ICCEPM 2020
The 8th International Conference on Construction Engineering and Project Management
Dec. 7-8, 2020, Hong Kong SAR

Real-time automated detection of construction noise sources based on convolutional neural networks

Seunghoon Jung[1], Hyuna Kang[2], Juwon Hong[3], Taehoon Hong[4]*, Minhyun Lee[5], Jimin Kim[6]

[1] *Department of Architecture and Architectural Engineering, Yonsei University, Seoul, Republic of Korea,* E-mail address: saber21@yonsei.ac.kr

[2] *Department of Architecture and Architectural Engineering, Yonsei University, Seoul, Republic of Korea,* E-mail address: hyuna_kang@yonsei.ac.kr

[3] *Department of Architecture and Architectural Engineering, Yonsei University, Seoul, Republic of Korea,* E-mail address: juwonae@yonsei.ac.kr

[4] *Department of Architecture and Architectural Engineering, Yonsei University, Seoul, Republic of Korea,* E-mail address: hong7@yonsei.ac.kr

[5] *Department of Architecture and Architectural Engineering, Yonsei University, Seoul, Republic of Korea,* E-mail address: mignon@yonsei.ac.kr

[6] *Department of Architecture and Architectural Engineering, Yonsei University, Seoul, Republic of Korea,* E-mail address: cookie6249@yonsei.ac.kr

Abstract: Noise which is unwanted sound is a serious pollutant that can affect human health, as well as the working and living environment if exposed to humans. However, current noise management on the construction project is generally conducted after the noise exceeds the regulation standard, which increases the conflicts with inhabitants near the construction site and threats to the safety and productivity of construction workers. To overcome the limitations of the current noise management methods, the activities of construction equipment which is the main source of construction noise need to be managed throughout the construction period in real-time. Therefore, this paper proposed a framework for automatically detecting noise sources in construction sites in real-time based on convolutional neural networks (CNNs) according to the following four steps: (i) Step 1: Definition of the noise sources; (ii) Step 2: Data preparation; (iii) Step 3: Noise source classification using the audio CNN; and (iv) Step 4: Noise source detection using the visual CNN. The short-time Fourier transform (STFT) and temporal image processing are used to contain temporal features of the audio and visual data. In addition, the AlexNet and You Only Look Once v3 (YOLOv3) algorithms have been adopted to classify and detect the noise sources in real-time. As a result, the proposed framework is expected to immediately find construction activities as current noise sources on the video of the construction site. The proposed framework could be helpful for environmental construction managers to efficiently identify and control the noise by automatically detecting the noise sources among many activities carried out by various types of construction equipment. Thereby, not only conflicts between inhabitants and construction companies caused by construction noise can be prevented, but also the noise-related health risks and productivity degradation for construction workers and inhabitants near the construction site can be minimized.

Key words: Construction noise, Real-time automated detection, Convolutional neural network, Short-time Fourier transform, Temporal image processing

1. INTRODUCTION

Environment pollutants (e.g., greenhouse gas, noise, odor, waste) from industrial activities cause critical social problems that undermine the quality of life and environmental rights of humans [1, 2].

Among the various environmental pollutants, especially, noise (i.e., unwanted, unpleasant, and loud sounds) generated from construction sites can directly and immediately affect the people around the noise source [3, 4]. Moreover, as construction projects have become more complex and larger, the construction noise from various types of heavy construction equipment throughout the construction period (i.e., several months to years) can damage health of construction workers and inhabitants near the construction site along with their working and living environment [5-7]. Despite the strengthened regulations and management guidelines on noise emissions to address these noise issues [8-10], construction noise has been provoking a stream of many complaints and conflicts between inhabitants and construction companies that cause delay in the construction project [11, 12]. In addition, construction noise causes health problems such as hearing loss, and productivity degradation through decreased concentration of construction workers located directly next to the noise source [7, 13]. Consequently, to minimize the deterioration in quality of life and potential conflicts arising from inhabitants, and to ensure the productivity and safety of the construction workers or labors, construction managers need to manage the level of construction noise as much as possible. However, until now, construction noise in South Korea has been managed and monitored only for a short period of time (e.g., three times of five-minute measurements) after the noise complaints occur. Therefore, during the rest period of time, the unmanaged construction noise can constantly affect to construction workers and inhabitants. Due to this poor management of construction noise, inhabitants near construction sites as well as construction workers, are still exposed to unnecessary and unpleasant noise during the entire construction period. To overcome the limitations of the current noise management methods, the level of construction noise and the activities of noise generating construction equipment, which is the main source of the construction noise, need to be monitored and managed throughout the construction period in real-time.

With the necessity of managing both construction noise and equipment in real-time, several previous studies proposed different methods for identifying construction equipment or noise using various deep learning methods based on the following two data format: (i) audio data; and (ii) visual data. First, some studies classified the activities of construction equipment and analyzed the effects of noise on humans by analyzing the characteristics of the audio data so as to recognize construction equipment or assess construction noise. Cheng et al. [14] converted the audio data into a time-frequency representation using a short-time Fourier transform (STFT), then classified the activity of each construction equipment using an support vector machine (SVM). Abdoli et al. [15] classified the environmental sounds (e.g., drilling, engine idling, jack-hammers) based on the audio data using a 1-dimensional convolutional neural network (CNN) without data preprocessing. Lee et al. [16] and Ballesteros et al. [17] analyzed the characteristics of construction noise (i.e., sound pressure level and frequency) according to the type of construction stage and equipment, and assessed the impact of construction noise (such as annoyance and stress) on construction workers or general public. Second, several studies identified the types and activities of construction equipment by analyzing visual data to measure and monitor the performance of construction equipment. Kim et al. [18] proposed a framework to identify the interactive operations between excavators and dump trucks based on a comprehensive visual dataset of activities by using a tracking-learning-detection method. Fang et al. [19] extracted feature maps from visual data and improved the accuracy of identifying construction equipment and workers in real-time using improved faster region-based CNN (Faster R-CNN). Golparvar-Fard et al. [20] collected spatio-temporal visual features from the visual data of a single piece of construction equipment, then identified the activity through the distributions of the spatio-temporal features using an SVM.

Although these previous studies identified the types and activities of construction equipment by analyzing audio or video data using deep learning methods, following limitations still remain. First, previous studies successfully analyzed the characteristics and effects of construction noise on construction equipment only using the audio data, but they did not classify or manage construction noise in real-time. Second, previous studies identified the types or activities of the construction equipment only using the visual data, but there are limitations in identifying multiple noises generating construction equipment in real-time on a large and complex construction site. In these studies, some of them were failed to identify the activity of construction equipment because they analyzed 2-dimensional data (i.e., single image) which is non-temporal. Others successfully identified the activity of construction equipment, however, they still had a limitation in identifying multiple construction equipment at the same time. Third, since previous studies only analyzed either the audio or video data through deep learning in order to judge the types and activities of the construction equipment, it was hard to judge the noise generating construction equipment practically at the actual construction sites. That is, when only

audio data is analyzed, the type and activity of the equipment can be identified, but the actual location of the equipment remains unknown. On the other hand, when only visual data is analyzed, the type, activity, and location of the equipment can be identified, but the magnitude of the noise remains unknown. Consequently, to manage the construction noise practically and effectively in real-time, it is necessary to simultaneously analyze the temporal audio and visual data to detect the multiple noises generating construction equipment at a construction site. Therefore, this study aimed to propose a framework for automatically and systematically detecting the noise source on the whole construction site in real-time by classifying the noise source based on the temporal audio data and detecting it based on the visual data through CNN-based deep learning.

2. METHODS AND MATERIALS

In this study, the proposed framework for real-time automated detection of construction noise sources consists of four steps (refer to Fig. 1). First, the noise sources in the construction site should be defined according to the types of equipment and their activities. Second, the audio and visual data on the construction equipment should be preprocessed so as to be input of the audio CNN and the visual CNN. Third, the construction noise should be automatically classified based on the audio input data by feeding them into the audio CNN in order to determine the current noise source in the construction site. Fourth, the construction equipment and activities which are determined to be noise sources should be detected on the video of the construction site by feeding visual input data into the visual CNN. As a result, by utilizing the video of the construction site, the proposed framework can make it possible to automatically detect where the construction noise is generated in real-time. Details of each step are described below.

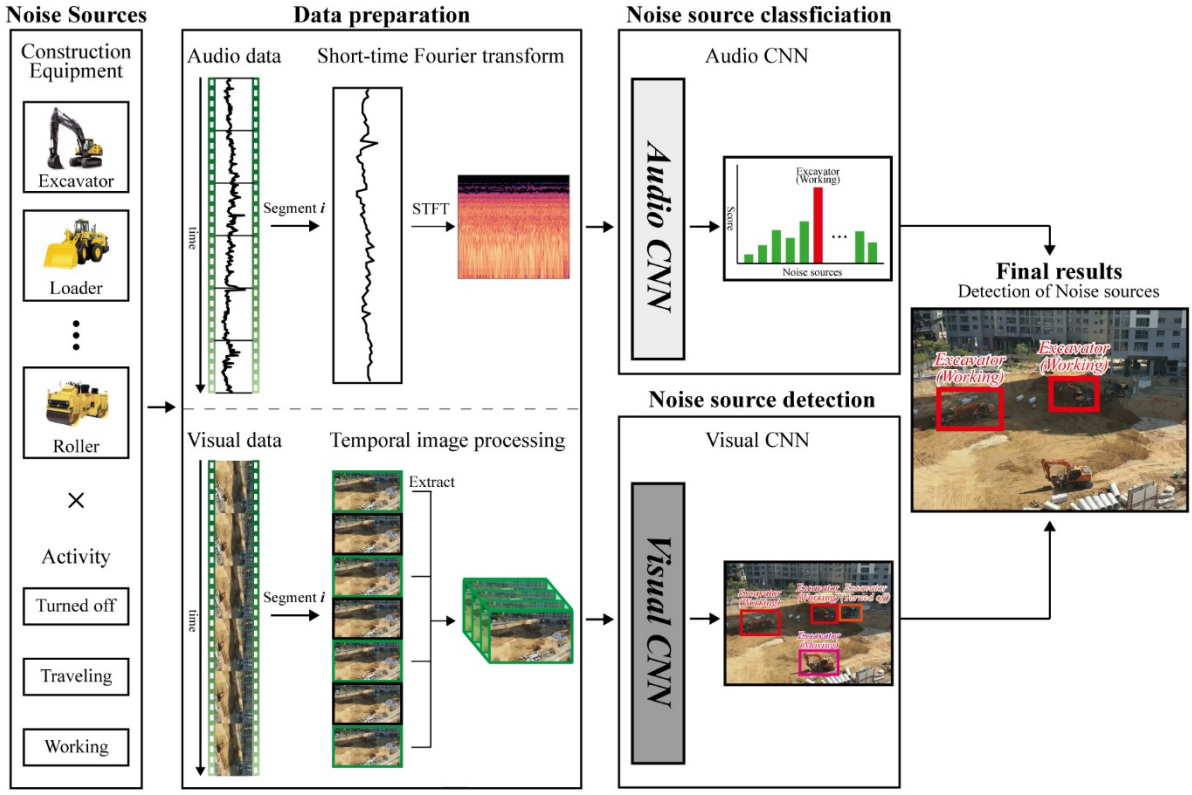

Figure 1. The proposed framework for a real-time automated detection of construction noise sources

2.1. Step 1. Definition of the noise sources

On construction sites, various types of heavy construction equipment generate noises which are different in amplitude, frequency, and wavelength while performing a particular activity [16]. Therefore, this study defined noise sources according to the types of heavy construction equipment and their activities. The types of heavy construction equipment can be defined in terms of the noise-generating construction equipment regulated under the *Noise and Vibration Control Act* of South Korea [21]:

excavator, roller, loader, breaker, earth auger, pile driver, and dump truck. In addition, the activities of each type of construction equipment can be defined by taking into account the different characteristics of the noises of each activity: turned off, traveling, and working. However, since a roller and dump truck among the defined types of heavy construction equipment, travel and work at the same time, their working activity is considered as a traveling activity. As a result, the noise sources in this study are defined by seven types of construction equipment with their activities, as shown in Table 1.

Table 1. The defined noise sources

Types	Activities		
	Turned off	**Traveling**	**Working**
Excavator	○	○	○
Roller	○	○	-
Loader	○	○	○
Breaker	○	○	○
Earth auger	○	○	○
Pile driver	○	○	○
Dump truck	○	○	-

2.2. Step 2. Data preparation

In order to extract the feature of the noise and movements of the noise sources, the audio and visual data on the noise sources should be collected and preprocessed. The audio and visual data on the noise sources can be collected from audio and video recordings on the construction site. When the noise is generated from the construction equipment, the audio signal of the generated noise changes over time according to the pattern of the activity. Moreover, the activity of the construction equipment cannot be detected when using a single image, as the defined types of construction equipment generate noises with movements or actions that can cause changes in each sequence of the video. Therefore, to include the information of those changes over time, each segmented frame from the audio and video recordings with the proper length of time should be preprocessed into 3-dimensional data including width, height, and color channels to then serve as input data for the CNNs. Towards this end, the proposed framework transforms the collected data using the STFT for the audio data and temporal image processing for the visual data.

Firstly, to extract the feature from the audio data, the audio signal should be converted into the form of a time-frequency domain using the STFT. The discrete Fourier transform (DFT), which is usually used for spectral analysis of the signal, can only extract the frequency domain information on the audio signals. On the other hand, the STFT, which is a sequence of the DFTs of the divided signals with a certain window size in time, can extract the time-frequency domain information that makes it possible to track the changes of the frequency properties of the audio signal over time. To perform the STFT on the audio data, first, the collected raw audio data should be sampled to change the continuous signal into discrete frames. After the sampling of the raw audio data, the discrete frames should be segmented to make the frameset of which to calculate the STFT. The STFT of the frameset can be obtained by dividing the frameset with the window size and then multiplying each divided frames by a window function (refer to Eq. (1)). As a result, the raw audio data can be transformed into 3-dimensional representation to be the input data for the audio CNN. As shown in Fig. 2, the audio input data indicates the amplitude of a certain frequency at a certain time that can be shown as the intensity of the color in the image. For the training dataset, the transformed frameset should be labeled with the defined noise sources.

$$X(k,t) = \sum_{m=0}^{M-1} w(m)x(m+Zt)e^{-2\pi jmk/M} \quad (1)$$

where $X(k,t)$ is the value at frequency k and time t, m is the signal frame, M is the window size, $w(m)$ is the window function, $x(m+Zt)$ is the original divided signal, and Z is the parameter of a window stride.

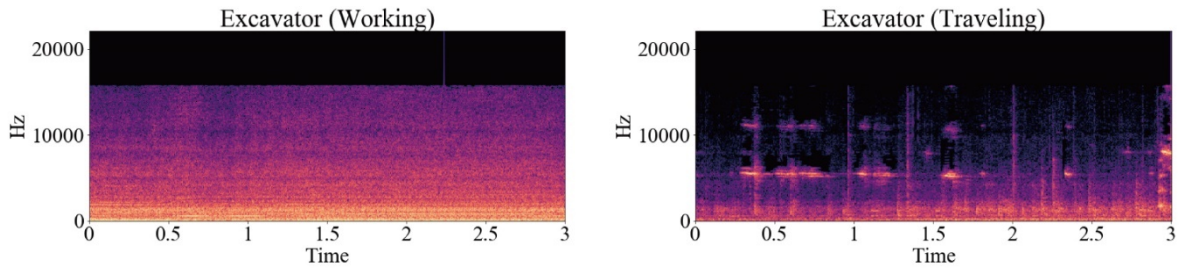

Figure 2. An image example for the STFT of the construction equipment

Secondly, the video data should be transformed to extract the feature. Video can be described as a set of image frames. Accordingly, actions in the video can be regarded as the temporal change of the color of the pixels. Therefore, in order to consider these changes in the input data for the visual CNN, the temporal image processing that concatenates multiple image frames into a single set of image frames is used for generating the visual data. The temporal image processing can start by sampling the image frames from the video. In addition, to recognize the actions from the sampled image frames, they need to contain enough length of time from the first to the last image frames while reflecting definite changes in pixels. If too many image frames are sampled, however, the CNN cannot operate in real-time due to excessive computational demands. Accordingly, after sampling the image frames, a subset of the sampled image frames should be extracted at particular frame intervals. Then, the extracted image frames from the subset should be concatenated into a single set of input data for the visual CNN.

For example, if the subset frames are extracted at eight intervals from 25 frames of the video whose width and height is 416×416, the visual input data will be the concatenated four frames (i.e., 1st, 9th, 17th, and 25th) with the size of 416×416×12 (i.e., width×height×(RGB channels×four frames)). As a result, the visual input data can reflect enough length of time during the activity of the construction equipment with less amount of data. Just like the audio input data, the visual input data can be also transformed into 3-dimensional data and each transformed data should be labeled with the defined noise sources for the training dataset.

2.3. Step 3. Noise source classification using the audio CNN

As CNNs have been proven to be effective in image classification, there have been increasing attempts to analyze audio signals using CNN algorithms (i.e., AlexNet, Inception, VGG, and ResNet) and the results were promising [22]. In order to determine what kind of activity of the construction equipment was the source of the current noise, this study proposed the audio CNN as a real-time automated classification algorithm by using the AlexNet which significantly outperformed the other prior algorithms.

In this study, the AlexNet which consists of five convolutional layers for feature extraction, three fully connected layers for classification, and three max-pooling layers for subsampling is applied to the audio CNN. To prevent the gradient vanishing which interrupts the training of the network, the ReLU (Rectified Linear Unit) is used as the activation function of overall neurons. On the other hand, the output of the last fully connected layer can be calculated using the softmax function as the activation function to represent a probabilistic distribution over the class labels. Since there are seven defined noise sources in this study, the output neurons of the last fully connected layer should be modified to seven for the audio CNN.

In order to train the audio CNN on the target audio data, transfer learning which is fine-tuning based on pre-trained AlexNet should be performed. The training process can be done by adjusting the weights of neurons using the backpropagation algorithm that is widely used for training the neural networks. As a result, by feeding the input variables with the size of 224×224×3 into the network, the probability score of each noise source can be calculated as an output. By using this audio CNN, any noise source whose score is above the certain criteria can be suggested as the current noise source in the construction site.

2.4. Step 4. Noise source detection using the visual CNN

Based on the results of the noise classification in Step 3, the suggested noise sources should be detected on the video of the construction site. Since many activities of different types of construction equipment are operated at the same time on construction sites, it should be possible to detect multiple activities on a single image in real-time. Therefore, object detection algorithms that can automatically detect multiple objects in real-time are implemented for developing the visual CNN. In this study, an object detector algorithm called You Only Look Once v3 (YOLOv3), which is well-balanced in terms of speed and accuracy, is modified to be applied to the visual CNN (refer to Fig. 3) [23-25]. The YOLOv3 that uses the Darknet-53 which consists of 53 convolutional layers predicts the bounding boxes on the objects and the class confidence of the objects to localize and classify the objects. Further information about the architecture of the YOLOv3 can be found in [26]. By using multiple convolutional and residual layers, YOLOv3 can divide the input image with the size of 416×416 into grid cells with 3 different scales (i.e., 13×13, 26×26, and 52×52) to consider the various size of the object. Each grid cell has prediction results encoding 3 bounding boxes with different scales, objectness confidence representing the probability of the object existence on the grid cell, and class confidence representing the probability of belonging to each class. As a result, the dimension of the output for each grid can be represented as Eq. (2).

$$Dimension_{output} = S \times S \times \left(3 \times (4 + 1 + C)\right) \tag{2}$$

where S is the width and height of the grid, 3 is the number of bounding boxes, 4 is the properties of the bounding box which are bounding box coordinate, width, and height, 1 is the objectness confidence, and C is the class confidence for each class.

Since the YOLOv3 is for object detection rather than action detection, it needs to be modified to detect the activities of the construction equipment from the pre-processed visual data. Therefore, an early fusion which combines the data immediately on the first convolutional layer is applied to detect noise sources while preventing excessive calculations for the visual CNN so as to operate in real-time [27]. As such, the input dimension of the visual CNN should be modified from the YOLOv3 so as to coordinate with the dimension of the visual input data.

Just like the noise classification algorithm (i.e., audio CNN), transfer learning should be performed for the visual CNN based on the pre-trained YOLOv3. As a result, the construction equipment of the activity determined to be the noise source by the audio CNN can be localized on the real-time video as an output of the visual CNN.

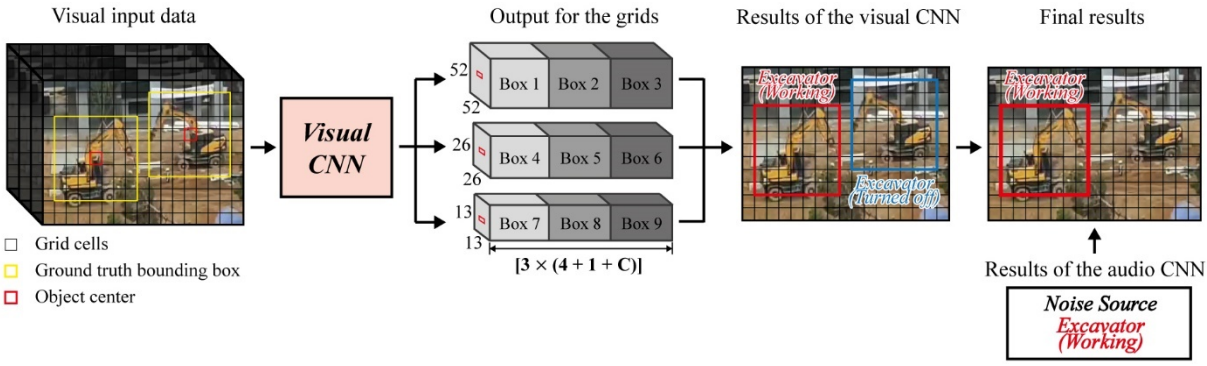

Figure 3. The process flow of the visual CNN

3. CONCLUSION

This study proposed a framework for an automated detection system of noise source on the whole construction site in real-time based on temporal audio and visual data using convolutional neural networks (CNNs). The proposed framework used the short-time Fourier transform (STFT) and temporal image processing to contain temporal features of the audio and visual data. In addition, the audio CNN and visual CNN were proposed with modifications of the AlexNet and the YOLOv3 in order to classify and detect the noise sources in real-time. The proposed framework could be helpful for environmental construction managers to efficiently identify the noise by automatically detecting the noise sources that

are generally complicated in the construction site due to many activities of various types of construction equipment. Furthermore, they can immediately control the noise level by providing information about the noise source in real-time that is essential to constantly and effectively reduce noise under a certain level. Thereby, complaints and conflicts between inhabitants and a construction company caused by construction noise can be effectively prevented. Moreover, noise-related health risks and productivity degradation of construction workers and inhabitants near the construction site can be minimized. With such contributions, the proposed framework could be applied to construction sites as a management system for securing the safety and productivity of the construction project.

However, this study only proposed a framework for the automated real-time detection system of noise sources in the construction site. Furthermore, some noise sources show various actions during a single working activity, such as digging and rotating for the excavator. Therefore, to effectively apply and validate the proposed framework, more detailed classes should be considered for developing the detection system along with the experimental study of the proposed framework with sufficient data.

ACKNOWLEGEMENTS

This work was supported by the National Research Foundation of Korea (NRF) grand funded by the Korea government (MSIT; Ministry of Science and ICT) (NRF-2018R1A5A1025137).

REFERENCES

[1] D. Shelton, “Human Rights, Environmental Rights, and the Right to Environment”, Stan. J. Int'l L, vol. 1, pp. 509–544, 1991.

[2] D. R. Boyd, “The constitutional right to a healthy environment”, Environment, vol. 54, no. 4, pp. 3–14, 2012.

[3] WHO, “Burden of disease from environmental noise: Quantification of healthy life years lost in Europe”, pp. 126, 2011.

[4] M. D. Seidman and R. T. Standring, “Noise and quality of life” International Journal of Environmental Research and Public Health, vol. 7, no. 10, pp. 3730–3738, 2010.

[5] H. Zhang, D. Zhai, and Y. N. Yang, “Simulation-based estimation of environmental pollutions from construction processes”, Journal of Cleaner Production, vol. 76, pp. 85–94, 2014.

[6] M. D. Fernández, S. Quintana, N. Chavarría, and J. A. Ballesteros, “Noise exposure of workers of the construction sector”, Applied Acoustics, vol. 70, no. 5, pp. 753–760, 2009.

[7] X. Li, Z. Song, T. Wang, Y. Zheng, and X. Ning, “Health impacts of construction noise on workers: A quantitative assessment model based on exposure measurement”, Journal of Cleaner Production, vol. 135, pp. 721–731, 2016.

[8] Ministry of Environment, “Guidance of Noise & Vibration Control under Construction”, Sejong-si, Republic of Korea, 2006.

[9] I. Bennett, “Control of noise at Work Regulations 2005”, Acoustic Bulletin, vol. 31, no. 3, pp. 35–37, 2006.

[10] Ministry of Environment, “Enforcement Rules of Noise vibration Control Act”, vol. 9, no. 1. Sejong-si, Republic of Korea, 2010.

[11] Central Environmental Dispute Mediation Committee, “Environmental Dispute Resolution Statistical Data”, Sejong-si, Republic of Korea, 2018.

[12] Department Environmental Protection, “Pollution Complaint Statistics 2017”, 2018. https://www.epd.gov.hk/epd/english/laws_regulations/enforcement/pollution_complaints_statistics_2017.html [accessed 29 July 2019].

[13] SP. Dozzi and SM. AbouRizk, “Productivity in Construction”, Ottawa: Institute for Research in Construction, National Research Council, 1993.

[14] C. F. Cheng, A. Rashidi, M. A. Davenport, D. V. Anderson, “Activity analysis of construction equipment using audio signals and support vector machines”, Automation in Construction, vol. 81, no. September 2016, pp. 240–253, 2017.

[15] S. Abdoli, P, Cardinal, A. L. Koerich, "End-to-End Environmental Sound Classification using a 1D Convolutional Neural Network", Expert Systems with Applications, vol. 136, pp. 252-263, 2019.

[16] S. C. Lee, J. Y. Hong, and J. Y. Jeon, “Effects of acoustic characteristics of combined construction noise on annoyance”, Building and Environment, vol. 92, pp. 657–667, 2015.

[17] M. J. Ballesteros, M. D. Fernández, S. Quintana, J. A. Ballesteros, and I. González, “Noise emission evolution on construction sites. Measurement for controlling and assessing its impact on the people and on the environment”, Building and Environment, vol. 45, no. 3, pp. 711–717, 2010.

[18] J. Kim, S. Chi, and J. Seo, “Interaction analysis for vision-based activity identification of earthmoving excavators and dump trucks”, Automation in Construction, vol. 87, no. January, pp. 297–308, 2018.

[19] W. Fang, L. Ding, B. Zhong, P. E. D. Love, and H. Luo, “Automated detection of workers and heavy equipment on construction sites: A convolutional neural network approach”, Advanced Engineering Informatics, vol. 37, no. May, pp. 139–149, 2018.

[20] M. Golparvar-Fard, A. Heydarian, and J. C. Niebles, “Vision-based action recognition of earthmoving equipment using spatio-temporal features and support vector machine classifiers”, Advanced Engineering Informatics, vol. 27, no. 4, pp. 652–663, 2013.

[21] Ministry of Environment, “Noise and Vibration Control Act.”, Sejong-si, Republic of Korea, 2017.

[22] S. Hershey et al., “CNN architectures for large-scale audio classification” 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 131–135, 2017.

[23] D. Kim, M. Liu, S. H. Lee, and V. R. Kamat, “Remote proximity monitoring between mobile construction resources using camera-mounted UAVs”, Automation in Construction, vol. 99, no. December 2018, pp. 168–182, 2019.

[24] F. Wu, G. Jin, M. Gao, Z. HE, and Y. Yang, “Helmet Detection Based On Improved YOLO V3 Deep Model”, 2019 IEEE 16th International Conference on Networking, Sensing and Control (ICNSC), pp. 363–368, 2019.

[25] H. Wang, Y. Yu, Y. Cai, X. Chen, L. Chen, and Q. Liu, “A Comparative Study of State-of-the-Art Deep Learning Algorithms for Vehicle Detection,” IEEE Intelligent Transportation Systems Magazine, vol. 11, no. 2, pp. 82–95, 2019.

[26] J. Redmon and A. Farhadi, “YOLOv3: An Incremental Improvement”, arXiv preprint arXiv: 1804.02767, 2018.

[27] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and F. F. Li, “Large-scale video classification with convolutional neural networks”, Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 1725–1732, 2014.