Proceedings of the 2022 Winter Simulation Conference
B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C.G. Corlu, L.H. Lee, E.P. Chew, T. Roeder, and P. Lendermann, eds.

CONSTRUCTING AN AUDIO DATASET OF CONSTRUCTION EQUIPMENT FROM ONLINE SOURCES FOR AUDIO-BASED RECOGNITION

Gilsu Jeong
Changbum R. Ahn
Moonseo Park

Department of Architecture and Architectural Engineering
Seoul National University
Gwanak-ro 1, Gwanak-gu
SEOUL, 08826, SOUTH KOREA

ABSTRACT

Monitoring equipment and constructing activity data in construction sites are essential to obtain reliable decision-making through simulation models. The audio-based equipment monitoring could provide critical information about the work process and site conditions. Although a large-scale dataset is essential for audio-based activity recognition, it is time consuming and labor intensive to collect data on site. Therefore, this study proposes a framework for constructing an audio dataset of equipment from online sources. The framework involved selecting appropriate audio using machine learning algorithms, audio denoising, and audio separation models. The validity of the constructed dataset was examined with six classifiers and compared with the benchmark models constructed using real-world equipment audio. The classification results provided 64%–93% accuracy, which demonstrates that the constructed dataset using the proposed framework is effective in recognizing real-world sounds. The outcomes are anticipated to improve audio-based activity recognition processes, potentially helping to monitor equipment productivity.

1 INTRODUCTION

As construction projects increasingly become bigger and more complicated, simulation models are used to make decisions for various resource operations that are difficult and time-consuming to evaluate on the real-world construction site (Matrinez et al. 1999; Hajjar et al. 1999). To provide the greatest results from the simulation model, it should accurately reflect the real information of construction projects by incorporating data that describes resources and processes (Gong et al. 2010; Akhavian et al. 2015). Unlike previous studies that have conducted simulation for construction resources using expert judgment, subjective assumptions and parameters from past projects, data-driven simulation models have been considered important to ensure the reliability of critical decisions such as equipment operation time (AbouRizk 2010; Akhavian et al. 2013). For high fidelity simulation, monitoring equipment and constructing fine-grained activity data obtained as a result of activity recognition in construction sites are essential.

In recent years, many researchers have tried to apply automated monitoring methods to reduce the time spent on monitoring and to improve accuracy and efficiency (Ahn et al. 2012; Fang et al. 2018; Xiao et al. 2021). Various data, including visual, kinematic, and audio (Sherafat et al. 2021), have been used in automated monitoring methods. Since environmental variables like lighting, dust, snow, and rain can affect visual data, there should be no obstructions between the camera and the resources. (Lee et al. 2020). Also, it needs light to record images and there are privacy issues when utilized on the jobsite (Cheng et al. 2017;

Kim et al. 2017; Angah et al. 2020). Kinematic data requires many devices, which makes this method costly, and signals could be interrupted or disrupted when multiple tags are present in construction site. Especially, installation of sensor devices on all deformable components of equipment and workers is difficult (Alshibani et al. 2016; Kim et al. 2021).

On the other hand, a construction site is an environment in which various sounds are generated, and in particular, the sound of construction equipment is the main audio source. These audio data can provide useful information indicating type of input equipment, work process, and site condition (Lee et al. 2020). Therefore, a deeper understanding and analysis of audio data for construction equipment could provide insightful field information for project managers and participants to intuitively assess site conditions and remotely manage site processes. To achieve the potential of the equipment audio data for audio-based activity recognition, various studies have used microphones for collecting audio data from construction sites and then conducted machine or deep learning techniques after data preprocessing (Cheng et al. 2016; Cho et al. 2017; Scarpiniti et al. 2021). To establish an accurate monitoring system with such a process, sufficient audio datasets are indispensable.

In order to obtain the desired audio dataset, the current state is based on manually recording and collecting audio data at the construction site, which is inefficient and time-consuming (Cheng et al. 2017). In the case of image and video data of construction equipment, there are online resources that can be applied to deep learning models. However audio data of equipment is rarely found and there is no data refined for use in audio-based activity recognition research. Recent papers have demonstrated the possibility of collecting high-quality and large-scale environmental and speech datasets in an automated manner by using object detection algorithms (Nagrani et al. 2017; Chen et al. 2020). Likewise, constructing audio datasets of construction equipment from online sources is promising and would provide an unconstrained dataset, but there exist several challenges: (1) such data from online sources often includes unwanted sounds (e.g., noise, speech, and music) in addition to the desired equipment sound; (2) audio of online videos often does not match with the visual scenes of those videos; and (3) they often include sounds from multiple pieces of equipment at the same time.

The primary objective of this study is to propose a framework for constructing an audio dataset of construction equipment from online videos. The framework includes collecting online videos of construction equipment, selecting positive clips that satisfy desired conditions, enhancing audio signals by denoising, and separation for audio-based activity recognition. This study focused on constructing audio datasets for four types of equipment (excavator, compactor, dozer, and breaker). Finally, we examined whether the datasets constructed from the proposed framework allow for construction of effective audio-based activity recognition models by training various classifiers using the constructed datasets and testing those classifiers on a real-world audio dataset.

2 RELATED WORK

2.1 Audio-Based Activity Recognition of Construction Equipment

There are various types of equipment and noise sources that produce different sounds at construction sites. By analyzing these audio signals, it is possible to monitor the progress of equipment and measure productivity (Sabillon et al. 2018). To achieve the potential of audio-based equipment activity recognition, various studies have been conducted to develop monitoring systems with the audio data. Those studies have used microphones to capture audio signal, and conducted signal processing and feature extracting to enhance the data (Sherafat et al. 2020). Also, machine or deep learning techniques are adopted to detect and classify activity types of equipment. There are studies that present attempts to use audio as a method for identifying activity of equipment and develop support vector machine (SVM) and hidden Markov models (HMM) to classify activities of equipment (Cheng et al. 2017; Sabillon et al. 2018; Zhang et al. 2018, Sherafat et al. 2022). Various deep learning (DL) techniques such as convolutional neural network (CNN), deep recurrent neural network (DRNN), long-short term memory (LSTM), and deep belief network

(DBN) are implemented to improve the performance of audio-based equipment recognition systems (Zhang et al. 2018; Scarpiniti et al. 2020; Lee et al. 2020; Scarpiniti et al. 2021).

In order to construct these models, it is a priority to prepare sufficient audio datasets for use in research. Most of the previous studies have used datasets recorded on the jobsite, and consequently, these datasets are not large enough and are limited to specific sites and equipment types/models. Also, since the aforementioned papers used audio datasets recorded in a restricted environment, the mixed audio data is not sufficiently considered when multiple types of equipment are operated simultaneously. Thus, motivated by these challenges, this study suggests a framework to construct a large-scale audio dataset of equipment from online sources and provides a comprehensive solution to cover the entire range of equipment.

2.2 Constructing Audio Datasets from Online Sources

While several studies constructed image datasets of construction equipment from online sources (Tajeen et al. 2014; Xiao et al. 2021), there has been no attempt to build an audio dataset of construction equipment from online sources. On the other hand, in other fields such as speech recognition, various audio datasets have been constructed from online videos on YouTube (Nagrani et al. 2017; Gemmeke et al. 2017; Chaudhuri et al. 2018). In addition, recent papers have proposed how to build a dataset automatically from online videos using computer vision algorithms for visual verification (Chen et al. 2020; Watanabe et al. 2020; Lee et al. 2021). These approaches could be used to construct an audio dataset of equipment, but several improvements are necessary. Since the previous methods of building datasets automatically focused on visual information, the dataset could not be guaranteed to contain the desired audio signal. In a study for curating datasets for speaker recognition, other vision-based approaches such as face tracking and mouth motion detecting are applied to determine which face belongs to the speaker (Nagrani et al. 2017). However, the models are trained for speech, so it is challenging to apply to construction equipment. In addition, the audio signal extracted from the online video may contain irrelevant audio because it is recorded from unattributable media. Although the unnecessary sound should be excluded, existing research has rarely suggested ways to improve audio quality. Therefore, it is necessary to confirm the correspondence between visual scene and audio of the equipment and eliminate unnecessary audio data to improve audio quality.

3 METHODOLOGY

The proposed framework to construct audio datasets of construction equipment from online videos in an automated manner included four main steps: collecting online videos, selecting positive clips through computer vision and audio classification algorithms, enhancing audio signals by denoising, and separating audio signals. Figure 1 shows the research framework for constructing audio datasets. The first step was to collect appropriate videos of equipment to be used in this study from online video sources and aimed to reduce irrelevant videos by using detailed search query. The second step was to select the detected section in which equipment appears visually that contained audio signals of equipment by applying a computer vision-based object detection model and the CNN-based audio classification algorithm to the videos. The third step aimed to enhance the audio signal by applying a denoising model to eliminate noise contained in the audio signal extracted from the selected section. Last, if two or more equipment sounds were mixed, audio separation was conducted to separate mixed audio signals to make an effective dataset for deeper audio-based activity recognition. By following these processes, audio datasets of construction equipment can be developed, including annotation.

3.1 Collecting Online Videos

The first step to develop construction equipment audio datasets extracted from online videos is to determine a tentative list of equipment classes for the dataset and to download videos through web crawling. Four types of construction equipment, namely excavator, compactor, dozer, and breaker, were chosen in this study. The selected types of equipment are considered to be common operating equipment on construction

sites. The videos were searched using the selected equipment category as a search query. In order to not search irrelevant videos such as animation, toys, images, and duplicate videos, the search query needed to be specific and include information such as equipment name, manufacturing brand, and even model name. For example, using *excavator caterpillar 311* as a search query rather than just *excavator* is more effective to obtain videos of the actual construction equipment desired in this study. Under this approach, videos of construction equipment were downloaded using web crawler. Audio signals were extracted from the downloaded videos, and each video and audio file were annotated based on the search query.

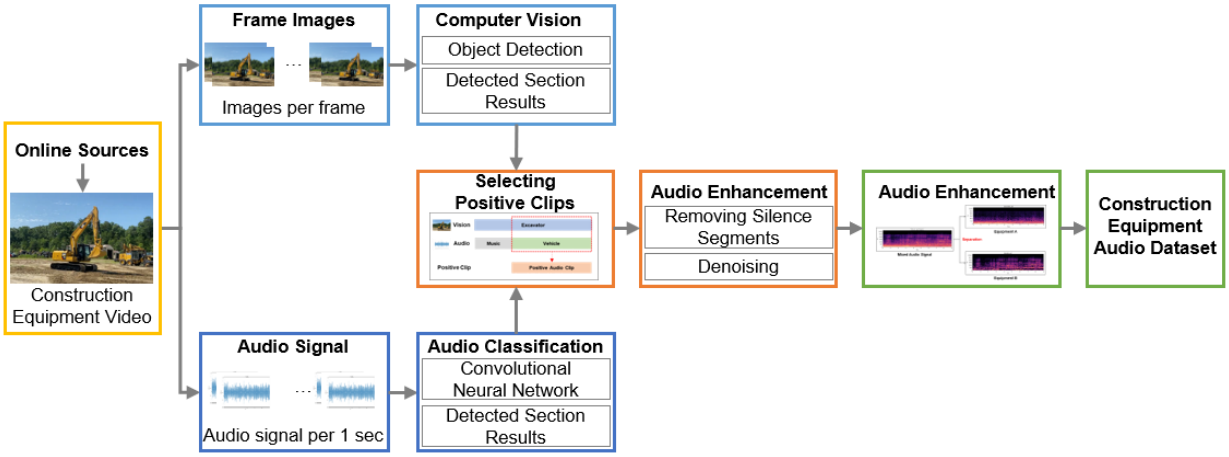

Figure 1: Research framework to construct equipment audio datasets.

3.2 Selecting Positive Clips

The purpose of this step was to select positive video and audio clips of construction equipment from among the downloaded videos. In detail, the first step was to find a section in which construction equipment is detected in the entire video through a computer vision algorithm. The next step was to find a section in which audio signals of the construction equipment are detected through audio classification algorithms in the same videos. Through these two steps, the final candidate clips were selected. Figure 2 shows the task definition to select positive clip sections by the results of object detection and audio classification. As shown in the figure, it is possible to derive the result of the section in which the excavator is detected through the object detection model. In addition, the audio classification model can identify which audio is included, and among them, the section result containing the audio signal of the explorer can be derived. Based on the results derived through the two models, the overlapping section is selected as a positive clip section.

The object detection algorithm was used to find the section in which the desired equipment appears in the video, and the YOLOv5(You Only Look Once) model was used in this study. The YOLOv5 model is a one-stage algorithm for object detection and this study applied it because of its superior accuracy, speed, and diverse application to detect and extract information of construction equipment from videos (Zhu et al. 2021). The dataset used to train the object detection model included online open image data and a large-scale construction equipment image dataset named the Alberta Construction Image Dataset (ACID) (Xiao et al. 2021). The pretrained model was applied to the previously downloaded videos of construction equipment. The detection results included equipment presence information, coordinate information for each frame, and time information.

It is not appropriate to select all sections as positive clips just because equipment has been detected by the object detection model. It can be detected incorrectly by an error of the computer vision algorithm. To control this issue, this study set criterions by controlling the confidence threshold. For the YOLOv5 model, to maximize probability-based detection quality (PDQ), the inflection points of the confidence score threshold was about 0.7 and this study applied the same value for the detection result with high quality and

accuracy (Wenkel et al. 2021). Furthermore, if the video is recorded too far from the equipment, there is a high possibility that the audio signal contains a lot of noise. This problem was solved by excluding cases in which the box size detected in the equipment is less than 1/3 of the frame size.

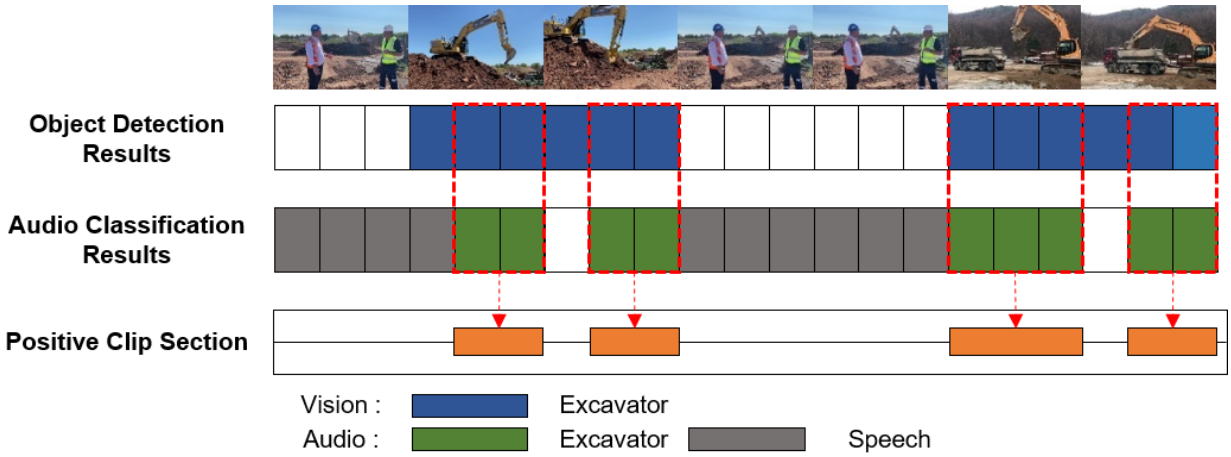

Figure 2: Selecting positive clip sections by the results of object detection and audio classification.

Figure 3 illustrates example images of detected equipment according to the object detection result. Figure 3(a) shows positive section clips selected through object detection models, Figure 3(b) shows excluded section clips due to detected box size condition, and Figure 3(c) shows section clips in which

Figure 3: Example images of detected equipment by object detection: (a) positive section clips; (b) excluded section clips due to box size condition, and (c) section clips with multiple equipment detected.

For the same video that was detected in object detection, an audio classification algorithm was used to find the sections that contain construction equipment audio signals. Even if the construction equipment appears on the screen in the previously downloaded construction equipment video, it cannot be guaranteed to include the audio signal of the equipment. There are many videos that contain other audio signals in the background, especially music or human voices. In order to extract the pure audio signal of construction equipment, it is necessary to find a section containing only the audio signal of equipment, except for the section containing other audio.

To make a distinction, the audio classification was conducted with a CNN-based PANNS(pretrained audio neural networks) pretrained model trained with a large-scale audio dataset (Kong et al. 2020). The audio signal was extracted separately from the downloaded videos, and the audio classification model was conducted to predict audio signal in every second. The critical issue of this process is to exclude sections

that contain other audio signals such as music and human speech. The post-processed prediction results of the audio classification model indicate a prediction confidence value for each class and if the value of music and speech exceeds 0.01, that audio signal was excluded. Figure 4 shows spectrograms of the audio classification results. Figure 4(a and b) are classified as music and speech without equipment sound, so those signals were excluded. If classified as equipment as shown in Figure 4(c), it was classified as a positive clip, but if it is mixed as shown in Figure 4(d), only the section with the equipment audio signal was selected by cutting the audio into seconds. Through this process, the positive audio clip containing audio signals of the construction equipment could be selected. Finally, based on the results derived through both object detection and audio classification models, the overlapping sections were selected as the final positive clips. The clips selected in this process have a high probability of audio-visual correspondence.

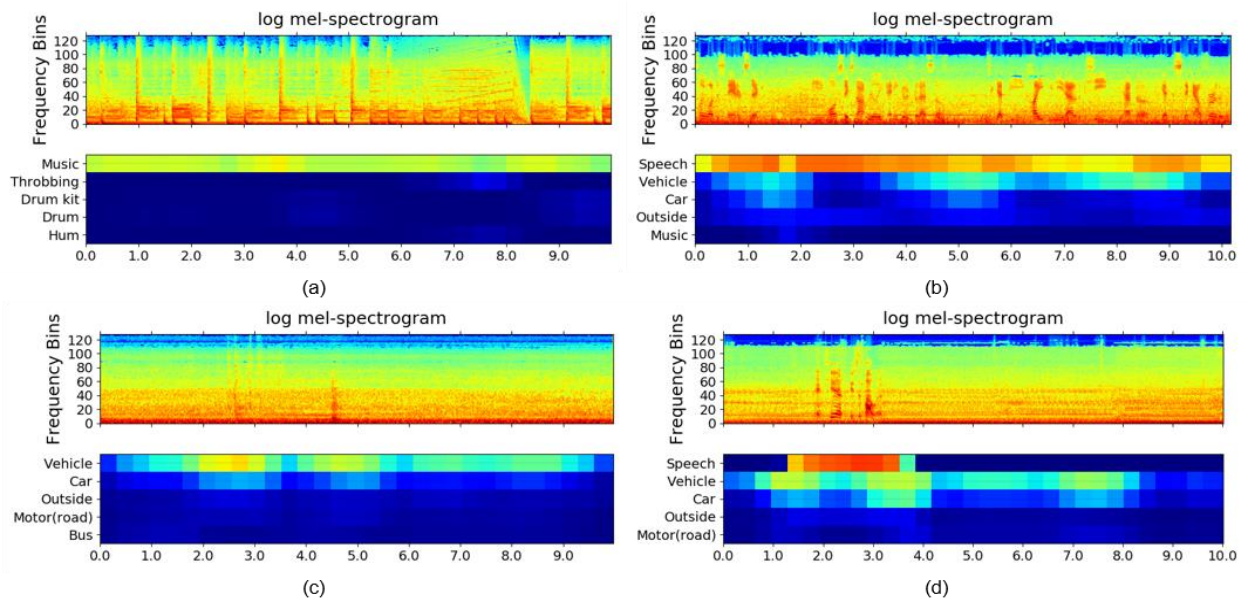

Figure 4: Example results of audio classification: (a) classified as music (negative); (b) classified as speech (negative); (c) classified as equipment sound without other sound (music or speech) (positive); and (d) partially classified as equipment sound (partially positive).

3.3 Enhancement of Audio Signal

Although positive clips containing audio signals of construction equipment are selected through object detection and audio classification, background noise (white noise, wind noise, etc.) may be included due to the nature of the construction site. Therefore, this study conducted audio enhancement to improve the quality of the audio signal by audio denoising. To provide adequate audio data, each audio clip was split into a fixed length frame and two steps of audio enhancement were performed.

During the first step, silence segments were removed with a bandpass filter from each audio clip. The bandpass filter effectively passed a specified band of frequencies and filtered out frequencies below and above this defined band. Any audio signal under the threshold of –30dB was removed in this step (Scarpiniti et al. 2021) The next step of audio enhancement was denoising the background noise in construction sites. The denoising method used in this study is a method of extracting features such as the short-time Fourier transform (STFT), mean, and standard deviation of background noise, computing a threshold noise level based upon those characteristics, generating a mask over the noise, and finally applying a noise mask to the original audio to be denoised (Sainburg et al. 2020). Ten audio signals for noise used in this step were manually sampled with only background noise, not including the audio signal of construction equipment in the construction site videos.

3.4 Separation of Audio Signal

On construction job sites, multiple pieces of equipment are operating simultaneously and their audio signals are mixed (Sherafat et al. 2022). This mixed audio signal is one of the significant challenges in audio-based activity recognition. Since the audio data extracted in the previous step may also have mixed audio signals, it was necessary to separate the audio through an audio source separation model and convert it into each audio signal. There is a hardware-based audio separation method in which microphone arrays are installed at the construction site to amplify the intended sound of equipment and remove noise from other directions (Sherafat et al. 2020). However, this method cannot be applied to audio data extracted from online videos because it cannot be determined how it was recorded, and a simpler method of software-based audio separation should be applied.

Therefore, this study suggests an effective way to separate mixed audio signals through an unsupervised audio separation method by applying a time-frequency masking approach. Masking means to assign each of the time-frequency bins to one source, in part or in whole (Manilow et al. 2018). In other words, if the audio signal of one equipment is known, a mask is generated with the audio signal and applied to the mixture phrase to separate and return other equipment audio signals. The known audio signal of equipment refers to the audio signal generated in a phrase in which only one type of equipment is operating in a video. The proposed separation model was applied to the videos with two types of equipment operating, with some sections containing only one audio signal. In the previous process, positive sections in which equipment was detected and the audio signal of that equipment was included are selected, and it was possible to select the section in which two equipment are detected in the video through the results of object detection. The separation model was applied to only the audio that satisfies the conditions.

Figure 5 shows the framework of audio signal separation proposed in this study. First, a mel-frequency spectrogram image was extracted by computing the STFT (short-time Fourier transform in an audio signal when there was only one type of equipment and an audio mask was generated. At the same time, the spectrogram of the audio signal of the mixture phase with two types of equipment was extracted in the same way and the obtained mask was applied to it. The mask is a matrix that is the same size as a spectrogram and contains values, and each value in the mask determines what proportion of energy of the original mixture that a source contributes (Luo et al. 2018). Applying a mask means multiplying the value of the mask by the spectrogram of the original mixture. If the mask is inverted and applied to the mixture, only the spectrogram of the other remains, and finally, two separate audio signals are returned.

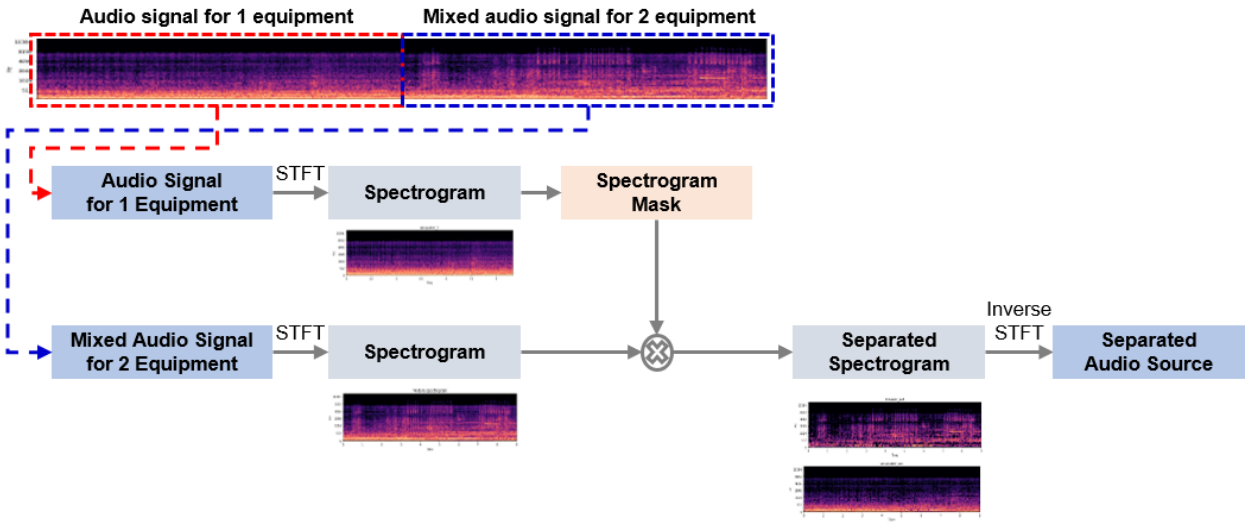

Figure 5: The framework of audio separation.

4 RESULTS

4.1 Experiment Results

Through the proposed method, an audio dataset was constructed for four types of construction equipment, namely excavator, compactor, dozer, and breaker. A total of 820 videos were downloaded for the four types and a selection process was applied to these videos. After proceeding according to the method proposed in this study, a total of 23,537 clips were selected as the audio dataset for equipment. Table 1 gives the details in terms of the duration of the selected clips and the number of clips for each type of equipment.

Table 1: Details of equipment video and audio dataset.

No.	Type of equipment	Number of videos	Duration of selected clips	Number of audio instances
1	Breaker	123	21:33	1,008
2	Compactor	119	56:13	1,217
3	Dozer	118	1:30:21	1,929
4	Excavator	460	3:44:10	5,464

4.2 Experiment Design to Test Dataset Validity

The experiment was conducted to examine the validity of the constructed dataset in training a model to recognize equipment types. We trained diverse classifiers using the constructed dataset and evaluated their performance against two types of the test datasets: (1) a dataset constructed from online sources (Test Dataset 1), and (2) a dataset collected from real-world operations (Test Dataset 2). Test Dataset 1 was randomly sampled from the dataset explained in the previous section (30% of the dataset from online sources; 2,886 instances), and the remaining data (70%; 6,732 instances) of the dataset constructed from online sources were used for training. Test Dataset 2 is sampled from Lee et al. (2020) and Sherafat el al. (2020), and is composed of a total 800 instances for each type of equipment. The performance evaluated on Test Dataset 2 is compared with the performance reported on Lee et al. (2020), which used a dataset collected from real-world operations for both training and testing of classifiers. Six classifiers including IBk (instance-bases learning with parameter k; kNN), KStar (K instant based learner), MLP(multilayer perceptron), PART(decision list), RandomForest, and RandomSubSpace, which were the best classifiers validated in Lee et al. (2020), were chosen and trained. The following set of features were extracted from audio data and used in the model: mel-frequency cepstral coefficients (MFCC), chroma energy normalized, spectral features, zero crossing rate, beat, and tempo. These features are representatively used in audio analysis (Lu et al. 2002; Patsis et al. 2008). WEKA version 3.8 software was used in this experiment.

4.3 Experiment Results and Discussion

Table 2 presents the experiment results. Against Test Dataset 1, all the classifiers showed considerable performance (accuracy: 81%–95%, F-1 score: 0.805–0.958), but against Test Dataset 2, great degradation of the performance were reported for all the classifiers except IBk (kNN) and MLP (accuracy: 64%–93%, F-1 score: 0.692–0.939). The reported performance against Test Dataset 2 in the top three classifiers (i.e., IBk, MLP, and KStar) are found to be comparable with the benchmark model (Lee et al. 2020), which used a dataset collected from real-world operations for both training and testing of classifiers. This demonstrates that the dataset constructed from online sources using the proposed framework is as effective as the dataset constructed in a traditional way in training a model to recognize real-world equipment sound.

Figure 6 shows the confusion matrices of the IBk (kNN) for four types of equipment with Test Dataset 1(a) and Test Dataset 2(b), respectively. The results show that the excavator has higher accuracy than other equipment and there are slightly unbalanced classes. However, this unbalanced data is not a critical concern

because the achieved accuracy is sufficient enough to demonstrate the applicability. In addition, since Test Dataset 2 does not have visual information, unlike Test Dataset 1 extracted from video, it is difficult to determine what conditions affected the audio signal during the collection process and the audio might be mixed with other audio sources. From this consideration, it is verified that the suggested approach had a potential capability to identify audio signals at construction sites.

Table 2: Accuracy results of classifiers.

No.	Classifier	Accuracy (%) (F-1 Score)		
		Test Dataset 1	Test Dataset 2	Benchmark model (Lee et al. 2020)
1	IBk (kNN)	95.71 (0.958)	93.75 (0.939)	85.28 (0.855)
2	KStar	89.93 (0.900)	78.63 (0.791)	80.37 (0.804)
3	MLP	88.04 (0.880)	83.87 (0.848)	91.06 (0.932)
4	PART	80.50 (0.805)	67.75 (0.692)	83.66 (0.837)
5	RandomForest	87.12 (0.880)	75.63 (0.812)	93.16 (0.932)
6	RandomSubSpace	83.09 (0.844)	64.02 (0.722)	81.81 (0.818)

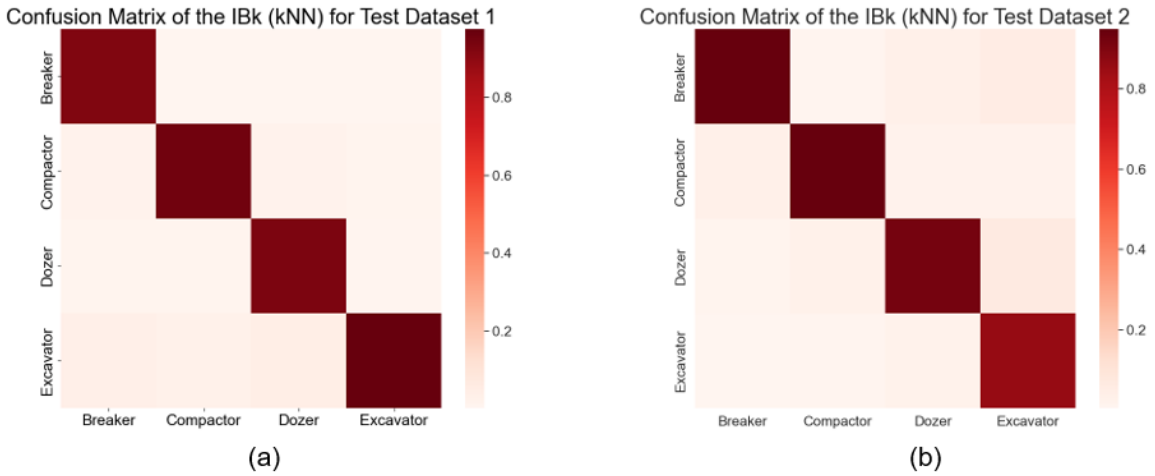

Figure 6: Confusion matrix of the IBk (kNN) for Test Dataset 1(a); and Test Dataset 2(b).

5 CONCLUSION

This study proposed a framework of constructing audio data at the action level of construction equipment from online videos for a deeper understanding of audio-based activity recognition. By applying the proposed audio separation approach, this study also introduced a method to deal with mixed audio signals of multiple types of equipment. The audio dataset of equipment constructed by the proposed framework was verified through audio classification and a case study was also conducted with audio data recorded at an actual construction site. The analysis results of the audio classification indicated that the proposed method had sufficient feasibility and potential to be applied to audio-based activity recognition and to contribute to the research community. By applying the suggested framework to construct a large-scale audio dataset from online sources, audio-based activity recognition could improve data-driven monitoring equipment and be more comprehensively understood. For further study, the proposed audio separation

model will be improved to analyze the audio signal at the action level, and it will be applied to more classes of equipment to increase the applicability on construction sites.

ACKNOWLEDGMENTS

This was supported by the Obayashi Corporation (No. 0583-20210055, Japan). Test Dataset 2 was provided by Louisiana State University research team and University of Utah research team. The authors express their gratitude for the support.

REFERENCES

AbouRizk, S. 2010. "Role of Simulation in Construction Engineering and Management". *Journal of Construction Engineering and Management*, 136(10), 1140-1153.

Ahn, C. R., S. Lee, F. Peña-Mora**,** and A. Schapiro. 2012. "Monitoring System for Operational Efficiency and Environmental Performance of Construction Operations using Vibration Signal Analysis". *In Construction Research Congress 2012*, West Lafayette, Indiana, United States, 1879-1888.

Akhavian, R., and A. H. Behzadan. 2013. "Knowledge-Based Simulation Modeling of Construction Fleet Operations Using Multimodal-Process Data Mining". *Journal of Construction Engineering and Management*, 139(11), 04013021.

Akhavian, R., and A. H. Behzadan. "Construction Equipment Activity Recognition for Simulation Input Modeling using Mobile Sensors and Machine Learning Classifiers". *Advanced Engineering Informatics*, 29(4), 867-877.

Alshibani, A., and O. Moselhi. 2016. "Productivity based Method for Forecasting Cost & Time of Earthmoving Operations using Sampling GPS Data". *Journal of Information Technology in Construction*, 21(3), 39-56.

Angah, O., and A. Y. Chen. 2020. "Tracking Multiple Construction Workers through Deep Learning and the Gradient based Method with Re-matching based on Multi-Object Tracking Accuracy". *Automation in Construction*, 119, 103308.

Chaudhuri, S., J. Roth, D. P. Ellis, A. Gallagher, L. Kaver, R. Marvin, C. Pantofaru, N. Reale, L. G. Reid, K. Wilson, and Z. Xi. 2018. "AVA-speech: A Densely Labeled Dataset of Speech Activity in Movies". *ArXiv Preprint ArXiv:1808.00606*.

Chen, H., W. Xie, A. Vedaldi, and A. Zisserman. 2020. "Vggsound: A Large-Scale Audio-Visual Dataset". In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 721-725.

Cheng, C. F., A. Rashidi, M. A. Davenport, and D. Anderson. 2016. "Audio Signal Processing for Activity Recognition of Construction Heavy Equipment". In *Proceedings of the International Symposium on Automation and Robotics in Construction*, 33, 1.

Cheng, C. F., A. Rashidi, M. A. Davenport, and D. V. Anderson. 2017. "Activity Analysis of Construction Equipment using Audio Signals and Support Vector Machines". *Automation in Construction*, 81, 240-253.

Cho, C., Y. C. Lee, and T. Zhang. 2017. "Sound Recognition Techniques for Multi-Layered Construction Activities and Events". *Computing in Civil Engineering 2017*, Seattle, Washington, 326–334.

Fang, Q., H. Li, X. Luo, L. Ding, T. M. Rose, W. An, and Y. Yu. 2018. "A Deep Learning-based Method for Detecting Non-certified Work on Construction Sites". *Advanced Engineering Informatics*, 35: 56–68.

Gemmeke, J. F., D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. 2017. "Audio Set: An Ontology and Human-Labeled Dataset for Audio Events". *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, 776–780.

Gong, J., and C. H. Caldas. 2010. "Computer Vision-Based Video Interpretation Model for Automated Productivity Analysis of Construction Operations". *Journal of Computing in Civil Engineering*, 24(3), 252-263.

Hajjar, D., and S. AbouRizk. 1999. "Simphony: an Environment for Building Special Purpose Construction Simulation Tools". *In Proceedings of the 31st Conference on Winter Simulation: Simulation---a bridge to the future*, 2, 998-1006.

Kim, J., and S. Chi. 2017. "Adaptive Detector and Tracker on Construction Sites Using Functional Integration and Online Learning". *Journal of Computing in Civil Engineering*, 31(5), 04017026.

Kim, J., S. Chi, and C. R. Ahn. 2021. "Hybrid Kinematic–Visual Sensing Approach for Activity Recognition of Construction Equipment". *Journal of Building Engineering*, 44, 102709.

Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., & Plumbley, M. D. 2020. "Panns: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition". *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28, 2880-2894.

Lee, S., J. Chung, Y. Yu, G. Kim, T. Breuel, G. Chechik, and Y. Song. 2021. "ACAV100M: Automatic Curation of Large-Scale Datasets for Audio-Visual Video Representation Learning". *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10274-10284.

Lee, Y. C., M. Scarpiniti, and A. Uncini. 2020. "Advanced Sound Classifiers and Performance Analyses for Accurate Audio-based Construction Project Monitoring". *Journal of Computing in Civil Engineering*, 34(5), 04020030.

Lu, L., H. J. Zhang, and H. Jiang. 2002. "Content Analysis for Audio Classification and Segmentation". *IEEE Transactions on Speech and Audio Processing*, 10(7), 504-516.

Luo, Y., and N. Mesgarani. 2018. "Tasnet: Time-domain Audio Separation Network for Real-time, Single-channel Speech Separation". In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP*), 696-700.

Manilow, E., P. Seetharaman, and B. Pardo. 2018. "The Northwestern University Source Separation Library". *In ISMIR*, 297-305.

Martinez, J. C., and P. G. Ioannou. 1999. "General-purpose Systems for Effective Construction Simulation". *Journal of Construction Engineering and Management*, 125(4), 265-276.

Nagrani, A., J. S. Chung, and A. Zisserman. 2017. "Voxceleb: a Large-scale Speaker Identification Dataset". *arXiv preprint arXiv*:1706.08612.

Patsis, Y., and W. Verhelst. 2008. "A Speech/Music/Silence/Garbage/Classifier for Searching and Indexing Broadcast News Material". *In 2008 19th International Workshop on Database and Expert Systems Applications*, 585-589.

Sabillon, C. A., A. Rashidi, B. Samanta, C. F. Cheng, M. A. Davenport, and D. V. Anderson. "A Productivity Forecasting System for Construction Cyclic Operations Using Audio Signals and a Bayesian Approach," in *Construction Research Congress 2018*, New Orleans, Louisiana, 295–304.

Sainburg, T., M. Thielk, and T. Q. Gentner. 2020. "Finding, Visualizing, and Quantifying Latent Structure Across Diverse Animal Vocal Repertoires". *PLoS Computational Biology*, 16(10), e1008228.

Scarpiniti, M., F. Colasante, S. Di Tanna, M. Ciancia, Y. C. Lee, and A. Uncini. 2021. "Deep Belief Network based Audio Classification for Construction Sites Monitoring". *Expert Systems with Applications*, 177, 114839.

Scarpiniti, M., D. Comminiello, A. Uncini, and Y. C. Lee. 2021. "Deep Recurrent Neural Networks for Audio Classification in Construction sites". *In 2020 28th European Signal Processing Conference (EUSIPCO)*, 810-814.

Sherafat, B., C. R. Ahn, R. Akhavian, A. H. Behzadan, M. Golparvar-Fard, H. Kim, Y.-C. Lee, A. Rashidi, and E. R. Azar. 2020. "Automated Methods for Activity Recognition of Construction Workers and Equipment: State-of-the-Art Review". *Journal of Construction Engineering and Management*, 146(6), 03120002.

Sherafat, B., A. Rashidi, and S. Asgari. 2020. "Comparison of Different Beamforming-Based Approaches for Sound Source Separation of Multiple Heavy Equipment at Construction Job Sites". In *Proceedings of the 2020 Winter Simulation Conference*, edited by K. H. B. Gae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Thiesing, 2435–2446. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Sherafat, B., A. Rashidi, and S. Asgari. 2022. "Sound-based Multiple-Equipment Activity Recognition using Convolutional Neural Networks". *Automation in Construction*, 135, 104104.

Tajeen, H., and Z. Zhu. 2014. "Image Dataset Development for Measuring Construction Equipment Recognition Performance". *Automation in Construction*, 48, 1-10.

Watanabe, N., S. Fukui, Y. lwahori, Y. Hayashi, W. Achariyaviriya, and B. Kijsirikul. 2020. "Automatic Construction of Dataset with Automatic Annotation for Object Detection". *Procedia Computer Science*, 176: 1763–1772.

Wenkel, S., K. Alhazmi, T. Liiv, S. Alrshoud, and M. Simon. 2021. "Confidence Score: The Forgotten Dimension of Object Detection Performance Evaluation". *Sensors*, 21 (13): 4350.

Xiao, B., S. C. Kang. 2021. "Development of an Image Data Set of Construction Machines for Deep Learning Object Detection". *Journal of Computing in Civil Engineering*, 35(2), 05020005.

Zhang, T., Y. C. Lee, M. Scarpiniti, and A. Uncini. 2018. "A Supervised Machine Learning-Based Sound Identification for Construction Activity Monitoring and Performance Evaluation". *Construction Research Congress 2018*, New Orleans, Louisiana, 358–366.

Zhu, X., S. Lyu, X. Wang, and Q. Zhao. 2021. "TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-captured Scenarios". In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, BC, Canada, 2778-2788.

AUTHOR BIOGRAPHIES

GILSU JEONG is a Ph. D candidate in the Department of Architecture and Architectural Engineering at Seoul National University. He received a bachelor's degree from SNU in 2018. In 2020, he graduated with a master's course for construction management at SNU. His research interests focus on automatic monitoring of construction resources with computer vision-based and audio-based to measure and improve productivity of construction activities. His e-mail address is jeonggskr@snu.ac.kr.

CHANGBUM RYAN AHN is an associate professor in the Department of Architecture and Architectural Engineering at Seoul National University. He earned a Ph.D. degree in Civil Engineering from the University of Illinois at Urbana-Champaign. His research interests focus on data-driven techniques and smart-sensing technologies for analyzing and identifying humans' collective behavior patterns derived from wearable sensors and/or crowdsourced dataset, in order to design and build (1) safe construction workplaces, (2) smart and connected urban communities, and (3) intelligent and energy-efficient building systems. His email address is cbahn@snu.ac.kr.

MOONSEO PARK is a professor at the Department of Architectural Engineering of Seoul National University. got into the Department of Architecture of Seoul National University in 1985, completed the courses for a bachelor's degree in 1989, and graduated master's course for City Planning at SNU in 1992. In 1998-2001, he received master's degree and doctor's degree for Project Management in MIT. After graduation, he worked for the Dept. of Building in National University of Singapore as an assistant professor. Currently, his major research area is systematic approach for construction, knowledge-based construction etc. His email address is mspark@snu.ac.kr.