Automated Vision-Based Activity Identification for Demolition Operations

Mohammad Javad Shooshtari and Juyeong Choi

Abstract Demolition projects involve various types of heavy equipment (e.g., excavators, dump trucks, loaders, etc.). As such, the success of demolition projects is significantly dependent on heavy equipment operations. Prior studies have investigated heavy equipment productivity within the context of construction operations (i.e., earthwork) by tracking machine productivity through traditional approaches (e.g., manually tracking the duration of heavy equipment activities, etc.), which is a time-consuming, labor-intensive, and error-prone job. To facilitate research on heavy equipment productivity, recent studies employ artificial intelligence-based methods to automatically identify heavy equipment activities and measure productivity in construction operations. However, unlike earthwork activities where most of the tasks are relatively simple and repetitive, demolition activities are more complex and dynamic (i.e., related to structural demolition and material separation). Due to the varied nature of demolition activities, applying existing approaches to identify demolition activities is questionable. This study presents an automatic vision-based activity identification model based on three-dimensional Convolutional Neural Networks (CNNs), which can extract spatial and temporal features simultaneously. The proposed approach can recognize three excavator activities related to material separation (i.e., grabbing, swinging, and dumping) used in demolition operations. To develop the model, small-scale excavators were used to simulate a real-world demolition operation (i.e., separating materials), while two cameras were used to record videos of such experiments. Recorded video datasets were manually labeled and used to train the proposed model. Compared to construction projects, demolition projects are not relatively common. Therefore, it would have taken a while to collect the data from real-world demolition sites for training and validating the activity identification model. Through small-scale demolition simulations, the feasibility of the vision-based activity identification model was validated, which will support its application for full-scale demolition productivity improvement (i.e., by reducing the

M. J. Shooshtari · J. Choi (✉)
Department of Civil and Environmental Engineering, FAMU-FSU College of Engineering, Florida State University, 2525 Pottsdamer Street, Tallahassee, FL 32310, USA
e-mail: jchoi@eng.famu.fsu.edu

S. Desjardins et al. (eds.), *Proceedings of the Canadian Society for Civil Engineering Annual Conference 2023, Volume 5*, Lecture Notes in Civil Engineering 499,
https://doi.org/10.1007/978-3-031-61503-0_8

time and labor required for manual tracking of heavy equipment activities, monitoring the productivity of demolition operations, and enabling the development of timely and effective demolition strategies and productivity improvement measures accordingly).

Keywords Demolition projects · Computer vision · Activity identification

1 Introduction

Demolition projects are usually followed by construction projects, which imposes a time constraint on the demolition project [1]. The transfer of demolished material to staging areas must be completed within a specified time frame to ensure the following construction project does not fall behind schedule. Given the importance of meeting this constraint, it is crucial to monitor the productivity of demolition projects to ensure that they are completed within the allotted time frame.

Demolition projects rely heavily on the use of heavy equipment, such as excavators. In order to effectively monitor and evaluate the productivity of demolition projects, it is important to identify and investigate the activities performed by heavy equipment. Traditionally, heavy equipment activity identification was performed through labor-intensive means, but these methods are costly, time-consuming, and prone to error. Automated activity identification models have been developed to address such challenges [2].

Previous studies have developed activity identification models for construction projects (e.g., earthwork) but have not yet developed such models for demolition projects. Unlike construction projects, Demolition projects involve a range of more complicated activities, focusing on building demolition and material separation and removal. These activities generate a large amount of waste, which requires proper management to avoid environmental pollution. Furthermore, demolition activities require careful planning and execution to ensure the safety of workers and the public, as well as the removal of existing structures, which can be challenging due to the presence of hazardous materials such as asbestos. The set of activities performed by excavators during demolition projects requires the operators to possess unique skills and expertise, as they involve tearing down and removing structures rather than building them up. In addition, demolition activities require specialized equipment and skilled labor, which can increase the cost of the project. Lastly, demolition activities can have a significant impact on the surrounding environment and community, which requires proper communication and consultation with stakeholders. Due to the varied and complex nature of demolition activities, the application of existing construction-focused activity identification models is inadequate. To effectively monitor and evaluate the productivity of demolition projects, it is important to develop specific automated activity identification models for demolition activities.

The focus of this study is to develop an automated vision-based activity identification model for recognizing excavator activities during small-scale demolition

simulations. The proposed model uses video footage to identify three demolition activities (grabbing, swinging, and dumping) through computer vision and deep learning algorithms. In this study, small-scale experiments were designed and implemented, enabling the investigation of demolition activities under various settings. Since demolition projects do not occur as frequently as construction projects, this approach saved substantial time and effort that would have otherwise been required in real-world demolition projects. Such experiments were recorded, labeled, and further used to train the vision-based activity identification model. This model paves the way for the productivity of excavators in demolition projects to be monitored in an automated manner, which facilitates research for improving the efficiency of equipment operations for demolition. The study represents an important step forward in the field and has the potential to impact future demolition projects.

This paper is structured in the following manner: First, a review of the literature on sensor-based and vision-based activity identification models is presented. Afterward, the methodology section elaborates on data collection, data processing, and model development. The results of data collection and model accuracy are reported in the results section, followed by a brief discussion on influencing factors on the model performance. The conclusion presents the limitations of the study and identifies areas for future research.

2 Background

Productivity monitoring and management of heavy equipment are crucial for the success of massive demolition projects. Demolition contractors often prioritize recycling materials as it can provide an additional source of revenue. However, the process of material separation for recovery is time-consuming and can impede sustainable practices [3]. Therefore, it is essential to develop specialized activity identification models to monitor and improve the efficiency of heavy equipment operations in demolition projects. By accurately tracking the activities of heavy equipment during demolition operations, project managers can evaluate productivity, optimize processes, and reduce waste.

In recent years, researchers have focused on developing sensor-based automated activity identification models for construction equipment. Ahn et al. evaluated the feasibility of measuring the operational efficiency of construction equipment using low-cost accelerometers by collecting acceleration data from the real-world operation of excavators and calculating features to classify the operation into engine-off, idling, and working modes [4]. Akhavian et al. investigated the potential of using built-in smartphone sensors (i.e., accelerometer and gyroscope) as multi-modal data collection nodes to detect detailed construction equipment activities [5]. Kim et al. examined the use of inertial measurement units (IMUs) embedded in a smartphone to identify the activities of construction equipment, demonstrating the potential for using smartphone IMUs for continuous and cost-effective activity identification [6].

Vision-based automated activity identification models have recently become the focus of attention to recognize construction activities. For example, Cheng et al. proposed a vision-based excavator activity identification and productivity measurement method using deep learning, which accurately recognized excavator actions and calculated activity times and average cycle times [7]. Another scholarly work developed a low-cost, vision-based method for analyzing excavator productivity in earthmoving tasks using zero-shot learning for activity recognition without pre-training or fine-tuning and has been tested on real construction site videos with high accuracy [8]. Chen et al. proposed a framework for automatically analyzing the activity and productivity of multiple excavators in construction sites using 3D Convolutional Neural Networks (3D CNNs) to detect, track, and recognize activities, which was tested on videos from real construction sites [9].

Sensor-based and vision-based models are both popular in construction activity identification, but each has its own advantages and disadvantages. Sensor-based models offer real-time data on equipment activities using sensors attached to the equipment body, but they can be costly and have limitations in data collection and accuracy, placement and calibration of sensors, and capturing certain activities [2]. Vision-based models use video footage for a comprehensive view of construction sites, but can be affected by lighting and occlusions, and have lower accuracy in dynamic environments. Despite these limitations, vision-based models are generally preferred because they are less intrusive, more flexible, cost-effective, and versatile for identifying a wider range of activities [10, 11].

3 Data Collection and Methodology

This section provides details about the data collection and model development processes (Fig. 1). As shown in Fig. 1, there are three primary phases in the developed framework: data collection and labeling, data processing, and model development. The first phase includes designing and performing small-scale experiments simulating demolition operations, which are recorded, labeled, and divided into video clips, each showing a single activity. Several data processing approaches were then applied to video clips to improve data quality and reduce noise. A 3D-CNN-based model was developed, trained, and evaluated using the processed video clips. Following sub-sections discuss the details of each step.

3.1 Data Collection and Labeling

In this study, small-scale experiments were designed and conducted to simulate real-world demolition operations. In these experiments, undergraduate students were asked to operate a remote-control excavator to perform a real-world demolition operation (i.e., separating different types of material from a mix). Material separation is

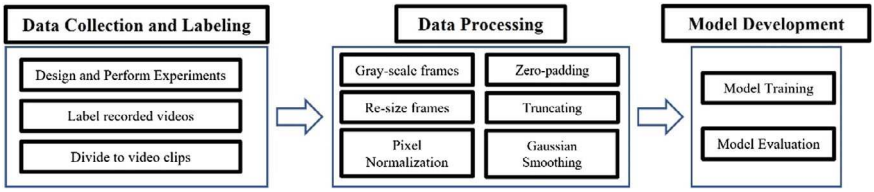

Fig. 1 Overview of data collection and labeling, data processing, and model development

crucial in demolition operations as it directly impacts the efficiency of the process. Proper material separation enables the recycling of valuable resources, reduces the environmental impact of demolition, and saves time and money by reducing the costs associated with waste disposal [12]. While conducting the experiments, two video cameras were used to record each experiment. Figure 2 shows the experiment settings, including the excavator, the material mix, two video cameras, and a student performing the experiment.

Figure 3 shows the excavator and video cameras used for small-scale experiments. The 20-lb excavator has a maximum carrying capacity of 180 Lbs and a digging power of 1.1 Lbs per cubic inch with a motor power of 110 Lbs. It comes with various attachments such as a fork, shovel, and drill, and has 2000 mAh battery providing 45 to 50 min of work time. Two Logitech C922 HD PRO webcams were used to record experiments. Each webcam has a Full HD 1080p video recording resolution at 30 frames per second with a 78° field of view.

Figure 4 shows three demolition activities (i.e., grabbing, swinging, and dumping), based on which undergraduate students labeled recorded videos by assigning a label to each activity and noting the start and end time in a spreadsheet. The labeling

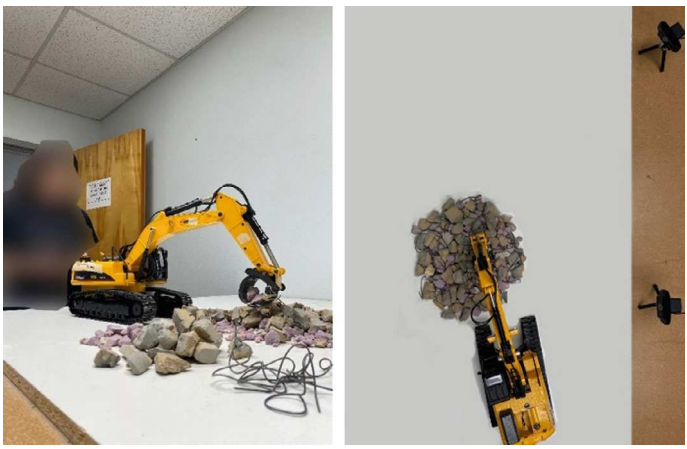

Fig. 2 An undergraduate student performing a small-scale experiment (left) and the face-on view of the experiment setting (Right)

Fig. 3 Top race TR-211M full functional remote-control excavator (left) and the Logitech C922 HD PRO Webcam (Right)

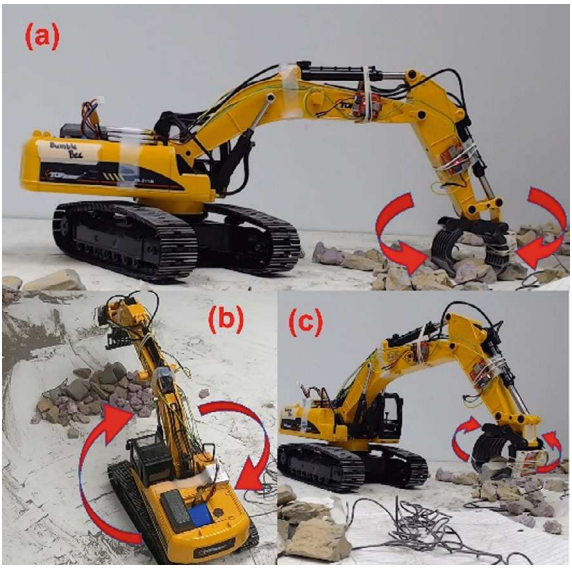

Fig. 4 Demolition activities: **a** grabbing, **b** swinging, and **c** dumping

process was supervised by graduate students to ensure validity. Recorded videos were then broken down into shorter clips, each showing a particular activity with a corresponding activity label (Fig. 1).

3.2 Data Processing

Preprocessing video data is vital for enhancing data quality and reliability, which facilitates accurate analysis and meaningful information extraction. In this study, we used a few video processing techniques, such as gray scaling, frame resizing, pixel normalization, Gaussian smoothing, zero-padding, and truncating.

Gray scaling converts colored images into grayscale, while frame resizing changes the size of video frames for faster processing or better visual output. Pixel normalization adjusts brightness and contrast by transforming pixel values to a fixed range, while Gaussian smoothing reduces noise and unwanted details. Zero-padding and truncating handle data of varying lengths, ensuring consistent dimensions for easier AI model processing. These pre-processing techniques can be used in combination for more effective handling of data [13, 14].

3.3 Model Development

3D CNNs are powerful tools for video classification as they capture spatial and temporal features and process video data in three dimensions, leading to a better understanding of relationships in video sequences. They outperform 2D CNNs and hand-crafted features and are designed to handle high-dimensional video data. Recent studies emphasize the importance of leveraging both spatial and temporal information for accurate video analysis and demonstrate the superiority of 3D CNNs in classification tasks [15, 16].

3D Convolutional and Pooling layers extract and downsize important features from 3D input data in a Convolutional Neural Network. The 3D Convolutional layers learn local features, while the 3D Pooling layers preserve only the most important ones, allowing the network to effectively identify and use relevant features for the task at hand. Repeated use of these layers enables the network to learn increasingly complex features at different levels of abstraction, improving overall performance [17].

Figure 5 shows the model architecture developed and used to identify excavator activities (i.e., grabbing, swinging, and dumping) performing demolition activities. The model consists of 3D Convolutional layers, followed by 3D pooling layers and dropout layers. It extracts spatial and temporal features present in the input video data. Extracted features were then fed into fully connected layers to predict the output.

Model Training. Consideration of hyperparameters is crucial before training deep learning models. Important hyperparameters include batch size, epochs, learning rate, optimizer, loss function, and activation function. The batch size affects generalization ability and computation time, while the number of epochs defines the number of times the dataset is used to train the model. The learning rate controls the magnitude of updates, with low rates leading to slow convergence and high rates leading to instability. The optimizer adjusts weights and biases based on loss function gradients, which is a measure of difference between predicted and true outputs. Activation functions determine neuron activation or output and contribute to final prediction. A well-chosen combination of hyperparameters leads to a well-trained model [18]. (See Table 1 for values of discussed hyperparameters).

Model Evaluation. Before training the model, 85 percent of video clips were used to train the model, while the rest were used for evaluating the model performance.

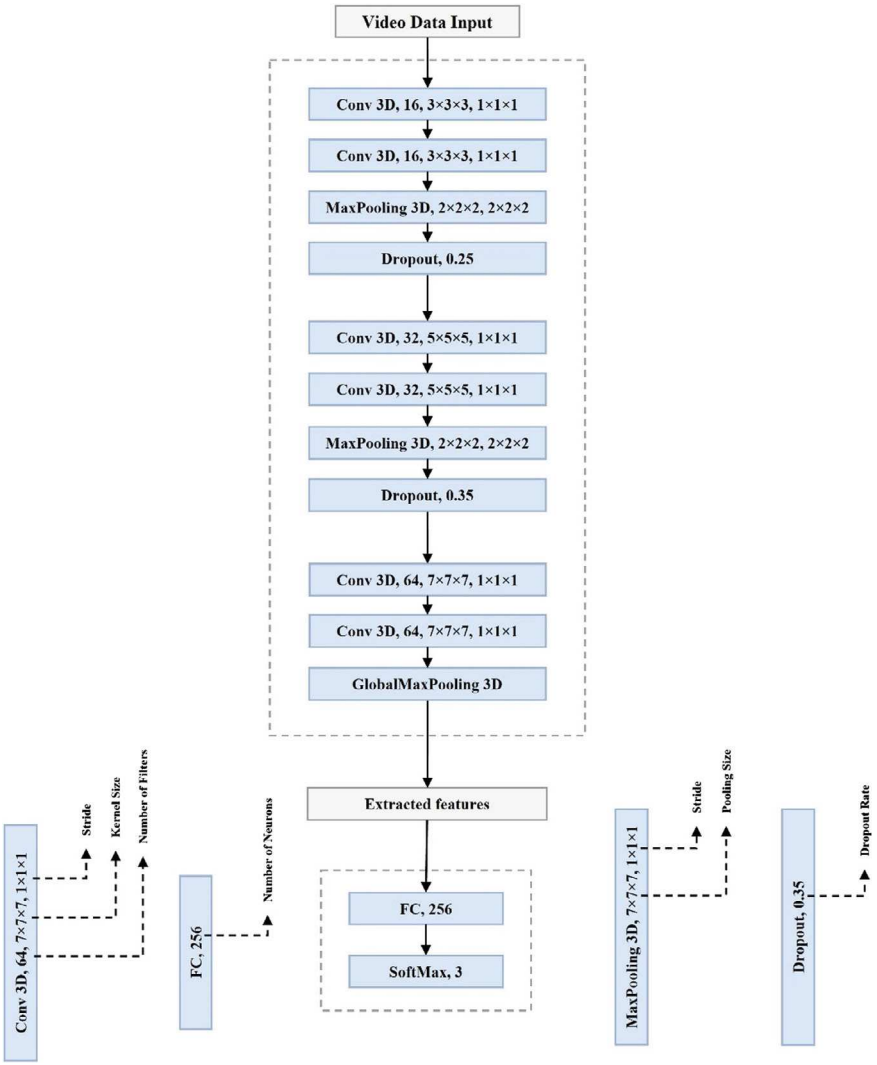

Fig. 5 Model architecture

The model has been developed using TensorFlow 2.10.0 and Python 3.10.0. The training has been performed with a personal computer that had an AMD Ryzen Threadripper PRO 3995WX processor with 64 cores @ 2.70 GHz, 256 GB of DDR4 RAM, and one NVIDIA RTX A4000 GPU.

Table 1 Selected hyperparameters and their corresponding values

Hyperparameter	Value
Learning rate	0.0001
Batch size	64
Epochs	50
Optimizer	Adam
Loss function	Categorical cross-entropy
Activation function	ReLU[a]
Metric	Accuracy

[a] Expect for the last layer, which has a SoftMax activation function to make a prediction

4 Results

This section presents the results of data collection efforts and the model development process.

4.1 Data Collection and Data Processing

The total duration of recorded video with both webcams from the experiments was 45 min and 51 s and, considering 30 frames per second, consisting of 82,535 frames.

As mentioned earlier, recorded videos were labeled and divided into shorter video clips, each presenting a single activity. Table 2 shows the distribution of video clips across different activities (e.g., 345 swinging activities were label). Overall, the recorded videos were divided into 889 video clips, of which 755 were used for training and 134 were used for evaluating model performance.

The lengths of video clips (Fig. 6) are observed to be significantly different, but model training requires video clips with equal lengths. Therefore, a threshold was determined, and zero-padding and truncating have been used at the same time to add zeros (i.e., black frames) at the end of shorter video clips (i.e., those with fewer frames than the threshold) and remove additional frames from longer video clips (i.e., those with a larger number of frames than the threshold). The mean of video clip durations is 2.45 s, which is used as the threshold for zero-padding and truncation. The red line shows the threshold in Fig. 6.

Table 2 Number of video clips for each activity

Activity	Grabbing	Swinging	Dumping
Total number of video clips	291	345	253
Number training video clips	252	291	212
Number of test video clips	39	54	41

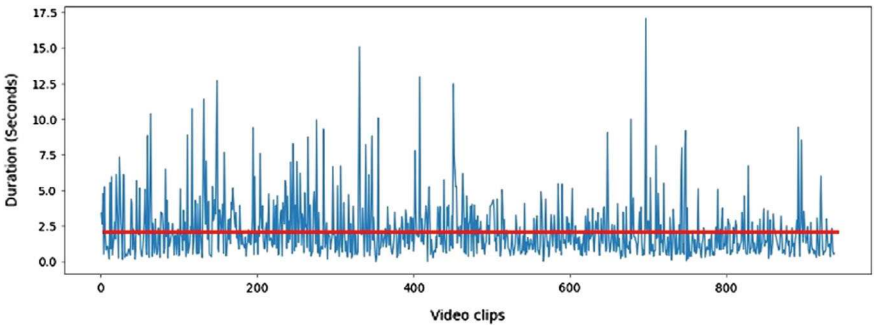

Fig. 6 Duration of video clips

4.2 Model Development

After processing the video clips and ensuring that they all had equal length, the model was developed using the hyperparameters discussed in Sect. 3.3. The model achieved a training accuracy of 0.85 and a test accuracy of 0.78. To determine if the model was properly trained, learning curves were employed. In the learning curve, a large gap between the training and test accuracies with relatively low scores for both indicates underfitting, while continued decrease in the training error and a decrease in the test error followed by an increase indicates overfitting. Overfitting implies that the model may have learned the training data too well and may not generalize well to new data.

The learning curves of the model are shown in Fig. 7, which illustrate the accuracy and loss values of the model at each epoch. It can be observed that both the training and test accuracy have been consistently increasing, while the training and test error have been decreasing. This indicates that the model was able to successfully learn patterns within the training data and was able to generalize well to the test data. However, after the 35th epoch, the model started to overfit as the gap between the training and validation accuracy and the test error began to increase.

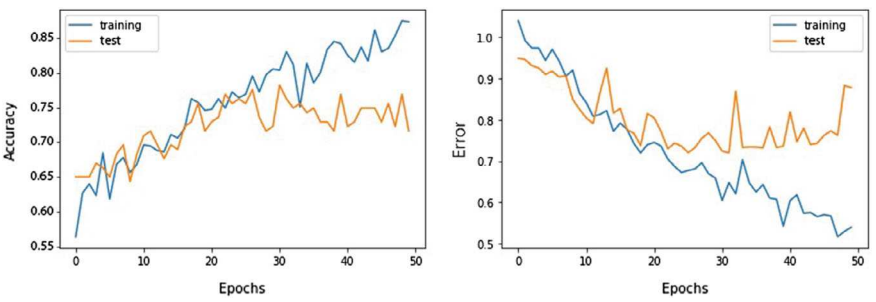

Fig. 7 Model learning curves

Table 3 Confusion matrix for model predictions

	Grabbing (Predicted)	Swinging (Predicted)	Dumping (Predicted)
Grabbing (True)	30	6	3
Swinging (True)	6	43	5
Dumping (True)	2	7	32

Table 3 shows the confusion matrix for the model predictions. Each row represents the number of test instances in a true class, and each column represents the number of instances in a predicted class.

To compute the class-wise accuracy, the number of true positives for each activity is divided by the total number of video clips for that activity. Therefore, the model accuracy in predicting grabbing, swinging, and dumping activities are 0.77, 0.79, and 0.78, respectively.

5 Discussion

The objective of our study was to develop an automated vision-based activity identification model for excavator activities during demolition operations, which differ from construction operations in terms of their complexity and variability. Traditional approaches for tracking heavy equipment productivity during demolition operations are costly, time-consuming, and prone to error. Therefore, we proposed an activity identification model that utilizes computer vision and deep learning algorithms to recognize three demolition activities (i.e., grabbing, swinging, and dumping) based on video footage.

We found that the proposed model achieved an overall accuracy of 78%, which demonstrates its potential for identifying excavator activities during demolition operations. However, we also observed some misclassifications, which can be attributed to the overlapping nature of the activities during material separation. For example, the grabbing activity may have been misclassified with swinging due to the simultaneous rotation of the excavator's body and bucket's closure during these activities. Similarly, the swinging activity may have been misidentified with grabbing and dumping due to the overlap between these activities, as the bucket may still be in motion while the excavator is swinging, leading to confusion between the three activities. Additionally, the grabbing activity may have been confused with the dumping activity, as these activities only differ in the opening and closing of the excavator's bucket.

Despite these limitations, our proposed model has several practical implications for real-world demolition projects. It can improve the accuracy and efficiency of identifying heavy equipment activities related to material separation during demolition operations, which will enable project managers to monitor and evaluate the productivity of demolition projects. Additionally, it can reduce the time and labor required to track and manually label heavy equipment activities, allowing workers

to focus on more critical tasks. Moreover, it can enhance safety on demolition sites by providing an automated and accurate monitoring system for heavy equipment activities, thus reducing the risk of accidents. Lastly, it can support the development of effective demolition plans and strategies by providing real-time data on equipment productivity, which can inform decision-making and improve project outcomes.

6 Conclusion and Future Research

The success of demolition projects relies heavily on the efficiency of heavy equipment operations, which have traditionally been tracked manually. Such a manual tracking approach is time-consuming, labor-intensive, and prone to errors. Recent studies have explored the use of artificial intelligence to automate this process, but the complex and dynamic nature of demolition activities presents challenges in accurately identifying heavy equipment activities. This study proposes a 3D CNN-based model that can automatically identify excavator activities related to material separation during demolition operations. The model extracts spatial and temporal features simultaneously and was trained using manually labeled video datasets of small-scale demolition experiments. The study validates the feasibility of the vision-based activity identification model, which has the potential to monitor the productivity of real-world demolition projects.

While providing an important contribution to demolition activity identification models, this study has certain limitations, including the limited set of materials used in small-scale experiments, the limited number of activities investigated, and the controlled nature of the experiments. Future research efforts could explore a dual-stream approach, using more cameras and evaluating their placement, incorporating other modules such as detection and tracking, and conducting hyperparameter optimization to improve the model's accuracy and robustness. The proposed vision-based activity identification model proposed in this study has several practical applications for real-world demolition projects, such as improving productivity monitoring, reducing time and labor costs, supporting effective demolition planning, and facilitating research for improving equipment efficiency. Its development represents an important step forward in the field and has the potential to positively impact the construction industry and society.

References

1. Thomsen FS, Kohler N (2011) Deconstruction, demolition and destruction. Build Res Inf 39(4):327–332. https://doi.org/10.1080/09613218.2011.585785
2. Sherafat B et al (2020) Automated methods for activity recognition of construction workers and equipment: state-of-the-art review. J Constr Eng Manage 146(6):03120002. https://doi.org/10.1061/(asce)co.1943-7862.0001843

3. Jalloul H, Pinto A, Choi J (2022) A pre-demolition planning framework to balance recyclability and productivity, pp 892–901. https://doi.org/10.1061/9780784483978.091
4. Ahn CR, Lee S, Peña-Mora F (2013) Application of low-cost accelerometers for measuring the operational efficiency of a construction equipment fleet. J Comput Civ Eng 29(2):04014042. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000337
5. Akhavian R, Behzadan AH (2015) Construction equipment activity recognition for simulation input modeling using mobile sensors and machine learning classifiers. Adv Eng Inform 29(4):867–877. https://doi.org/10.1016/J.AEI.2015.03.001
6. Kim H, Ahn CR, Engelhaupt D, Lee SH (2018) Application of dynamic time warping to the recognition of mixed equipment activities in cycle time measurement. Autom Constr 87:225–234. https://doi.org/10.1016/J.AUTCON.2017.12.014
7. Cheng MY, Cao MT, Nuralim CK (2022) Computer vision-based deep learning for supervising excavator operations and measuring real-time earthwork productivity. J Supercomput 79(4):4468–4492. https://doi.org/10.1007/S11227-022-04803-X/TABLES/11
8. Chen C, Xiao B, Zhang Y, Zhu Z (2023) Automatic vision-based calculation of excavator earthmoving productivity using zero-shot learning activity recognition. Autom Constr 146:104702. https://doi.org/10.1016/J.AUTCON.2022.104702
9. Chen C, Zhu Z, Hammad A (2020) Automated excavators activity recognition and productivity analysis from construction site surveillance videos. Autom Constr 110:103045. https://doi.org/10.1016/j.autcon.2019.103045
10. Roberts D, Golparvar-Fard M (2019) End-to-end vision-based detection, tracking and activity analysis of earthmoving equipment filmed at ground level. Autom Constr 105:102811. https://doi.org/10.1016/j.autcon.2019.04.006
11. Kim J (2020) Visual analytics for operation-level construction monitoring and documentation: state-of-the-art technologies, research challenges, and future directions. Front Built Environ 6:202. Frontiers Media S.A. https://doi.org/10.3389/fbuil.2020.575738
12. Yeheyis M, Hewage K, Alam MS, Eskicioglu C, Sadiq R (2013) An overview of construction and demolition waste management in Canada: a lifecycle analysis approach to sustainability. Clean Technol Environ Policy 15(1):81–91. Springer. https://doi.org/10.1007/s10098-012-0481-6
13. Sharma V, Gupta M, Kumar A, Mishra D (2021) Video processing using deep learning techniques: a systematic literature review. IEEE Access 9:139489–139507. https://doi.org/10.1109/ACCESS.2021.3118541
14. Kang B (2007) A review on image and video processing. Int J Multimedia Ubiquitous Eng 2(2):49–64
15. Kurmanji M, Ghaderi F (2019) A comparison of 2D and 3D convolutional neural networks for hand gesture recognition from RGB-D data. In: ICEE 2019—27th Iranian conference on electrical engineering, pp 2022–2027. https://doi.org/10.1109/IRANIANCEE.2019.8786671
16. Hara K, Kataoka H, Satoh Y (2017) Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp 6546–6555. https://doi.org/10.48550/arxiv.1711.09577
17. Diba A et al (2017) Temporal 3D ConvNets: new architecture and transfer learning for video classification. https://doi.org/10.48550/arxiv.1711.08200
18. Yu T, Zhu H (2020) Hyper-parameter optimization: a review of algorithms and applications. https://doi.org/10.48550/arxiv.2003.05689