Leadership in Sustainable Infrastructure
Leadership en Infrastructures Durables

Vancouver, Canada
May 31 – June 3, 2017/ *Mai 31 – Juin 3, 2017*

ANNOTATION OF HEAVY CONSTRUCTION VIDEOS USING OBJECT DETECTION AND BAYESIAN NETWORKS

Rezazadeh Azar, Ehsan[1,2]
[1] Lakehead University, Canada
[2] eazar@lakeheadu.ca

Abstract: Emergence of low-cost videotaping devices and storage systems, including hardware and cloud-based, have resulted in rapid increase of the recorded construction videos. In return, several vision-based systems were developed to detect and track resources in videos to extract productivity and safety metrics. There are, however, limited efforts to semantically annotate and retrieve videos of the interest from the construction video archives. This paper introduces a semantic annotation framework which uses object recognition to locate objects, and then applies Bayesian Belief networks to annotate the objects and their actions. Finally, it employs fuzzy logic to retrieve the indexed videos. The developed system was evaluated using videos from various sources, such as a video hosting website, which provided promising performance in retrieval of the videos and also highlighted the areas for the future improvements.

1 Introduction

Construction jobsites are distinguished for their temporary and dynamic environment, because equipment and landscape of the work zones continuously change as the project progresses. Digital imagery and videotaping have been widely used in construction industry to address the need for visual documentation of the project progress, equipment, and methods used during construction. Progress monitoring, productivity measurement, quality control, and claims are some application examples for the construction images and videos. These valuable resources, however, are not used to their full potential, because images and videos are manually analysed, which is labor-intensive and expensive (Liu and Golparvar-Fard 2015; Golparvar-Fard et al. 2013). Thereby, several research projects investigated methods to automate data extraction form digital image and videos. These efforts can be classified into three main groups: 1) visual monitoring of civil infrastructure or building elements; 2) visual monitoring of construction equipment and workers; 3) activity recognition (Yang et al. 2015). Methods in the first group mainly process still images and 4D BIM snapshots to detect building elements and monitor project's progress. But the methods in the other two groups depend on the stream of videos captured by site cameras. Object detection, tracking, and action recognition algorithms are typically used in these systems to monitor workers and equipment in construction sites. Vision-based monitoring is a highly active field in the construction research community and a number of research efforts investigated recognition (Yuan et al. 2016; Tajeen and Zhu 2014; Memarzadeh et al. 2013; Rezazadeh Azar and McCabe 2012a; Rezazadeh Azar and McCabe 2012b; Chi and Caldas 2011) and tracking of equipment (Zhu et al. 2016; Park et al. 2012; Brilakis et al. 2011; Gong and Caldas 2011). The extracted spatiotemporal data can be used to estimate productivity (Bügler et al. 2016; Golparvar-Fard et al. 2013; Rezazadeh Azar et al. 2013; Gong and Caldas 2011) and to assess site safety (Kim et al. 2015; Chi and Caldas 2012).

These methods, however, are only developed to analyze videos for a certain task, such as estimation of loading cycles (Bügler et al. 2016; Rezazadeh Azar et al. 2013), and are not able to detect various types of actions that might appear in construction videos. This issue limits their ability to provide a comprehensive

annotation for an efficient video retrieval. Therefore, efforts are needed toward development of methods for semantic organization and retrieval of construction videos.

Conventional video retrieval typically relies on associated metadata, such as manual tags, descriptions, or keywords, which are usually subjective and limited. This limitation has encouraged research efforts to develop content-based video annotation, in which the algorithms use image and video processing techniques to interpret contents of the videos and annotate them. Annotation systems might use low-level features such as patterns, color values, and shapes, medium-level data such as material and objects, and high-level semantics, including actions and scenes, to analyze and annotate contents. The search engine retrieves videos of interest by matching the user queries with the machine-generated annotations. Annotation algorithms usually try to follow human cognitive perception of the video content and present annotations in a verbal format (Altadmri and Ahmed 2014). These systems typically extract low-level visual features or medium-level data, i.e. objects or material, and then employ decision-making methods such as Markovian models (Windridge et al. 2015), kernel methods (Jiang et al. 2013), and Bayesian networks (Tavassolipour et al. 2014, Kolekar 2011) to interpret extracted data and provide semantic labels. Video annotation algorithms have two main classes, generic and domain-specific methods. Algorithms in the first group provide simple and nonspecific representation of the video contents and might not be able to address professional applications. On the other hand, domain-specific algorithms are designed to annotate actions and/or events in a certain category of videos. For example, several research projects developed methods to annotate events in soccer videos (Tavassolipour et al. 2014; Saba and Altameem 2013). Efforts to address this problem in construction domain, however, are limited and a few studies used contribution of members of the public (Liu and Golparvar-Fard 2015) or manual segmentation of initial frames (Kim et al. 2016) for annotation.

This research aims at developing an innovative system to semantically annotate videos of heavy construction operations. This paper first describes the architecture of the system, and then explains methods used to develop the framework. Next, the experimental results section provides performance of the system using a number of test videos. Lastly, results, shortcomings, and future research directions for this system are discussed.

2 System Architecture

A straight-forward approach for annotation of construction videos is to use an object recognition algorithm to identify appearing equipment and index videos, but such system will have critical limitations. First, the output will be limited to equipment type and can not provide high-level information about the appearing operations. In particular, most of the earthwork equipment are multipurpose and could carry out different operations. Second, state-of-the-art object detection methods fail to provide consistent and robust precision and recall rates in uncontrolled environments (Andreopoulos and Tsotsos 2013), which could negatively affect annotation results. Therefore, this system uses object recognition as a part of the framework. Detected objects are analyzed by a probabilistic reasoning algorithm to identify the most probable action(s), which are then recorded as the annotation. Figure 1 shows a schematic view of this system.

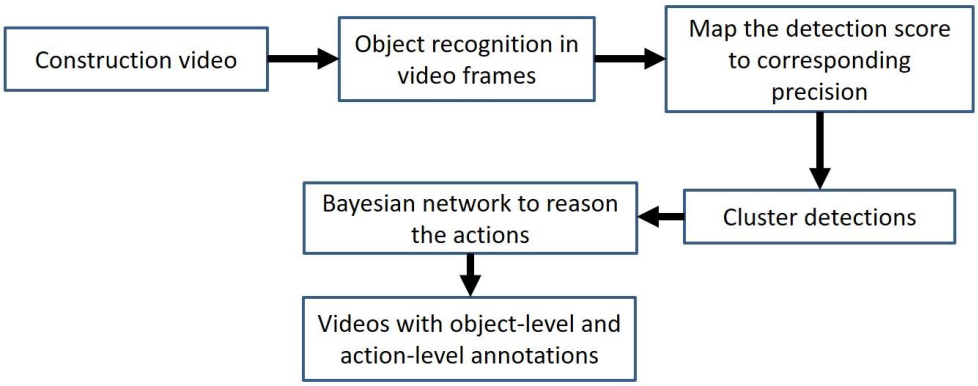

Figure 1: Schematic representation of the framework

3 System Modules

3.1 Object Recognition

There are several object recognition algorithms and some of them, such as Bayes and neural network (Chi and Caldas 2011), Histogram of Oriented Gradients (HOG) (original method developed by Dalal and Triggs (2005)) (Memarzadeh et al. 2013; Rezazadeh Azar and McCabe 2012 a and b), and latent SVM (originally method developed by Felzenszwalb et al. (2010)) (Tajeen and Zhu 2014) were tested for recognition of construction equipment. This research, however, does not aim at evaluation of object recognition algorithms and any method could be employed to detect earthmoving equipment in the video frames. The HOG algorithm was used in this research. Implementation of the method using parallel computing on graphics processing unit for accelerated performance and the availability of training datasets were the main motivations to use this algorithm in this system. Details about development of HOG classifiers are discussed in Rezazadeh Azar and McCabe (2012a and b). It should be noted that HOG could be replaced with any other recognition method in this framework. The HOG detectors were trained for recognition of five main types of equipment, including dump trucks, rollers, bulldozers, graders, and excavators, and the precision and recall diagrams of the results on test samples are presented in Figure 2.

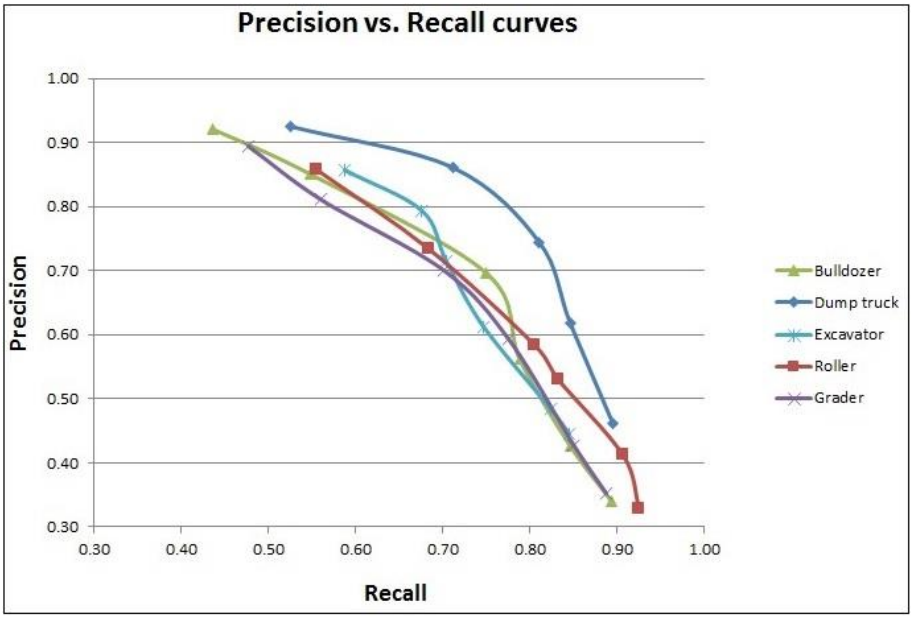

Figure 2: Precision vs recall rates for five equipment types

Bulldozers have two end-effectors, blade and ripper, with different functionalities, thus a secondary detector was added to check whether a detected bulldozer is equipped with a shank in the rear area of the machine. The HOG method was trained using images of the ripper shanks in both raised and inserted configurations. This process indicates if the detected dozer is equipped with a ripper shank system. Then a further HOG detector was used to locate the ripper. Successful detection of a ripper is interpreted as the ripper is not inserted into the ground (Figure 3).

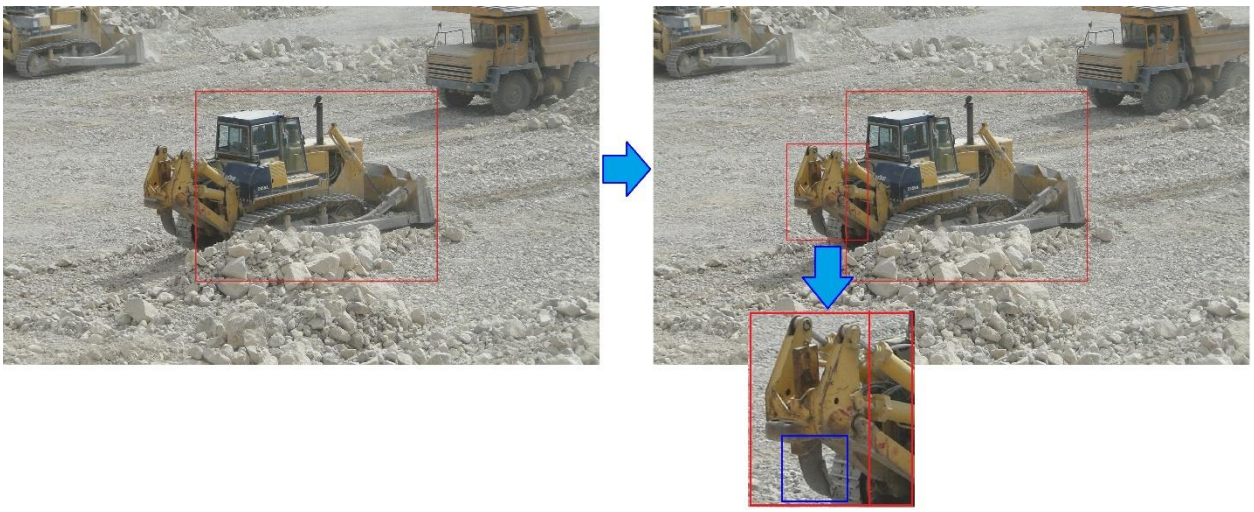

Figure 3: Detection of the shank and then ripper; the method concludes that the ripper is not in use

A 4-second interval was used to grab a frame from the video stream for object recognition (Rezazdeh Azar et al. 2013; Rojas 2008). Although setting a low threshold in the binary classifier of the HOG will result in rather high recall rate, it will also produce excessive false positives. The HOG method uses an sliding box to search a frame for target, and the windows with a score greater than the threshold are accepted as a positive test. The score of each positive window was mapped to the matching precision in the precision vs recall diagram (see Figure 2). This way, the system is able to determine the precision of each detection. In this approach, the threshold of the binary classifier for the HOG detector is set to a low-level value to gain a high recall rate, with no less than 88% in test dataset (see Figure 2). A greater precision score means that there is a higher probability that the detection is a true positive. Then, the module records the metadata of each detection, including the frame number and time into video, equipment type, detection score, and corresponding precision.

Because construction videos could be long and contain various operations, they were divided into two-minutes sections and each section was separately annotated. In this approach, the detections in each section were consolidated using a k-means algorithm (MacQueen 1967). Processing of the frames in each section of a video could produce a set of n detections (x1, x2, …, xn), where each detection is a two-dimensional real vector (equipment type and precision), of k (≤ n) types of equipment. The k-means algorithm aims to cluster the n observations into k sets S = {S1, S2, …, Sk} so to minimize the point-to-cluster-centroid distances of all observations to each centroid (see MacQueen 1967 for details). If two or more consecutive video parts have the same clusters, then those sections will be merged.

3.2 Probabilistic Reasoning

This module employs belief network to link the detected objects to higher-level semantics, i.e. earthwork operations. Belief network uses a directed acyclic graph to represent the conditional probabilistic relationships between a set of random variables and their dependencies. The conditional probabilities are based on Bayes's theorem, which are presented in Equations 1 and 2. The | character represents the conditional probability (A is true given that B is true); ∧ characterises and; ¬ represents not.

[1] $P(B \mid A) = (P(A|B) \times P(B))/(P(A))$

[2] $P(A \wedge B) = P(B|A) \times P(A)$

The belief network technique was used to analyze equipment operations for each series of frames. This probabilistic reasoning method models the semantic relationships between the equipment and their actions in two main layers. The first step in development of a belief network is to determine the relationship between the variables in two layers. The graph of this network was developed through collaboration of three site

managers with at least 15 years of experience in heavy civil engineering. Since the aim of this project was to develop a research prototype, only a limited type of earthmoving equipment, including excavators, dump trucks, bulldozers, rollers, and graders, were selected. Some of these machines can carry out different tasks, but the network focuses on the most common operations of these machines. For example, an excavator is occasionally used for lifting objects or a bulldozer could be used for pushing a scraper. But this research effort aims at proposing an innovative video annotation framework, rather than developing a generic software product, thus only a limited number of actions were modeled.

The next step in development of a belief network is to determine the probabilities of the parent nodes, i.e. P(A). Detected objects are the parent nodes in this network and the child nodes are the actions, given the object is detected. Therefore, the precision of the detected object is set as the probability of the parent node. The third phase is to determine the conditional probability between the parent and child nodes. These probabilities are typically estimated based on the expert knowledge or historical data of the domain. The application of historical data, however, has a major shortcoming, because the equipment utilization data are subjective and could differ considerably based on the nature of the projects. Expert subjective probability elicitation is the other promising approach to complete the conditional probabilities of the belief network (Tang and McCabe 2007). The conditional probabilities were solicited form nine experts with at least five years of experience in heavy construction (six site superintendents and three managers of heavy operations). The solicitation was carried out using a survey containing 10 questions, and Table 1 provides a sample question for the spreading task. A probability scale was provided (see Tang and McCabe (2007) for details) in the questionnaire to facilitate consistent probability determination, which is shown in Table 2.

Table 1: Question eliciting the probability of spreading observation

Observation	Appearance of Bulldozer	Appearance of Grader	Probability
Spreading	True	True	
	False	True	
	True	False	
	False	False	

Table 2: The scale used to facilitate consistent probability elicitation

Impossible	Seldom	Sometimes	50-50	Often	Usually	Always
1%	10% - 20%	30% - 40%	50%	60% - 70%	80% - 90%	99%

The solicited probabilities were combined to determine the conditional probability values for each child node in the Bayesian Network. This consolidation was carried out using a Bayesian updating technique, which considers the heterogeneity of expert knowledge and the quality of expert responses (Clemen and Winkler 1999). This issue is critical in solicitation from small groups, because the overconfidence of the respondents could have a substantial impact on the integrated probabilities (Clemen and Winkler 1999). Figure 4 illustrates the consolidated probabilities of the experts' responses. In general, there are two types of equipment in this figure: single-purpose and multipurpose. For example, roller is a single-purpose machine, but there are other machines with a variety of functions. Bulldozer is a special type in the second group, because it has two end-effectors: blade and ripper. As described in the object recognition section, a secondary detector was developed to check whether a ripper is inserted to ground. This outcome of this test could change the outcome of the probabilistic reasoning for bulldozers (see Figure 4).

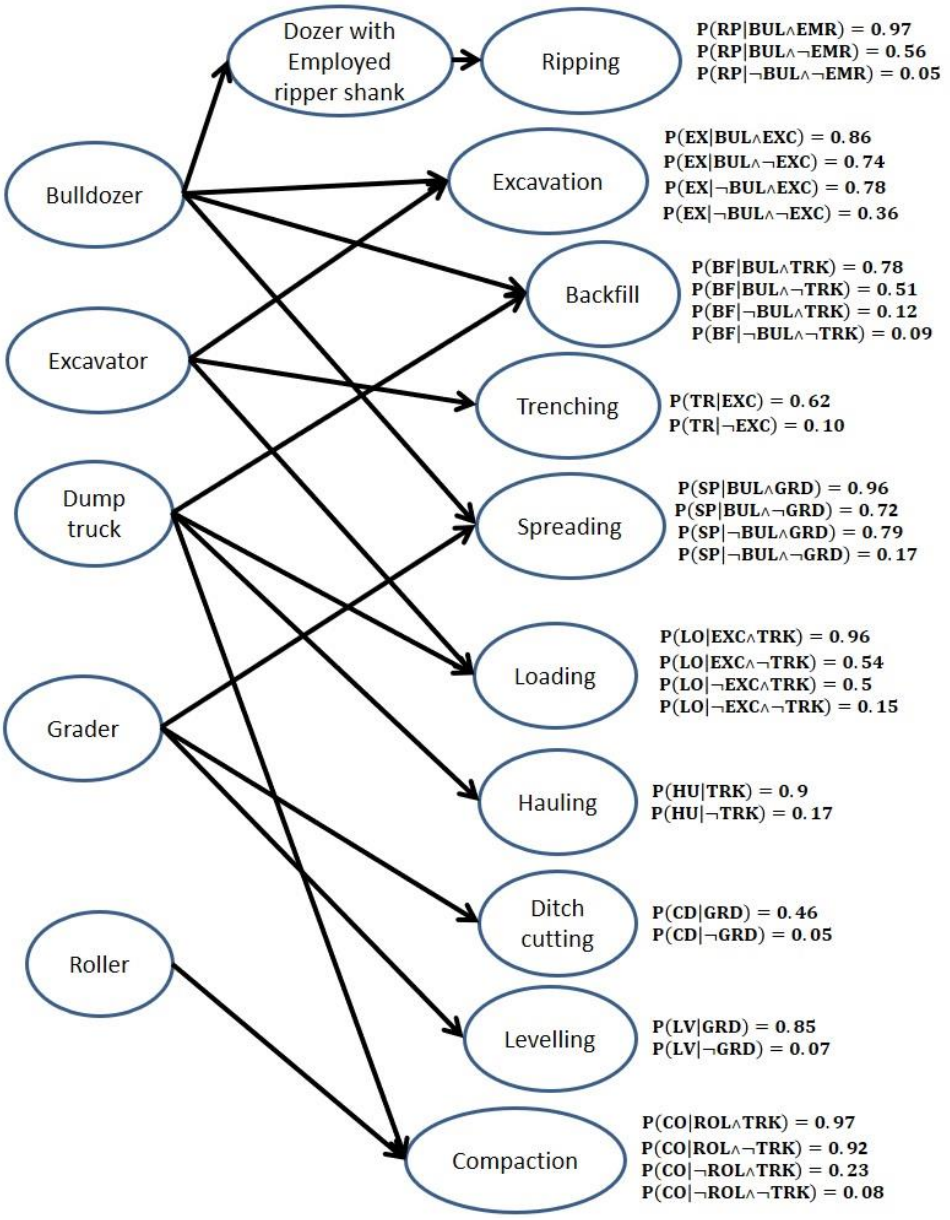

Figure 4: Consolidated conditional probabilities of equipment and their actions based on the experts' judgment

Given the P(action | detected equipment in the video section) and P(detected equipment in the video section), which is the corresponding precision of the clustered detections, the probability of the possible actions was calculated (see Equation 2), and the resulting annotations (including equipment, actions, and their probability) and the time of the observations were recorded. The system retrieves the candidates by matching the user's keyword(s) and annotations. Moreover, this system uses three fuzzy membership functions to address the differences in the probability of annotated actions, in which the user can set the level of confidence for the video retrieval. These functions include high, medium, and low levels of confidence. High confidence level would only retrieve videos with high-probability annotations, and medium and low cut-off levels would theoretically result in larger numbers of retrieved videos, including more false positives. These fuzzy membership functions are presented in Figure 5.

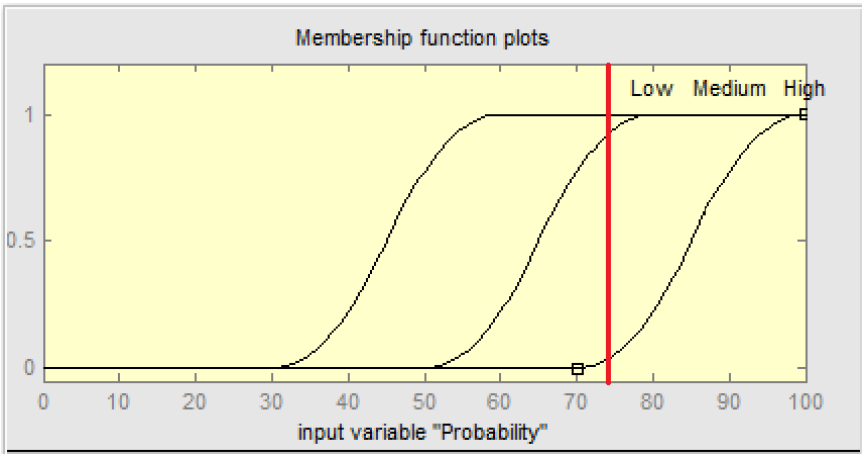

Figure 5: Fuzzy membership functions used for retrieval

4 Experimental Results

This system was implemented in Visual Studio express 2012 environment and OpenCV 2.4.12 library (OpenCV 2015) was used for computer vision processes. Performance of the developed framework was tested using 49 videos, including 38 from a publicly available video-sharing website (YouTube), and 11 from author's archive. The test videos contained various viewpoints and lighting conditions (excluding nighttime videos). A laptop with a 3.5 GHz Intel Core i7 CPU, an NVIDIA GTX 960M GPU, and 8 GB RAM was used to process test videos. The precision and recall curves of the retrievals for the actions of each type of equipment, using high, medium, and low confidence levels, are presented in Figure 6. Moreover, Figure 7 shows some sample video frames with true positive annotations.

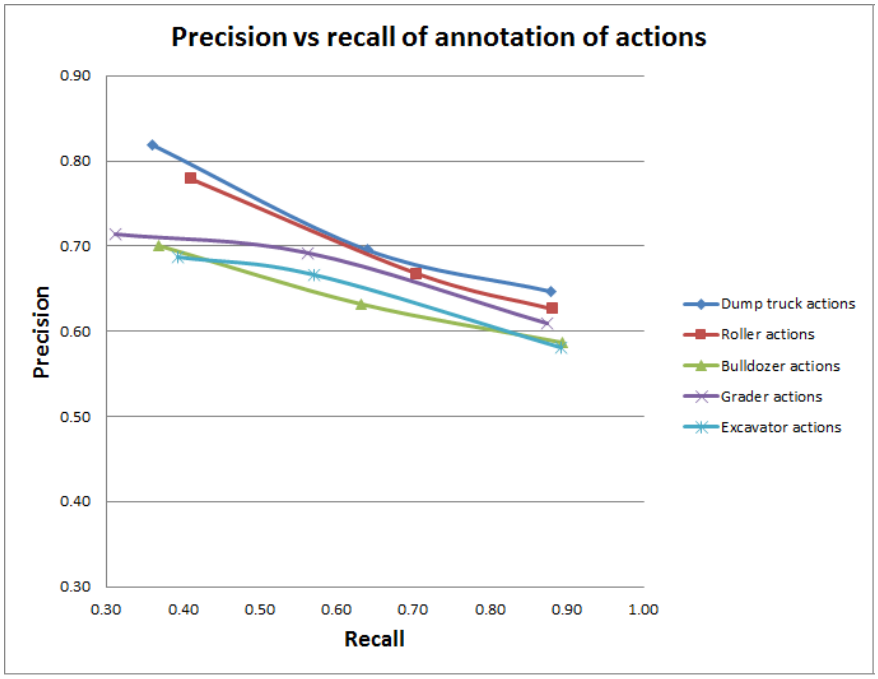

Figure 6: Precision vs recall rates of the annotations of the actions on three confidence levels

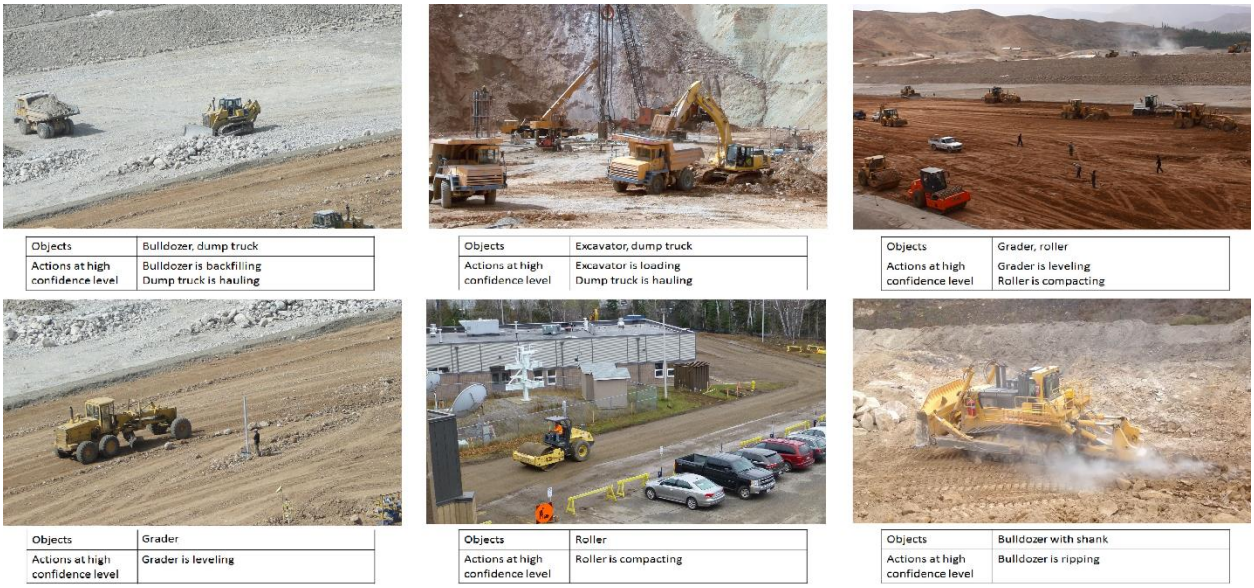

Figure 7: Sample video frames and corresponding annotations

5 Discussion

Findings of the experiments show a lower precision and recall rates in annotation of the bulldozer and excavator operations compared to dump trucks, rollers and graders. These two types of equipment can carry out multiple operations, which could contribute to this imprecision. In particular, manual review of the results revealed that the object detector successfully identified the machine type in some instances, but the annotations of the corresponding actions were not correct. Because the conditional probability, P(action | identified equipment), of the incorrect operation was greater than probability of the correct one. For example, a single excavator was located (no other equipment was present) while it was trenching, but the system recognized excavation, because the conditional probability of the excavation was greater than trenching (0.78 vs 0.62).

The performance of the systems also depends on the content of the test videos. For examples, evaluation of systems using a diverse set of videos, similar to the test dataset used in this research, could negatively affect the results. Because the probabilistic reasoning module of the system uses expert judgment, thus it might fail to successfully index occasional actions. For example, conditional probability of ditch cutting given the detection of a grader was determined lower than levelling. Thereby, the system might not perform satisfactory on the videos captured during a road maintenance and ditch cleaning operations. A paratactical solution is to adjust the conditional probabilities for the specialty projects.

Figure 8 provides more detailed results in processing of the frames with static and dynamic backgrounds. It is evident that the system had a better performance on the videos captured by stationary cameras, because they had stable views and provided longer chances to detect earthmoving plants and to reason concurrences of different types of equipment in a certain work zone.

This annotation approach, however, has a main shortcoming, because it does not consider the visual elements other than equipment on the jobsite environment. Application of material and texture recognition methods can enhance understanding of construction sites, which could improve the annotation results. In addition, metadata of the videos, including geotagging and time, could be analyzed together with the annotations to improve the video retrieval results.

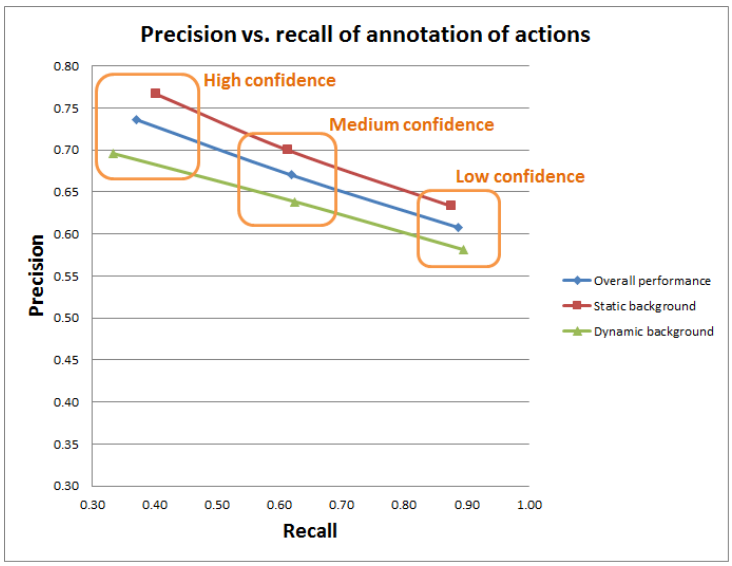

Figure 8:Precision and recall rates in static and dynamic viewpoints

6 Conclusion

An innovative video annotation system is introduced in this paper, which employs object detection and probabilistic reasoning to annotate heavy construction videos. This domain-specific video annotation system uses expert knowledge to develop a Bayesian network to represent possible events in the videos of equipment-intensive operations. This probabilistic reasoning module links the detected equipment to the most probable action(s) and calculates their probability. A prototype system was developed for annotation of actions of five types of earthmoving equipment, including bulldozers, dump trucks, graders, excavators, and rollers. A fuzzy-based function was used to enable users set the level of confidence to search for semantic concepts in the videos. The results showed promising performance on annotation of test videos, but efforts are needed to improve precision and recall rates. The future work will aim at analysing other visual information, such as the scene and material, and metadata of the videos (e.g. geotagging) to include other information in reasoning, which will result in improved annotations.

References

Altadmri, A. and Ahmed, A. 2014. A framework for automatic semantic video annotation. *Multimedia tools and applications*, **72**(2): 1167-1191.

Andreopoulos, A. and Tsotsos, J. K. 2013. 50 years of object recognition: Directions forward. *Computer Vision and Image Understanding*, **117**(8): 827-891.

Brilakis, I., Park, M.W. and Jog, G. 2011. Automated Vision Tracking of Project Related Entities, *Advanced Engineering Informatics*, **25**(4): 713-724.

Bügler, M., Borrmann, A., Ogunmakin, G., Vela, P. A. and Teizer, J. 2016. Fusion of Photogrammetry and Video Analysis for Productivity Assessment of Earthwork Processes. *Computer-Aided Civil and Infrastructure Engineering*, **32**(2): 107-123.

Chi, S. and Caldas, C.H. 2011. Automated Object Identification Using Optical Video Cameras on Construction Sites. *Journal of Computer-Aided Civil and Infrastructure Engineering*, **26**(5): 368–380.

Chi, S. and Caldas, C. H. 2012. Image-based safety assessment: Automated spatial safety risk identification of earthmoving and surface mining activities. *Journal of Construction Engineering and Management*, **138**(3): 341–351.

Clemen, R. T. and Winkler, R. L. 1999. Combining probability distributions from experts in risk analysis. *Risk analysis*, **19**(2): 187-203.

Dalal, N. and Triggs. B. 2005. Histograms of Oriented Gradients for Human Detection. Conference on Computer Vision and Pattern Recognition, IEEE, San Diego, CA, USA, 2: 886 – 893.

Felzenszwalb, P.F., Girshick, R.B., McAllester, D. and Ramanan, D. 2010. Object Detection with Discriminatively Trained Part Based Models. *Journal of Pattern Analysis and Machine Intelligence*, IEEE Transactions on, **32**(9): 1627 – 1645.
Golparvar-Fard, M., Heydarian, A. and, Niebles, J.C. 2013. Vision-based action recognition of earthmoving equipment using spatio-temporal features and support vector machine classifiers. *Advanced Engineering Informatics*, **27**(4): 652–663.
Gong, J. and Caldas, C.H. 2011. An object recognition, tracking, and contextual reasoning-based video interpretation method for rapid productivity analysis of construction operations. *Automation in Construction*, **20**(8): 1211–1226.
Jiang, Y. G., Bhattacharya, S., Chang, S. F. and Shah, M. 2013. High-level event recognition in unconstrained videos. *International Journal of Multimedia Information Retrieval*, **2**(2): 73-101.
Kim, H., Kim, K. and Kim, H., 2015. Vision-Based Object-Centric Safety Assessment Using Fuzzy Inference: Monitoring Struck-By Accidents with Moving Objects. *Journal of Computing in Civil Engineering*, 04015075.
Kim, H., Kim, K. and Kim, H. 2016. Data-driven scene parsing method for recognizing construction site objects in the whole image. *Automation in Construction*, **71**(2): 271–282.
Kolekar, M. H. 2011. Bayesian belief network based broadcast sports video indexing. *Multimedia Tools and Applications*, **54**(1): 27-54.
Liu, K. and Golparvar-Fard, M. 2015. Crowdsourcing construction activity analysis from jobsite video streams. *Journal of Construction Engineering and Management*, **141**(11): 04015035.
MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, **1**(14): 281-297.
Memarzadeh, M., Golparvar-Fard, M. and Niebles, J. C. 2013. Automated 2D detection of construction equipment and workers from site video streams using histograms of oriented gradients and colors. *Automation in Construction*, **32**: 24–37.
OpenCV. 2015. Open Source Computer Vision. < http://opencv.org/ > (August 25, 2015).
Park, M., Koch, C. and Brilakis, I. (2012). Three-Dimensional Tracking of Construction Resources Using an On-Site Camera System. *Journal of computing in civil engineering*, **26**(4): 541–549.
Rezazadeh Azar, E. and McCabe, B. 2012a. Automated visual recognition of dump trucks in construction videos. *Journal of computing in civil engineering*, **26**(6): 769-781.
Rezazadeh Azar, E. and McCabe, B. 2012b. Part based model and spatial-temporal reasoning to recognize hydraulic excavators in construction images and videos. *Automation in constr.*, **24**: 194-202.
Rezazadeh Azar, E., Dickinson, S. and McCabe, B. 2013. Server-Customer Interaction Tracker: Computer Vision–Based System to Estimate Dirt-Loading Cycles. *Journal of Construction Engineering and Management*, **139**(7): 785–794.
Rojas, E.D. 2008. Construction Productivity: A Practical Guide for Building and Electrical Contractors. J. Ross Publishing, Fort Lauderdale, Florida.
Saba, T. and Altameem, A. 2013. Analysis of vision based systems to detect real time goal events in soccer videos. *Applied Artificial Intelligence*, **27**(7): 656-667.
Tajeen, H. and Zhu, Z. 2014. Image dataset development for measuring construction equipment recognition performance. *Automation in Construction*, **48**: 1–10.
Tang, Z. and McCabe, B. 2007. Developing complete conditional probability tables from fractional data for Bayesian belief networks. *Journal of Computing in Civil Engineering*, **21**(4): 265-276.
Tavassolipour, M., Karimian, M. and Kasaei, S. 2014. Event detection and summarization in soccer videos using bayesian network and copula. *IEEE transactions on circuits and systems for video technology*, **24**(2): 291-304.
Windridge, D., Kittler, J., de Campos, T., Yan, F., Christmas, W. and Khan, A. 2015. A novel markov logic rule induction strategy for characterizing sports video footage. *IEEE MultiMedia*, **22**(2): 24-35.
Yang, J., Park, M. W., Vela, P. A. and Golparvar-Fard, M. 2015. Construction performance monitoring via still images, time-lapse photos, and video streams: Now, tomorrow, and the future. *Advanced Engineering Informatics*, **29**(2): 211-224.
Yuan, C., Li, S. and Cai, H. 2016. Vision-Based Excavator Detection and Tracking Using Hybrid Kinematic Shapes and Key Nodes. *Journal of Computing in Civil Engineering*, 04016038.
Zhu Z., Ren X. and Chen Z. 2016. Visual Tracking of Construction Jobsite Workforce and Equipment with Particle Filtering. *Journal of Computing in Civil Engineering*, 04016023.