Automation in Construction 94 (2018) 282–289

Contents lists available at ScienceDirect

Automation in Construction

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

Convolutional neural networks: Computer vision-based workforce activity assessment in construction

Hanbin Luo[a,b], Chaohua Xiong[a,b], Weili Fang[a,b,*], Peter E.D. Love[c], Bowen Zhang[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:
Activity analysis
Convolutional neural networks
Computer vision
Construction
Video interpretation

ABSTRACT

Computer vision approaches have been widely used to automatically recognize the activities of workers from videos. While considerable advancements have been made to capture complementary information from still frames, it remains a challenge to obtain motion between them. As a result, this has hindered the ability to conduct real-time monitoring. Considering this challenge, an improved convolutional neural network (CNN) that integrates Red-Green-Blue (RGB), optical flow, and gray stream CNNs, is proposed to accurately monitor and automatically assess workers' activities associated with installing reinforcement during construction. A database containing photographs of workers installing reinforcement is created from activities undertaken on several construction projects in Wuhan, China. The database is then used to train and test the developed CNN network. Results demonstrate that the developed method can accurately detect the activities of workers. The developed computer vision-based approach can be used by construction managers as a mechanism to assist them to ensure that projects meet pre-determined deliverables.

1. Introduction

Having in place an efficient and effective workforce is key a determinant of time and cost performance of construction projects [1]. Statistical evidence demonstrates that productivity in the construction industry, worldwide, has been declining over several decades [2–4]. An issue that has stymied the ability to engender and enact a program of productivity improvement during construction is a lack of data to establish a 'base-line' of worker performance. If, however, the productivity of workers' is to be accurately monitored in real-time, then construction managers and their teams need to directly put in place mechanisms to address those issues that impact operations to ensure a project's desired performance levels are sustained. The corollary being the ability to control and maintain a project's predetermined deliverables and acquire a much-needed understanding and knowledge of issues that adversely affect productivity.

There has been a plethora of studies that have sought to monitor and analyze worker activity on-site using a variety of methods (e.g., direct observation, surveys, and interviews) [5–8]. While such methods have been useful in creating a body of knowledge that has been used to monitor worker activity, they are time consuming and labor intensive to undertake and have tended to produce subjective results [9–11]. In addressing these limitations, there has been a shift toward monitoring individual worker's activity and tracking their location and equipment by using non-visual sensors such as radio frequency identification (RFID) tags [12], ultra-wide band [13–15], and global positioning system (GPS) sensors [16, 17]. Several existing methods based on non-visual sensors generally track a worker's location and therefore do not measure key operational parameters of a process of such as the working sequence and cycle time. Moreover, the accuracy of the data required obtained to determine the location and productivity levels of workers often varies and can contain considerable noise rendering it difficult for performance assessments to be undertaken.

To address the limitations of location-based methods that have been used for activity recognition, computer vision has been widely used to automatically monitor workers on-site [18–20]. Computer vision essentially enables rich information (e.g., locations and behaviors of project entities and site conditions) to be extracted from images and videos. A large family of video action recognition methods is based on shallow high-dimensional encodings of local spatio-temporal feature,

* 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.2018.06.007
Received 14 June 2018; Accepted 14 June 2018

such as Histogram of Oriented Gradients (HOG) and Histogram of Optical Flow (HOF). Despite the success of the research undertaken by Yang et al. [21], several challenges remain unresolved in construction engineering, which include the recognition of:

- *Activities in complex and changing conditions*: Worker activities are typically recorded in various and changing backgrounds that are subjected to occlusions, illumination variance, and viewpoint changes.
- *Multi-subject interactions and group activities*: Workers perform interactive activities with one or more people and objects (e.g., materials). Therefore, more powerful methods to understand the construction scene would still be an issue.

To overcome these issues, a convolutional neural network (CNN) can be applied to automatically recognize the activities of workers [22]. The deep three-stream CNN can accommodate complex activities, as it can simultaneously capture static spatial features, short-term and long-term motion in a video [23]. Against this contextual backdrop, the research presented in this paper develops a deep three-stream CNN that integrates Red-Green-Blue (RGB), optical flow, and gray stream CNNs to automatically recognize worker activity on construction sites. At the same time, we use a reinforcement fusion strategy to fuse the results of the three stream CNNs. The technical challenges of the developed three-stream CNN and implications for future research are then identified. Prior to introducing the three-stream CNN, a review of extant literature on activity recognition methods in construction is presented.

2. Related Work

2.1. Vision-based method for activity recognition

The advent of high-resolution video cameras, augmented storage capacity of databases and increasing accessibility of the Internet has transformed the ability to document construction operations. As a result, research in the area of computer vision has become increasingly popular to continuously monitor activities on construction sites [9,24–28]. Depending on the complexity and duration, the activities that have been monitored can be classified as: (1) gestures; (2) actions; (3) interactions; and group activities [29]. However, research has tended to focus recognizing gestures and unsafe action recognition [24,30,31]. A detail review of the extant literature reveals that vision-based worker activity recognition has been reliant on the use of traditional handcrafted representation-based approaches, which tend to follow a bottom-up strategy for worker activity recognition.

Yang, et al. [21], for example, established a comprehensive database of worker actions (e.g., nailing and plastering) and developed a dense trajectory method to recognize them. Similarly, Gong, et al. [32] utilized the three-dimensional (3D) Harris detector, Histogram of oriented Gradients (HOG) and Histograms of Oriented optical Flow (HOF), and Bayesian network models as a learning method to recognize worker and backhoe activities. Handcrafted feature methods are heavily reliant on the manual extraction of hand-engineered features from inputs, which are derived from conventional machine learning and pattern recognition techniques. However, the construction of a machine or pattern recognition learning system requires careful engineering and considerable domain knowledge. Expertise is needed to design a feature extractor that can transform raw image data into feature vectors using classifiers such as a support vector machine (SVM) and k-nearest neighbors (k-NN) to detect or determine patterns from the inputs [33,34].

While previous research has demonstrated acceptable levels of activity recognition performance, difficulties are encountered on-site due to spatial conflicts, lighting, and occlusions [26,32]. Moreover, the ability to recognize workers activities is further hampered by the color and shape of their clothing and a site's topography, which can impact the identification of plant [35]. In accommodating these limitations, it has been suggested that deep learning can be used to provide improved levels of accuracy and bypass the process of customizing features and using classifiers to recognize workers [36].

2.2. Convolutional neural network

Deep learning methods that incorporate CNNs are effective for computer vision and pattern recognition [37,38]. LeCun, et al. [39] developed the LeNet-5 (a CNN model), which recognizes handwritten numbers, based on a dataset created by the Mixed National Institute of Standards and Technology. CNN models can effectively and automatically recognize features from static images by stacking multiple convolutional and pooling layers.

CNN methods have been widely applied to a variety of problems that are encountered in construction [22,40–43]. For example, Cha, et al. [40] developed a deep architecture for a CNN to detect concrete cracks without extracting defect features. Similarly, Feng, et al. [41] constructed a deep active learning system to identify defects (e.g. cracks, deposits and water leakages) and then classify them by their image. Contrastingly, Roberts, et al. [42], however, developed a CNN to detect and classify cranes for monitoring safety hazards using unmanned aerial vehicles. Likewise, Ding, et al. [22] proposed a hybrid learning model that integrated CNNs and long short-term memory (LSTM) to detect worker unsafe behavior.

Video recognition research has been largely driven by advances in image recognition approaches, which were often adapted to deal with video/images data. A number of studies have attempts to develop a deep architecture to recognize human actions from video [44–46]. In the majority of these studies, a stack of consecutive video frame forms the input of a CNN for activity recognition. Thus, a CNN model is expected to implicitly learn spatio-temporal motion-dependent features in its first layers and is reliant on the use of a large image/video database to recognize the activities of workers [47]. For example, Taylor, et al. [45] applied Gated Restricted Boltzmann Machines (GRBMs) to extract and learn human motion features in an unsupervised manner and then resorted to convolutional learning to fine tune its parameters. S, et al. [46] extended the 2D ConvNet to video domain for recognizing human actions using relatively small datasets. Conversely, Karpathy, et al. [48] tested the ConvNets with deep structures on a large dataset, referred to as the Sports-1 M. However, these deep models achieved a lower performance compared with shallow hand-crafted representation approaches [44]. This may have been due to the (1), available action datasets being too small for the purposes of deep learning; and (2) complexity of motion patterns to enable learning to take place. Though, it should be noted that access to databases of this nature have been difficult to create in construction and as a result this has hindered the development of intelligent monitoring systems [19].

A popular state-of-the-art activity recognition network is the two-stream convolutional network, which incorporates spatial and temporal networks [49]. Essentially, the two-stream network can achieve ‘good performance’ (i.e., a high degree of accuracy in detecting objects) in spite of limited training data and is comparable to state-of-the-art methods [49], such as improved dense trajectories [44] and spatial-temporal HMAX (Hierarchical Model and ‘X’) network [50]. However, clutter, occlusions, and changing lighting conditions are prevalent on construction sites and therefore can influence the ability to recognize worker' activities.

While extracting features from images is a relatively straightforward process, obtaining them from videos is more complex, as consideration needs to be given to spatial and temporal components. The spatial aspect, which is in the form of individual frames, possesses information about scenes and objects that are depicted in the video. While the temporal aspect, which is in the form of motion across frames, conveys the movements of the observer (the camera) or the two-stream model, which contains optical flow stream and spatial stream, has proven to be