Automation in Construction 81 (2017) 240–253

Contents lists available at ScienceDirect

Automation in Construction

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

Activity analysis of construction equipment using audio signals and support vector machines

Chieh-Feng Cheng[a], Abbas Rashidi[b,*], Mark A. Davenport[a], David V. Anderson[a]

[a] Georgia Institute of Technology, Atlanta, GA, USA
[b] University of Utah, UT, USA

ARTICLE INFO

Keywords:
Construction heavy equipment
Audio signals
Support vector machines
Activity analysis
Productivity

ABSTRACT

In the construction industry, especially for civil infrastructure projects, a large portion of overall project expenses are allocated towards various costs associated with heavy equipment. As a result, continuous tracking and monitoring of tasks performed by construction heavy equipment is vital for project managers and jobsite personnel. The current approaches for automated construction equipment monitoring include both location and action tracking methods. Current construction equipment action recognition and tracking methods can be divided into two major categories: 1) using active sensors such as accelerometers and gyroscopes and 2) implementing computer vision algorithms to extract information by processing images and videos. While both categories have their own advantages, the limitations of each mean that the industry still suffers from the lack of an efficient and automatic solution for the construction equipment activity analysis problem. In this paper we propose an innovative audio-based system for activity analysis (and tracking) of construction heavy equipment. Such equipment usually generates distinct sound patterns while performing certain tasks, and hence audio signal processing could be an alternative solution for solving the activity analysis problem within construction jobsites. The proposed system consists of multiple steps including filtering the audio signals, converting them into time-frequency representations, classifying these representations using machine learning techniques (e.g., a *support vector machine*), and window filtering the output of the classifier to differentiating between different patterns of activities. The proposed audio-based system has been implemented and evaluated using multiple case studies from several construction jobsites and the results demonstrate the potential capabilities of the system in accurately recognizing various actions of construction heavy equipment.

1. Introduction

It is an unfortunate fact that the construction industry suffers from lower productivity rates as compared to most manufacturing industries. According to the Lean Construction Institute [1], the productive time in the construction industry is only 43%, compared to 88% in manufacturing. One major factor contributing to this issue is that construction projects are all unique and it is typically very difficult to find completely similar projects/operations. As a consequence, in contrast to the manufacturing industries which includes highly repetitive processes, a single project management technique or a single set of fixed performance measures are rarely available in the construction industry. Thus, the first step towards improving productivity within the construction industry is to develop efficient techniques for assessing the performance and productivity of key resources that is sufficiently flexible to handle the widely varying conditions that arise across different jobsites. Since a large portion of total project costs typically derive from the costs of renting/owning/leasing/maintaining construction heavy equipment, we will focus on recognizing and tracking activities of construction heavy equipment. Note that beyond productivity assessment and analysis, monitoring of construction equipment is useful for emission control and monitoring [2], safety management [3], and analyzing/reducing idle times [4,5].

The common practice for recognizing and monitoring activities at construction jobsites is through manual data collection and direct observations. This process is known to be time consuming and labor intensive. In recent years a number of methods for automatically recognizing and tracking locations and actions of construction heavy equipment have been introduced. These methods include using active sensors (GPS (Global Positioning System), accelerometers, RFID (Radio-Frequency Identification) tags, etc.) or passive sensors (processing videos using computer vision algorithms). These methods are capable of producing promising results; however, as explained below, several constraints limit their application in real world construction jobsites. As

* Corresponding author.
E-mail address: abbas.rashidi@utah.edu (A. Rashidi).

http://dx.doi.org/10.1016/j.autcon.2017.06.005
Received 28 September 2016; Received in revised form 20 March 2017; Accepted 10 June 2017
Available online 22 June 2017

a result, the construction industry still lacks a practical, real-time, and non-intrusive method for performance characterization and monitoring of jobsite operations.

To address the drawbacks of existing methods for activity analysis of construction equipment, and in order to further advance the knowledge in this area, this paper presents a novel activity analysis based on intelligent processing of audio signals. Audio signals, recorded and collected at construction jobsites, reflect corresponding activities that take place and can provide very useful information for project managers. The proposed system begins with recording the sounds generated by construction equipment during their routine operations by using commercially available microphones. The recorded audio file then goes through a signal enhancing and feature extraction algorithm to reduce unwanted background noise. The result is then in the form of a time-frequency decomposition which can be fed into a machine learning algorithm such as a *support vector machine (SVM)* for the purpose of classifying the time-frequency features as being characteristic of particular machines/activities.

The remainder of the manuscript is organized as follows. Section 2 outlines current states of practice and research for action/location tracking of construction heavy equipment. This is followed by the research methodology in Section 3. Section 4 summarizes necessary steps for running experiments to evaluate the performance of the proposed system and finally, conclusions and future research directions are presented in Section 5.

2. Location and action tracking of construction heavy equipment

2.1. Location tracking of construction heavy equipment

The current state of practice for monitoring construction operations at jobsites is based on manually collecting and analyzing the necessary data. The manual data collection process could take place either through direct observations or by watching real time video streams. Throughout this process, the operator must manually classify and record productive versus non-productive (or idle) times. The results, along with other useful information such as maintenance notes and qualities of accomplished work, are manually entered into timesheets or other types of records. The ultimate goal of this process is calculating the percentage of equipment time spent on value-adding activities and thus, evaluating the productivity rates of the machine which will be eventually used for time and cost analysis purposes.

The manual process of monitoring construction equipment activities is labor intensive, time consuming, and error prone [6]. As a result, both practitioners and researchers in the area of construction engineering and management have recognized a substantial need for real time, accurate, and dynamic systems for activity analysis and monitoring purposes. Spatial location tracking of construction machines was the first attempt to tackle this issue. The localization and tracking of construction equipment (and other construction resources) is now a common practice within the construction industry. Currently, a number of companies (Giga Trak, Navman wireless, Fleetmatics, Linxup, Fleetilla, LiveViewGPS, etc.) offer commercial packages and services for location tracking of construction machinery. Accurately localizing and tracking construction equipment enables project managers and machine owners to better manage their assets in terms of fuel consumptions, security concerns, and assessing the performance of operators.

Several remote sensing technologies such as GPS (Global Positioning System), RFID (Radio-Frequency Identification), and Ultra-Wideband (UWB) sensors can be used for location tracking of construction machines. These technologies are all based on the time-of-arrival principle: “the propagation time of a signal can be directly converted into distance if the propagation speed in known” [7]. The most popular localization system is GPS which provides location and time information anywhere on the earth if there is an unobstructed line of sight to four or more earth orbiting satellites [8]. Although the service is available worldwide for free, attaching a high-precision GPS receiver to every worker or equipment could become costly [9]. An alternative approach is to use active RFID tags. Each tag includes an on-board power source and a locally installed antenna for the signaling electronics. These battery-powered tags can effectively communicate with a receiver for distances of up to 100 m and their unit cost is in the order of tens of dollars. Another advantage of these tags is the ability to operate without line of sight at long distances which makes them a good candidate for dynamic and crowded construction sites. The main issue, however, is the identification problem: “a reasoning mechanism is required in order to locate-tagged construction items on jobsites” [9]. UWB sensing is a special form of active RFID that can locate objects from communications between multiple tags and receivers. Tags emit signals that can be captured and processed by fixed receivers. UWB has shown several advantages in comparison to other active sensing technologies: longer range, higher measurement rates, improved measurement accuracy, and immunity to interference from rain, fog, or clutter [7]. However, the presence of a dense and expensive network of stationary receivers (i.e., measurement infrastructure) is necessary for this purpose.

The remote sensing technologies described here, facilitate 3D location tracking; however, the primary issue that still remains challenging is to recognize the equipment actions from indirect data (i.e., accurately distinguish various productive actions from non-value-adding ones). Moreover, the technologies described above are often viewed as intrusive by equipment operators as a sensor needs to be installed on each machine to record their every movement. The issue becomes even more problematic for rental equipment due to the effort and cost of repeatedly installing and removing sensing units from the equipment and, consequently, the need to continuously update the monitoring software database [10].

2.2. Action recognition of construction equipment using active sensors

Location tracking of heavy equipment provides very useful information; however, projects managers are typically more interested in monitoring various operations performed by the machine over time, instead of merely tracking locations. For this reason, researchers have recently investigated several approaches to recognizing various actions of equipment using active sensors. The application of sensors such as gyroscopes and accelerometers for activity analysis purposes has been studied in a number of other contexts, including healthcare, sport management, and computer science [11–14]. Such sensors are able to accurately provide acceleration and rotation rates in three dimensions (on the x, y, and z axes). This information is particularly helpful for activity tracking of construction equipment as the acceleration rates differ for various tasks performed by the machines. In the study conducted by Ahn et al. [15], accelerometer data was used to differentiate between three modes of an excavator operation: engine-off, idling, and working. Their study illustrated very promising results; however, the level of detail in describing activities was limited to these three categories. In addition, the necessity of mounting active sensors on machines is still a serious problem. To overcome this issue, Akhavian and Behzadan [6] investigated the use of built-in smartphone sensors to extract accelerations and rotation information. Using smartphones greatly facilitate the application of active sensors for activity recognition. However, the installation setting is not always feasible: The sensor (s) need to directly be attached to the equipment (or place in the operator's cabin). This setting is not always practical especially for some smaller equipment/construction tools such as jack hammers, concrete cutting saw, small concrete mixers and so on. For those cases, a passive or indirect sensing system, without the need for being directly attached to the machine, is more desired.

2.3. Action recognition of construction equipment using computer vision

Within the last two decades, advances in high resolution cameras and rapid increases in the computational capacity of computers has led to the emergence of an alternative solution for the activity analysis and tracking problems: computer vision. It is now possible for every individual working in a construction jobsite to carry an inexpensive camera or a smartphone so that recording a video of the scene and processing these videos using computer vision algorithms is a potentially feasible solution for activity analysis of construction machines. An off-the-shelf camera can be used to recognize the activities of construction machines and potentially rectify the need for mounting sophisticated sensors on each piece of equipment.

A basic computer vision algorithm for performance monitoring of construction heavy equipment consist of four major steps. First, the desired object (one or multiple pieces of machines in this case) should be recognized in the initial video frames using distinct features such as color values or geometric and topological features (existence of Sharpe edges, existence of convey vs. convex objects, etc.). The current 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 attain the desired performance in complex scenes by quantizing the descriptors of positive and negative samples in order to be used for training a classifier.

As the second step, the precise location of the recognized object should be tracked across the consecutive frames. The major advantage of the object tracking step is to limit the search space to certain regions in video streams and hence reduce the impact of noise caused by lateral movements of the camera and dynamic foregrounds. Contour-based tracking, kernel-based tracking, and feature matching are the three popular tracking algorithms in the literature [16].

Action recognition is the third – and probably the most challenging – step of a performance monitoring algorithm. It is worth noting that there is an important distinction between the concept of *actions* and *movements*. According to Bobick [17], for example, digging a foundation by an excavator is a general action and includes different activities such as digging, swinging, or dumping. Each activity itself consists of a sequence of various movements like raising the arm, swinging the bucket, or pushing the soil. In recent years, several researchers have conducted studies on computer vision based methods for action recognition of different types of construction equipment, especially those involved in earthmoving projects. Zou and Kim [18] proposed an image processing-based method for automatic quantification of idle times of hydraulic excavators. Their suggested approach is limited to classifying only productive versus idle states of hydraulic excavators. Rezazadeh and McCabe [19] introduced a more general framework that is able to combine object recognition, tracking, and rational events to recognize dirt loading to a dump truck by a hydraulic excavator. Golparvar-Fard et al. [20] 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. [21] combined photogrammetry and video analysis for tracking the progress of earthwork processes. They implemented photogrammetry to calculate the volume of the excavated soil in regular intervals while the video analysis was used for generating statistical data regarding the construction activities.

The forth, and the final, step of a vision based monitoring algorithm is the performance assessment step. During this step, the inferred actions are automatically logged along with temporal information. Such a chronological list of actions could be used to assess the equipment's well-being or abnormality over time and take preventive measures when necessary to minimize the operating/repair cost and the downtime. It is also a beneficial tool to measure the amount of accomplished work. Finally, the overall performance of the equipment/workers (individually or as a system) is evaluated and areas for improvement are identified from detailed field data. This enables analyzing the site as a system and optimizing its overall performance.

2.4. Gaps in knowledge: why use sound?

Even though the results of applying computer vision algorithms for recognizing various activities of construction equipment are promising; a number of important constraints still limit their practical application. Most importantly, computer vision algorithms are very sensitive to environmental factors such as occlusions, lighting and illumination conditions. It is very difficult to imagine a cluttered, busy jobsite without occlusions arising which would eventually disrupt accurate and automatic activity analysis results. Also important is that an appropriate level of illumination (not too dark, yet not in direct sunlight) is required to captured high quality video streams [22–24]. This, at the very least, makes it difficult to use computer vision for performance monitoring of the operations that are performed after sunset (which is especially popular in urban projects).

The idea of exploiting audio signals as an alternative source of information provides an alternative solution not subject to the limitations described above. There are some existing works [25–27]. Audio signals are a potentially rich source of information since construction heavy equipment often generates unique sound patterns while performing various tasks, and hence by processing this data we can potentially extract a great deal of information regarding the underlying activities. In addition, using audio signals provides a number of advantages over images/video and/or traditional location tracking devices:

1. Existence of background noise might be considered as a negative factor for audio signal processing algorithms; however, there are several methods to address this issue and to remove/reduce unwanted background noise.
2. It is well known that some construction operations are easier to recognize with sound than with other sensors. An excellent example for this case is a hydraulic rock breaker. The two major activities of a hydraulic rock breaker are a) maneuvering and positioning and b) breaking up rock in a cycle. This is an extremely challenging scenario for active spatial remote sensing and/or computer vision-based approaches as the hammer only moves slightly while performing the operation. However, the distinct sound of the hammer and engine could be used to determine whether the equipment is maneuvering, breaking rock, or idling.
3. It is also well known that a single camera can only cover a limited field of view. To fully cover a large size construction jobsite, it would be necessary to install a network of cameras in various locations, elevations, and/or angles, as shown in Fig. 1. This is not a constraint for audio signals as most of microphones are able to operate with 360° of sensitivity.
4. The audio pattern generated by each individual machine is often independent of the operator and the specific way that the task is performed. Operators can perform a task in several ways. For example, imagine a hydraulic excavator is digging a trench. This operation might include a series of movements such as diggings, rotating, swinging, and loading. These tasks could be handled in various ways such as different angles, swinging to left or right, etc. A computer vision algorithm would likely need to consider all these scenarios separately, while the audio signal analysis would always yield the same result.
5. Majority of construction operations consist of several repetitive cycles and the entire operation might take several days or weeks to complete. A hydraulic excavator may spend several hours/days to complete an excavation task. As a result, a robust recognition/tracking algorithm should also be able to perform the recognizing/tracking tasks for several hours/days. In other words, size and total amount of the collected data is an important constraint. Table 1 summarizes the size of the data and computational requirements for

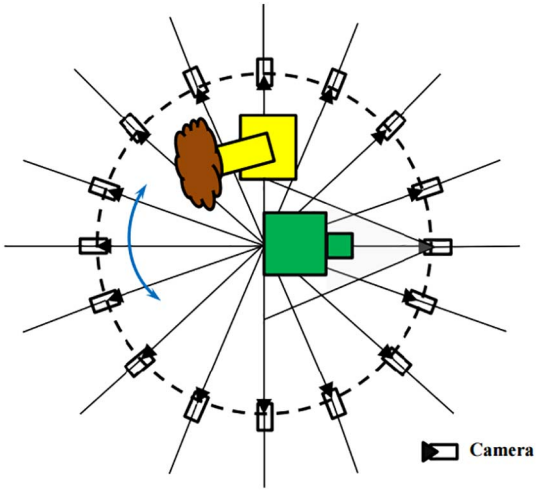

Fig. 1. Obtaining full coverage of a jobsite using computer vision techniques will require a network of cameras with overlapping fields of views, such as the example shown here.

Table 1
Approximate data rates for different sensor modalities: sound, location (NMEA format), and video.

Input type	#Sensors	Sampling rate	Resolution	Data rate
Sound	1	10 kHz	8 bit	10 kB/s
Location	10	12 kHz	800 bit	12 kB/s
Video	1	4608 kHz	8 bit	4608 kB/s

different types of recognition methods. As shown in the table, storing and processing audio files is computationally less expensive than video streams and even the data collected by using on-board location/acceleration sensors.

Compared to MEMS devices, a system based on processing audio files provide the following advantages:

1. Although MEMS devices perform very well for activity recognition of several different types of equipment (mainly heavy equipment such as earthmoving machines), the sensor(s) need to directly attach to the equipment (or place in the operator's cabin). This installation setting is not always practical especially for some smaller equipment/construction tools such as jack hammers, concrete cutting saw, and small concrete mixers and so on. For those cases, a passive or indirect sensing system, such as the proposed audio based system, is a more feasible option since the microphones could only be installed somewhere in the jobsite and in the proximity of the machine.
2. MEMS devises could be mainly used for activity recognition of single machines. For single machines, MEMS devices and the audio based system both require similar hardware settings (one sensor or microphone per machine). When it comes to modeling a construction jobsite as a whole system, and for the cases that several machines operate simultaneously, an audio based system is a better option. There is no need to attach one microphone to each single machine and it is possible to handle the job by placing a couple of microphones for the entire jobsite (depending on the size of the jobsites, number of machines, etc.) and using robust source separation algorithms and phase information of audio signals. Unlike MEMS devices, an audio based system could also consider the possible interactions between different machines as an extra source of information for analyzing the jobsite system. For example, when a loader and a truck are getting close to each other to perform the loading operation, the audio pattern/level is changing.

Finally, it is important to emphasize that this study is novel from the perspective of the selected sensing data. For the first time, this study introduces applying audio signals in Construction Engineering and Management domain.

2.5. Audio based classification models for engineering systems

Audio signal classification is a broad research area. Audio signal classification is part of the research aspect for auditory scene analysis, which is the study of how we decompose an auditory scene into its component auditory events. For any classification scheme to work on a sound containing more than one auditory event, some auditory scene analyses must be performed. Some classical applications for audio signal classification include pitch detections, automatic music transcriptions, speech and language applications, and multimedia databases. Speech has been one of the fundamental audio research topics for decades, thus a lot of audio signal classification works are conducted for speech recognition [25,26]. In recent years, speech recognition gained dramatic improvements in acoustic modeling yielded by machine learning based models [28–30]. More details about audio signal classification can be found in David [27].

Other audio based classification models are also widely used in engineering aspects. Bengtsson et al. [31] defined a three-module process to diagnose audio recordings. The sound is first recorded into a computer and then fed into the processing module. In this processing module, the recording is pre-processed to remove unwanted signals such as noise. After the pre-processing procedure, feature extraction is conducted to identify characteristic features of the audio sample. Once features for the audio sample is extracted, the condition monitor and diagnosis module can compare the observed audio sample to an existing library and provide the diagnosis for it, which can be classification or other inspections for the audio recording. The system can be used to identify faults based on sound recordings in industrial robot fault diagnosis [31].

In the manufacturing industry, ultrasonic sensors are widely used in industrial automation for presence and distance detection of solids and fluids. Ultrasonic sound is usually in the range of 20 kHz to 100 kHz, which is far beyond the human ear audible range but it can still be applied to audio based analysis systems. The classical application for ultrasonic signal processing in the industrial environment is on Condition Based Maintenance (CBM), which includes the following: bearing inspection; testing gears/gearboxes; pumps; motors; steam trap inspection; valve testing; detection/trending of cavitation; compressor valve analysis; leak detection in pressure and vacuum systems such as boilers, heat exchangers, condensers, chillers, tanks, pipes, hatches, hydraulic systems, compressed air audits, specialty gas systems and underground leaks; and testing for arcing and corona in electrical apparatus [32].

3. Research methodology

As noted above, construction heavy equipment often presents distinct sound patterns while performing certain activities, and thus it is possible to extract useful information about their activities by recording and processing audio at construction jobsites. The proposed audio-based framework begins with recording construction equipment using commercially available audio recorders. The captured audio is then fed into a signal enhancement algorithm to reduce background noise commonly found at construction jobsites. After the enhancement, the audio is then converted into a time-frequency representation using the *short-time Fourier transform (STFT)*. A *support vector machine (SVM)* is then used to classify the STFT features at each time to identify different sound patterns (corresponding to the various activities of each machine). The output of this classifier is then further processed using a window filtering approach to identify windows of time corresponding to different activities (Fig. 2). Each of these steps are described in

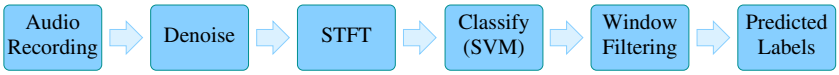

Fig. 2. Overall audio-based activity classification system.

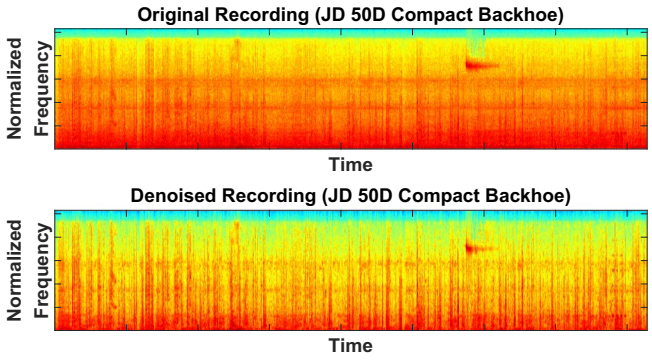

Fig. 3. Comparison between original recording and denoised recording.

further detail below.

3.1. Signal enhancement

The captured audio is assumed to contain the signals of interest along with environmental and other background noise sources. This noise will have a negative effect on identifying certain activity patterns; thus, effective signal enhancement can have a significant impact on the proposed procedure. If the level of enhancement is too low, the background noise will interfere with our signal of interest, while if the enhancement is too aggressive, our desired signal might be distorted. We apply a signal enhancement algorithm developed by Rangachari and Loizou [33] because it has been proven to be both effective and efficient. Its performance is confirmed by various listening tests that indicated significantly higher preference for this algorithm compared to the other existing noise-estimation algorithms [33]. As shown in Fig. 3, we can observe that the frequency pattern is more distinct in the denoised recording than the original recording. 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 jobsite. 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 spectrum. More details about this algorithm can be found in Rangachari and Loizou [33].

3.2. Time-frequency representation

The denoised audio signal is then converted to a time-frequency representation using the *short-time Fourier transform (STFT)*. The STFT reveals the frequency content of a signal of local windows in time and allows us to track this content as it changes over time. We use a Hanning window with size 512, a 1024-point DFT (discrete Fourier transform), and a 50% overlap (256 overlapped samples). The window size is not critical but must be long enough to provide sufficient frequency resolution; however, if it is too long, the temporal aspects of the signal are blurred. The 512 sample met this criteria so we choose it. The 50% overlapping for Hanning window has the advantage that the sum of the overlapping window functions is exactly one everywhere. The constant sum implies a proper reconstruction of the signal after inverse Fourier transform and summation of the overlapping windows. The output of the STFT consists of both magnitude and phase, but for single microphone recordings we discard the phase and consider only the magnitude.

3.3. Classification via support vector machines

To identify various activities within a captured audio file, we use techniques from machine learning to obtain models that characterize the audio for different tasks for each piece of machinery. The learning algorithm which we use in this research is the *support vector machine (SVM)* algorithm [34]. SVMs works well for high dimensional data and relatively few labelled data points [35]. The essential idea is to use training data belonging to each of the two different classes (e.g., productive activities versus non-productive activities) and to use this to build a simple decision rule for classifying future data.

More concretely, we let $(\mathbf{x}_i, y_i)$, $i = 1,2,\ldots,n$ denote the training data where $\mathbf{x}_i \in \mathbb{R}^d$ is a d-dimensional feature vector and $y_i \in \{+1, -1\}$ indicates the class of $\mathbf{x}_i$. The SVM training procedure involves solving the following optimization problem [36]:

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \quad \frac{1}{2}\left\|\mathbf{w}\right\|^2 + C\sum_{i=1}^{n}\xi_i$$

$$\text{s.t.} \quad y_i(k(\mathbf{w}, \mathbf{x}_i)+b) \geq 1-\xi_i \quad \text{for} \quad i=1, 2, \ldots, n$$
$$\xi_i \geq 0 \quad \text{for} \quad i=1, 2, \ldots, n.$$

Above, $C \geq 0$ is a tradeoff parameter that controls overfitting, and $k(\mathbf{w},\boldsymbol{\xi})$ represents a *kernel* function which mathematically captures the similarity between the vectors $\mathbf{w}$ and $\boldsymbol{\xi}$. The solution to the SVM optimization problem yields a vector $\mathbf{w}$ and an offset b from which we can make future predictions via the simple decision rule given by

$$f_{\mathbf{w},b}(\mathbf{x}) = \text{sgn}(k(\mathbf{w}, \mathbf{x}) + b).$$

In our experiments, we use the *radial basis function (RBF)* kernel. The most common choices for SVMs are linear kernel and RBF kernel. We tested both kernels and found that the performance with RBF kernel is better than linear kernel. The *RBF* kernel is given by

$$k(\mathbf{x}, \mathbf{x}') = \exp(-\gamma \|\mathbf{x}-\mathbf{x}'\|^2),$$

where $\gamma \geq 0$ is a bandwidth parameter which we must select.

To solve the SVM optimization problem, we use the LIBSVM package in MATLAB [37]. To generate training data, we extract 10 to 20 s for each activity of interest (assigned labels of ± 1 depending on the activity). The parameters C and γ are selected by considering a log-scale range from 2^{-6} to 2^5. (Note that we select the parameters independently for each machine.) We use 10-fold cross validation to select the appropriate values of C and γ. After training the SVM models for each machine, we extract other segments from the audio files (selected at random) as the testing data.

3.4. Window filtering

After training an SVM for each machine, we can produce a predicted class for any window of the audio signal. This results in each time bin being assigned a label — but in practice what is truly needed is the total time period of the different activities. The time period for a specific activity can last for seconds, but each second will have hundreds of time bins. In addition, some activities manifest only briefly in the audio signal, so a simple voting or median filtering could fail to catch such activities. Therefore, to identify the time period and to smooth out local fluctuations in the output labels, we apply a simple window filtering to the predicted labels. The window filtering algorithm starts with a small window scanning through the predicted labels in the time domain and calculating the percentage for each activity. If the percentage for a specific activity is higher than a set threshold, the window will be labelled with that activity. Thus, a threshold may be set to capture a brief

Fig. 4. The microphone array used for collecting audio files for this project (left) and placing audio recorders at a jobsite (right).

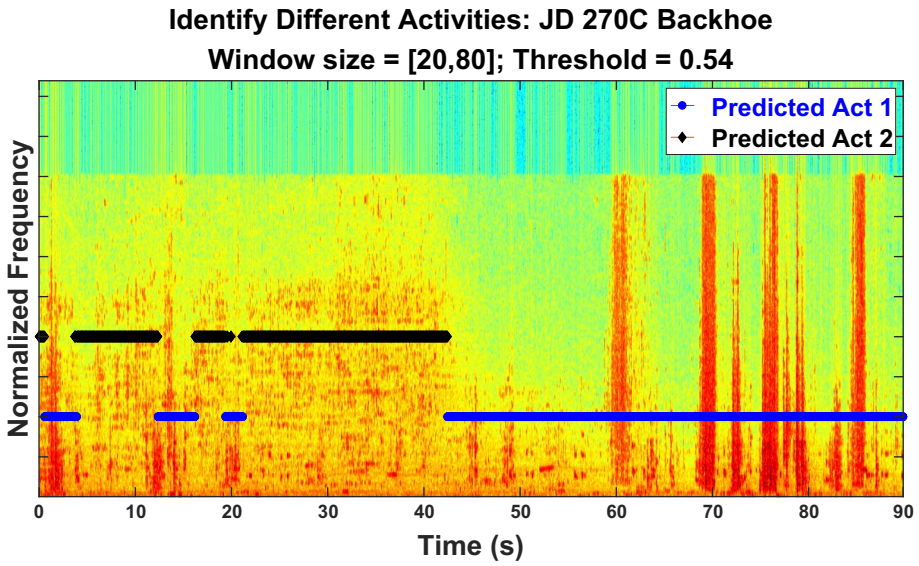

Fig. 5. Predicted Label for JD 270C Backhoe.

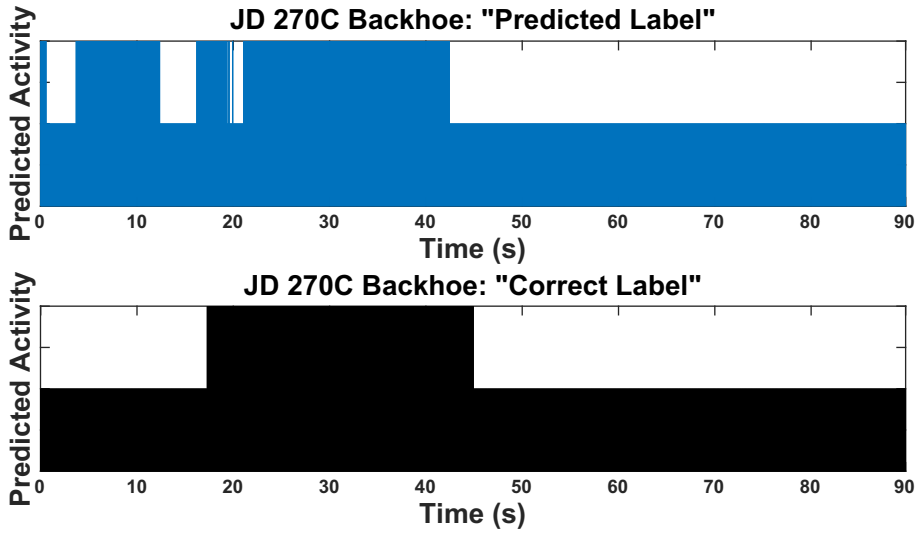

Fig. 6. Comparison between Predicted label and Correct label for JD 270C Backhoe.

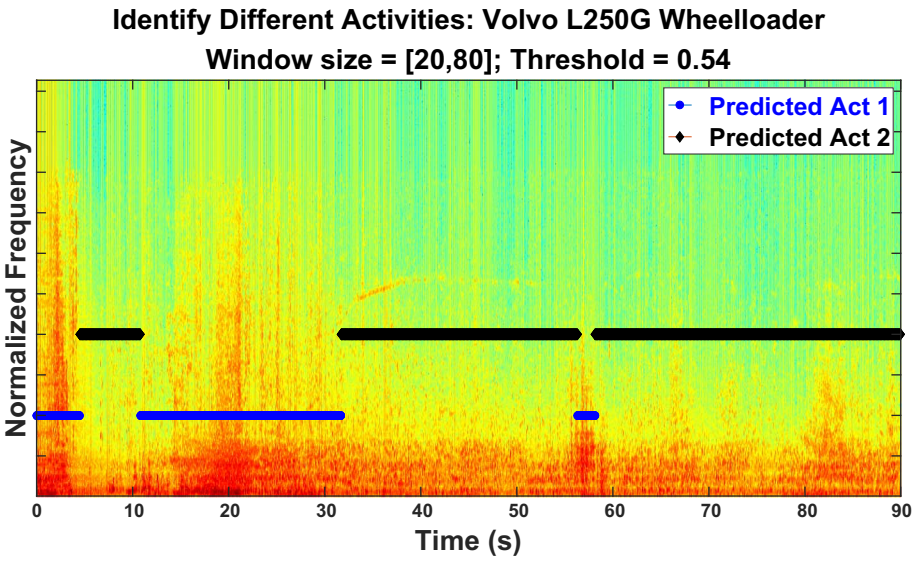

Fig. 7. Predicted Label for Volvo L250G Wheelloader.

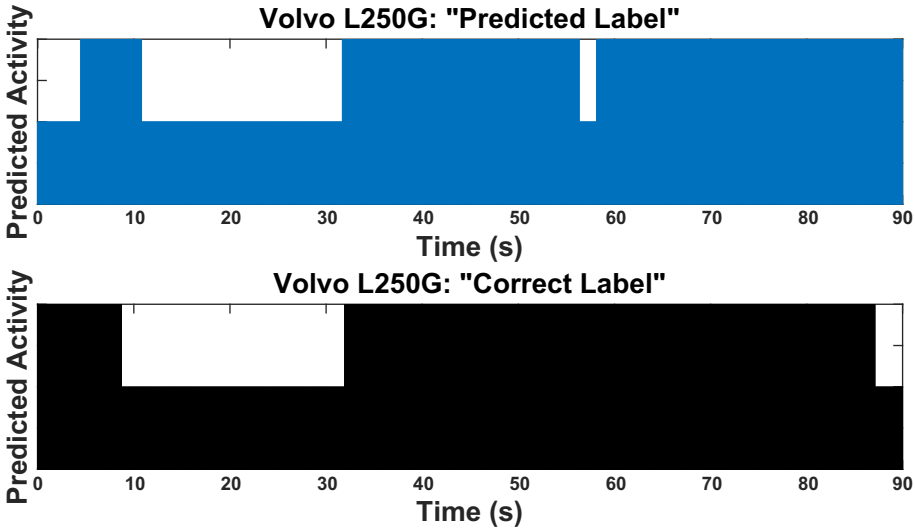

Fig. 8. Comparison between Predicted Label and Correct Label for Volvo L250G Wheelloader.

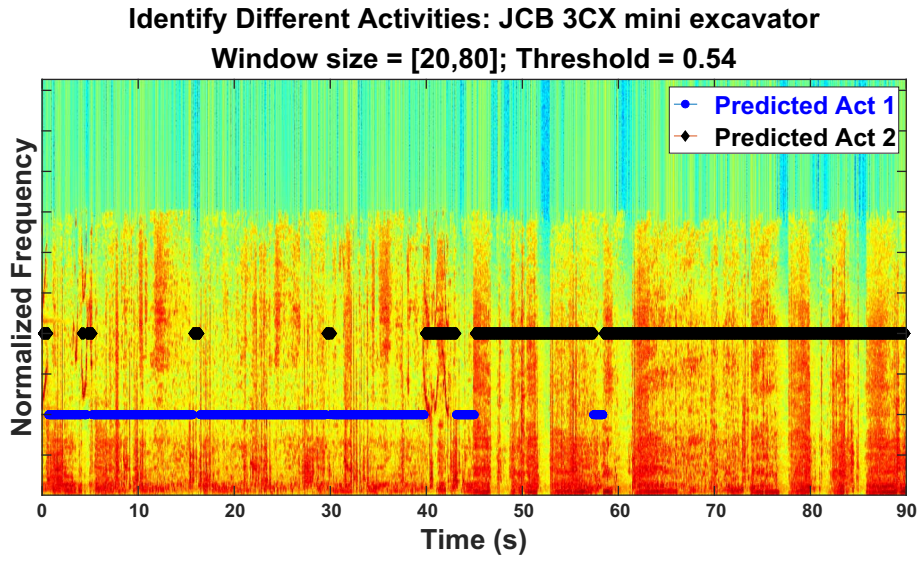

Fig. 9. Predicted Label for JCB 3CX mini excavator.

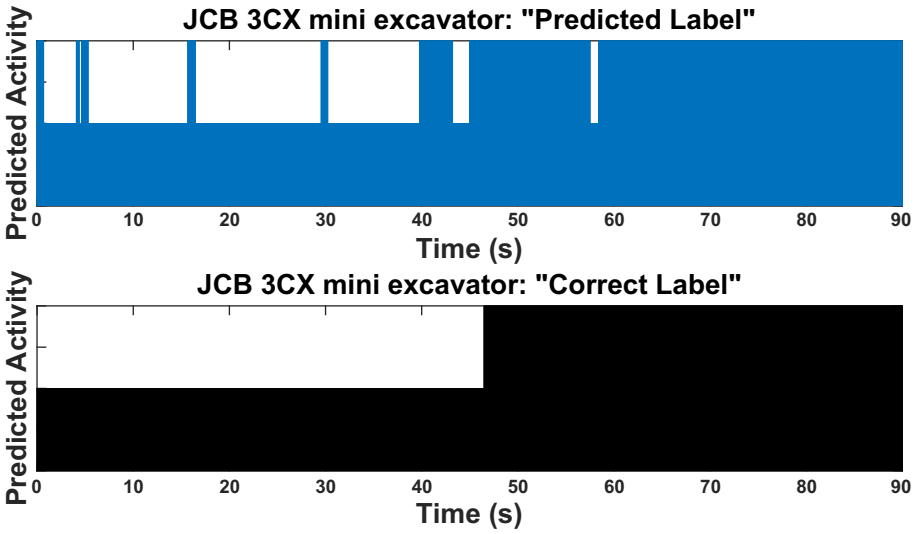

Fig. 10. Comparison between Predicted Label and Correct Label for JCB 3CX mini excavator.

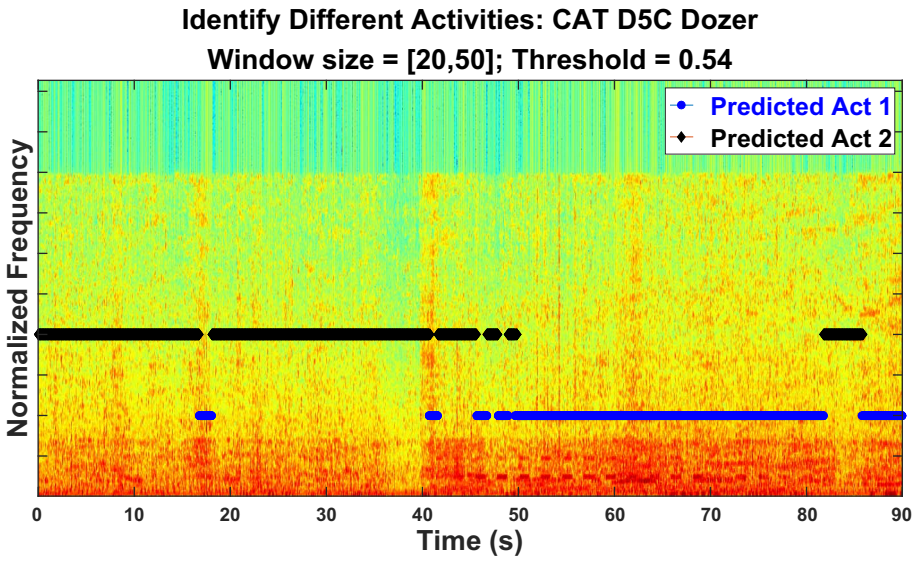

Fig. 11. Predicted Label for CAT D5C Dozer.

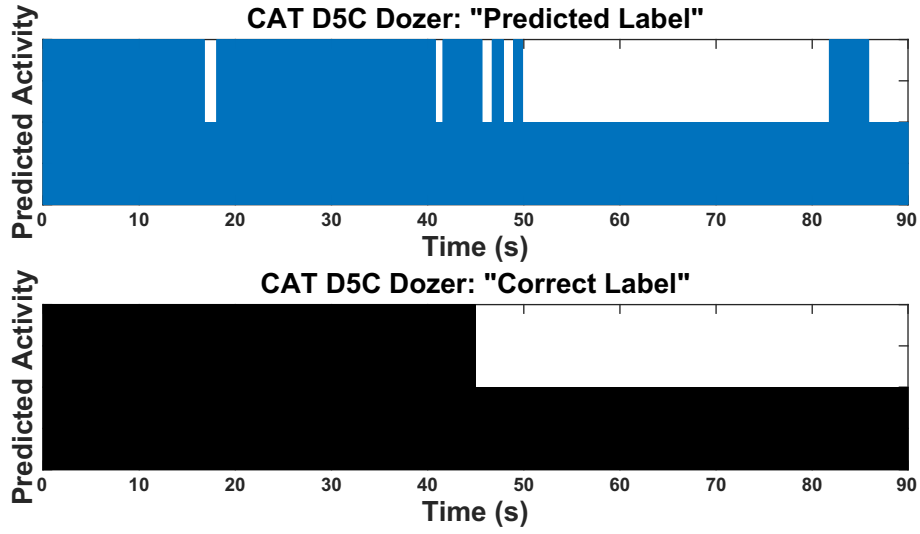

Fig. 12. Comparison between Predicted Label and Correct Label for CAT D5C Dozer.

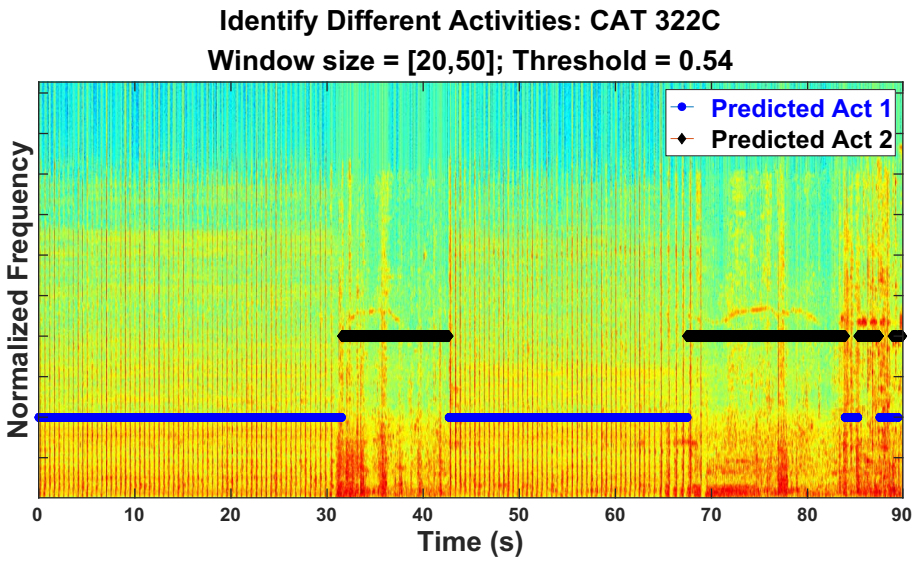

Fig. 13. Predicted Label for CAT 322C with Hydraulic Hammer.

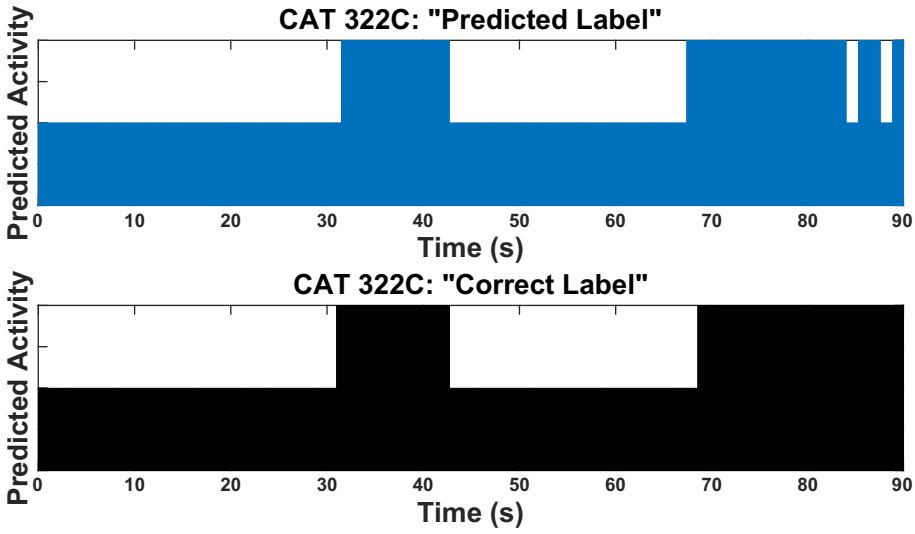

Fig. 14. Comparison between Predicted Label and Correct Label for CAT 322C with Hydraulic Hammer.

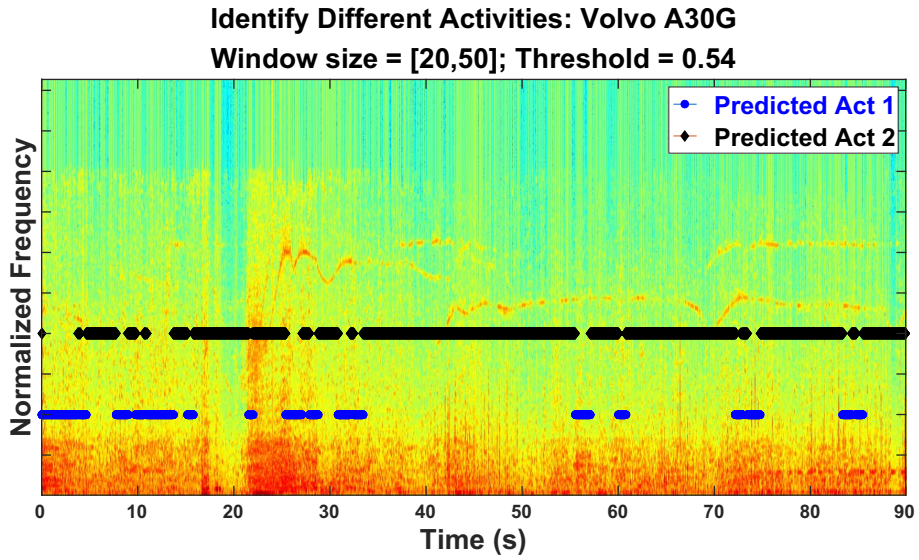

Fig. 15. Predicted Label for Volvo A30G Dumper.

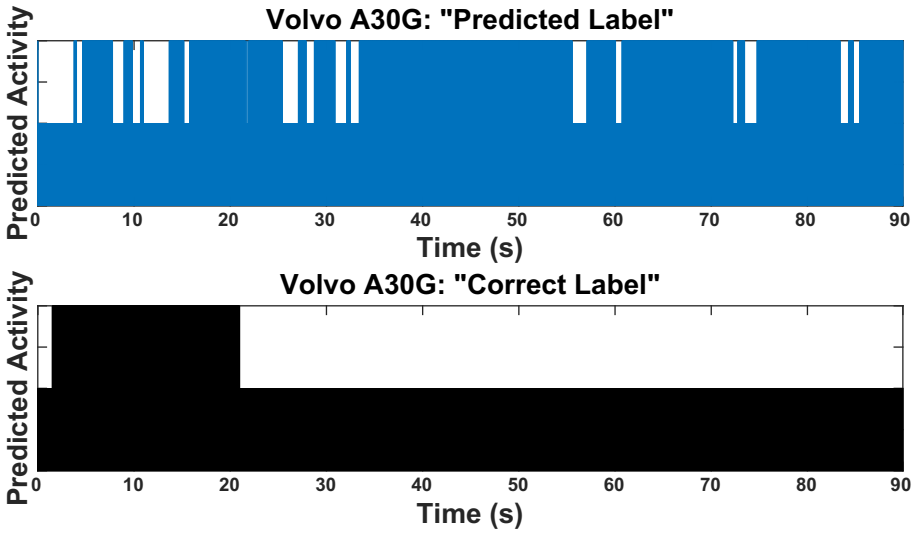

Fig. 16. Comparison between Predicted Label and Correct Label for Volvo A30G Dumper.

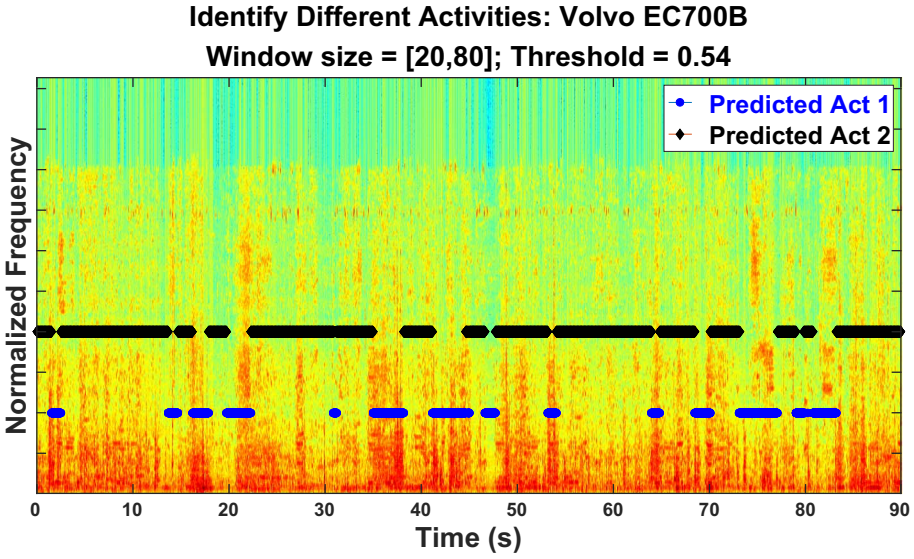

Fig. 17. Predicted Label for Volvo EC700B breaking up asphalt.

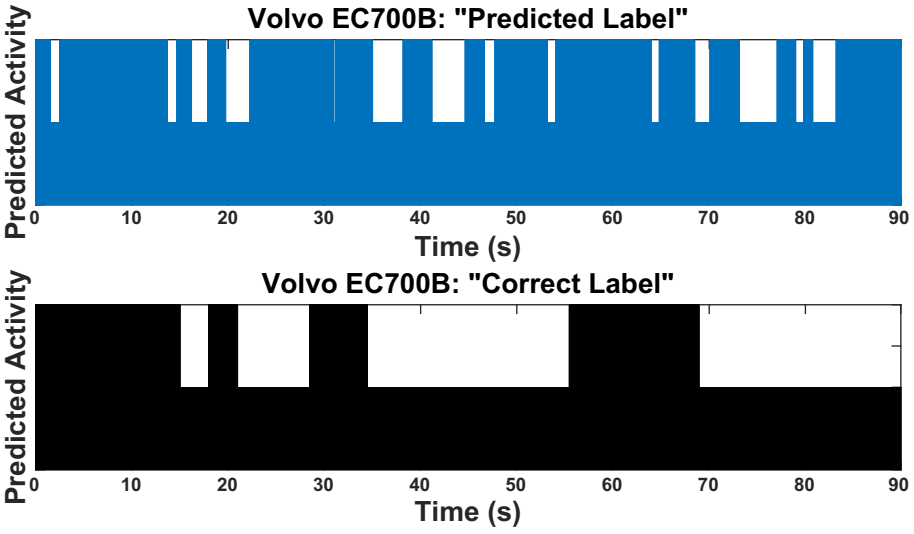

Fig. 18. Comparison between Predicted Label and Correct Label for Volvo EC700B breaking up asphalt.

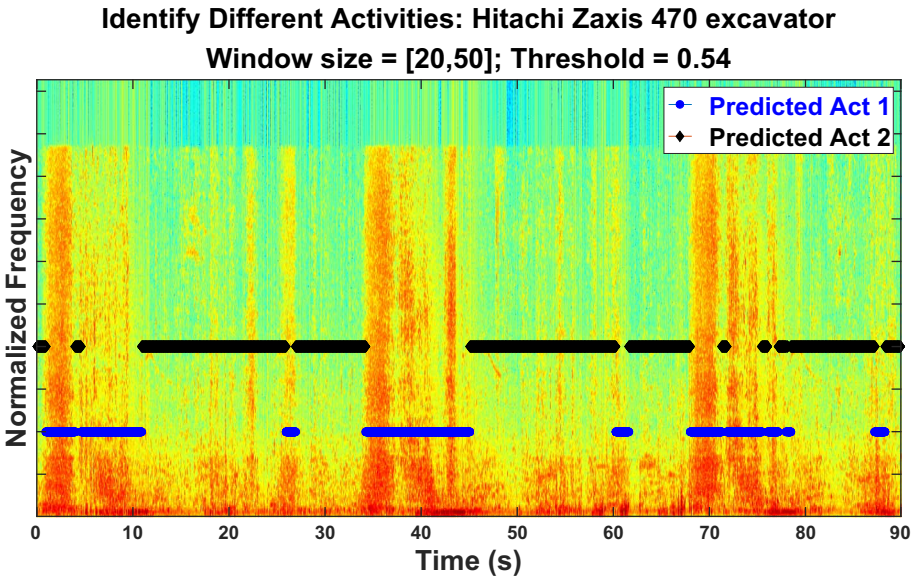

Fig. 19. Predicted Label for Hitachi Zaxis 470 excavator.

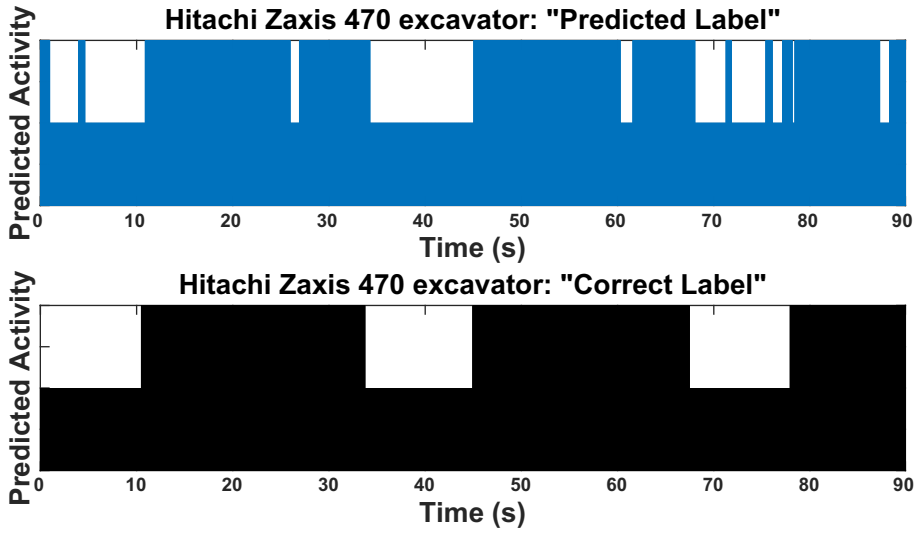

Fig. 20. Comparison between Predicted Label and Correct Label for Hitachi Zaxis 470 excavator.

Table 2
Confusion matrix. JD 270C Backhoe.

		Actual label	
		Act 1	Act 2
Predicted label	Act 1	0.8341	0.1679
	Act 2	0.1659	0.8521

impact sound or a sustained sound according to the activity. The choice of a threshold is done by experiments, and we found that 0.52 to 0.54 will be a practical choice for every case. Using different threshold will affect the performance of window filtering procedure. After scanning through all the predicted labels, the procedure is repeated using the larger window applied to the outputs of the smaller windows. The size of the window will vary in different cases, but in general the small window can be set as a quarter second and the large window can be a second or 2 s As shown in Fig. 5, "Window size [20, 80]" indicates we use a small window that contains 20 time bins and a larger window with 80 time bins. Filtering by the larger window can improve the output result smoothed by the small window. After performing the window filtering, we will make predicted labels labelled in windows rather than time bins.

3.5. Scope of research and limitations

While implementing the proposed audio-based action analysis system, the following assumptions/limitations need to be taken into account:

1. The proposed system is only applicable for construction machines that generate discrete sound patterns during their routine operations. Obviously certain types of equipment, such as tower cranes and graders, do not fall into this category and the system is not able to analyze their activities.
2. At this stage, the system is capable of classifying operations into two categories: active or productive (known as major activity within the paper) and passive or non-productive (known as minor activity). Minor activities include items such as swinging, maneuvering and moving. In order to fully analyze the operations, minor activities should also be classified into sub-categories. The current system is

Table 3
Confusion matrix. Volvo L250G Wheelloader.

		Actual label	
		Act 1	Act 2
Predicted label	Act 1	0.8003	0.0981
	Act 2	0.1997	0.9019

Table 4
Confusion matrix. JCB 3CX mini excavator.

		Actual label	
		Act 1	Act 2
Predicted label	Act 1	0.8326	0.0218
	Act 2	0.1674	0.9782

Table 5
Confusion matrix. CAT D5C Dozer.

		Actual label	
		Act 1	Act 2
Predicted label	Act 1	0.8490	0.0477
	Act 2	0.1510	0.9523

Table 6
Confusion matrix. CAT 322C with Hydraulic Hammer.

		Actual label	
		Act 1	Act 2
Predicted label	Act 1	0.9807	0.0996
	Act 2	0.0193	0.9004

Table 7
Confusion matrix. Volvo A30G Dumper.

		Actual label	
		Act 1	Act 2
Predicted label	Act 1	0.1775	0.4322
	Act 2	0.8225	0.5668

Table 8
Confusion matrix. Volvo EC700B breaking up asphalt.

		Actual label	
		Act 1	Act 2
Predicted label	Act 1	0.3817	0.1277
	Act 2	0.6183	0.8723

Table 9
Confusion matrix. Hitachi Zaxis 470 excavator.

		Actual label	
		Act 1	Act 2
Predicted label	Act 1	0.8569	0.0738
	Act 2	0.1431	0.9262

not able to differentiate between those sub-categories and a further study regarding this issue will be part of the future research plan.
3. Although we believe that sound recording devices (microphones) are more robust against environmental factors (compared to cameras), a number of factors still can affect the performance of the system. Existence of sound barriers or objects such as massive RC blocks might affect the audio recording process. Studying the impacts of those elements is not within the scope of this work. In this research, we try to record audio files in open construction jobsite areas with minimal acoustical barriers.
4. The focus of this research is on activity analysis of single machines. To achieve this goal, smaller jobsites with only one active machine, or single machines operating relatively far from other machines generating noise are ideal. In the future, the authors will study the case of activity analysis of several machines operating simultaneously and in a noisy construction jobsite.
5. In this study, microphones are installed in the jobsite and in the proximity of the equipment. The microphone is assumed to be approximately "close" enough to the target machine. A more detailed study on the impacts of hardware settings and layout (location, distance, orientation, and number of microphones) is not within the scope of this research.

4. Experimental setup and results

4.1. Experimental setup

In order to evaluate the performance of the proposed system, 11 different pieces of construction machines operating at various jobsites has been selected as case studies: 1) JCB 3CX mini excavator, 2) JD 270C Backhoe, 3) Volvo L250G Wheel Loader, 4) CAT D5C Dozer, 5) CAT 322C with Hydraulic Hammer, 6) Volvo A30G Dumper, 7) Volvo EC700B breaking up asphalt, 8) Hitachi Zaxis 470, 9) CAT 320E Excavator, 10) Ingersoll Rand SD-25F Compactor, and 11) John Deere 700J Dozer. We present our experiment results for first eight machines in this paper.

Each piece of machine was carefully monitored and the generated sounds while performing routine tasks were captured using a commercially available recorder (Fig. 4: Tascam DR-05 2 GB; costs: around $100). In parallel to recording generated sound patterns, a smart phone was used to video tape the entire scene. The captured video files will be used later to manually label the audio file and classify different activities and thus, generate the validation benchmark (or ground truth data). Next, each audio file was manually labelled based on various activities took place during the recording time. The label here will be used as correct label in the experimental results. 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 file 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 file was then sent through the audio processing pipeline and divided into activities 1 and 2. Finally, the performance of the algorithm for each case study has been compared to manually labelled files. The comparison results are depicted in the following tables and figures.

4.2. Experimental results

In the tables and figures, the label "Act 1" represents the major activity, while "Act 2" illustrates the minor activities. Figs. 5–20 illustrate the predicted activities based on the STFT spectrogram and the comparison between predicted and actual labels. To quantify the accuracy of the proposed system, the confusion matrices for each equipment were formed (Tables 2–9). The confusion matrix shows how the predicted labels match with the actual labels. The figures illustrate that our proposed SVM-based approach can detect specific audio patterns in

the STFT domain. For example, the blue line in Fig. 5 indicates label for "Act 1" predicted by SVMs, which is the major activity for JD 270C Backhoe, while the black line indicates the predicted minor activity. We plotted two lines with different height only to make different predicted activities labels visually separated. It clearly shows that "Act 1" and "Act 2" present a large difference in the STFT domain. The similar results can be seen in the figures for other construction equipment, such as Figs. 7, 13, and 19. Another way to visually evaluate the performance of the system is throughout the comparison charts as shown in Figs. 6, 8, 10, 12, 14, 16, 18, and 20. Tables 2–9 indicate that the performance of the proposed system for automatically recognizing activities of single machines is very promising. Generally, the proposed system can have over 80% even 85% accuracy identifying different activities. For JCB 3CX mini excavator and CAT 322C with hydraulic hammer, as shown in Tables 4 and 6, "Act 1" and "Act 2" have strongly different patterns in their STFTs, thus the identification accuracy can be over 90%.

The proposed audio-based framework provides an efficient way to identify different activities for construction equipments. The machine learning algorithm needs only a few seconds of recording to obtain sufficient data to construct a model for learning audio patterns. We do not need a large database compared to neural network based machine learning method to implement the identification. Also, the machine learning model can learn the audio patterns for each activity without manual analysis of the audio recording. Thus, it will not be highly dependent on the database of pure activity recordings. We only need relatively small amount of recordings to construct the identifying system.

5. Discussion and future work

Recognizing, tracking and monitoring of construction equipment activities is the first step for productivity assessment of construction jobsites. Despite the significance of this task, there is no automated system for efficiently recognizing and monitoring operations of construction machines at jobsites. To tackle this issue, the authors proposed an innovative audio-based system for activity analysis of single machines performing various tasks. As illustrated in the previous sections, collecting and processing audio files are computationally efficient, non-intrusive, and inexpensive. The presented system was implemented and evaluated using several datasets from various construction jobsites and the obtained results are very satisfying. In addition, the following lessons have been learned from this study:

1. The presented framework is novel as it is the first research plan that attempts to introduce audio as an alternative source of information to recognize and log construction activities at a jobsite.
2. The results of implementing the suggested framework for activity analysis of construction heavy equipment are very promising. In particular, the implemented signal enhancement algorithm was able to efficiently separate the major audio signal and remove the background noise.
3. Although the system works acceptably well for the majority of construction heavy equipment tested, there are still a few machines that do not generate distinct sound patterns while performing routine tasks. Graders are examples of this category of machines.
4. For specific types of equipment, including Volvo A30G Dumper (Fig. 15) and Volvo EC700B breaking up asphalt (Fig. 17), the major activity does not have a dominant and specific pattern in STFT domain, thus the performance of identifying "Act 1" and "Act 2" is low. To achieve better identification performance – or, to avoid the bad result presented in Volvo A30G (Figs. 15 and 16 and Table 7) – we believe that we need to also include phase information (which was not used in this system). This enhanced algorithm will be left for future work.
5. Assuming that the same models of equipment, generates very similar sound patterns during similar operations, we believe that the results could be generalized and the system works for the other same model machines. At this stage, we do not have enough datasets to prove this claim though and this case will be explored in more details in the future.

Future extension of this research study will include activity analysis of multiple machines using single or multiple microphones, and eventually, acoustical modeling of the entire construction jobsite. The authors also intend to use the audio signal processing system parallel to other activity analysis techniques (e.g., computer vision) to enhance the generated results and cover a broader range of applications. A detailed, quantitative comparison study between the proposed audio-based model and other available methods (MEMS sensors and computer vision techniques) is another item needs to be considered as part of the future research plan.

Acknowledgments

This research project 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 manuscript are those of the authors and do not reflect the views of the funding agency. The authors also appreciate the assistance of Mr. Chris Andres Sabillon Albarran, graduate student at Georgia Southern University, with data collection and audio recordings.

References

[1] L.C. Institute, What is Lean Construction? (2004) http://www.leanuk.leanconstruction.org/whatis.htm.
[2] C. Ahn, S. Lee, F. Peña-Mora, Monitoring system for operational efficiency and environmental performance of construction operations using vibration signal analysis, Construction Research Congress, 2012, pp. 1879–1888, http://dx.doi.org/10.1061/9780784412329.189.
[3] T. Cheng, J. Teizer, Real-time resource location data collection and visualization technology for construction safety and activity monitoring applications, Autom. Constr. 34 (2013) 3–15.
[4] R. Akhavian, A. Behzadan, Remote monitoring of dynamic construction processes using automated equipment tracking, Construction Research Congress, 2012, pp. 1360–1369, http://dx.doi.org/10.1061/9780784412329.137.
[5] R. Akhavian, A. Behzadan, Knowledge-based simulation modeling of construction fleet operations using multimodal-process data mining, J. Constr. Eng. Manag. 139 (11) (2013).
[6] R. Akhavian, A. Behzadan, Construction activity recognition for simulation input modeling using machine learning classifiers, Proceedings of the Winter Simulation Conference (WSC 2014), 2014, pp. 3296–3307.
[7] T. Cheng, M. Venugopal, J. Teizer, P. Vela, Performance evaluation of ultra-wideband technology for construction resource location tracking in harsh environments, Autom. Constr. 20 (2011) 1173–1184.
[8] N. Pradhananga, J. Teizer, Automatic spatio-temporal analysis of construction site equipment operations using GPS data, Autom. Constr. 29 (2013) 107–122.
[9] D. Torrent, C. Caldas, Methodology for automating the identification and localization of construction components on industrial projects, J. Comput. Civ. Eng. 23 (1) (2011) 1–13.
[10] E. Rezazadeh Azar, S. Dickinson, B. McCabe, Server–customer interaction tracker: a computer vision-based system to estimate dirt loading cycles, J. Constr. Eng. Manag. 139 (7) (2013) 785–794.
[11] L. Bao, S. Intille, Activity Recognition from User-Annotated Acceleration Data, in: A. Ferscha, F. Mattern (Eds.), In Pervasive Computing, Springer, Berlin Heidelberg, 2004, pp. 1–17.
[12] N. Ravi, N. Dandekar, P. Mysore, M.L. Littman, Activity Recognition from Accelerometer Data, Proceeding of the AAAI, 2005, pp. 1541–1546.
[13] Q. Li, J.A. Stankovic, M.A. Hanson, A.T. Barth, J. Lach, G. Zhou, Accurate, Fast Fall Detection Using Gyroscopes and Accelerometer-Derived Posture Information, Proceeding of the Sixth International Workshop on Wearable and Implantable Body Sensor Networks, IEEE, Piscataway, New Jersey, 2009, pp. 138–143.
[14] K. Altun, B. Barshan, Human Activity Recognition Using Inertial/Magnetic Sensor Units, in: A. Salah, T. Gevers, N. Sebe, A. Vinciarelli (Eds.), In Human Behavior Understanding, Springer, Berlin Heidelberg, 2010.
[15] C.R. Ahn, S. Lee, F. Peña-Mora, Application of low-cost accelerometers for measuring the operational efficiency of a construction equipment fleet, J. Comput. Civ. Eng. 04014042 (2013).
[16] M. Park, A. Makhmalbaf, I. Brilakis, Comparative study of vision tracking methods for tracking of construction site resources, Autom. Constr. 20 (7) (2011) 905–915.
[17] A. Bobick, Movement, activity and action: the role of knowledge in the perception of motion, Philos. Trans. R. Soc. Lond. B Biol. Sci. 352 (1358) (1997) 1257–1265.

[18] J. Zou, H. Kim, Using hue, saturation, and value color space for hydraulic excavator idle time analysis, J. Comput. Civ. Eng. 21 (4) (2007) 238–246.
[19] E. Rezazadeh Azar, B. McCabe, Part based model and spatial–temporal reasoning to recognize hydraulic excavators in construction images and videos, Autom. Constr. 24 (2012) 194202.
[20] M. Golparvar-Fard, A. Heydarian, J. Niebles, Vision-based action recognition of earthmoving equipment using spatio-temporal features and support vector machine classifiers, Adv. Eng. Inform. 27 (4) (2013) 652663.
[21] M. Bugler, G. Ogunmakin, J. Teizer, P. Vela, A. Borrmann, A Comprehensive Methodology for Vision-Based Progress and Activity Estimation of Excavation Processes for Productivity Assessment, Technische Universität München, 2014, http://www.cms.bgu.tum.de/publications/2014_Buegler_EG-ICE.pdf.
[22] I. Brilakis, H. Fathi, A. Rashidi, Progressive 3D reconstruction of infrastructure with videogrammetry, Autom. Constr. 20 (7) (2011) 884895.
[23] M. Golparvar-Fard, J. Bohn, J. Teizer, S. Savarese, F. Pea-Mora, Evaluation of image-based modeling and laser scanning accuracy for emerging automated performance monitoring techniques, Autom. Constr. 20 (8) (2011) 1143–1155.
[24] F. Dai, A. Rashidi, I. Brilakis, P. Vela, Comparison of image-based and time-of-flight-based technologies for three-dimensional reconstruction of infrastructure, J. Constr. Eng. Manag. 139 (1) (2013) 69–79.
[25] G. Guodong, Stan Z. Li, Content-based audio classification and retrieval by support vector machines, IEEE Trans. Neural Netw. 14 (1) (2003).
[26] L. Lie, Z. Hong-Jiang, J. Hao, Content analysis for audio classification and segmentation, IEEE Trans. Speech Audio Process. 10 (7) (2002).
[27] G. David, Audio Signal Classification: History and Current Techniques, Technical Report TR-CS 2003-07, (2003).
[28] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Rahman Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, B. Kingsbury, Deep neural networks for acoustic modeling in speech recognition, IEEE Signal Process. Mag. 29 (6) (2012) 82–97.
[29] A. Graves, A.-R. Mohamed, G.E. Hinton, Speech Recognition with Deep Recurrent Neural Networks, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2013, pp. 6645–6649.
[30] W. Chan, N. Jaitly, Q. Le, O. Vinyals, Listen, Attend and Spell: A Neural Network for Large Vocabulary Conversational Speech Recognition, 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2016, pp. 4960–4964.
[31] M. Bengtsson, E. Olsson, P. Funk, M. Jackson, Technical Design of Condition Based Maintenance System. A Case Study Using Sound Analysis and Case-Based Reasoning. (2004).
[32] P. NaikR., Ultrasonic: A New Method for Condition Monitoring, (2009).
[33] S. Rangachari, P. Loizou, A noise estimation algorithm for highly nonstationary environments, Speech Comm. 28 (2006) 220–231.
[34] A. Rashidi, M.H. Sigari, M. Maghiar, D. Citrin, An analogy between various machine-learning techniques for detecting construction materials in digital images, KSCE J. Civ. Eng. 20 (2016) 1178–1188.
[35] A. Ben-Hur, J. Weston, A users guide to support vector machines, Data Min. Techniques Life Sci. (2010) 223–239.
[36] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297.
[37] C.C. Chang, C.J. Lin, LIBSVM: A Library for Support Vector Machines, (2001) Software Available at http://www.csie.ntu.edu.tw/cjlin/libsvm.