Advanced Engineering Informatics 25 (2011) 771–782

Contents lists available at ScienceDirect

Advanced Engineering Informatics

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

Learning and classifying actions of construction workers and equipment using Bag-of-Video-Feature-Words and Bayesian network models

Jie Gong [a,*], Carlos H. Caldas [b,1], Chris Gordon [a,2]

[a] Department of Construction, Southern Illinois University Edwardsville, Edwardsville, IL 62026, USA

[b] Department of Civil, Architectural, and Environmental Engineering, The University of Texas at Austin, 1 University Station C1752, Austin, TX 78712-0273, USA

A R T I C L E I N F O

Article history:
Received 3 February 2011
Received in revised form 21 April 2011
Accepted 3 June 2011
Available online 2 July 2011

Keywords:
Automated data collection
Computer vision
Productivity analysis
Action recognition
Bag-of-Words
Bayesian network models

A B S T R A C T

Automated action classification of construction workers and equipment from videos is a challenging problem that has a wide range of potential applications in construction. These applications include, but are not limited to, enabling rapid construction operation analysis and ergonomic studies. This research explores the potential of an emerging action analysis framework, Bag-of-Video-Feature-Words, in learning and classifying worker and heavy equipment actions in challenging construction environments. We developed a test bed that integrates the Bag-of-Video-Feature-Words model with Bayesian learning methods for evaluating the performance of this action analysis approach and tuning the model parameters. Video data sets were created for experimental evaluations. For each video data set, a number of action models were learned from training video segments and applied to testing video segments. Compared to previous studies on construction worker and equipment action classification, this new approach can achieve good performance in recognizing multiple action categories while robustly coping with the issues of partial occlusion, view point, and scale changes.

1. Introduction

Advanced sensing and information technologies are increasingly used on construction jobsites for collecting and analyzing a variety of project information that traditionally relied on manual methods [1–11]. Among these technologies, video becomes an easily captured and widely spread media, serving the purposes of construction method analyses, progress tracking, and worker ergonomic studies in the construction industry [11–13]. The associated demand for reducing the burden of manual analyses in retrieving information from video motivates further research in automated construction video understanding.

Recent studies have focused on leveraging computer vision algorithms to automate the manual information extraction process in analyzing recorded videos [4,11,14–16]. However, despite considerable progress in construction object tracking, classifying the action of construction workers or construction equipment in single view video, especially in beyond simple categories like working and not working, remains a hurdle for reaping the full benefits of video-based analysis in method studies and worker ergonomic studies. Robust action analysis algorithms that are capable of differentiating subtle action categories and handling scene clutter, occlusion, and view point changes are essential to overcome such a hurdle.

In this paper, we aim to explore the potential of an emerging visual learning approach in classifying subtle action categories in a variety of construction video segments. By action, we consider the combination of rigid and non-rigid motions. This visual learning approach is composed of four major steps including feature detection, feature representation, feature modeling, and model learning. More specifically, it utilizes 3D-Harris detector as the feature detector, local histograms as the feature representation, Bag-of-Words as the feature model, and Bayesian network models as the learning mechanism for action learning and classification. For simplicity purpose, we refer this approach as the Bag-of-Video-Feature-Words in the remaining part of this paper. We developed a test bed in MATLAB to evaluate the performance of this new approach in learning and classifying action categories in construction videos. At the same time, this study also aimed to tune a set of model parameters for the model to perform well in construction scenario. Two video data sets, including backhoe actions and worker actions in a formwork activity, are constructed from a large number of construction videos as the evaluation data sets. As the main contributions of this paper, we demonstrate that the Bag-of-Words model with local action feature representations and Bayesian learning methods have a great potential in significantly

* Corresponding author. Tel.: +1 618 650 2498.
E-mail addresses: jgong@siue.edu (J. Gong), caldas@mail.utexas.edu (C.H. Caldas), cgordon@siue.edu (C. Gordon).
[1] Tel.: +1 512 471 6014; fax: +1 512 471 3191.
[2] Tel.: +1 618 650 2867.

1474-0346/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aei.2011.06.002

advancing automated construction video understanding as it performs well in learning subtle action categories in challenging construction videos. We also characterized the impact of model parameters on the model performance; therefore, a set of good choices of model parameter values are identified.

The rest of the paper is organized as follows. Section 2 briefly reviews the relevant literature in computer vision-based construction video analysis and the background of action analysis. Section 3 explains the Bag-of-Video-Feature-Words model. Section 4 evaluates the performance of the Bag-of-Video-Feature-Words model on two video data sets. Section 5 concludes the paper.

2. Research background

Recently, extensive research studies have been devoted to developing automated data collection methods for material management [2,17], productivity monitoring [11,15,16], project status updating [8], and quality control [1]. Many of those studies have been inspired and driven by the emergence and rapid development of advanced sensing technologies such as real-time localization and/or identification technologies, and 3D imaging systems. Typical examples of real-time localization and/or identification technologies include Global Positioning System (GPS), Radio Frequency Identification (RFID), and Ultra Wide Band (UWD). Terrestrial laser scanners, Flash LADAR, and stereo vision cameras are examples of 3D imaging systems that have attracted increased attentions from the construction industry. With the rapid development of technologies, it is generally agreed that the ability of processing vast volume of data collected by new technologies is a major obstacle to gain the full benefit of these technologies [18].

2.1. Computer vision for construction activity analysis

Computer vision algorithms can be widely used in construction to improve a variety of manual processes if the problem of reliable recognition and tracking of objects on construction jobsites can be solved. In this regard, many recent studies have focused on evaluating the performance of existing vision recognition and tracking algorithms in construction environments [4,14,15]. In lieu of automated productivity measurement using videotaping, there are so far three main approaches. They include detecting the movement of construction resources [19], recognizing and tracking the trajectories of construction resources [11], and recognizing worker gestures [15].

2.2. The general approaches used in computer vision-based human action classification

Research in human action analysis quickly evolves in the computer vision domain as described in a series of comprehensive reviews [20–24]. While the automatic capture and analysis of human action has been a highly active research area for decades, there is still no silver bullet type of algorithm that can be directly applied in different applications. It is widely recognized that inferring the pose and action of humans from images or videos is a hard and often ill-posed problem.

If action analysis is the only concern, there are three types of high-level methodologies that can potentially be used (Fig. 1). In Methods I & II, it is intuitive to start with the detection of humans in the images, or more precisely, segmenting humans from the background scenes. Then, specific types of action features of detected humans will be computed. Lastly, these action features will be used to classify the actions of humans, either actions at a single moment as depicted in an image or actions in a period of time as shown in a sequence of images, into different categories. The difference between Method I and Method II is that Method I uses a single image to infer the actions, while Method II use a sequence of images. Analyzing actions from a single image is highly dependent on the accuracy of human detection since it relies on human pose analysis. Method III reflects a recent approach for analyzing human actions in images. In principle, it uses local action features to infer global action characteristics. Computing local action features in video images often doesn't require segmentation. Several studies have argued that segmentation itself is a difficult task that often fails the rest of processing steps, and directly analyzing the actions in video images at the global level can be a viable alternative [25–26].

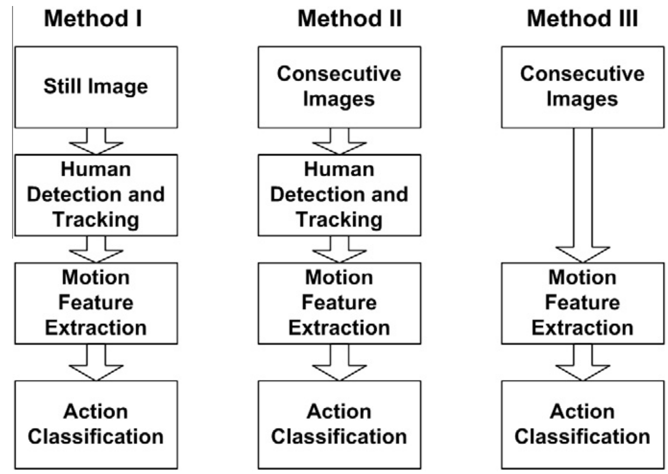

Fig. 1. Three general approaches to analyze human actions.

2.3. Local feature-based human action recognition

Local features in images have recently been extensively studied in computer vision because these features allow finding correspondences between images in spite of large changes in viewing conditions, occlusions, and image clutter as well as yield interesting descriptions of image contents [30,39]. Local features in images are often generated in two steps including detecting interest points in images and computing descriptors to describe the support region surrounding each detected interest point.

There are two main categories of interest point detectors: corner detectors and blob detectors. Commonly used corner detectors are Harris detector, SUSAN detector, and Harris-Laplace/Affine; Hessian detector, Hessian-Laplace/Affine, and Salient regions are examples of blob detectors. A comprehensive review of local features can be found in [30]. When interest points are detected in images, they typically represent significant changes in gradients along two directions (x-row and y-column) in an image. Thus, they represent changes in a spatial domain. Depending on the type of detectors used, these interest points might be invariant to scales or view point changes or both. Intuitively, these features can be extended into temporal domain by incorporating significant changes between consecutive video frames. In the context of human action analysis, a popular approach to facilitate the computation of such features in consecutive video frames is to treat human action in video sequences as silhouettes of a moving torso and protruding limbs undergoing articulated action and that such silhouettes can be described using three-dimensional shapes or volumes [25,31].

After the interest points are detected, local descriptors are used to describe the support regions surrounding interest points. Commonly used local descriptors include shape context [41], SIFT [39], steerable filters [42], Histogram of Gradient (HoG) [32], and