Automation in Construction 86 (2018) 118–124

Contents lists available at ScienceDirect

Automation in Construction

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

A deep hybrid learning model to detect unsafe behavior: Integrating convolution neural networks and long short-term memory

Lieyun Ding[a,b], Weili Fang[a,b,*], Hanbin Luo[a,b], Peter E.D. Love[c], Botao Zhong[a,b], Xi Ouyang[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 Construction, China
[c] Dept. of Civil Engineering, Curtin University, Perth, Western Australia, Australia
[d] School of Electronic Information and communications, Huazhong University of Science and Technology, Wuhan, Hubei, China

ARTICLE INFO

Keywords:
Deep learning
Convolution neural network
Long short-term memory
Unsafe actions
Safety
Video surveillance

ABSTRACT

Computer vision and pattern recognition approaches have been applied to determine unsafe behaviors on construction sites. Such approaches have been reliant on the computation of artificially complex image features that utilize a cumbersome parameter re-adjustment process. The creation of image features that can recognize unsafe actions, however, poses a significant research challenge on construction sites. This due to the prevailing complexity of spatio-temporal features, lighting, and the array of viewpoints that are required to identify an unsafe action. Considering these challenges, a new hybrid deep learning model that integrates a convolution neural network (CNN) and long short-term memory (LSTM) that automatically recognizes workers' unsafe actions is developed. The proposed hybrid deep learning model is used to: (1) identify unsafe actions; (2) collect motion data and site videos; (3) extract the visual features from videos using a CNN model; and (4) sequence the learning features that are enabled by the use of LSTM models. An experiment is used to test the model's ability to detect unsafe actions. The results reveal that the developed hybrid model (CNN + LSTM) is able to accurately detect safe/unsafe actions conducted by workers on-site. The model's accuracy exceeds the current state-of-the-art descriptor-based methods for detecting points of interest on images.

1. Introduction

Ensuring the safety of people is a pervasive and challenging task in construction due the dynamic and complex work conditions that exist on-site [1–3]. Accidents and fatalities during construction have been and remain a worldwide problem. This is despite regulatory reforms, legislation, and efforts by industry associations and extensive research being undertaken to redress this problem [4–7]. According to Heinrich [8], approximately 88% of all accidents that occur during construction materialize as a consequence of unsafe behavior. If unsafe behavior can be reduced or even prevented, then safety performance will naturally improve. According to Fam, et al. [9] unsafe behavior is enacted when an employee does not respect safety rules, standards, procedures, instructions, and specified project criteria. Such actions can adversely influence an employee's performance and/or endanger others within the workplace.

Conventional methods to determine workers' behavior have been predominately based upon observational methods. While such methods may provide useful information, they are time-consuming, labor-intensive and are subjective in nature. Due to these limitations, computer vision technology, which has been used for object recognition [10–12], can be applied to identify workers' unsafe actions on-site [13–18]. Human behavior recognition has been typically based on the use of depth sensors (Kinect™), and collection of motion data from stereo videos that are reconstructed to build a three-dimensional (3D) skeleton model [19–25]. For example, multiple video cameras have been used to monitor the behavior of workers by estimating the positioning of an individual's joints in 3D [20–24]. This method provides a useful way to obtain accurate motion data. But more specifically, it provides the ability to record, model, and analyze the human motions that have resulted from committing an unsafe action. However, monitoring the positioning of workers within a 3D environment may require lengthy computational periods and the depth sensor's line of motion may also be subjected to sensitivities in lighting [26,27].

Against this contextual backdrop, a novel hybrid deep learning model that integrates a convolution neural network (CNN) and long short-term memory (LSTM) to automatically recognize workers' unsafe actions is developed. The hybrid model is used to: (1) identify unsafe actions; (2) collect motion data and site videos; (3) extract visual features from videos using a CNN model; and (4) sequence the learning

* 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.autcon.2017.11.002
Received 23 March 2017; Received in revised form 30 October 2017; Accepted 3 November 2017
0926-5805/ © 2017 Published by Elsevier B.V.

features that are enabled by the use of LSTM models.

Video cameras are used to collect motion data. Then, a deep learning technique is applied to detect unsafe actions. Deep learning is essentially a branch of machine learning based on a set of algorithms that attempt to model high level abstractions in data. The videos of human actions contain spatial and temporal information, and a deep CNN model is trained and learns from multiple frames and spatial features contained within them. The features generated from the CNNs are fed into the LSTM models so that they can learn from peoples' actions over a period of time. The CNN is akin to a feed-forward Neural Network; it has an end-to-end structure with an automatic feature extraction. The LSTM, however, is similar to Recurrent Neural Network (RNN) models, which can enable long-range temporal interval learning to occur. The CNN and LSTM are merged together so that a sequence of feature representations of unsafe acts derived from action videos can be automatically extracted. An experiment is used to demonstrate the effectiveness of the developed deep learning hybrid model.

2. Prior research

The orthodoxy of computer vision has been reliant on extracting handcrafted features from inputs. Hand-crafted feature-based methods usually employ a three-stage procedure, which consists of: (1) extraction; (2) representation; and (3) classification. Image representation that is used to recognize human actions can extract features such as shapes and temporal motions from images. Action recognition features, however, need to contain rich information so that a wide range of them can be identified and analyzed. Techniques that can be used to analyze such features include: (a) classifier tools (e.g. Support Vector Machine (SVM)); (b) temporal state-space models (e.g., Hidden Markov models (HMM); (c) conditional random fields (CRF)); and (d) detection-based methods (e.g., bag-of-words coding). Gong, et al. [27], for example, applied the space–time interest point detector to identify interest points on the images of workers and equipment [28]. Then, using the Histogram of Oriented Gradients (HOG) [29] and Histogram of Optical Flow (HoF) descriptors to determine these interest points. This method, however, is unable to capture a worker's motions that are contained in a video or displayed on an image, which hinders its ability to accurately detect actions.

Research focusing on detecting and recording unsafe actions has tended to be based on the use of depth sensors (Kinect™) or multiple cameras to extract 3D skeleton models of a worker. For example, Han and Lee [20] utilized stereo cameras to collect motion data to construct a 3D skeleton model and used pattern recognition to identify common unsafe actions. Similarly, Liu, et al. [21] used two smartphones as stereo cameras to acquire motion data to extract 3D human skeletons to track people. Alternatively using a depth (RGB-D) sensor, Han, et al. [22] developed a modeling methodology to recognize and classify unsafe behaviors. Yet, the aforementioned methods have been demonstrated in a controlled indoor environment and therefore have not accommodated the nuances of a construction site. In addition, the requirement for lengthy computation periods, low levels of accuracy, the presence of occlusions, and high levels of illumination have stymied their capacity to effectively learn and therefore used in a real-life setting.

The use of CNN + LSTM to examine spatial and temporal information is an area that has received a considerable amount of interest within the field of computer vision [30,31]. For example, an action in a video may span different granularities. Therefore, to better recognize such actions Li et al. [29] modelled each granularity as a single stream by 2D (for frame and motion streams) and 3D (for clip and video streams) using convolutional neural networks (CNNs). In this instance, the CNN is able to learn from spatial and temporal representations. However, to address the issues associated long-term temporal dynamics Li, et al. [30] employed LSTM networks to the frame, motion and clip streams. The use of LSTMs have also been applied to deal with unsegmented videos and improve the training ability of temporal deep learning models to detect activity progression [31].

In consideration of this earlier work, the framework developed in this paper uses an Inception-v3 rather than a Visual Geometry Group (VGG) deep network, which previous studies have tended to use. The Inception-v3 deep neural network, which was developed by Google®, has been demonstrated to achieve 76.88% Top-1 and 93.344% Top-5 accuracy at the ImageNeT Large Scale Visual Recognition Challenge 2012 (ILSVRC2012) [32]. The structure of the Inception-v3 is presented in Szegedy, et al. [33]. There are four convoluting modules in each basic configuration of the Inception-v3. The CNN comprises of a 42 layer deep convolutional network over a 299 × 299 receptive field, containing over 130 layers [33]. It is a complex neutral network, that consists of many convolutional and pooling layers, batch normalization and special modules. The ILSVRC2012 used a subset of ImageNet with approximately 1000 images in each of 1000 categories. In total, 1.2 million training, 50,000 validation, and 150,000 testing images are available. Top-1 is the conventional level of accuracy that is required: the model answer with the highest probability needs to match the expected answer. A level of Top-5 accuracy refers to any of the model's five highest probability answers that correspond to the expected answer.

3. Deep learning

Deep learning methods incorporating CNN have been demonstrated to be an effective method for computer vision and pattern recognition [34,35]. Lecun, et al. [36] developed the LeNet-5 (a CNN model) based on the Mixed National Institute of Standards and Technology (MNIST) dataset to recognize hand-written numbers. Existing limitations in computing power have hindered the potential of CNNs, but they have been successfully applied to small datasets such as the MNIST, and the Canadian Institute for Advanced Research (CIFAR-10). Improvements in hardware have provided an ability to effectively train large CNN networks by stacking multiple convolutional and pooling layers to not only recognize features from static images, but also those from videos [37,38] The CNN is configured using a graphics processing unit (GPU). This is a specialized electronic circuit designed to rapidly manipulate and alter memory to accelerate the creation of images in a frame buffer, which is intended for output in a display device [39].

Video recognition, however, is a more challenging task when compared to static images due to the difficulties associated with capturing consecutive spatial and temporal information. Xu, et al. [40] proposed a 3D convolution to compute features from both spatial and temporal dimensions. Additionally, Karpathy, et al. [41] increased the speed of training using two stream CNNs (a low-resolution context stream and a high-resolution fovea stream) to implement a largescale video classification. Notably, deep learning significantly outperforms traditional methods in video recognition. For example, it has been demonstrated that two stream CNNs [42] outperform the Interrupt Descriptor Table (iDT) method [43].

Building upon research presented above, the next section of this paper introduces a deep hybrid model that combines CNN and LSTM network to collect motion data from a video camera and learn from representations that are acquired. The CNN model is applied to each frame to capture the spatial features from the videos, while the LSTM network is used to understand the temporal information from the consecutive frames that are produced.

3.1. Action video representation with deep models

Fig. 1 presents the workflow of the proposed action recognition approach. Initially, the deep models are trained to compute feature representations from the action videos, which are structured using a combination of CNNs and LSTM models. Then, the sequences of feature vectors generated from the second LSTM layer are inserted into the