Advanced Engineering Informatics 30 (2016) 327–336

Contents lists available at ScienceDirect

Advanced Engineering Informatics

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

Vision-based action recognition of construction workers using dense trajectories

Jun Yang [a,*], Zhongke Shi [b], Ziyan Wu [a]

[a] School of Mechanics, Civil Engineering and Architecture, Northwestern Polytechnical University, China
[b] School of Automation, Northwestern Polytechnical University, China

ARTICLE INFO

Article history:
Received 12 July 2015
Received in revised form 25 April 2016
Accepted 26 April 2016

Keywords:
Worker
Action recognition
Construction
Computer vision
Dense trajectories

ABSTRACT

Wide spread monitoring cameras on construction sites provide large amount of information for construction management. The emerging of computer vision and machine learning technologies enables automated recognition of construction activities from videos. As the executors of construction, the activities of construction workers have strong impact on productivity and progress. Compared to machine work, manual work is more subjective and may differ largely in operation flow and productivity among different individuals. Hence only a handful of work studies on vision based action recognition of construction workers. Lacking of publicly available datasets is one of the main reasons that currently hinder advancement. The paper studies worker actions comprehensively, abstracts 11 common types of actions from 5 kinds of trades and establishes a new real world video dataset with 1176 instances. For action recognition, a cutting-edge video description method, dense trajectories, has been applied. Support vector machines are integrated with a bag-of-features pipeline for action learning and classification. Performances on multiple types of descriptors (Histograms of Oriented Gradients – HOG, Histograms of Optical Flow – HOF, Motion Boundary Histogram – MBH) and their combination have been evaluated. Discussion on different parameter settings and comparison to the state-of-the-art method are provided. Experimental results show that the system with codebook size 500 and MBH descriptor has achieved an average accuracy of 59% for worker action recognition, outperforming the state-of-the-art result by 24%.

1. Introduction

Productivity in the construction industry has been declining during the past few decades [1]. Since labor accounts for 33–50% of the total cost of a project, their productivity is a key factor in schedule and budget control [2]. One efficient way to manage workers' performance is to monitor their activity on site, analyze the operation in real time, optimize the work flow dynamically [3–6]. Historical observation can also benefit future worker training and education.

To monitor worker activities, current efforts usually lean on foremen collecting information from construction site by means of onsite observations, survey or interview [7]. Post processing is often required to analyze the collected data manually. The entire procedure is labor intensive, cost sensitive and can be prone to error. As reported in [8], for a case study of 870 m^2 tiling trade, 336 manual observations are required to measure the six workers' productivity. The observation has to be made four rounds a day, lasting for 14 days. Not to mention each observation has to record the specific task in detail, as well as the active and inactive time. There is an urgent need of automated activity analysis of construction workers.

Recent years, with the prevalence of cameras in construction sites, images and videos become low-cost and reliable information resources. The emerging of computer vision and machine learning technologies enables analyzing construction activities automatically. In the past decade, many researchers have dedicated to this field and made remarkable achievements [9–11]. However, some open challenges remain unsolved. For example, the behavior of construction workers needs to be further explored.

Recognition of worker behavior can be performed at various levels of abstraction. As suggested by Moeslund et al. [12], there are "action primitives", "actions", and "activities". An action primitive is an atomic movement usually in limb level, e.g., pick up a brick. An action is composed by a series of action primitives, either sequential different primitives or repetitive single primitive, e.g.,

* Corresponding author.
E-mail addresses: junyang@nwpu.edu.cn (J. Yang), zkshi@nwpu.edu.cn (Z. Shi), zywu@nwpu.edu.cn (Z. Wu).

http://dx.doi.org/10.1016/j.aei.2016.04.009
1474-0346/

laying a brick contains steps of "pick up a brick", "get mortar with a trowel", "smear the mortar", "place the brick", and "knock the brick with the trowel to fasten". An activity is in the highest level, involving in a number of subsequent actions, e.g., building a wall requires measuring, alignment, and laying bricks.

In this paper, we focus on worker action recognition from pre-segmented video clips. If integrating with action detection or segmentation in longer videos, worker productivity can be assessed automatically. Furthermore, action recognition can form initial steps towards worker activity analysis.

The contributions of this paper are twofold. First, a large scale dataset of worker actions covering a wide range of trades has been constructed. Existing human action data sets mainly focus on general body movements (walking, waving, turn around) [13–15] or common daily activities (sports [16,17], cooking [18–22], etc.). Datasets on specialty activities are rare, which by their nature have smaller inter-class difference and introduce difficulties in recognition. Second, how existing action recognition algorithm will perform on a large scale construction dataset is unknown, especially when both coarse-grained and fine-grained actions coexist. By adopting a cutting-edge video representation method – dense trajectories and evaluating on various feature descriptors, we achieve an average accuracy of 59% for worker action recognition, outperforming the state-of-the-art result by 24%.

The proposed worker action dataset is available upon request. A preliminary version of this article has appeared in [23].

The rest of the paper is organized as follows. Section 2 reviews the related literatures and discusses existing challenges. Section 3 describes the methodology in detail by illustrating dense trajectories algorithm and related feature descriptors, as well as the classification method. Section 4 presents the new data set. Section 5 gives out experimental results with discussion on parameters setting and comparison against state-of-art results. Section 6 concludes the paper.

2. Related work

This section introduces the state-of-the-art human action recognition from different aspects and discusses open challenges in worker action recognition.

2.1. Action recognition in computer vision field

Action recognition has gained plenty of interest in computer vision field due to its potential in a wide range of applications, such as robotics, video surveillance, and human–computer interface [24,25]. During the past decades, numerous approaches have been proposed for human action recognition. One of the most successful line of work is the Bag-of-Feature (BoF) [26], which detects local features in video frames, represents videos with feature descriptors, generates codebook by clustering on features and obtains a sparse histogram representation over the codebook for learning and classification. Action is spatial movement across time. Local spatio-temporal features encode video information at a given location in space and time [27]. Therefore they are suitable for action recognition. Feature detection approaches range from extended Harris detector [28], Gabor filter-based detector [29] to Hessian matrix based detector [30]. Some widely used feature descriptors are higher order derivatives, gradient information, optical flow and brightness [14,26,29]. Other researchers extend successful image descriptors to spatio-temporal domain for action recognition, such as 3D-SIFT [31], HOG3D [32], extended SURF [30], and Local Trinary Patterns [33]. Instead of representing features in the joint 3D space–time domain (wherein spatial information in images is 2D), a more intuitive option is to track feature points across time. Wang et al. [34] proposed to track the densely sampled feature points across the optical field and represent features combining multiple descriptors. Their method achieved a state-of-the-art performance on several common datasets. However, how it will score on specialty activities is still unknown.

2.2. Vision-based construction operation analysis

During the past decade, many researchers have applied computer vision technologies for construction operation analysis. For more comprehensive reviews, please refer to [9–11]. One main stream method is to detect, track workers and equipment and analyze their activities by poses or trajectories combining prior knowledge. Zou and Kim [35] track the excavator by appearance and judge the idle time through its movement status. Azar et al. [6] detect and track the excavator and dump truck simultaneously to analyze the dirt loading cycle. Gong and Caldas [36,37] detect a concrete bucket in video streams through machine learning and estimate its travel cycles based on the prior knowledge of construction site layout. Yang et al. [38] perform similar work of monitoring concrete placement activity by tracking the crane jib through 3D pose estimation. Peddi et al. [39] track workers tying rebar through blob matching, extract skeletons for pose estimation and classify their working status into effective, ineffective and contributory by poses. Gong and Calda [40] evaluate several popular algorithms for construction object recognition and tracking and develop a prototype system for construction operation analysis. Bugler et al. [41] propose a novel scheme to combine tracking based activity monitoring with photogrammetry based progress measurement for excavation process analysis.

However, in cluttered construction scenarios, it is difficult to detect and track construction entities stably through a long duration [42]. Errors from previous stages (detection and tracking) might accumulate and affect the activity analysis adversely. To solve this problem, a recent trend is to adopt the Bag-of-Feature pipeline for action recognition without detecting or tracking any construction entities explicitly. Gong et al. [43] utilize the 3D-Harris detector [28] as the feature detector, HoG (Histogram of Gradient) and HoF (Histogram of Optical Flow) as the feature descriptor, and Bayesian network models as the learning method for worker and backhoe action recognition. Golparvar-Fard et al. [44] focus on action recognition of earthmoving equipment. They use Gabor filter as feature detector [29], HoG and HoF as descriptor and Support Vector Machines for action learning. Both the above mentioned works [43,44] are tested on relatively small datasets. The average numbers of action types per each dataset are four and three separately. What is more, they all adopt a joint spatio-temporal feature description. The space domain and the time domain in videos have different characteristics naturally. It may not be reasonable to simply join them together.

Apart from obtaining videos by common cameras, adopting RGB-D cameras becomes a new trend in construction operation analysis [4,3,45,46]. Since RGB-D cameras can capture depth information, skeleton information is usually extracted to infer body poses related to various worker actions.

2.3. Datasets for action recognition

As a prerequisite for evaluation and comparison, a large amount of human action datasets have been created [47]. The complexity of existing datasets increases as that of the corresponding algorithms. Early age data sets concern more for full body actions and are usually captured under control environments. Typical examples are the Weizmann dataset [13], the KTH dataset [14] and the UIUC dataset [15]. Soon after there comes a need for real-world videos with less limitation on environment,