Advanced Engineering Informatics 37 (2018) 139–149

Contents lists available at ScienceDirect

Advanced Engineering Informatics

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

Automated detection of workers and heavy equipment on construction sites: A convolutional neural network approach

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

[a] *School of Civil Engineering and Mechanics, Huazhong University of Science and Technology, Wuhan 430074, PR China*
[b] *Hubei Engineering Research Center for Virtual, Safe and Automated Construction (ViSAC), HUST, PR China*
[c] *Deptment of Civil Engineering, Curtin University, Perth, Western Australia 6023, Australia*

ARTICLE INFO

Keywords:
Deep learning
Image
Improved Faster R-CNN
Object detection
Construction site

ABSTRACT

Detecting the presence of workers, plant, equipment, and materials (i.e. objects) on sites to improve safety and productivity has formed an integral part of computer vision-based research in construction. Such research has tended to focus on the use of computer vision and pattern recognition approaches that are overly reliant on the manual extraction of features and small datasets (< 10k images/label), which can limit inter and intra-class variability. As a result, this hinders their ability to accurately detect objects on construction sites and generalization to different datasets. To address this limitation, an Improved Faster Regions with Convolutional Neural Network Features (IFaster R-CNN) approach is used to automatically detect the presence of objects in real-time is developed, which comprises: (1) the establishment dataset of workers and heavy equipment to train the CNN; (2) extraction of feature maps from images using deep model; (3) extraction of a region proposal from feature maps; and (4) object recognition. To validate the model's ability to detect objects in real-time, a specific dataset is established to train the IFaster R-CNN models to detect workers and plant (e.g. excavator). The results reveal that the IFaster R-CNN is able to detect the presence of workers and excavators at a high level of accuracy (91% and 95%). The accuracy of the proposed deep learning method exceeds that of current state-of-the-art descriptor methods in detecting target objects on images.

1. Introduction

The use of computer-vision to automatically detect the presence of people, plant, materials and equipment (i.e., objects) from images or videos on construction sites to improve safety and productivity has received widespread attention in the extant literature [1–3]. Specific applications of computer-vision include: monitoring of progress [1,4,5]; tracking of workers [6,7]; occupational health assessments [8,9]; quality management [10–12], and tracking the use of personal protection equipment (PPE) [2,13,14].

Several studies have utilized video cameras in conjunction with methods such as the Histograms of Oriented Gradients and Colors (HoG + C) features descriptor, Histogram of Optical Flow (HOF), and Scale Invariant Feature Transform (SIFT), to detect the presence of objects on construction on-sites [15–17]. Methods of this nature have been overly reliant upon manually extracting hand-crafted features from inputs that are derived from conventional machine-learning and pattern recognition. Constructing a pattern-recognition or machine-learning system requires careful engineering and considerable domain expertise to design a feature extractor that is able to transform raw images data into a feature vectors that can be used by a classifier (e.g., Support Vector Machine (SVM) and k-Nearest Neighbors (k-NN)) to detect or determine patterns in the input [18,19].

With such methods, the system's detection performance is highly dependent upon designing an effective algorithm that is able to select and describe the appropriate candidate region. Moreover, the underlying machine learning models are typically trained using small datasets (< 10k images/label), which can limit inter and intra-class variability. As a result, this hinders their ability to accurately detect objects on construction sites and generalization to different datasets [20].

To address the limitations of applying hand-engineered features, Convolutional Neural Network (CNN) can be applied to automatically detect objects on construction sites [21]. While several studies in construction have used CNNs for a variety of purposes such as detecting defects in structures [11,22], recognizing unsafe behavior [23] and determining the pose of workers on-site [8], there has been limited research that has examined their use to identify workers and heavy plant. By being able to identify workers and heavy equipment unsafe conditions and behavior can be detected and therefore provide managers with a mechanism to improve their safety performance.

[*] Corresponding author at: School of Civil Engineering and Mechanics, Huazhong University of Science and Technology, Wuhan 430074, PR China.
E-mail addresses: weili_f@hust.edu.cn (W. Fang), dly@hust.edu.cn (L. Ding), dadizhong@hust.edu.cn (B. Zhong), plove@iinet.net.au (P.E.D. Love), luohbcem@hust.edu.cn (H. Luo).

https://doi.org/10.1016/j.aei.2018.05.003
Received 10 November 2017; Received in revised form 2 May 2018; Accepted 7 May 2018

Table 1
Prior work on vision-based object detection.

Target objects	Feature descriptors	Detection approach	**Author**
Hydraulic excavators	HOG	Part-based object recognition model	Azar and Mccabe [18]
Workers	Background subtraction, HOG, HSV color histogram	k-NN classifier	Park and Brilakis [26]
Excavator, truck, and worker	HOG + C	Multiple Binary SVM	Memarzadeh et al. [31]
Concrete, red brick, and OSB boards	HGB Histogram, HSV Histogram, Histogram of Dominant edge	Multilayer Perceptron (MLP), Radial Basis Function (RBF), and SVM.	Rashidi et al. [32]
Workers wear hardhat	Background subtraction and HOG	Support Vector Machine SVM	Park et al. [2]
Various objects (e.g., plant)		Nearest neighbors, and SIFT	Kim et al. [33]
Defects in pavements		Semantic texton forests (STFs)	Radopoulou and Brilakis [34]

CNNs have been found to be highly effective in discovering intricate structures within high-dimensional datasets and therefore can be used for object detection in a number of domains. With this in mind, the research presented in this paper aims is to develop a CNN that can be used to automatically detect the presence of objects in real-time on construction sites. To accommodate the existence of occlusions and the varying size (i.e. scale) of objects that occur on construction sites, an Improved Faster R-CNN (IFast-R-CNN) is introduced and trained to detect workers and plant items. Then, the technical challenges of the developed IFaser R-CNN and the implications for future research are identified. Prior to introducing the CNN, the paper commences with a review of the extant literature of vision-based detection methods in construction.

2. Vision-based detection of objects

Vision-based detection of objects has become fertile area of research in construction as a result of recent advancements in technology and computing [1,24]. Table 1 presents a summary of vision-based research methods that have been used to detect objects on construction sites. Chi and Caldas [25], for example, applied a background subtraction algorithm to extract features from images. Then, using a naïve Bayes classifier and neural network, workers, loaders, and backhoes were identified. Contrastingly, Park and Brilakis [26] and Rezazadeh Azar and McCabe [27] relied on a Histogram of Oriented Gradient (HOG) and Haar-like features to perform worker and equipment detection. Similarly, Memarzadeh et al. [28] combined a HOG and color features with a new multiple binary SVM classifier to automatically detect and distinguish workers and equipment using videos.

Despite previous research demonstrating acceptable levels of detection performance, difficulties have been encountered on-site due to issues associated with spatial conflicts, obstructions, lighting, and an object's size (Fig. 1). The ability to detect objects has also been hampered by the color and shape of workers clothing and the topography of the site, which impacts the identification of plant, as denoted in Fig. 2. To bypass the process of the customizing features and using classifiers for visual detection, it has been suggested that the use of deep learning can result in levels of accuracy being improved, while accommodating several of the limitations of conventional computer-vision based approaches [22,29,30].

3. Convolutional neural networks

CNN can be used to overcome the limitations associated with manually extracting features due to their ability to stack multiple convolutional and pooling layers. Essentially, the convolution layer is used to detect the local conjunctions of features from the previous layer (i.e., by merging semantically similar features into one).

CNNs methods incorporate multiple levels of representation. These representations are obtained by composing simple but non-linear modules that are transformed from one level (starting with the raw input) into a representation at a higher one. An image, for example, is obtained from an array of pixel values, with the first layer automatically extracting the object's features only at its edges. The second layer can assemble motifs into larger combinations that correspond to parts of familiar objects with subsequent layers detecting objects as arrangements of them. The key feature of CNNs-based methods is that layers of features are automatically extracted, which is a general learning procedure for extracting feature from images data.

Deep CNN methods have been demonstrated to be effective method for object detection, with a significant amount of research being undertaken to improve accuracy of this process [21,35,36]. The Faster R-CNN, You Only Look Once (YOLO) and Single Shot Multibox Detector (SSD) are the most widely used deep methods for object detection [36]. While the Faster R-CNN is not the quickest computation approach, it is the most accurate [37]. However, the Faster R-CNN is unable to accurately detect small scale objects, as the RPN network can only extract those with a larger pixel size due to the anchor being oversized [33]. To achieve a higher level of accuracy in the detection of an object's size and shape, an improved Faster R-CNN that utilizes a large dataset is proposed. Here five different scales with box areas (e.g., 32^2, 64^2, 128^2, 256^2, and 512^2) and four aspect ratios (e.g., 1:1, 1:2, 1.5:1, 2:1) of anchors instead of three scales with box areas (e.g., 128^2, 256^2, and 512^2) and three aspect ratios (e.g., 1:1, 1:2, 2:1) are used.

4. Research approach

A design science research approach is adopted to design and develop a deep model that can automatically detect objects. Design science research focuses on the development and performance of (designed) artifacts with the intention of improving their functional performance. It is typically applied to categories of artifacts such as algorithms, human/computer interfaces, process models, languages and the design and development of technology [38,39]. Moreover, design science can be used to develop the corresponding knowledge and applications to design and implement a product that has value to an organization [38,39]. The research process used to design and develop the improved CNN for automatically detecting objects on a construction site is presented in Fig. 3.

4.1. Design approach for detection CNNs

The Faster R-CNN architecture and parameters are aligned with the default setting presented in Ren et al. [21]. It consists of two modules. The first is the RPN, which is a fully convolutional network for generating object proposals that are fed into the second module. The second module is the Fast R-CNN (FRCN) detector that aims to refine the region proposals. The key aspects of the Faster R-CNN is briefly introduced, and both the RPN and Fast R-CNN detector use the Zeiler and Fergus (ZF) network [41]. The procedure to implement the IFaster R-CNN model is described in the following three steps:

Step 1: Any size of original image is inputted into the CNN, and it will be reshaped into a fixed size. Based on the convolution layers