Automated Visual Recognition of Construction Equipment Actions Using Spatio-Temporal Features and Multiple Binary Support Vector Machines

Arsalan Heydarian¹, Mani Golparvar-Fard², and Juan Carlos Niebles³

¹ Graduate Student, Vecellio Construction Engineering and Management, Via Dept. of Civil and Env. Eng., and Myers-Lawson School of Construction, Virginia Tech, Blacksburg, VA; PH (540) 383-6422; FAX (540) 231-7532; email: aheydar@vt.edu

² Assistant Professor, Vecellio Construction Engineering and Management, Via Dept. of Civil and Env. Eng., and Myers-Lawson School of Construction, Blacksburg, VA; PH (540) 231-7255; FAX (540) 231-7532; email: golparvar@vt.edu

³ Assistant Professor, Electrical and Electronic Eng., Universidad del Norte, Barranquilla, Colombia, PH +57 (5) 350-9270; FAX (540) 231-7532; email: njuan@uninorte.edu.co

ABSTRACT

Video recording of construction operations provides an understandable data that could be used to analyze and improve construction performance. Despite the benefits, manual stopwatch study of previously recorded videos can be labor-intensive, may suffer from biases of the observers, and impractical after substantial period of observations. To address these limitations, this paper presents a new vision-based method for automated action recognition of construction equipment from different camera viewpoints. This is particularly a challenging task as construction equipment can be partially occluded and they usually come in wide variety of sizes and appearances. The scale and pose of the equipment action can also significantly vary based on the camera configurations. In the proposed method, first a video is represented as a collection of spatio-temporal features by extracting space-time interest points and describing each feature with a histogram of oriented gradients (HOG). The algorithm automatically learns the probability distributions of the spatio-temporal features and action categories using a multiple binary Support Vector Machine (SVM) classifier. This strategy handles noisy feature points arisen from typical dynamic backgrounds. Given a novel video sequence, the multiple binary SVM classifier recognizes and localizes multiple equipment actions in long and dynamic video sequences containing multiple equipment actions. We have exhaustively tested our algorithm on 1,200 videos from earthmoving operations. Results with average accuracy of 85% across all categories of equipment actions reflect the promise of the proposed method for automated performance monitoring.

INTRODUCTION

Equipment activity analysis, the continuous and detailed process of benchmarking, monitoring, and improving the amount of time construction equipment spends on different construction activities can play an important role in improving construction productivity and minimizing construction carbon footprint. It examines the proportion of time equipment spend on specific construction activities. Combination of detailed assessment and continuous improvement can help minimize the idle time, improve productivity of operations (Gong and Caldas 2011), save time and money (Zou and Kim 2007), and result in reduction of fuel use, construction...
emissions and carbon footprint (Lewis et al. 2011, EPA 2010). It can also extend equipment engine life and provide safer environment for operators and workers.

Despite the great benefits that activity analysis provides in identifying areas for improvement, implementation, and reassessments, an accurate and detailed assessment of work in progress requires an observer for equipment involved in every construction activity which can be prohibitively expensive. In addition, due to the variability on how construction tasks are carried out, or in the duration of each work step, it is often necessary to record several cycles of operations. Not only are the traditional time-studies labor intensive, but also significant amount of information that needs to be manually collected and analyzed can affect the quality of the process. Furthermore, without a detailed and continuous activity analysis, it is not possible to investigate the relationship between the activity duty cycles versus fuel use and emissions (Frey et al. 2010). There is a need for a low-cost, reliable, and automated method that can be widely applied across all projects. The method needs to remotely and continuously track construction operations, analyze their productivity, and provide detailed field data. Such data along with non-road diesel construction equipment fuel use and emission field data can be used for detailed construction carbon footprint monitoring (Lewis et al. 2011).

Over the past few years, cheap and high resolution digital cameras, extensive data storage capacities, in addition to the availability of internet on construction sites have enabled capturing and sharing of construction video streams on a truly massive scale. Detailed and dependent video streams provide a transformative potential of gradually and inexpensively sensing action and location of construction equipment, enabling construction firms to remotely analyze progress, safety, quality, productivity, and carbon footprint (Heydarian and Golparvar-Fard 2011). Using a network of high definition video cameras, this research proposes a new method for automated detection of construction equipment actions. Our proposed method can reliably detect construction equipment actions from limited line of sight, with static and dynamic occlusions, and varying illuminations which are typical in unstructured and dynamic construction sites. In the following sections, first the state of knowledge in the areas of automated tracking and action recognition are briefly reviewed. Next, the research objectives and the proposed vision-based method are discussed. Finally the comprehensive data and experimental results are presented.

BACKGROUND AND RELATED WORK

Current Research in Automated Construction Resource Tracking

In recent years, several groups have focused on creating techniques that can automatically assess construction productivity and facilitate operation idle reductions. Gong and Caldas (2011, 2009), Grau and Caldas (2009), Haas and Goodrum (2009), and Su and Liu (2007) emphasize on the importance of a real-time construction operation tracking of resources. Gong and Caldas (2009) presented a vision-based tracking model for monitoring a bucket in concrete placement operations. Zou and Kim (2007) presented an image-processing approach that automatically quantifies the idle time of hydraulic excavator. This approach uses color information for detecting 2D motion of equipment, and may not be robust to changes of illumination, viewpoint, occlusions, or scale. In order to eliminate deficiencies in existing manual data collection, El-Omari and Moselhi (2009), and Ergen et al. (2007) introduced an
automated locating and tracking of construction equipment using Radio-Frequency Identification (RFID) combined with GPS technology. Nonetheless, RFID tags still require a comprehensive infrastructure to be installed on the jobsite, and in the case of GPS, the line of sight in many locations may be adversely impacting their benefits.

In most of these state-of-the-art approaches, data collection and analysis are not fully automated. This significant amount of information required to be manually collected (1) may adversely affect the quality of the analysis, and make it subjective (Golparvar-Fard et al. 2009, Gong and Caldas 2009, Grau and Caldas 2009), and (2) minimizes opportunities for continuous monitoring which is a necessary step for performance improvement (NIST 2011). Hence, many critical decisions may be made based on inaccurate or incomplete information, ultimately leading to project delays and cost overruns. Several research studies (e.g., Brilakis et al. 2011, Gong and Caldas 2011) have proposed vision-based methods for tracking project entities that has the potential in addressing some of the needs for inexpensive tracking mechanisms. Among these techniques, Gong and Caldas (2011) proposed a new action recognition method and recommended application of unsupervised probabilistic Latent Semantic Analysis learning method. Given limited line of sight, static and dynamic occlusions, and varying illuminations, applicability of unsupervised learning models in unstructured and dynamic construction sites still requires further verifications.

Current Research in Action Recognition in Computer Vision Community

A large body of work in vision community has studied the human action recognition. Although, these methods are promising and could help develop automated algorithms to improve the construction industry, in most cases they are only applied to controlled settings, and may not directly apply to typical dynamic construction environments. Furthermore, several assumptions such as known starting points for actions in videos, or minimal variation in duration of activities can significantly impact their performance with less controlled video streams. Nonetheless, certain elements of these works can be effectively used for construction action recognition.

A number of approaches have adopted the bag of spatio-temporal interest point representation for human action recognition (Dollar 2005, Laptev et al. 2005). This can be combined with discriminative classifiers (e.g., SVMs) (Marszalek et al. 2009, Laptev et al. 2008), semi-latent topic models (Wang and Mori 2009), or unsupervised generative models (Niebles et al. 2008, Wong et al. 2007). Such approaches are effective but ignore temporal ordering of features in the sequence. Other methods have shown the use of temporal structures for recognizing human actions using dynamical Bayesian networks and Markov models (Ikizler and Forsyth 2008, Laxton et al. 2007). These methods require manual design and detailed training data which is time consuming, or can only work in simplified environments.

RESEARCH OBJECTIVES AND ASSUMPTIONS

This research explores application of a network of cameras that can visually document and assess construction operations at high fidelity. Given a collection of pre-labeled videos, the goal is to automatically learn different classes of equipment
actions per equipment and apply the learned model to perform action categorization in new video sequences. This work does not intend to assess all types of construction equipment. Rather the focus is on earthmoving equipment. The videos are assumed to contain camera motion which is typical due to lateral forces such as wind on construction sites. Also dynamic construction background in the videos is included to generate motion clutter. Finally, it is assumed each video only contains one major action for single equipment. These assumptions are relaxed during the testing stage to validate the robustness of our approach against cases of severe occlusion and multiple instances of equipment. The objectives are: (1) Create a comprehensive video dataset of all possible types of equipment actions; (2) automatically classify different earthmoving equipment from video frames at a high accuracy; (3) automatically recognize different actions of construction equipment at reasonable accuracy; and (4) test and validate the proposed research framework on long sequences of video with multiple equipment and multiple action categories. In the following, the research methodology and validation to address these objectives are detailed.

PROPOSED METHOD

Automated Action Recognition Model

An overview of our approach is presented in Figure 1. Given a set of labeled video sequences, the objective is to identify an action class for each sequence. First each video is summarized as a set of spatio-temporal features. A Histogram of Oriented Gradients (HOG) feature descriptor is then learned for each interest point and a codebook is generated to describe the video in form of a set of distinct visual codes using K-means clustering. Next, each video is formed into a mixture of these distinct visual codes and is represented with a bag of words histogram. Finally, we use a multiple binary discriminative Support Vector Machine (SVM) classifier to classify action categories. In the following, each step is discussed in more detail.

Figure 1. An Overview of Our Feature Detection, Supervised Learning, and Testing

Spatio-Temporal Feature Representation - There are several choices in the selection of features to describe actions of equipment. In general, there are three popular types of features: static features based on edge and shape of individual elements (Feng and Perona 2002), dynamic features based on optical flow measurements (Dalal et al. 2006), and spatio-temporal features obtained from local video patches (Blank et al. 2005, Cheung et al. 2005, Dollar et al. 2005, Laptev 2005). Spatio-temporal features are shown to be useful in the human action categorization (Niebles et al. 2008). Hence, in our method videos are represented as collections of spatio-temporal features by extracting space-time interest points. To do so, 2D Gaussian and 1D Gabor filters (Dollar et al. 2005) are applied to the video to obtain the responses. In order to create codebooks from the spatio-temporal features and their descriptors, the feature descriptors of all training video sequences are initially clustered. This is due to the large size of the total number of features. Cuboids, containing the spatio-temporally windowed pixels with respect to the local maxima of the response functions are then extracted. Based on a given temporal
scale, the size of the cuboid is set to contain the volume of data that contributed to the response function at that interest point. For each cuboid, a HOG histogram (Laptev et al. 2008) is then computed and a set of training codebooks are created. Assuming orientation of the camera is fixed during video recording, separable linear filters are applied as follow:

$$R = (I \otimes g \otimes h_{ev})^2 + (I \otimes g \otimes h_{od})^2$$

(1)

Where $I$ is a video frame, $g(x, y, \sigma)$ is the 2D Gaussian kernel, applied along the spatial dimensions, $h_{ev}(t; \tau, \omega) = -\cos(2\pi t \omega) \times \exp(-t^2/\tau^2)$ and $h_{od}(t; \tau, \omega) = -\sin(2\pi t \omega) \times \exp(-t^2/\tau^2)$ are a quadrature pair of 1D Gabor filter which is applied temporally. The two parameters $\sigma$ and $\tau$ correspond to the spatial and temporal scales of the detectors respectively. Fig. 2 shows an example of interest-points detection in excavator’s actions. Each red box represents a detected spatio-temporal interest point.

**Codebook Formation** - In order to learn the distributions of spatio-temporal features from the training dataset, a set of HOG descriptors corresponding to all detected interest points is generated. The codebook is then constructed by clustering these descriptors using the $K$-means clustering algorithm and Euclidean distance as the clustering metric. This process assigns a unique cluster membership to each detected interest point, such that each video can be represented as a probability distribution of the spatio-temporal interest points belonging to different cluster centers. Each probability distribution is then structured in form of a histogram.

**Learning Action Models Using a Multiple Binary Support Vector Machine classifier**

To train a supervised learning model for our action categories, the method of Support Vector Machine (SVM) classification is used. SVM is a discriminative learning method which is based on the structural risk minimization induction principle. In the case where samples are limited, traditional classifiers such as Naïve Bayes or KNN may not work well since they may result in over-fitting. For each action category ($i$), an individual binary linear-kernel SVM classifier ($\Psi_i$) is formed. In order to extend the binary classification decision of each SVM classifier to multiple classes (i.e., to classify multiple actions per equipment), we adopt the one-versus-all multiclass classification scheme. When training the SVM classifier that corresponds to each action class, we set all the examples from that class as positive and the examples from all other classes as negatives. The result of the training process is one binary SVM classifier per action of interest (e.g., digging vs. non-digging). Given a novel testing video, we apply all binary classifiers and select the action class corresponding to the classifier with the highest score.

**EXPERIMENTS AND DISCUSSION OF THE RESULTS**

**Creating a Comprehensive Video Collection of Earthmoving Operations**

Due to the lack of existing databases for benchmarking visual actions of different construction equipment, at first it was necessary to create a complete...
comprehensive database. In order to prove the concept, we initially focused on only one type of construction operation (e.g., excavation). Eventually this technique could be implemented for all equipment on the job site to automatically visualize different actions. Given the variety of construction equipment, their forms and shapes, different representation from different viewpoints, lighting and weather conditions, and static and dynamic occlusions it is very important to assemble a comprehensive datasets of possible actions. As shown in Fig. 3, cameras were initially setup around a semi-circle, roughly 45° apart from one another. The cameras were placed in multiple distances from the equipment to ensure video streams are collected at different scales of equipment in 2D.

This structured data collection helped with creating a comprehensive dataset of different equipment actions from different viewing points and scales. The videos were broken down into several groups with respect to the type of equipment used. Particularly, we formed the following groups of earthmoving equipment: (1) an excavator and several dump trucks; (2) a backhoe and several dump trucks; (3) a scraper, an excavator, and several dump trucks; (4) a scraper, a dozer, several dump trucks; and finally (5) a loader, and several dump trucks. During the validation process based on the above categories, it was further assumed that the types of equipment and the possible categories of actions are known. For this purpose, over 100 hours of videos from five different construction sites were recorded. The recorded video database contained four types of excavator’s actions: digging, hauling, dumping, and swinging along with three types of truck’s actions: filling, moving, and dumping. From these videos, approximately 400 training videos were annotated for each actions of equipment (overall ~2,800 videos). The collected data is shared at: (www.raamac.cee.vt.edu/equipmentactionrecognition).

Experimental Results on Automated Action Recognition

In this paper, we focused on earthmoving operation of an excavator and multiple trucks. The training dataset of actions for the excavator consists of four actions: digging, hauling, swinging, and dumping (Fig. 4a, b, c). For the dump trucks, the actions consist of dumping, filling, and moving actions (Fig. 4d, e, f). Since “hauling” and “swinging” have similar visual characteristics, in some of the experiments these categories were integrated into the same class as which is named “combined”.

First, the spatio-temporal feature points were extracted from the 250×250 training videos, and 4D cuboids were formed. After several analyses of the outcomes of different values for $\sigma$ and $\tau$, the spatial and temporal scales of the detectors were set to 2 and 3 respectively. This defines the sizes of the cuboids and the
windows which were selected based on the local maxima of the response functions. The HOG descriptors and the probability distribution of the spatio-temporal features were then automatically learned through the $K$-means clustering algorithm. To test the accuracy of the proposed algorithm, the annotated videos from the excavator and truck databases were divided into training and testing groups. To visually understand the best training/testing ratio, five different experiments were performed (50/50, 60/40, 70/30, 75/25, and 80/20 percent). In each case, the decision values (Fig. 5), precision recalls (Fig. 6), and confusion matrices (Fig. 7) of each action were analyzed.

Once the histogram representation for all input sequences is computed, the algorithm automatically learns the classification parameters through the multiple one-against-all binary SVM classifiers (Fig. 5a, b, and c). For instance, the “digging” videos which are above the red classification line in Fig. 5a were separated from the “non-digging” videos actions below the red line. This process was performed for all the training videos for every actions of the given equipment. Based on the learned decision values, the multiple binary SVM classifier categorize the correct action class for the testing videos (Fig. 5d, e, and f). As shown in Fig. 5f, approximately 75 of the 150 testing videos were put in the “combined” action category (i.e., hauling and swinging). Fig. 6 illustrates the precision and recall – the fraction of retrieved instances that are relevant and the fraction of relevant instances that are retrieved – that are computed for both training and testing categories. To measure the accuracy of our machine learning model, the confusion matrices for each equipment action were formed. The confusion matrix shows how many actual results match the predicted classification results. After several experiments, it was realized the “hauling” and “swinging” actions’ features are similar and could possibly be set as one category. Fig. 7a shows that in the case of combining these categories, a higher accuracy can be achieved. Given all possible variations such as viewpoint, lighting, and scale, the results of our truck action classification is promising (Fig. 7c).

To automatically and accurately recognize the starting point and the duration of each equipment action with a confidence level of 95 percent on long video sequences, a new temporal sliding window algorithm is developed. In this algorithm, the average ($\mu$) and standard deviation ($\sigma$) of the expected duration of each action is calculated using the training

![Figure 5. Excavator's Training/Testing Decision Value](image_url)

![Figure 6. Training/Testing Precision-Recall for Excavator](image_url)
dataset. Due to the need for accurate prediction of the actions and their durations, each temporal sliding window (overall duration of \( \mu \pm 2.05\sigma \) representing the 95 percent confidence interval) is divided into separate time frames. These time intervals, all starting from frame \( (t_i) \) have sequential terminating points. For example, if four separate time frames are selected, then the first time frame ends at \( [t_i+(\mu-2.05\sigma)] \), the second one ends at \( [t_i+(\mu-1.025\sigma)] \), the third ends at \( [t_i+(\mu+1.025\sigma)] \), and finally the last one ends at \( [t_i+(\mu+2.05\sigma)] \). For each time frame, the spatio-temporal features are first extracted and the probability of their distribution for each time frame is automatically learned by clustering their HOG descriptors using \( K \)-means clustering algorithm. The outcome is a histogram for each time frame. The histograms for all time frames are placed into a secondary multiple binary SVM classifier, and for each time frame, the action category plus their classification scores are stored. The time frame which its detected action category has the highest score will be used to determine the duration of the detected action and identify the starting point for the next iteration. This process is repeated until the last frame is visited.

![Figure 7. Confusion Matrix for Excavator and Truck Action Categories](image)

With the proposed new approach (1) accurate duration of each equipment action is recognized with 95 percent confidence level addressing one of the major limitations of most vision-based algorithms; (2) 2D tracking of construction equipment is automated (Fig.8); (3) multiple action recognition for multiple operating equipment is automated (yet another limitation in previous studies) (see Fig.9); (4) the algorithm is trained to automatically detect the temporal starting point of all action sequences without \textit{a priori}; (5) a new “idle-recognition” procedure is added to the algorithm for detection of the idle time given possible lateral motions of the camera (e.g., impact of wind) or in presence of visual noise (e.g., background clutter); (6) and most importantly the results of action recognition on a long video sequence can be used as the tool for productivity and carbon footprint analysis of operations.

![Figure 8. Equipment Action Recognition](image)

![Figure 9. Multiple Action Recognition for Multiple Operating Equipment](image)
CONCLUSION AND FUTURE WORK

This paper presents a new method for automated action recognition of earthmoving equipment from a network of fixed video cameras. The experimental results show the robustness of the proposed approach to variation in size and type of construction equipment, camera configuration, lighting condition, and presence of occlusions. Compared to other sensing technologies (e.g., GPS, wireless trackers), the application of the proposed method is deemed practical as it does not require additional hardware for tagging construction entities. It can also minimize the challenges associated with the detection and reduction of construction equipment idle time which has been identified as a challenge by EPA and also the Construction Industry Institute. Successful execution of the proposed research will transform the way construction operations are currently being monitored. Construction operations will be more frequently assessed through an inexpensive and easy to install solution, thus relieving construction companies from the time-consuming and subjective task of manual method analysis of construction operation, or installation of expensive location tracking and telematics devices.

The current model is capable of automatically recognizing the actions of the construction equipment on a given video. In order to provide a comprehensive productivity and carbon emission analysis, future work will include real-time tracking of construction equipment in 3D, equipment recognition (brand and model), and real-time action recognition of equipment under various degrees of occlusion.

ACKNOWLEDGEMENT

The authors would like to thank the Virginia Tech Facilities, Holder and Skanska Co. for providing access for a comprehensive data collection. The authors also like to acknowledge the support of the following members of the RAAMAC lab with data collection: Chris Bowling, David Cline, Hooman Rouhi, Hesham Barazi, Daniel Vaca, Marty Johnson, Nour Dabboussi, and Moshe Zelkowicz.

REFERENCES


U.S. Environmental Protection Agency (EPA) (2010). “Climate change indicators in the united states.” USEPA #EPA 430-R-10-00.


