Audio Signal Processing for Activity Recognition of Construction Heavy Equipment

Chieh-Feng Cheng[1], Abbas Rashidi [2], Mark A. Davenport[3], and David Anderson[4]

[1] PhD student, School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA
Email: ccheng71@gatech.edu
[2] Visiting Assistant Professor, College of Engineering and Information Technology, Georgia Southern University, Statesboro, GA
Email: arashidi@georgiasouthern.edu
[3] Assistant Professor, School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA
Email: mdav@gatech.edu
[4] Professor, School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA
Email : anderson@gatech.edu

Abstract

Action recognition and tracking of construction heavy equipment is the first step for benchmarking and analyzing the performance of individual machines and evaluating the productivity of a jobsite as a whole. Aside from direct observations, the current approaches for automatically recognizing and tracking various actions of construction heavy equipment includes: 1) using active sensors such as RFID tags, GPS and accelerometers or 2) computer vision-based activity analysis (processing images or videos).

In this paper, we present a novel audio-based approach for activity recognition of construction heavy equipment. Construction machines often produce distinct sound patterns while performing certain activities and it is possible to extract useful information by recording and processing those audio files at construction jobsites. The proposed audio-based framework begins with recording generated sound patterns of construction equipment using commercially available audio recorders. The recorded signal is then fed into a signal enhancement algorithm to reduce background noise commonly found at construction jobsites. The modified audio signal is then converted into a time-frequency representation using the Short-Time Fourier Transform (STFT). A Support Vector Machine (SVM) is then trained to differentiate between the acoustic patterns of the various activities of each machine. The processed audio signal is then finally divided and classified into various activities using a window filtering approach and by setting proper thresholds. We implemented the presented audio-based system at several jobsites as case studies and the results illustrate the efficiency of the system in automatically recognizing various actions of construction heavy equipment.

Keywords–

Audio signals; Construction; Heavy equipment; Activity recognition; Short-Time Fourier Transform

1 Introduction

According to the Lean Construction Institute, productive time in the construction industry is only 43% while it is 88% in manufacturing [1]. For many years, the construction industry has suffered from the lack of real-time performance monitoring, holistic project management, labor efficiency, and waste prevention tools. This has led to cost overruns in almost 90% of construction projects with an average of 28% higher than forecast costs [2]. For these reasons, “effective monitoring and feedback process” has been identified as the third most important attribute (after scheduling deficiencies and construction methods) that affects cost escalations within construction projects [3].

Action recognition and tracking of construction heavy equipment and personnel is the first step for benchmarking, analyzing, and evaluating the productivity of a construction jobsite. A number of monitoring techniques, such as direct observations, activity sampling, computer vision-based activity analysis (images or videos), etc., have been proposed to address this problem, but these proposals are either too inefficient to be practically useful or have other limitations that have limited their adoption by practitioners; hence, the construction industry is still overwhelmed by delays and suffers from frequent cost and time overruns.

To address these existing issues, we propose a novel audio-based activity recognition and performance monitoring system for construction heavy equipment. Construction heavy equipment often generates distinct acoustic patterns while performing various operations and it is feasible to extract useful information by processing the recorded audio data. In addition, a monitoring system based on processing audio data is computationally efficient, non-intrusive, and inexpensive.

The rest of the paper is organized as follows: Section 2 briefly reviews the state of practice and research for activity recognition and performance monitoring of construction equipment. Details of the proposed audio-based performance monitoring system are presented in Section 3. Necessary experimental setup for evaluating the proposed system and the obtained results are summarized and discussed in Section 4. Finally, Section 5 concludes the paper by interpreting the results and providing recommendations for future work.

2 State of Knowledge: Activity Recognition of Construction Equipment

The state-of-practice is based on manual data collection and analysis (watching the operation directly or through real-time video streams). It involves manually filling out timesheets for different equipment, collecting paper documentation like maintenance notes, measuring the amount of accomplished work, etc. The assessment is generally done through the work sampling concept. The concept relies on the principle that the percentage of equipment time spent on value-adding activities is an indirect measure of productivity. The objective is to approximate what is taking place in the field as accurately as possible. An observer records activities at regular intervals and then categorizes them to assess where the time is utilized [4]. This translates to generic and infrequent control activities that are performed off-line and consequently do not enable timely corrective measures to mitigate damage to an ongoing project. The performance is often demonstrated using a combination of graphical tools that represent the operation and the resources involved; for example, equipment balance charts, process charts, and flow diagrams [5].

A substantial need for real-time, automated, and dynamic control systems for on-site equipment has been recognized in both the industry and academia [6-10]. Such systems which provide relevant information to the user by exploiting context are called context aware systems in robotics and telecommunications.

2.1. Activity Recognition of Construction Equipment Using Active Sensors

Location tracking of heavy equipment has been one the first attempts for automating the performance monitoring of construction operations.

The problem of location tracking of construction machines can be currently addressed by several active spatial remote sensing technologies such as GPS, RFID, and Ultra-Wideband (UWB) sensors. They work based on the time-of-arrival principle; specifically, that the propagation time of a signal can be translated directly into distance from the source to the receiver [11]. These sensors are attached to equipment or workers and provide their 3D location in real-time. However, they are not capable of providing any information regarding the type of the ongoing actions. This is addressed in two ways: 1) human observation and work sampling; and 2) using extra sensors such as laser scanners, depth cameras (RGB-D), and a system of RGB cameras that are able to provide a 3D point cloud of the equipment (Fig. 4). This makes the problem more complex. The choice of a particular sensor depends on several factors including the application type, line-of-sight between objects and sensors, required signal strength, calibration requirements, permitted bandwidth, implementation costs, and environmental conditions [11].

Besides location tracking of heavy equipment, certain types of active sensors might be able to provide useful information about the performed operations. For example, in [12] Ahn et al. proposed an activity recognition system based on using an accelerometer. This approach is particularly efficient for earthmoving machines; however it is not general enough to cover a vast range of different types of construction equipment.

2.2. Computer Vision-Based Activity Recognition of Construction Equipment

Video-based techniques present an alternative and appealing solution for the aforementioned active positioning sensors. They are passive (i.e., do not emanate any sorts of energy and only use the natural light in the environment; [13-15]) and capable of providing semi-real-time information at relatively low-cost. A single inexpensive camera could be used to recognize actions of multiple pieces of equipment and minimize the need for sophisticated on-board sensors for each piece of equipment [16]. The recorded videos also offer a reliable documentation for future reviews. The entire process commonly consists of 4 steps, as illustrated in Figure 1: equipment recognition, equipment tracking, action recognition, and estimation of activity performance indicators.

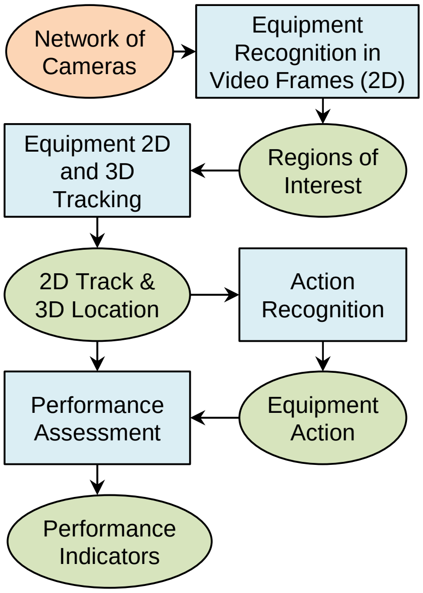

Figure 1. Overall flowchart for computer vision-based performance monitoring of construction equipment.

Equipment recognition, as a subset of object recognition, is a well-studied but challenging task due to several difficulties that arise from variations in illumination, viewpoint, occlusion, scale, articulated shapes, background clutter, and different equipment color [17]. The existing object recognition methods can be classified into three categories: recognition by parts, appearance-based recognition, and feature-based methods. It has been shown that the feature-based approach can achieve the desired performances in complex scenarios by quantizing the descriptors of positive and negative samples in order to be used for training a classifier.

Precise localization and tracking of multiple pieces of equipment in video streams is required for successful action recognition. Tracking the 3D trajectory of the equipment allows one to limit the search space to certain regions in video streams and hence minimizes the effect of noise caused by lateral movements of the camera and dynamic foreground. Contour-based tracking, kernel-based tracking, and feature matching are proposed in the literature for this purpose. Several comparative studies have indicated that the Mean-Shift algorithm outperforms other methods in visually noisy construction environment [18, 19] while the addition of the Kalman filter can further stabilize its performance.

Action recognition consists of identifying the various actions performed by a piece of equipment over time and is the most challenging step for video-based equipment performance monitoring [16]. A more precise terminology has been proposed by Bobick [20] that differentiates between movements, activities, and actions. In this terminology, for example, the action of digging a foundation by an excavator includes different activities like digging, swinging, or dumping. Each activity consists of a sequence of different movements like raising the arm or swinging the bucket. Several video-based equipment and worker action recognition studies have been performed in the construction community. Rezazadeh and McCabe [6] introduced a logical framework that combines object recognition, tracking, and rational events to recognize dirt loading to a dump truck by a hydraulic excavator. Kim and Caldas [21] presented an action recognition method for observing construction workers using interactions between actions and related objects which can be used to measure work rates for labor productivity monitoring. Golparvar-Fard et al. [16] proposed a method that initially represents a video stream as a collection of spatio-temporal visual features and then automatically learns the distributions of these features and action categories using a multi-class support vector machine classifier. More recently, Bugler et al. [22] combined photogrammetry and video analysis for tracking the progress of earthwork processes. They used photogrammetry to determine the volume of the excavated soil in regular intervals while the video analysis was used for generating statistical data regarding the construction activities.

The continuous recording of the data from previous steps enables the documentation of the various equipment actions over a period of time. A statistical analysis of the data would result in activity performance indicators such as active and idle time, cost, quality, productivity, ratio of completed work, etc. These indicators provide proactive resource monitoring capabilities and enable project managers to take corrective actions on performance deviations.

Despite the advantages that video-based performance monitoring provides over the techniques that use active sensing, the vision-based techniques still need to address significant technical and practical challenges. The first is the necessity for an appropriate level of illumination in the scene (neither dark nor direct sunlight). At the very least, this makes it extremely challenging to use video streams for performance monitoring of the tasks that are performed after sunset, which is not negligible especially in urban projects. In the absence of significant artificial lighting, a video-based approach would be nearly impossible. Another challenge is posed by the fact that cameras have a limited field of view and require a clear line-of-sight between the camera and any piece of equipment that is to be characterized. A related challenge is posed by the fact that construction sites are generally crowded with a large number of workers and equipment. Such congestion creates noisy and occluded areas in video streams which eventually disrupts accurate and automatic recognition of equipment actions. An additional key challenge is that the required video processing techniques are generally quite computationally expensive, which hinders their real-time performance and scalability. The last, and perhaps the most important, challenge is privacy, as not everyone is willing to be permanently monitored and recorded by cameras in a jobsite.

2.3. Gaps in Knowledge: Audio vs. Video

Audio processing provides four advantages over the location-based and/or computer vision-based techniques:

1) Some activities or actions are easier to recognize with sound than with other sensors. As a simple example, consider an excavator that is excavating a pit. In general, this action is not associated with a single characteristic gesture. Rather, it involves a series of movements such as digging, rotating, swinging, dumping, etc. that could be performed in a variety of ways and sequences (e.g., swinging in different directions and angles). This is extremely difficult to recognize via computer vision-based systems.
2) The sound produced by each piece of equipment is independent of the operator. An operator can perform an activity in a variety of ways. An action recognition system based on location sensors needs to take this into account, while an audio analysis always yields the same result.
3) The limited field of view of cameras used in vision-based action recognition methods constraints their range of operation from a single spot. A network of cameras is therefore needed to be placed in a circular pattern to cover all the angles in a jobsite (Figure 2). This is not a limitation for audio recording because a recorder can operate in 360 degrees. Moreover, because the sound produced by each piece of equipment can be heard even in the absence of a direct line-of-sight, an audio-based approach is less sensitive to occlusions.

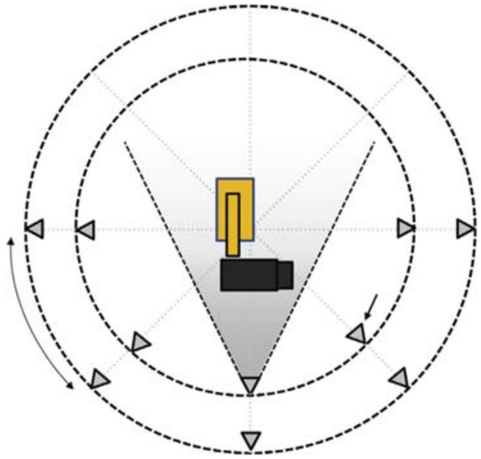

Figure 2. Camera layout in computer vision-based equipment action recognition [16].

4) The data rates that need to be analyzed in each scenario are different for every sensor modality. Table 1 demonstrates the order of magnitude for data rates in different scenarios. Even though the data rates could differ if a specific application is considered, it is obvious that processing audio data is computationally less demanding than video data.

Table 1. Approximate data rates for different sensor modalities

Input Type	Location (NMEA format)	Video	Sound
# Sensors	10	1	1
Sampling Rate	12 Hz	4608 kHz	10 kHz
Resolution	800 bit	8 bit	8 bit
Data Rate	12 kB/s	4608 kB/s	10 kB/s

3 Proposed Methodology: Audio-Based Activity Recognition of Heavy Equipment

To address the challenges associated with active sensors and computer vision-based approaches for activity recognition of construction equipment, we present an innovative activity recognition framework. This framework is based on processing audio signals generated at construction jobsites. The overall procedure is illustrated in Figure 3.

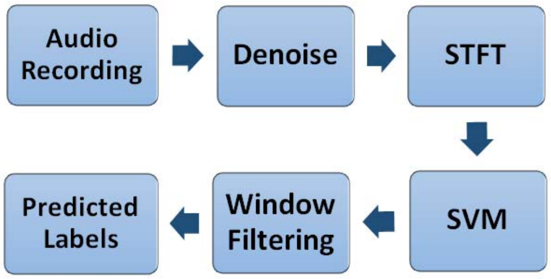

Figure 3. Overall framework for the proposed audio-based activity recognition algorithm.

While performing various tasks, the generated acoustic patterns of construction equipment can be recorded using commercially available microphones (Figure 4). We note that the current study is based on recognizing activities of a single machine and the more general and realistic setting (assuming the entire jobsite as a whole: multiple machines plus multiple microphones) will be considered as part of the future research plans.

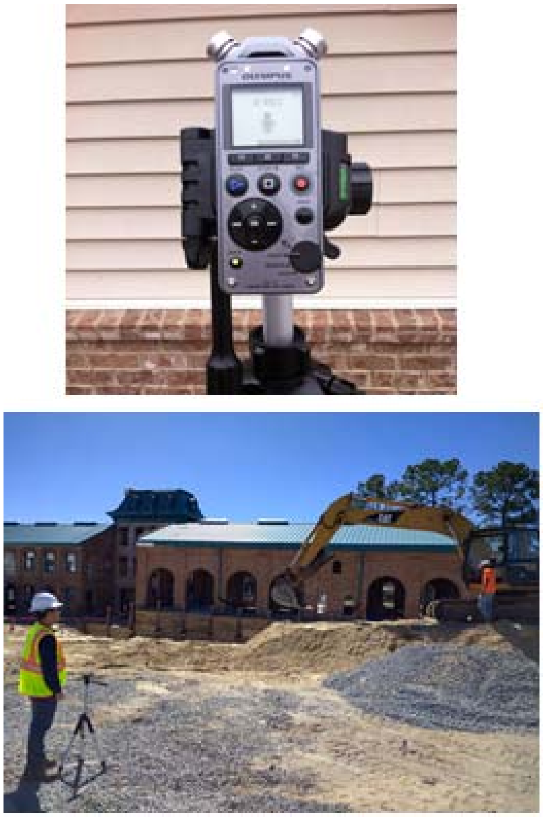

Figure 4. An audio recorder (top) and placing the recorder at a jobsite (bottom).

As the next step, each audio recording will go through a number of processing stages as explained below.

3.1. Denoising

The recorded audio is assumed to contain the signal of interest along with environmental and other noise sources. We apply a signal enhancement algorithm developed by Rangachari & Loizou [24] because it has proven to be both efficient and effective. A key aspect of this algorithm is that it can perform noise-estimation in highly non-stationary noise environments such as what might be encountered at a construction job site. An estimate of the noise is continuously updated in every frame using time–frequency smoothing factors computed based on signal-presence probability in each frequency bin of the noisy recording spectrum. More details about this algorithm can be found in [24].

3.2. STFT

The captured audio is then converted to a time-frequency representation using the Short-Time Fourier Transform (STFT). We use Hanning window with size 512, a 1024-point DFT (discrete Fourier transform), and a 50% overlap (256 overlapped samples). The output of the STFT consists of both magnitude and phase matrices, but for single microphone recordings, we only consider the magnitude part.

3.3. SVM

To identify various activities within a captured audio recording, we train a classifier for each piece of machinery. The learning algorithm which has been used in this research is the popular Support Vector Machine (SVM) algorithm. SVM training and operation phases were carried out using the MATLAB SVM package, LIBSVM [23], with a radial basis function kernel. We extract 10 to 20 seconds for each activity as the training data to construct our SVM model. The SVM parameters C and γ were tuned using a grid search over a log-scale ranging from 2^{-6} to 2^{5}, and set differently in the models for each machine. We use 10-fold cross validation to select each combination of C and γ. After building the SVM models for each machine, we randomly extract other periods of the audio recordings to use as the testing data.

3.4. Window Filtering

After building SVM models for each machine, we can identify activities from any part of the audio. However, a label will be assigned for each time bin. In practice what we would prefer is a higher-level labeling of the full period for each of the different activities. Thus, we apply window filtering on the predicted label. Using a small window the algorithm scans through the SVM labels in the time domain and calculates the percentage for each activity. If the percentage for a particular activity is higher than a set threshold, the window is labeled with that certain activity. This procedure is then repeated using a larger window.

4 Experimental Setup and Results

In order to evaluate the performance of the proposed system, 4 different pieces of construction machines operating at various jobsites were selected as case studies: 1) JCB 3CX mini excavator 2) JD 270C Backhoe 3) Volvo L250G Wheel Loader and 4) CAT D5C Dozer.

Each piece of machine was carefully monitored and the generated sounds while performing routine tasks were captured using a commercially available recorder (Tascam DR-05 2 GB). Next, each audio recording was manually labeled based on the various activities that took place during the recording time. Construction heavy equipment usually perform one major task (digging, loading, breaking, etc.) and one or more minor tasks (maneuvering, swinging, moving, etc.) in each cycle so we classified each audio recording based on two activities: Major and minor (or activity 1 and activity 2). For example, the major activity for JD 270C Backhoe is crushing.

Each audio recording was then sent through the audio processing pipeline and divided into activities 1 and 2 considering their manifestations in the STFT domain.

Finally, the performance of the algorithm for each case study was compared to manually generated labels. The comparison results are depicted in Table 3 and Figures 5 and 6.

In the table and figures, the label "Act 1" represents the major activity, while "Act 2" illustrates the minor activities. Table 3 indicates that the performance of the proposed system for automatically recognizing activities of single machines is very promising. For JCB 3CX mini excavator and CAT D5-C Dozer , "Act 1" and "Act 2" have strongly dissimilar patterns in the STFT domain, thus the identification performance can be over 90 percent. Another way to visually evaluate the performance of the system is throughout the comparison charts as shown in Figure 6.

5 Conclusion Remarks and Future Research Plans

In this paper, an innovative system for automatically recognizing construction heavy equipment activities is presented and evaluated. This system utilities audio signals as the major data source. Many tools and equipment used in construction activities produce a distinct sound. Even more, many of the actions are accompanied by a clear and dissimilar sound, be it digging soil or pouring concrete.

Processing audio signals to extract semantic information requires several steps as described in the previous sections. It is necessary to implement a signal enhancement algorithm to reduce the background noise which can potentially produce false activity recognition results. A machine learning procedure (SVM) is also required to train the sounds of each operation and machine.

The presented framework was tested using a number of audio files collected from various jobsites and the results are promising. As the first step of this research project, the proposed system has been able to accurately recognize sound patterns of single machines operating various tasks. As the plan for future research, the authors intend to study a challenging, yet more realistic setting by considering multiple machines operating simultaneously.

Acknowledgments

The presented research has been funded by the U.S. National Science Foundation (NSF) under Grants CMMI-1606034 and CMMI-1537261. The authors gratefully acknowledge NSF's support. Any opinions, findings, conclusions, and recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the NSF. The authors would also like to thank Chris Andres Sabillon Albarran, graduate student at the College of Engineering and Information Technology at Georgia Southern University, for his assistance in collecting and recording audio files.

References

[1] Aziz, R.F. and Hafez, S.M. Applying lean thinking in construction and performance improvement, *Alexandria Engineering Journal*, 52(4): 679-695, 2013.

[2] Azis, A.A.A., Memon, A.H., Rahman, I.A., and Karim, A.T.A. Controlling cost overrun factors in construction projects in Malaysia. *Research Journal of Applied Sciences, Engineering and Technology*, 5(8):2621-2629, 2012.

[3] Doloi, H. Cost overruns and failure in project management: Understanding the roles of key stakeholders in construction projects. *Journal of Construction Engineering and Management*, 139(3):267-279, 2013.

[4] Joshua, L. and Varghese, K. Accelerometer-based activity recognition in construction. *Journal of Computing in Civil Engineering*, 25(5):370-379, 2011.

[5] Su, Y.Y. Construction crew productivity monitoring supported by location awareness technologies. *PhD dissertation, University of Illinois at Urbana-Champaign*, 2010.

[6] Rezazadeh Azar, E. and McCabe, B. Part based model and spatial–temporal reasoning to recognize hydraulic excavators in construction images and videos. *Automation in Construction*, 24: 194–202, 2012.

[7] Ahn, C., Lee, S., and Peña-Mora, F. Application of low-cost accelerometers for measuring the operational efficiency of a construction equipment fleet. *Journal of Computing in Civil Engineering*, 29(2): 04014042-11, 2015.

[8] Akhavian, R. and Behzadan, A. Knowledge-based simulation modeling of construction fleet operations using multimodal-process data mining. *Journal of Construction Engineering and Management*, 139(11): 04013021-11, 2013.

[9] Akhavian, R. and Behzadan, A. Client-Server Interaction Knowledge Discovery for Operations-level Construction Simulation Using Process Data. *In Proceedings of the Construction Research Congress*, pages 41-50, Atlanta, GA, USA, 2014.

[10] Khosrowpour, A., Nieblesb, J. C., and Golparvar-Fard, M. Vision-based workface assessment using depth images for activity analysis of interior construction operations. *Automation in Construction*, 48:74–87, 2014.

[11] Cheng, T., Venugopal, M., Teizer, J., and Vela, P.A. Performance evaluation of ultra wideband technology for construction resource location tracking in harsh environments. *Automation in Construction*, 20(8):1173-1184, 2011.

[12] Ahn, C.R., Lee, S.H., and Pena-Mora, F. Construction equipment activity recognition from accelerometer data for monitoring operational efficiency and environmental performance. *In Proceedings of the International Conference on Construction Engineering and Project Management*, Anaheim, CA, USA, 2013.

[13] Dai, F., Rashidi, A., Brilakis, I., and Vela, P. Comparison of image-based and time-of-fight-based technologies for three dimensional reconstruction of infrastructure, *Journal of Construction Engineering and Management*, 139(1): 69-79, 2013.

[14] Brilakis, I., Fathi, H., and Rashidi, A. Progressive 3D reconstruction of infrastructure with videogrammetry, *Automation in Construction*, 20(7): 884–895, 2011.

[15] Rashidi, A., Fathi, H., and Brilakis, I. Innovative stereo vision-based approach to generate dense depth map of transportation infrastructure, *Journal of the Transportation Research Board*, 2215:93-99, 2011.

[16] Golparvar-Fard, M., Heydarian, A., and Niebles, J.C. Vision-based action recognition of earthmoving equipment using spatio-temporal features and support vector machine classifiers. *Advanced Engineering Informatics*, 27(4):652–663, 2013.

[17] Rezazadeh Azar, E. 2013. Computer vision-based solution to monitor earth material loading activities. *PhD Dissertation, Department of Civil Engineering, University of Toronto*, 2013.

[18] Gong, J., Caldas, C.H., and Gordon, C. Learning and classifying actions of construction workers and equipment using Bag-of-Video-Feature-Words and Bayesian network models. *Advanced Engineering Informatics*, 25(4):771–782, 2011.

[19] Park, M.W, Makhmalbaf, A., and Brilakis, I. Comparative study of vision tracking methods for tracking of construction site resources. *Automation in Construction*, 20(7): 905-915, 2011.

[20] Bobick, A. and Davis, J. Real-time recognition of activity using temporal templates. *In Proceedings of the Applications of Computer Vision, WACV '96., Proceedings 3rd IEEE Workshop on*, Pages 39-42, Sarasota, FL, USA, 1996.

[21] Kim, J.Y. and Caldas, C.H. Vision-based action recognition in the internal construction site using interactions between worker actions and construction objects. *International Association for Automation and Robotics in Construction. In Proceedings of the 30th ISARC*, Pages 661-668, Montréal, Canada, 2013.

[22] Bügler, M., Ogunmakin, G., Teizer, J., Vela, P.A., and Borrmann, A. A comprehensive methodology for vision-based progress and activity estimation of excavation processes for productivity assessment. *In Proceedings of the EG-ICE Workshop on Intelligent Computing in Engineering*, Pages 1-10, Cardiff, Wales, UK, 2014.

[23] Chang, C.C. and Lin, C.J. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3): 1-27, 2011.

[24] Rangachari, S. and Loizou, P.C. A noise-estimation algorithm for highly non-stationary environments. *Speech Communication*, 48:220-231, 2006.

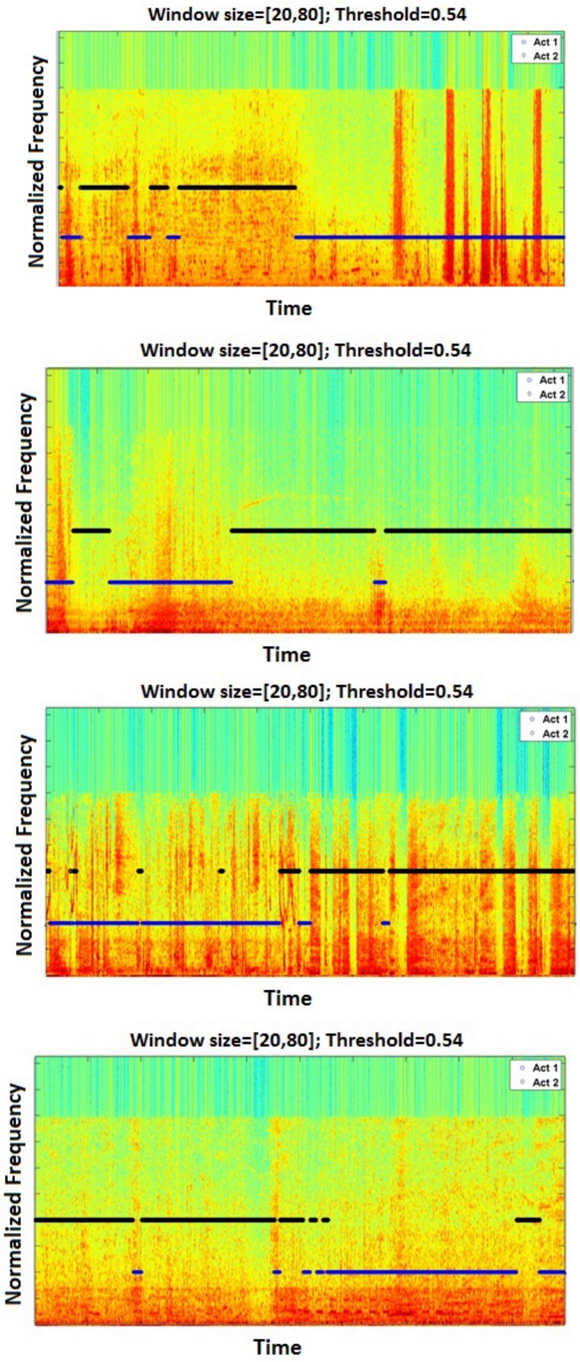

Figure 5. Visual representations of signal spectrums in frequency domain; From top to bottom: JD 270C Backhoe; Volvo L250G Wheel Loader; JCB 3CX mini excavator; CAT D5C Dozer.

Table 3. Comparison results for manually labeled audio files (correct label) versus the results obtained from implementing the audio system (predicted label).

JD 270C Backhoe		Correct Label	
		Act. 1	Act. 2
Predicted	Act. 1	0.8375	0.1728
Label	Act. 2	0.1625	0.8272
Volvo L250G Wheel loader		Correct Label	
		Act. 1	Act. 2
Predicted	Act. 1	0.8003	0.1136
Label	Act. 2	0.1997	0.8864
JCB 3CX mini excavator		Correct Label	
		Act. 1	Act. 2
Predicted	Act. 1	0.8326	0.0218
Label	Act. 2	0.1674	0.9782
CAT D5C Dozer		Correct Label	
		Act. 1	Act. 2
Predicted	Act. 1	0.8490	0.0477
Label	Act. 2	0.1510	0.9523

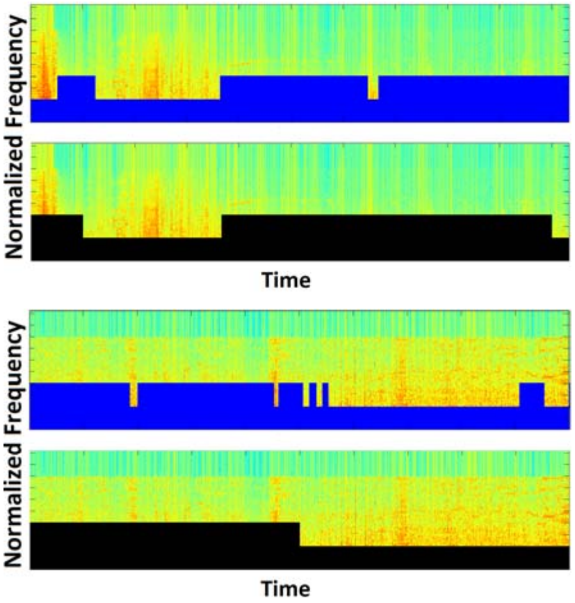

Figure 6. Comparison results between manually labeled audio files (blue) versus the results obtained from implementing the audio system (black).