being able to identify a person performing an unsafe act on construction sites has not been addressed in the normative construction and engineering literature.

The major contributions of this paper are twofold. First, to fulfill the prevailing research gap, we have applied a temporal attention mechanism to recognize peoples identify by removing the redundant information that resides within a video. The experimental results demonstrate that by combining an attention spatial network and attention temporal model we are able to focus on the spatiotemporal information and therefore avoid the redundant material contained within the video;

Second, in comparison with publicly available databases such as the

Fig. 8. Examples of error detection of a person's identity (0 frames to 180 frame).

iLIDS-VID, the database we have developed and used in this research has its own characteristic with lighting, occlusion, varying poses, and scales. We used the iLIDS-VID database to pre-train our ASTPN model and then used our database to fine tune the pre-trained ASTPN model. A comparison of the results that we obtained is presented in Table 4.

7. Limitations

Several limitations need to be acknowledged and addressed in the future if computer vision is to be effectively applied to identify individuals performing unsafe acts on-site. To improve the process of recognizing a person's identity we need to create a larger video dataset for training and testing. Our research focused on a limited number of activities (e.g., people walking). Therefore, the scope of our research needs to be extended by examining a wider range of activities that are typically performed on-site (e.g., bending, working at height).

A delay in recognizing a person's identity in real-time may occur due to the computation requirements that are placed on the attention network to extract representations from videos. The representation from the testing database also needs to be computed, which may reduce its speed to identify a person's identity. Thus, to enable our model to be able to determine a person's identity in real time we need to develop an optimal method to obtain an attentive representation without the repetitive computation of videos in the testing database.

8. Conclusions

In this paper, we have applied computer vision-based approach that utilizes the deep learning features of attention spatial and pooling networks to recognize the identity of a person on a construction site. Using a three-fold cross-validation process we demonstrated that our proposed approach can achieve a 79.2% level of accuracy in recognizing a person's identity. Considering the accuracy of our approach, we believe that it can be combined with existing computer vision-based approaches that have been developed to automatically detect individuals performing unsafe actions and instantly recognize them. Being able to identify who is performing an unsafe behavior during construction can lead to perfunctory interventions by management, which can result in immediate behavior modification to improve safety performance in construction sites. In addition, the obtained detection results can be used to provide direct visual feedback to the person who has performed the unsafe behavior.

Acknowledgments

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, No. 71821001), the China Scholarship Council.

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.

References

[1] E. Ahmed, M. Jones, T.K. Marks, An improved deep learning architecture for person re-identification, Proc., IEEE Comput. Vis. Pattern Recogn. (2015) 3908–3916.
[2] J.E.V. Aken, Management research based on the paradigm of the design sciences: The quest for field-tested and grounded technological rules, J. Manage. Stud. 41 (2) (2004) 219–246.
[3] A. Bedagkar-Gala, S.K. Shah, A survey of approaches and trends in person re-identification, Image Vision Comput. 32 (4) (2014) 270–286.
[4] L. Ding, S. Guo, H. Luo, X. Jiang, A Big-Data-based platform of workers' behavior: Observations from the field, Accident Anal. Prev. 93 (2016) 299–309.
[5] I. Briakis, H. Fathi, A. Rashidi, Progressive 3D reconstruction of infrastructure with videogrammetry, Autom. Constr. 20 (7) (2011) 884–895.
[6] M. Bügler, G. Ogunmakin, J. Teizer, P.A. Vela, A. Borrmann, A comprehensive methodology for vision-based progress and activity estimation of excavation processes for productivity assessment, in: Proc., Proceedings of the 21st International Workshop: Intelligent Computing in Engineering (EG-ICE), Cardiff, Wales.
[7] W. Chen, X. Chen, J. Zhang, K. Huang, Beyond triplet loss: A deep quadruplet network for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 1320–1329.
[8] S. Chi, C. Caldas, Image-based safety assessment: Automated spatial safety risk identification of earthmoving and surface mining activities, J. Const. Eng. Manage. 138 (3) (2011) 341–351, https://doi.org/10.1061/(ASCE)CO.1943-7862.0000438.
[9] R.M. Choudhry, Behavior-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] M. Chu, J. Matthews, P.E.D. Love, Integrating mobile building information modeling and augmented reality systems: An experimental study, Autom. Const. 85 (2018) 305–316.
[11] L. Ding, W. Fang, H. Luo, P.E.D. Love, B. Zhong, X. Ouyang, A deep hybrid learning model to detect unsafe behavior: Integrating Convolution Neural Networks and Long Short-Term Memory, Autom. Constr. 86 (118) (2018) 118–124.
[12] L. Ding, X. Jie, A review of metro construction in China: Organization, market, cost, safety, and schedule, Front. Eng. Manage. 4 (1) (2017) 4–19.
[13] S. Ding, L. Lin, G. Wang, H. Chao, Deep feature learning with relative distance comparison for person re-identification, Pattern Recogn. 48 (10) (2015) 2993–3003.
[14] S. Du, M. Shehata, W. Badawy, Hard hat detection in video sequences based on facial features, motion and color information, in: Proc., ICCRD2011-2011 3rd International Conference on Computer Research and Development, pp. 25-29, 10.1109/ICCRD.2011.5763846.
[15] D. Fang, H. Wu, Development of a safety culture interaction (SCI) model for construction projects, Saf. Sci. 57 (8) (2013) 138–149.
[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.
[17] 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. Inform. 37 (2018) 139–149.
[18] 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. 97 (2019) 170–177.
[19] B. Ferrer, J.C. Pomares, R. Irles, J. Espinosa, D. Mas, Image processing for safety assessment in civil engineering, Appl. Opt. 52 (18) (2013) 4385–4390.
[20] M. Frese, N. Keith, Action errors, error, and learning in organizations, Ann. Rev. Psych. 66 (1) (2015) 661–687.
[21] G.L. Geerts, A design science research methodology and its application to accounting information systems research, Int. J. Account. Inf. Syst. 12 (2) (2011) 142–151.
[22] M. Golparvar-Fard, F. Peña-Mora, C. Arboleda, S. Lee, Visualization of construction progress monitoring with 4D simulation model overlaid on time-lapsed photographs, J. Comput. Civ. Eng. 23 (6) (2009) 391–404, https://doi.org/10.1061/(ASCE)0887-3801(2009) 23:6(391).
[23] M. Golparvar-Fard, F. Peña-Mora, S. Savarese, D4AR–a 4-dimensional augmented reality model for automating construction progress monitoring data collection, processing, and communication, J. Inf. Tech. Constr. 14 (13) (2009) 129–153.
[24] M. Golparvar-Fard, F. Peña-Mora, S. Savarese, Automated progress monitoring using unordered daily construction photographs and IFC-based building information models, J. Comput. Civ. Eng. (2012), https://doi.org/10.1061/(ASCE)CP.1943-5487.0000205, 04014025.
[25] J. Gong, C.H. Caldas, Computer vision-based video interpretation model for automated productivity analysis of construction operations, J. Comput. Civ. Eng. (2009), https://doi.org/10.1061/(ASCE)CP.1943-5487.0000027.
[26] J. Gong, C.H. Caldas, An intelligent video computing method for automated productivity analysis of cyclic construction operations, in: International Workshop on Computing in Civil Engineering, Austin, Texas, United States, 2009b, pp. 64–73.
[27] S. Gong, M. Cristani, S. Yan, C.C. Loy, Person Re-Identification, Springer Publishing Company, Incorporated, 2014, 1447162951, 9781447162957.
[28] S. Guo, L. Ding, Y. Zhang, M.J. Skibniewski, K. Liang, Hybrid recommendation approach for behavior modification in the Chinese Construction Industry, J. Constr. Manage. 145 (6) (2019), https://doi.org/10.1061/(ASCE)CO.1943-7862.0001665.
[29] H. Luo, C. Xiong, W. Fang, P.E.D. Love, B. Zhang, X. Ouyang, Convolutional neural networks: Computer vision-based workforce activity assessment in construction, Automat. Constr. 94 (2018) 282–289.
[30] H. Guo, Y. Yu, Q. Ding, M. Skitmore, Image-and-skeleton-based parameterized approach to real-time identification of construction workers' unsafe behaviors, J. Constr. Eng. Manage. (2018), https://doi.org/10.1061/(ASCE)CO.1943-7862.0001497.
[31] S. Guo, P. Zhang, L. Ding, Time-statistical laws of workers' unsafe behavior in the construction industry: A case study, Phys. A 515 (2019) 419–429.
[32] R. Hadsell, S. Chopra, Y. Lecun, Dimensionality reduction by learning an invariant mapping, in: Proc., IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1735–1742.
[33] K.K. Han, M. Golparvar-Fard, Appearance-based material classification for monitoring of operation-level construction progress using 4D BIM and site photologs, Autom. Constr. 53 (2015) 44–57.
[34] S. Han, S. Lee, A vision-based motion capture and recognition framework for behavior-based safety management, Autom. Constr. 35 (2013) 131–141.

[35] S. Han, S. Lee, F. Peña-Mora, Vision-based detection of unsafe actions of a construction worker: Case study of ladder climbing, J. Comput. Civ. Eng. (2013), https://doi.org/10.1061/(ASCE)CP.1943-5487.0000279, 635-644.
[36] S. Han, S. Lee, F. Peña-Mora, Comparative study of motion features for similarity-based modeling and classification of unsafe actions in construction, J. Comput. Civ. Eng. (2014), https://doi.org/10.1061/(ASCE)CP.1943-5487.0000339, A4014005.
[37] S.U. Han, M. Achar, S.H. Lee, F. Peña-Mora, Empirical assessment of an RGB-D sensor on motion capture and action recognition for construction worker monitoring, Vis. Eng. 1 (1) (2013) 1–6.
[38] H.W. Heinrich, D. Petersen, N.R. Roos, Industrial Accident Prevention: A Safety Management Approach, McGraw-Hill, New York, 1980, pp. 457–461.
[39] A. Howard, Some improvements on deep convolutional neural network based image classification, 2013. https://arxiv.org/pdf/1312.5402.pdf.
[40] H. Ishimoto, T. Tsubouchi, Stereo vision based worker detection system for excavator, in: International Symposium on Automation and Robotics in Construction and Mining, 2013, pp. 1004–1012.
[41] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Proceedings of the 25th International Conference on Neural Information Processing Systems, 1097-1105. Lake Tahoe, Nevada, December 03-06, 2012. https://doi.org/10.1145/3065386.
[42] J.R. Kwapisz, G.M. Weiss, S.A. Moore, Cell phone-based biometric identification, in: Proc., IEEE International Conference on Biometrics: Theory Applications & Systems, pp. 1–7.
[43] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436, https://doi.org/10.1038/nature14539.
[44] W. Li, R. Zhao, T. Xiao, X. Wang, DeepReID: Deep filter pairing neural network for person re-identification, in: Proc., IEEE Conference on Computer Vision and Pattern Recognition, pp. 152–159.
[45] W. Liao, M.Y. Yang, N. Zhan, B. Rosenhahn, Triplet-based deep similarity learning for person re-Identification, 22-29 October, Venice, Italy, 2017, pp. 385–393.
[46] M. Liu, S. Han, S. Lee, Tracking-based 3D human skeleton extraction from stereo video camera toward an on-site safety and ergonomic analysis, Constr. Innov. 16 (3) (2016) 348–367.
[47] P.E.D. Love, J. Smith, P. Teo, Putting into practice error management Theory: Unlearning and learning to manage action errors in construction, Appl. Ergonom. 69 (2018) 104–114.
[48] P.E.D. Love, P. Teo, Statistical analysis of injury and nonconformance frequencies in construction: negative binomial regression model, J. Constr. Eng. Manage. (2017), https://doi.org/10.1061/(ASCE)CO.1943-7862.000132605017011.
[49] P.E.D. Love, P. Teo, F. Ackermann, J. Smith, J. Alexander, E. Palaneeswaran, J. Morrison, Reduce rework, Improve safety: An empirical inquiry into the precursors to errror in construction, Prod. Plan. Control 29 (5) (2018) 53–67.
[50] 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.
[51] 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, J. Constr. Eng. Manage. 143 (2) (2017), https://doi.org/10.1061/(ASCE)CO.1943-7862.0001224.
[52] B.D. Lucas, T. Kanade, An iterative image registration technique with an application to stereo vision, in: Proc 7th Intl. Joint Conf. on Artificial Intelligence (IJCAI) 1981, August 24-28. Vancouver, British Columbia, 1981, pp. 674–679. https://ri.cmu.edu/pub_files/pub3/lucas_bruce_d_1981_2/lucas_bruce_d_1981_2.pdf.
[53] T.C. Lukins, E. Trucco, Towards an automated visual assessment of progress in Construction Projects, in: Proc., BMVC, pp. 1–10.
[54] X. Luo, H. Li, T. Huang, T. Rose, A field experiment of workers' responses to proximity warnings of static safety hazards on construction sites, Saf. Sci. 84 (2016) 216–224.
[55] X. Ma, X. Zhu, S. Gong, X. Xie, J. Hu, K.M. Lam, Y. Zhong, Person re-identification by unsupervised video matching, Pattern Recogn. 65 (C) (2016) 197–210.
[56] N. Mclaughlin, J.M.D. Rincon, P. Miller, Recurrent convolutional network for video-based person re-identification, in: Proc., Computer Vision and Pattern Recogn., pp. 1325–1334.
[57] N. Mclaughlin, J.M.D. Rincon, P.C. Miller, Person Reidentification Using Deep Convnets With Multitask Learning, IEEE Press, 2017.
[58] B.E. Mneymneh, M. Abbas, H. Khoury, Evaluation of computer vision techniques for automated hardhat detection in indoor construction safety applications, Front. Eng. Manage. 5 (2) (2018) 227–239.
[59] M.M. Najafabadi, F. Villanustre, T.M. Khoshgoftaar, N. Seliya, R. Wald, E. Muharemagic, Deep learning applications and challenges in big data analytics, J. Big Data 2 (1) (2015).
[60] H. Pan, T. Guan, Y. Luo, L. Duan, Y. Tian, L. Yi, Y. Zhao, J. Yu, Dense 3D reconstruction combining depth and RGB information, Neurocomputing 175 (PA) (2016) 644–651.
[61] 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.
[62] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in PyTorch, in: 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA. https://openreview.net/pdf?id=BJJsrmfCZ.
[63] S.J. Ray, J. Teizer, Real-time construction worker posture analysis for ergonomics training, Adv. Eng. Inform. 26 (2) (2012) 439–455.
[64] J. Seo, S. Han, S. Lee, H. Kim, Computer vision techniques for construction safety and health monitoring, Adv. Eng. Inform. 29 (2) (2015) 239–251.
[65] K. Shrestha, P.P. Shrestha, D. Bajracharya, E.A. Yfantis, Hard-hat detection for construction safety visualization, J. Constr. Eng. (2015) 1–8.
[66] D. Stavens, The OpenCV library: computing optical flow, 2007. https://www.damienleroux.com/download/stavens_opencv_optical_flow.pdf.
[67] M. Srivastava, Smart kindergarten: sensor-based wireless networks for smart developmental problem-solving environments, in: Proc., International Conference on Mobile Computing and Networking, pp. 132–138.
[68] J.A. Van Aken, Management research as a design science: Articulating the research products of Mode 2 knowledge production in management, Brit. J. Manage. 16 (1) (2005) 19–36.
[69] D. Wang, F. Dai, X. Ning, Risk assessment of work-related musculoskeletal disorders in construction: State-of-the-art review, J. Constr. Eng. Manage. 141 (6) (2015), https://doi.org/10.1061/(ASCE)CO.1943-7862.0000979, 04015008.
[70] T. Wang, S. Gong, X. Zhu, S. Wang, Person re-identification by video ranking, in: Proc., European Conference on Computer Vision, pp. 688–703.
[71] X. Wang, P.E. Love, M.J. Kim, C.-S. Park, C.-P. Sing, L. Hou, A conceptual framework for integrating building information modeling with augmented reality, Autom. Constr. 34 (2013) 37–44.
[72] Z. Wang, J. Yu, Y. He, T. Guan, Affection arousal based highlight extraction for soccer video, Multimedia. Tools Appl. 73 (1) (2014) 519–546.
[73] O. Wirth, S.O. Sigurdsson, When workplace safety depends on behavior change: Topics for behavioral safety research, J. Safety Res. 39(6) 589–598.
[74] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, P. Zhou, Jointly attentive spatial-temporal pooling networks for video-based person re-identification, in: IEEE International Conference on Computer Vision (ICCV). 22-29 October, Venice, Italy, 2017, pp. 4743–4752. https://doi.org/10.1109/ICCV.2017.507.
[75] J. Yang, P. Vela, J. Teizer, Z. Shi, Vision-based tower crane tracking for understanding construction activity, J. Comput. Civ. Eng. 28 (1) (2012) 103–112, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000242.
[76] J. You, A. Wu, X. Li, W.S. Zheng, Top-push video-based person re-identification, in: Proc., Computer Vision and Pattern Recognition, pp. 1345–1353.
[77] Y. Yu, H. Guo, Q. Ding, H. Li, M. Skitmore, An experimental study of real-time identification of construction workers' unsafe behaviors, Autom. Constr. 82 (2017) 193–206.
[78] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, Q. Tian, MARS: A video benchmark for large-scale person re-Identification, 2016, pp. 868–888.
[79] N. Mclaughlin, J.M. Rincon, P. Miller, Recurrent con-volutional network for video-based person re-identification, IEEE Conference on Computer Vision and Pattern Recognition (2016) 1325–1334.
[80] S. Xu, D. Liu, L. Bao, W. Liu, P. Zhou, MHP-VOS: Multiple hypotheses propagation for video object segmentation, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 314–323.