applied sciences

MDPI

Article

Vision-Based Activity Classification of Excavators by Bidirectional LSTM

In-Sup Kim [1], Kamran Latif [1], Jeonghwan Kim [2], Abubakar Sharafat [3], Dong-Eun Lee [3,*,†] and Jongwon Seo [1,*,†]

[1] Department of Civil & Environmental Engineering, Hanyang University, Seoul 04763, Republic of Korea
[2] Department of Civil Engineering, Korea National University of Transportation, Chungbuk 27469, Republic of Korea
[3] School of Architectural, Civil, Environment and Energy Engineering, Kyungpook National University, Daegu 41566, Republic of Korea
* Correspondence: dolee@knu.ac.kr (D.-E.L.); jseo@hanyang.ac.kr (J.S.)
† These authors contributed equally to this work.

Abstract: Advancements in deep learning and vision-based activity recognition development have significantly improved the safety, continuous monitoring, productivity, and cost of the earthwork site. The construction industry has adopted the CNN and RNN models to classify the different activities of construction equipment and automate the construction operations. However, the currently available methods in the industry classify the activities based on the visual information of current frames. To date, the adjacent visual information of current frames has not been simultaneously examined to recognize the activity in the construction industry. This paper proposes a novel methodology to classify the activities of the excavator by processing the visual information of video frames adjacent to the current frame. This paper follows the CNN-BiLSTM standard deep learning pipeline for excavator activity recognition. First, the pre-trained CNN model extracted the sequential pattern of visual features from the video frames. Then BiLSTM classified the different activities of the excavator by analyzing the output of the pre-trained convolutional neural network. The forward and backward LSTM layers stacked on help the algorithm compute the output by considering previous and upcoming frames' visual information. Experimental results have shown the average precision and recall to be 87.5% and 88.52%, respectively.

Keywords: computer vision; activity recognition; convolution neural network (CNN); long short-term memory (LSTM); Googlenet; visual features

Citation: Kim, I.-S.; Latif, K.; Kim, J.; Sharafat, A.; Lee, D.-E.; Seo, J. Vision-Based Activity Classification of Excavators by Bidirectional LSTM. *Appl. Sci.* **2023**, *13*, 272. https://doi.org/10.3390/app13010272

Academic Editor: Oscar Reinoso García

Received: 15 November 2022
Revised: 18 December 2022
Accepted: 20 December 2022
Published: 26 December 2022

1. Introduction

There are a lot of factors affecting the working efficiency of construction projects, such as human factors, equipment maintenance, weather conditions, variable route plans and road conditions [1]. The site manager has to decide the key parameters affecting the working efficiency of a construction project. For that, they are required to have accurate, efficient, and cost-effective methods to meet the requirements of the construction project [2,3]. The site manager analyzes the working efficiency, productivity, optimum cost, and time required for an earthwork operation based on the information on construction equipment [4,5]. The completion time of an earthwork project depends upon the cycle time of the excavation, transportation time, and the amount of the soil needed to be transferred [6]. Understanding the construction equipment's working status and the effect of different environmental conditions on the construction process helps the manager minimize the idle time and non-value-added works to enhance the productivity of the construction process [7]. The complexity of the working schedule, the number of input variables, and the risk of decision variables increase with the increase in the size of a construction project [8]. Irrespective of the size of a construction project, the excavator remains the most commonly used equipment

to carry out multiple sorts of operations in varying conditions [9]. It is very important to continuously monitor and recognize the excavator's different activities such as dumping, hauling, swinging, moving, or stopping. However, these tasks cannot be achieved by traditional ways of monitoring and analyzing the construction cycles because they are a time-consuming and labor-intensive process, and prone to error [10]. Vision-based methods provide ease in the remote, accurate and continuous monitoring of construction equipment in real-time [11].

Cheap and high-resolution cameras, extensive data storage capacity, and the availability of the internet render the applicability of vision-based methods easy for the monitoring of construction sites continuously [10]. Vision-based artificial neural networks have been used for the detection of objects such as Histograms of oriented gradients (HOG)- classifier-based equipment detection, faster R-CNN-based detection of non-hardhat-using workers in construction sites [12], and Faster R-CNN with enhanced VGG-16 for object detection in optical remote sensing images [13]. CNN-based LSTM has been used for the detection of unsafe behavior of construction workers [14]. Additionally, computer vision has been used for pose estimation in construction sites [9], using a Cascaded Pyramid Network [15], a Stacked Hourglass Network and a method based on the integration of both for the pose estimation of construction equipment [16].

Similarly, the following vision-based algorithms have been reported for activity recognition in the construction industry such as histograms of oriented gradients (HOG) integrated with multi-class Support Vector Machines (SVM) [10], Bag-of-Video-Feature-Words integrated with the Bayesian learning model [17], 3D ResNet [11], and convolutional neural networks (CNN) [18]. A CNN is designed to exploit "spatial correlation" in data, whereas long short-term memory (LSTM) is designed to process and make predictions given sequences of data. Furthermore, a CNN model can be integrated with LSTM by providing spatial correlation of input data as sequenced data to the LSTM model. Long-term Recurrent Convolutional Network (LRCN) on the UCF101 dataset, [19], CNN-based LSTM model for unsafe human actions [14], and CNN-based double-layer long short-term memory (CNN-DLSTM) for the excavator [6] have been applied for the activity classification. These studies have not considered the pre-and post-frame while categorizing the individual frame of the scene. It is important to understand the scene with a context; a frame can be interpreted as both a return to a digging area and non-value-added swinging unless the context of the task performed is given. To better understand the scene it is thus necessary to consider multiple frames. The BiLSTM provides an opportunity to categorize the input information at any time t considering the information at time $t - 1$ and $t + 1$. CNN-based deep bidirectional long short-term memory (CNN-DBLSTM) was used for human action recognition on UCF-101, YouTube 11 Actions, and HMDB51 [20]. Based on the above literature, a pre-trained CNN model "Googlenet" and bidirectional long short-term memory (BiLSTM) were not used for the activity recognition of construction equipment. Therefore, the authors have investigated the integrated method based on a pre-trained CNN, Googlenet and a bidirectional long short-term memory (BiLSTM) for the activity classification of the excavator.

This study implemented a pre-trained convolutional neural network "Googlenet" with the recurrent neural network "LSTM" for classification based on the sequential pattern of video frames. The goal of this study is to practically demonstrate the applicability of Googlenet for feature extraction in the field of the construction industry. Additionally, the application of BiLSTM for the different activities of the excavator and its response is observed. This research's main objective is to evaluate the effectiveness of a pre-trained CNN-based bidirectional long short-term memory (CNN-BiLSTM) for the activity recognition of an excavator in a construction site. Excavators are the most common equipment for earthwork and construction sites with a cycle of activities: excavation, hauling, dumping, and swinging. The surveillance videos of the excavator are used as input to a pre-trained convolutional neural network to learn the sequential pattern of video frames. Based on the findings of the pre-trained model, BiLSTM classifies the different activities of the excavator.

The rest of this paper is organized as follows. The first section of this paper will identify the key information relevant to the vision-based activity classification and provide an overview of the existing studies on vision-based activity recognition in the next section. The next section will then elaborate on the architecture of the proposed research following by the training of the model with the video input of the earthwork site in successive sections. The model is trained with the video input of the construction site to validate the effectiveness of the framework. The experimental results will be analyzed in the second-to-last section, and finally, the last section will provide the conclusion and the research contribution with the future direction.

Related Works

Many efforts have been made to recognize the activities of construction equipment with the help of vision-based techniques. Recently, different vision-based techniques have been utilized to investigate the productivity, safety, cost and automation of the construction process. Surveillance cameras have been installed to record the activities at construction sites [21]. Cameras should be placed at a high and appropriate locations so that the occlusion may be minimal [22]. Visual data collected from these surveillance cameras is used for the activity recognition and safety of the workers and construction equipment such as excavators, dump trucks and dozers.

There are numerous vision-based methods available for the detection of objects, humans or construction equipment. SURF is a local feature detector and descriptor of the interesting points used to construct the feature vector. Additionally, SVM classifies different classes by constructing hyperplanes in a multidimensional space. A technique based on the Speeded-up Robust Features (SURF) and Support Vector Machine (SVM) was presented for facial recognition on Yalefaces and the UMIST dataset with an accuracy of 97.78% and 97.87%, respectively [23]. A part-based object recognition model was presented to detect the excavator at different poses using a discriminately trained HOG classifier [9]. A convolution neural network IFaster R-CNN method, consisting of a Region Proposal Network (RPN) and an R-CNN, was presented to detect workers and excavators in real-time [24]. The accuracy of the IFaster R-CNN model to detect the workers and the excavator was 91% and 95%, respectively. A deep learning method, Faster R-CNN, was used to detect non-hardhat-use at construction sites by surveillance videos [12]. The precision and recall for the test dataset were 95.7% and 94.9%, respectively.

To date, various methods have been developed and introduced to detect construction equipment and recognize the activities of excavators in earthwork sites. Motion-feature-based activity recognition methods extract the features from the consecutive frames of the video and convert the spatial and temporal information of the features into feature vectors. A computer-vision-based algorithm, Support vector machine (SVM), was presented to classify the single action of the excavator using spatiotemporal visual features [10]. The average accuracy of action recognition for the excavator and dump truck was 86.33% and 98.33%, respectively. Similar work was presented by 3D-Haris and local histograms to extract the features, and a Bayesian network was used to classify the excavator activities (relocating, excavating, swinging) instead of a Support vector machine (SVM) [17].

Computer vision and modern pattern recognition algorithms have outperformed the traditional approaches (support vector machines, linear regression, and Bayesian networks) in the field of image classification [25], Object detection [26], and action recognition [5,18]. The complex relationship of input and output classes of dynamic processes is solvable with sophisticated neural networks such as a CNN and recurrent neural network (RNN). CNN-based methods were found to be the best method for detecting the knife among the Bag of Words (BOWs), Hog-SVMs, pre-trained Alexnet and SVMs, and CNNs for safety purposes [26]. It was also found that Alexnet with an SVM provides the best accuracy and the time for training the algorithm was much higher for the CNN; however, the predicted time is higher for Alexnet than the CNN. In another study, a pre-trained CNN network VGG-16 was used to integrate the RGB, optical flow, and grey stream input

data collected from 12 different construction sites in Wuhan city to automatically identify the activities (walking, transporting, and steel bending) of the worker accurately, with a precision of 91%, 92%, and 100%, respectively [18]. For the fine-tuning of the RGB input data in this study, the dropout ratio was set at 0.9 to avoid overfitting. The overall accuracy and recall were 85% and 100%, respectively, to detect and classify the worker's activity. Furthermore, a framework (R-CNN, SORT, 3D ResNet) was presented to recognize the activities automatically (digging, loading, and swinging) and estimate the productivity of the excavator [11]. The average accuracy of activity recognition for the input data collected from the 21 construction sites was 87.6%, and the precision and recall for the digging, loading and swinging was 95% and 86%, 86% and 93%, and 84% and 80%, respectively. The accuracy of the productivity calculation was 83%. CNN has been used widely for the detection of objects and the classification of activities based on this literature. Pre-trained CNNs have also been reported for feature extraction, such as Alexnet, VGG-16, and 3D ResNet. However, the application of Googlenet has not been reported yet in the field of the construction industry, providing a potential application of Googlenet for the extraction of features.

Recurrent neural networks (RNNs) are well considered for time series classification problems, especially when assisted with convolutional neural networks (CNNs)/pre-trained CNNs. Spatial correlations of visual features extracted by CNNs are provided to the RNN models, and, hence, an integrated system of CNN and RNN can provide better performance. LSTM is one of the RNN models and has been reportedly used for time series classification problems. A comparison of the behavioral analysis between the LSTM and BiLSTM was conducted to evaluate their structural differences, effect on accuracy, and time required to reach equilibrium [27]. It was noted that the accuracy of the BiLSTM has increased and reduced errors by 37.78%. However, the time required to reach equilibrium was more for the BiLSTM. Furthermore, various studies have investigated the effect of BilSTM in other fields. A CNN- and Recurrent neural network (RNN)-based algorithm was presented to classify the emotion of the Word2vec database [28]. In the mentioned study, a hybrid model based on CNNs and BiLSTM was compared with standalone CNNs, LSTM, and BiLSTM. CNN-BILSTM models performed better than all standalone models with a precision, recall, and accuracy of 94.3%, 94.6%, and 94.2%, respectively. A bidirectional dilated LSTM (BiDLSTM)-based emotion classification method was presented for two datasets of tweets: WASSA Implicit Emotion Shared Tasks (IEST) and a new dataset Ekman's Emotion keyword (EEK). Both methods' accuracy was 72.83% and 80.79%, respectively [29]. A BiLSTM-based model was presented for sentiment classifications of a SemEval 2013 and IMDB movie review dataset with an accuracy of 85.02% and 95%, respectively [30].

A vision-based action recognition framework based on R-CNN and DLSTM was presented to classify actions based on the sequential pattern of earthmoving excavators [6]. In this study, R-CNN extracted the visual features and provided them to the first layer of LSTM to analyze the sequential pattern. The interim results of the first LSTM layer were analyzed again in the second layer to classify the activities of the excavator. The average precision and recall for all 3 classes, CNN, CNN_LSTM, and CNN-DLSTM, were 77.5% and 71.6%, 87.3.5% and 86.1%, and 90.9% and 89.2%, respectively. Additionally, the accuracy of these classes was 79.8%, 90.9% and 93.8%. The benefit of using BiLSTM over LSTM was the long-term bidirectional dependencies. It considered pre-and post-frames for the categorization of the frame. A convolution neural network (CNN) and deep bidirectional long short-term memory (DB-LSTM)-based method was presented to classify different actions on the UCF-101, HMDB51, and YouTube action video datasets [20]. Convolution neural networks extracted the features and the sequenced features were fed into the long short-term memory (LSTM). The average accuracy for the classification of the UCF-101, HMDB51, and YouTube action video datasets was 91.21%, 87.64%, and 92.84%, respectively. The different algorithms and techniques used in previous studies have been summarized briefly in Table 1.

Table 1. Summary of techniques used in the excavator detection and activity recognitions.

Author	Year	Accuracy	Classifier	Goal
Gong et al. [17]	2011	79	3D-Harris feature + Bayesian learning classifier	Classifying actions of construction workers and equipment
Golparvar et al. [10]	2013	86.33	3D HOG feature + SVM classifier	Activities of excavator and dump truck status
		98.33		
Yang et al. [31]	2016	57	HOG, HOE, MBH features + SVM classifier	Action recognition of construction worker
Luo et al. [32]	2018	80.5	Faster R-CNN detector + Relevance network	Recognizing diverse construction activities
Ding et al. [13]	2018		Faster R-CNN	Reduce the test time and memory requirements, enhanced VGG-16 net precision
Fang et al. [12]	2018	P = 95.7, R = 94.9	Faster R-CNN	Detect non-hardhat-use
Amin Ullah [20]	2017	91.21	CNN-BDLSTM	Novel CNN-BDLSTM method for activity recognition (human activities)
		92.84		
		87.64		
Zhou et al. [33]	2019	up to 99.88	SVM, ANN, Decision tree	Detecting excavator anomalies
Kim et al. [6]	2019	79.8	CNN	Excavator detector
		90.9	CNN-LSTM	Excavator tracking
		93.8	CNN-DLSTM	Excavator activity recognition
Quan Liu et al. [34]	2019	up to 98	Different pre-trained CNN model and Transfer learning	Classification of full/empty-load trucks in earthmoving operations
Chen et al. [11]	2020	87.6	Faster R-CNN + Deep SORT tracker + 3.D ResNet classifier	Excavator detection
				Tracking, activity recognition of excavators
Bhokare et al. [35]	2021	78	YOLOV3	Activity detection and classification
Cheng et al. [36]	2022	99.7	YOWO	Vision-based autonomous excavator productivity
Chen et al. [37]	2022	86	Zero-shot learning method CLIP	Productivity analysis in earthmoving

Several studies have demonstrated the use of bidirectional LSTM for speech, tweet, and human action recognition. However, there has been no attempt to examine the impact of bidirectional LSTM to automate earthwork operations. Furthermore, the application of Googlenet, standalone or integrated with other RNNs, was also not reported well in the construction industry. The goal of this study is to practically demonstrate the application of Googlenet for feature extraction in the field of the construction industry. Additionally, the application of BiLSTM for different activities of excavators and its response is observed. Thirdly, the integrated model of Googlenet and BiLSTM have yet to be adopted for the activity recognition of excavators in the earthworking field. Therefore, this paper considers the potential implications of the integrated CNN-based bidirectional long short-term memory (CNN-BiLSTM) model for the activity recognition of the excavator.

2. Research Framework/Methodology

Excavators are used for performing four basic activities: excavation, hauling, dumping, and returning, usually. These activities are defined as filling the bucket with dirt/rock, moving the filled bucket to the dump point, emptying the bucket into the truck/dump point, and moving back to the excavation point with an empty bucket, respectively. The excavator usually shows two types of sequential patterns. The first type of sequential pattern is one in which the excavator shows its sequential visual features pattern during a single activity. For example, when the excavator is lifting-up during the digging and hauling while the bucket is moving downward. Similarly, for the hauling and swinging, the main body of the excavator rotates, keeping the bucket and arm at a specific speed and position. The other sequential pattern is the operation cycle of the excavator, in which the excavator starts a specific activity and then works in a loop. In most cases, the first activity

of an earthwork operation is "digging" which is followed by the "hauling", "dumping", and "swing". Therefore, most probably, the last activity of the excavator would be the "dumping". The general views of the four activities are shown in Figure 1.

Figure 1. Activities of an excavator operation.

This study classifies the basic activities of excavators based on the sequential pattern of video frames. The pre-trained CNN used in this research is "Googlenet", the winner of the ILSVRC (ImageNet Large Scale Visual recognition Competition) 2014 [38]. Video clips of each activity are fed into a pre-trained CNN to extract the features from video frames. The sequenced feature vectors are the input for the bidirectional long short-term memory (BiLSTM) to classify the activities. The BiLSTM have multiple LSTM layers processing the data in both forward and backward directions simultaneously. The overall research framework is shown in Figure 2.

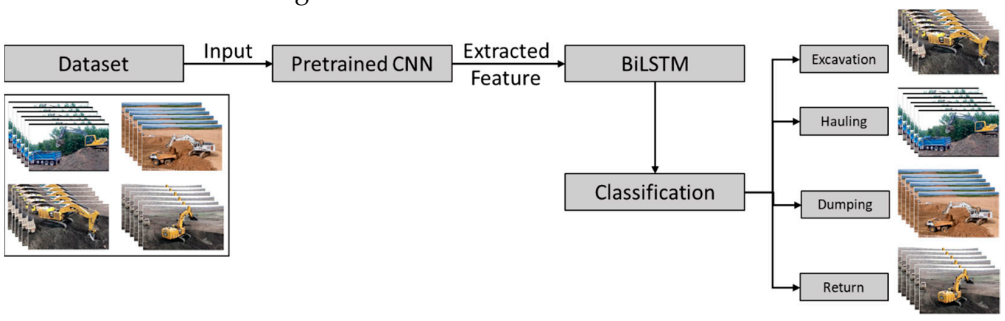

Figure 2. Overview of the research framework.

2.1. Convolutional Neural Network (CNN)

The CNN is a multi-layer architecture that automatically extracts features and facilitates the classifier by mapping the feature vectors [14]. It employs a convolution operation and activation function in the forward propagation phase on the output of the previous layer, as shown in Equation (1), where f is the activation function, b_k is the bias for this feature map, and W^k is the kernel value connected to the kth feature map.

$$H_{ij}^{k} = f\left(\left(W^{k} * x\right)_{ij} + b_{k}\right) \quad (1)$$

The process to recognize the activity type of an excavator by the sequential pattern starts with the extraction of features from the video frames. Video frames are the input for the hybrid CNN-BiLSTM model, and a pre-trained CNN "Googlenet" is used for the extraction of features. The architecture of the "Googlenet" is shown in Figure 3. It consists

of 148 layers. The layers consisted of the convolution layers, max-pooling layers, and inception layers. The CNN network consists of 22 layers of deep architecture which also contain 2 auxiliary layers connected to the output of Inception(4a) and inception(4d) layers. The architecture of 2 auxiliary classifiers consists of a 1 × 1 convolution of 128 filters, fully connected layers with 1025 outputs with Rectified Linear Units (ReLU) activation, a stride of 3, a dropout ratio of 0.7, and an average polling of filter size 5 × 5 and SoftMax classifier. The pre-trained model extracts the sequential patterns of key features from the video frames. The key features of the sequential pattern extracted from the video frames are fed into the BiLSTM model for the sequential analysis.

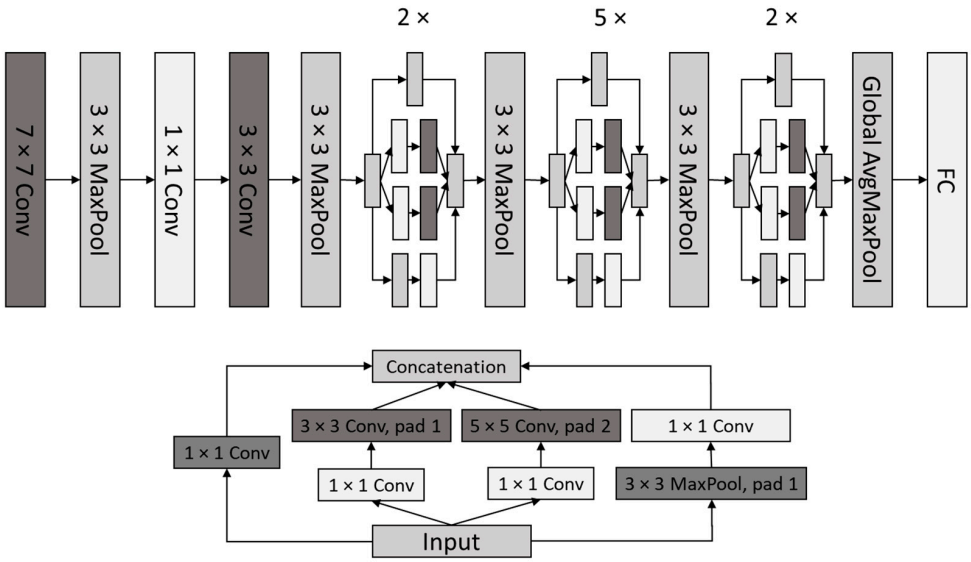

Figure 3. The GoogleNet architecture and Structure of the Inception block.

2.2. Long Short-Term Memory (LSTM)

Recurrent neural networks (RNNS) are the building blocks of neurons that connect the inputs, hidden layers, and outputs and process the selective parts of sequence data by an activation function at a time t. It processes the sequenced data by getting the previously hidden state ht-1 and new input data x_t, and multiplying it with the weights of inputs, adding up the biases as a feed for the activation function [20]. The architecture of an LSTM model consists of an input gate, output gate, memory gate, and a forget gate, as shown in Figure 4. These gates update the information flow through each block of LSTM. These gates include the activation functions such as the sigmoid function and tanh function, and operations such as addition and multiplication. The forget gate decides how much information needs to be kept or discarded. The memory gate decides the amount of information needed to be stored in the cell state, and the input gate updates the previous information as an input to analyze for the output. The output gate provides the output for the current block by analyzing the cell state. The LSTM model decides which information to keep or discard with the help of these gates. Hence, keeping the information of the previous cell/ frame can overcome the problem of the gradient descent.

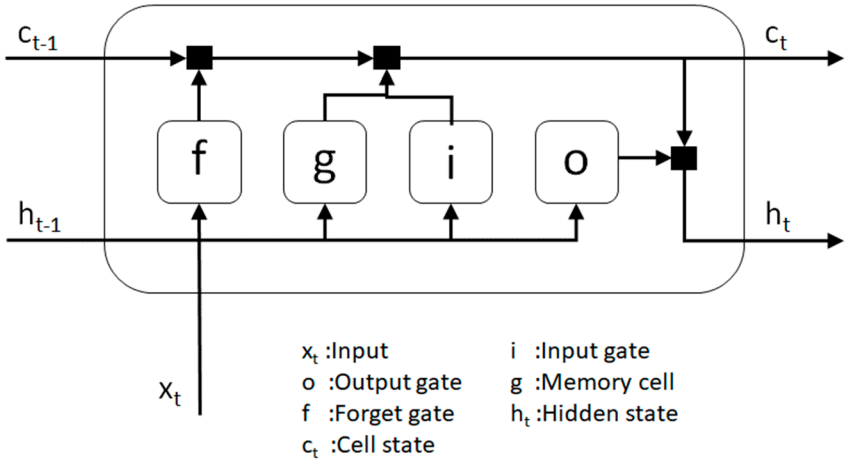

Figure 4. Basic architecture of LSTM cell.

For a time t, the updated cell state can be calculated from Equations (2) to (7) [14]. In these equations, δ is the activation function sigmoid (defined as $\delta(x) = (1 + e^{-x})^{-1}$), ⊗ is the pointwise multiplication operation, i_t, f_t, o_t, g_t, and c_t are the input gate, forget gate, output gate, and cell state, respectively, and Ws and Vs are the coefficient matrixes.

$$i_t = \delta(W_{xi}x_t + V_{hi}h_{t-1} + b_i) \quad (2)$$

$$f_t = \delta(W_{xf}x_t + V_{hf}h_{t-1} + b_f) \quad (3)$$

$$o_t = \delta(W_{xo}x_t + V_{ho}h_{t-1} + b_o) \quad (4)$$

$$g_t = \tanh(W_{xc}x_t + V_{hc}h_{t-1} + b_c) \quad (5)$$

$$c_t = f_t \otimes c_{t-1} + i_t \otimes g_t \quad (6)$$

$$h_t = o_t \otimes \tanh(c_t) \quad (7)$$

In general, the input is fed into a single layer LSTM model for the activation and processing of the output. The multi-layer LSTM models are used to boost the process and get better performance for time series problems. In this way, the layers of LSTM are stacked on each other, and each layer received the hidden state of the previous as an input and processed the frame in the same direction [14]. While in bidirectional LSTM (BiLSTM), the output of the current cell depends upon the previous frames as well as the upcoming frame. In BiLSTM, there are two single LSTM stacked on each other with the reverse direction of information exchange. LSTM and bidirectional LSTM (BiLSTM) layers learn unidirectional and bidirectional long-term dependencies, respectively, between time steps in time series and sequence data. The overview of the CNN-BiLSTM structure is shown in Figure 5. The cell's output is computed based on the hidden state of both forward and backward LSTM layers. The output of a frame at a time t is computed from the two consecutive frames, $t - 1$ and $t + 1$ [28].

2.3. Hybrid (CNN-BiLSTM) Network

LSTM layers learns the key information of time-series data and long-term dependencies between the input frames' time steps and sequence data. The hybrid model of CNNs and LSTM is used to recognize the activities of the excavator where the CNN interprets sequences of input to provide sequence data to the LSTM model to analyze the data using the sequences of the features.

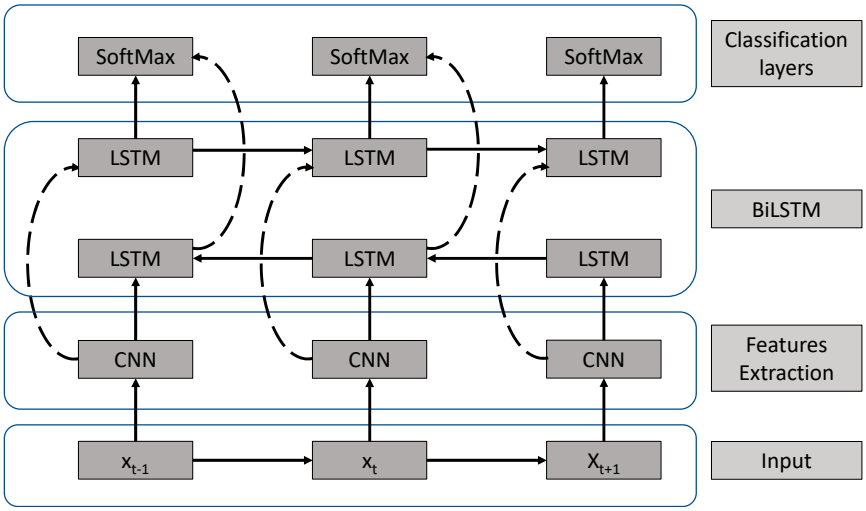

Figure 5. Structure of CNN-BiLSTM.

The sequence input layer provides sequence inputs or time-series data. This data has been converted into batches of image sequences in the sequence-folding layer, then convolution operations on the time steps of image sequences are performed on these batches independently. After the convolution, a sequence-unfolding layer restores the sequence structure of the input data. This data is flattened to collapse the spatial dimensions of the input into the channel (one dimensional) dimension. From this layer, the data is fed into the bidirectional LSTM layer and fully connected layers. After that, a SoftMax function is applied for the classification of the activities. The overall pattern of information flow in CNN-BiLSTM is shown in Figure 6.

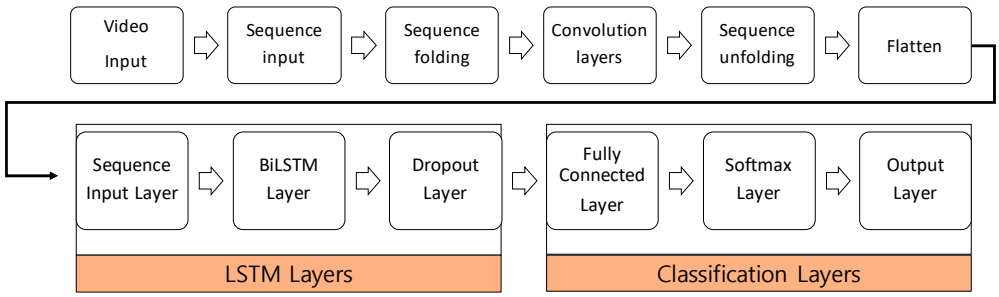

Figure 6. Information flow of CNN-BiLSTM model.

2.4. Standard Deep Learning Pipeline

The authors have followed the standard deep learning model pipeline for n-epochs for the model's training and tuning. The first step of every standard deep learning model pipeline is to fetch the raw data. The raw data in the current study is the YouTube videos of the excavator during earthwork activity at a construction site or earthwork site. The raw data is analyzed for the key parameters and then converted into the required form for data to run in the deep learning model. The input data is sorted into three categories: training, validation, and a testing dataset. The pre-processing of the data is performed very carefully because the availability of noise or outliers will affect the training of the model. The important thing is to check of the features and whether the key features are available in the processed data or not. This study has used the pre-trained CNN model "Googlenet" for the feature extraction. The sorted data is provided to the deep learning network to form a trained model.

The mini-batch is a subset of the input data, processed through all the neurons from the first to the last layer to predict the values at a time in the forward pass. In comparison, the backward pass calculates errors in the predicted and ground truth values to get the

best-fit hyperparameters and minimum loss function. The loss function is then utilized to update the weight of the input using the gradient descent approach in backpropagation. Backpropagation includes forward and backward passes to optimize the hyperparameters to minimize the error. Gradient-based optimization algorithms of the neural network decide how much to change the weight and the network's learning rate to reduce the loss. The learning rate, which is 0.0001 in this research, is a hyperparameter responsible for changes in the weight according to the estimated loss. Adam, a first-order gradient-based optimization of the stochastic objective function, was used in this study [39].

The validation data set is a separate dataset which is not used for the learning and training purposes of the model and processes as a forward pass from the first to the last layer. It helps the network to evaluate the effectiveness of the trained model by testing the validation dataset. The trained model processed the test data set, evaluated the important features, and classified or predicted the different categories of the test dataset based on the features extracted in the processing phase. The trained CNN-BiLSTM model in this research analyzed the sequential pattern of visual features in both directions while predicting a specific cell state's output. In total, 2 layers of LSTM stacked on at a time t provide the information from both consecutive time steps $t - 1$ and $t + 1$ to understand the change in the features and make an accurate prediction. In another study, a CNN-based deep bidirectional long short-term memory (CNN-DBLSTM) was used for human action recognition in the UCF-101, YouTube 11 Actions, and HMDB51 [20], and considered every 6th frame for the video sequences for the recognition propose. Meanwhile, this study implemented the pre-trained CNN model "Googlenet" for feature extraction and considered each frame of the video sequence. Furthermore, this study focuses more on the application of this algorithm in earthwork construction industries.

3. Experimental Implementation

3.1. Datasets

The earthwork operation videos were collected from the YouTube of Volvo, Caterpillar, Liebherr, Komatsu, and Hitachi excavators. The dataset consisted of 400 video clips of the 4 activities. Each clip contains only one of the activities of dumping (empty the bucket), hauling (swing with filled bucket), excavation (filling the bucket), and return (swing with empty bucket). In total, 80% of the data were randomly selected for the training purpose. The training data is further divided into two parts: training data and validation data. A total of 80% will be the training data and 20% is for validation purposes. The rest, a 20% data set, was kept as the test set.

Hmdb51 functions were used to label the videos. The size of the video was defined as H-by-W-by-C-by-S, where H, W, C, and S stand for Height, Width, Number of channels, and the number of frames in the video, respectively. These video clips were placed in separate folders with the folder name as the label. The resolution of the video was maintained at 720P for the whole dataset. The frames are RGB images from the video dataset. The color images (RGB) are the combination of the 3 basic colors of red, blue, and green, with each color containing 8 bits, which results in 24 bits for color images. The pixel values of the image frame are the input data for the convolution network.

3.2. Video Classification Training Process

The video input is fed into the convolutional neural network (CNN) to extract the features. A pre-trained CNN network, "Googlenet", is used to extract the feature vectors from the videos. The sequenced feature vectors from the video input are the output from the activation function on the last pooling layer of the convolution neural network "Googlenet". It consists of 148 layers. The layers consist of the convolution layers, max-pooling layers, and inception layers. The size of the convolution filter is kept 7-by-7, and 3-by-3, while max-pooling of 3-by-3 is used. The activation function used here is ReLU. After the convolution and max-pooling layers, the inception layers are used to stack all the processed information at the output. The add-in feature of the Googlenet is the inception layer, which

deals with the convolution of multiscale input frames. The filters and the weight of the Googlenet layers are determined using the error backpropagation. The inception layers in the architecture are 1-by-1, 3-by-3, and 5-by-5 and are convoluted by the information from previous frames layers, respectively, by the application of max-pooling and the ReLU activation function. An average pooling of 7-by-7 is used instead of the max-pooling feature vectors at the end of the architecture.

The minimum batch size for the training of sequences is kept at 16 for each iteration. The sequence length of each batch size has been measured and truncated to an optimum length, 400 in our case. The learning rate is the measure of the change in the weight while training the algorithm. The learning rate varies from 0 to 1, which is an important parameter to tune the algorithm in terms of efficiency and processing time. The processing time will increase by using a shorter learning rate. Additionally, a shorter learning rate will increase the number of epochs and iterations. The initial learning rate is kept constant at 0.0001 with the gradient threshold of 2. The forward and back pass in one iteration computes the output and loss function for that pass. The maximum number of epochs is 30, and the number of iterations per epoch is 15. After passing all data through the neural network, the input data is shuffled for the next epoch. The number of hidden units for the BiLSTM is kept at 2000 to remember the time steps. The dropout probability is set at 0.5, which will truncate the data to half. The learning rate of the algorithm has been selected based on the hit-and-trial methods. The results are shown in Table 2. The learning rate ranges between 0.1 and 0.00001 for the validation of optimum results. The performance of the algorithm is found to be optimal at a learning rate of 0.0001.

Table 2. Validation accuracy at different learning rate.

Learning Rate	0.001	0.0001	0.0005	0.00001
Validation Accuracy	58.73%	93.55%	69.84%	76.20%
Test Accuracy	55%	88%	51%	73%

4. Results and Discussion

4.1. Performance Metrics

Precision, recall, and accuracy are the parameters that describe the correctness of classification problems. Precision measures all the real positive outputs from the positively predicted ones, and recall is the measure of correctly predicted positive outputs from all positive outputs. The formula of the precision, recall and accuracy are shown in Equations (8)–(10), respectively.

$$\text{Precision} = {}^{\text{TP}}/_{\text{FP}+\text{TP}} \tag{8}$$

$$\text{Recall} = {}^{\text{TP}}/_{\text{FP}+\text{FN}} \tag{9}$$

$$\text{Accuracy} = {}^{(\text{TP}+\text{TN})}/_{(\text{TN}+\text{FP}+\text{TP}+\text{FN})} \tag{10}$$

The output that is predicted as positive and is positive is known as true positive (TP), and the output predicted as positive and is negative is known as false positive (FP) and vice versa. There are four activities included in the current study. A confusion matrix to describe all the possible outcomes for the third activity, "hauling" is shown in Table 3.

The outcomes in the diagonal are always true negative except for one. The third row shows the actual count of the "hauling" while the third column shows the predicted count of "hauling". Therefore, all of the which values fall in the third row, except diagonal, are wrongly classified as negative and so are named as "false negative". Similarly, all of the values which fall in the third column, except diagonal, are wrongly classified as positive and so are named as "false positive".

Table 3. Confusion matrix for activity "hauling".

Known	Predicted			
	Dumping	**Excavation**	**Hauling**	**Return**
Dumping	TN	TN	FP	TN
Excavation	TN	TN	FP	TN
Hauling	FN	FN	TP	FN
Swing	TN	TN	FP	TN

4.2. Evaluating Results

The implementation of the algorithm, training and testing of the visual data is carried out by a MATLAB R2020a environment in Window10, a 64-bit operating system with the hardware configuration of the intel(R) Core(TM) i7-7700 CPU @ 3.60 GHz, RAM 8 gigabyte. The training progress of the activity recognition system is showing in Figure 6. The number of epochs for the training process is selected as 30, and each epoch consists of 15 iterations. The training and validation error is updated after every epoch. The loss of information during the training is known as training loss, whereas the loss of a validation set tested by the previously trained neural network is known as validation loss. The loss of training and validation decreases with the increase in training and validation accuracy. The training and validation accuracy increases with the increase in the number of epochs, and the training loss is lower than the validation loss, as shown in Figure 7. Hence, it is clear that the system is reading the sequenced visual feature to classify the activities of the excavator.

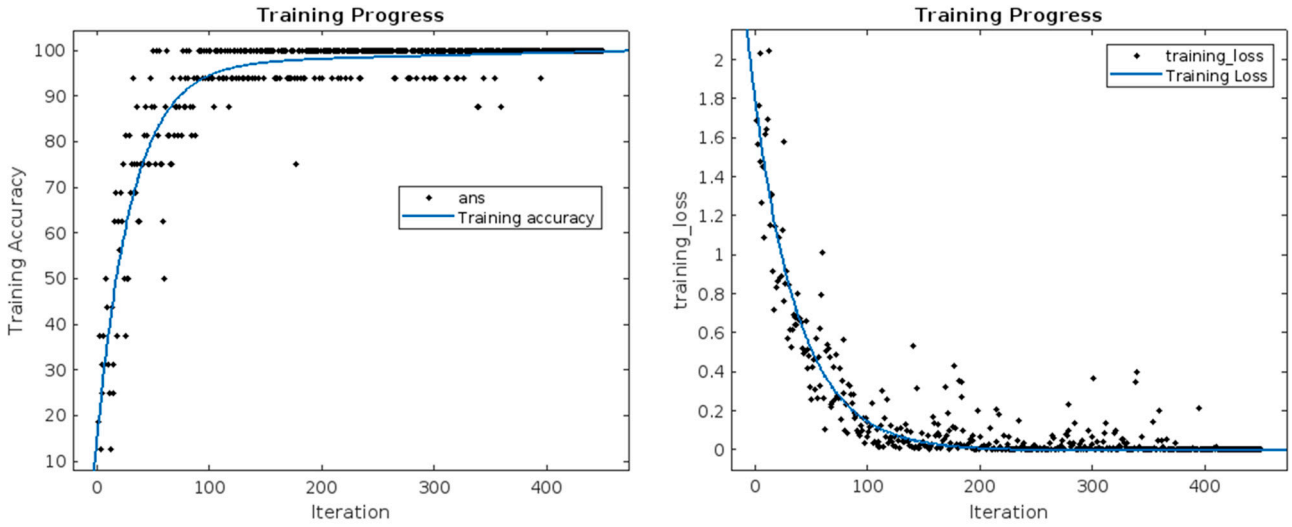

Figure 7. Training progress, and loss of CNN-LSTM network.

The validation accuracy progress throughout the training can be visualized through Figure 7. It can be observed that the validation accuracy increased after each iteration. Similarly, with the increase in the training/validation accuracy, the training loss reduces. The work presented is this study is novel in terms of the integrated application of CNN-BiLSTM for earthwork construction applications. A general approach to evaluate the performance of overall systems can be performed through a Receiver Operating Characteristic (ROC) Curve. A ROC curve demonstrates the true positive rate (TPR, or sensitivity) versus the false positive rate (FPR, or 1-specificity) for different classification scores. Each point on the ROC curve is a pair value of a TPR and TNR. A TPR in one class is a TNR of another class such as the plot between TPRs, and TNR is the same as the graph between 1-TPR and 1-TNR. By reading the plot at any point, the TPR and TNR can be calculated. The ROC curve has been shown in Figure 8 for each activity of the excavator. The behavior of the training performance can be visualized in it.

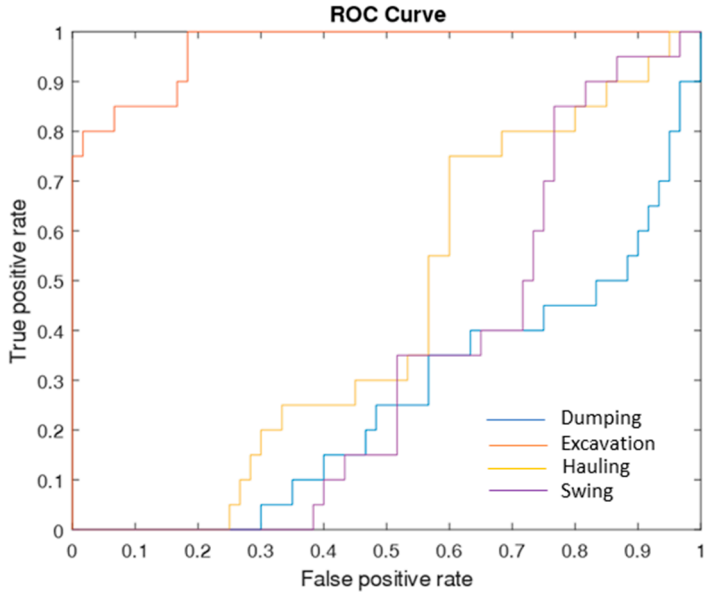

Figure 8. ROC curve for different activities of the excavator.

The validation dataset is part of the training set and takes part in the model building process. However, the validation data is different from the test dataset as well as the training dataset. The validation dataset evaluates the training performance by parameter selection and avoiding overfitting. A good validation accuracy with a good testing accuracy verified the correctness of the model training process. The validation accuracy of the model is calculated after completing the last iteration, as shown in Table 4. Twenty (20)% of the training dataset is used as a validation process, and the video clip is tested by the trained model to test and tune the model. All the clips are classified well in the respected classes. Only three (3) clips were misclassified: one of a swing falling into the dumping category, and two clips of excavations falling into the category of hauling and dumping. The validation precision of the 4 activities, dumping, excavation, hauling, and swinging, is 86.7%, 100%, 94.4%, and 100%, respectively. The overall precision, recall, miss rate and accuracy of the activity recognition system were 95.04%, 95.28%, 4.96%, and 95.2%. The result shows that the second activity, "excavation," misclassified the most among other activities. We obtained a true positive rate (TPR) of above 85% plus for all the attack categories and normal connections. For excavation attacks, there were 0.0% false predictions made by the model.

The performance of the activity recognition algorithm is shown in Table 5. A total of 80 video clips were run through the trained model to recognize each activity. As the test set contains 4 activities, each activity has 20 video clips. From 80 test videos, 70 were correctly recognized and the rest were misclassified. The model identified 18 out of 20 videos correctly for dumping and misjudged 2 videos as excavation and swinging, 1 of each. For the second activity, "excavation" was recognized in 17 out of 20 video clips correctly and misjudged in 3 as "dumping". For the third activity, "hauling" has shown the least favorable results compared with the rest of the categories. Out of 20, 16 videos were correctly identified while 4 video clips were misjudged as excavations. For the last activity, "swing" was classified correctly in 19 out of 20 video clips and misclassified in only 1 video clip as dumping.

Table 4. Confusion matrix of the validation dataset.

Confusion Matrix					
Dumping	13 21.00%	1 1.60%	0 0.00%	1 1.60%	86.70% 13.30%
Excavation	0 0.00%	12 19.40%	0 0.00%	0 0.00%	100% 0.00%
Hauling	0 0.00%	1 1.60%	17 27.40%	0 0.00%	94.40% 5.60%
Swing	0 0.00%	0 0.00%	0 0.00%	17 27.40%	100% 0.00%
Output Class	100% 0.00%	86% 14.30%	100% 0.00%	94% 5.60%	95.20% 4.80%
Target Class	Dumping	Excavation	Hauling	Swing	

Table 5. Confusion matrix of the test set.

Confusion Matrix					
Dumping	18 22.50%	3 3.80%	0 0.00%	1 1.30%	81.80% 18.20%
Excavation	1 1.30%	17 21.30%	4 5.00%	0 0.00%	77.30% 22.70%
Hauling	0 0.00%	0 0.00%	16 20.00%	0 0.00%	100.00% 0.00%
Swing	1 1.30%	0 0.00%	0 0.00%	19 23.80%	95.00% 5.00%
Output Class	90% 10.00%	85% 15.00%	80% 20.00%	95% 5.00%	87.50% 12.50%
Target Class	Dumping	Excavation	Hauling	Swing	

The average precision of the test dataset for the 4 activities of dumping, excavation, hauling, and swing, is 81.8%, 77.3%, 100%, and 95%, respectively. The overall precision, recall, miss rate and accuracy of the activity recognition system are 87.5%, 88.52%, 12.5%, and 87.5%, respectively. The result shows that the third activity, "hauling" was misclassified the most among the other activities.

Twenty (20)% of the total dataset is selected as a test dataset. The sequential pattern of the activities of the excavator usually starts from the excavation to hauling, dumping and swing. The 4 activity categories of excavators contain 80 video clips in total. The CNN-BiLSTM algorithm tested each video clip from the test set to verify the effectiveness of the trained model. The four activities of excavation, hauling, dumping, and swinging from the test data set are tested. The screenshots during the prediction of activity are shown in Figure 9a–d, respectively. The predicted label and ground truth have also been highlighted in Figure 9. The CNN-BiLSTM classification algorithm has classified the activities into their respective categories. The predicted and ground truth labels are compared to calculate the precision, recall, and accuracy of the algorithm. The average precision of the test dataset for dumping, excavation, hauling, and swing are 81.8%, 77.3%, 100%, 95%, and 18.2%, respectively. Similarly, the miss rates of these classes are 18.2%, 22.7%, 0.0%, and 5%, respectively. The average precision, recall and miss rates for the test dataset are 87.5%, 88.52%, and 12.5%, respectively.

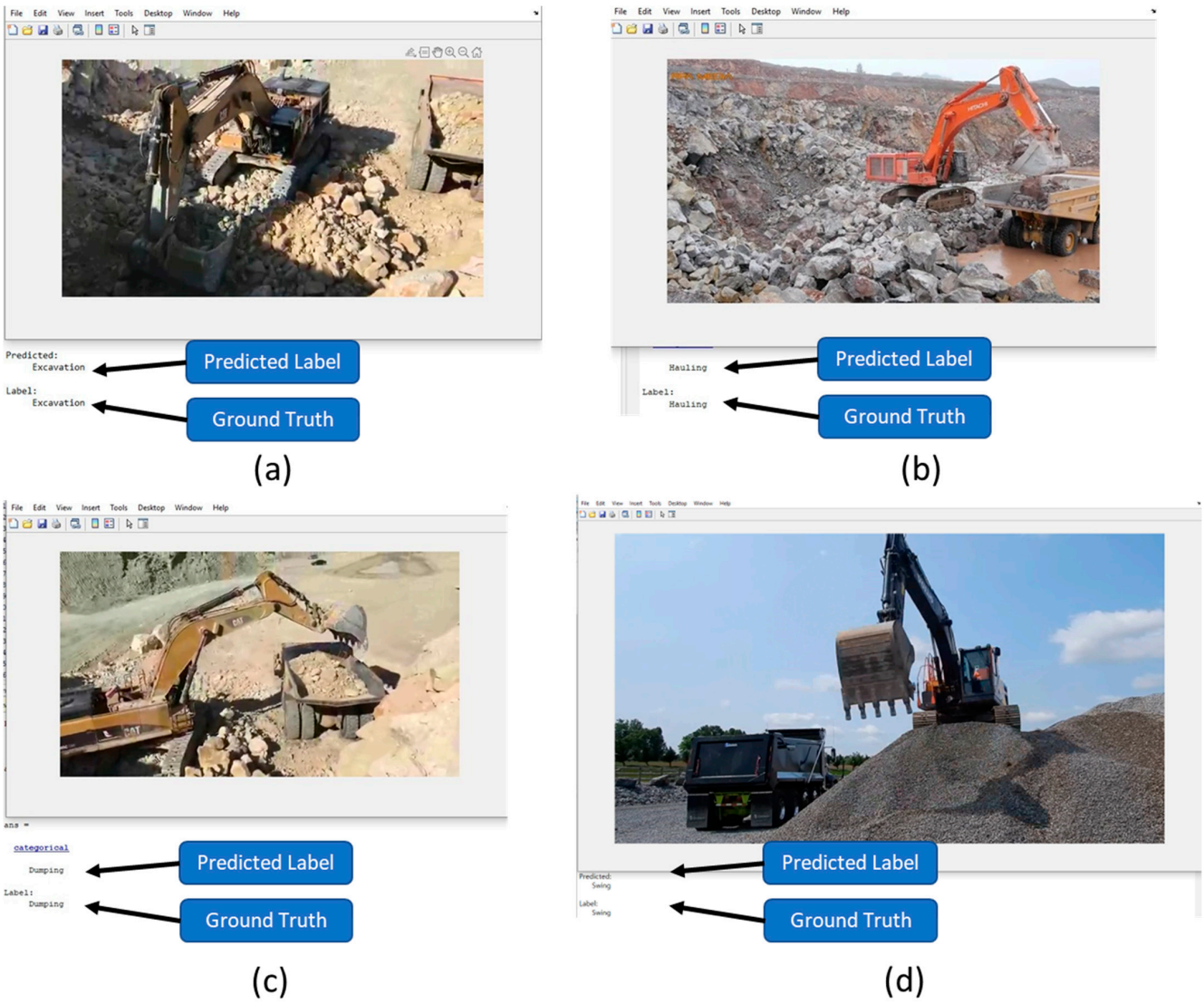

Figure 9. Example of successful activity recognition from the test dataset (**a**) Excavation (**b**) Hauling (**c**) Dumping (**d**) Swing.

The results of our study have been compared with some previous research studies in Table 6. These studies have used different algorithms and datasets. A CNN-based deep bidirectional long short-term memory (CNN-DBLSTM) was used in the study [20] for human action recognition on the UCF-101, YouTube 11 Actions, and HMDB51 datasets and similar results are reported for activity recognition on the HMDB51 dataset. In another study [17], the pre-determined starting point and duration of each activity has been used for classifying actions of construction workers and equipment. The study [11] used a Faster R-CNN detector + Deep SORT tracker + 3.D ResNet classifier for the excavator activity recognition and productivity analysis from construction and showed a similar kind of result. Furthermore, the study considered hauling and swinging as one activity. Meanwhile, this study implemented the pre-trained CNN model "Googlenet" for feature extraction and considered each frame of the video sequence. Furthermore, this study considered hauling and swing as different activities. Compared with these studies, the results of the proposed studies are promising.

Table 6. Performance comparison with previous studies.

Reference	Test Accuracy (%)	Method
Gong et al. [17]	79	3D-Harris feature + Bayesian learning classifier
Yang et al. [31]	57	HOG, HOE, MBH features + SVM classifier
Golparvar-Fard et al. [10]	86.33	3D HOG feature + SVM classifier
Luo et al. [32]	80.5	Faster R-CNN detector + Relevance network
Kim et al. [6]	90.9	Faster R-CNN detector + Tracking-Learning-Detection tracker + CNN-LSTM
Proposed framework	87.5	CNN(GoogleNet) and BLSTM

5. Conclusions

There have been many studies carried out for the vision-based activity classification of construction equipment. Most of them use the traditional BOW and SVM approaches. However, the response of a pre-trained CNN model "Googlenet" and bidirectional long short-term memory (BiLSTM) had not yet been used for the activity recognition of construction equipment. The goal of this study was to practically demonstrate the application of Googlenet for feature extraction, and of BiLSTM for the training and execution of the classification of excavator activities. Additionally, the performance of integrated models of Googlenet and BiLSTM have to be tested for the activity recognition of excavators in the earthwork field. Therefore, the authors have implemented an integrated method based on a pre-trained convolution neural network (CNN) "Googlenet" and bidirectional long short-term memory (BiLSTM) for the activity classification of excavators. First, the pre-trained CNN model extracts the sequential pattern of visual features from the video frames of construction equipment and then feeds it into the BiLSTM. The BilSTM consists of two LSTM layers, stacked on top of each other in both directions, with a forward direction as well as a backwards direction. The benefit of BiLSTM is that it not only includes the information from the previous frame but also from the upcoming frame. BiLSTM recognizes the different activities of the excavator after analyzing the input from the pre-trained convolution neural network. The bidirectional LSTM (BiLSTM) framework shows promising results on the YouTube-based excavator dataset. The experimental results show accuracies of 93.55% and 87.5% for the validation and test datasets, respectively. The experimental results of the CNN-BiLSTM framework show that the algorithm is capable of recognizing the different activities of construction equipment by using single-action videos datasets. The contribution to the knowledge of this research is the implementation and evaluation of the pre-trained CNN model "Googlenet" and bidirectional LSTM for the activity recognition of construction equipment. Furthermore, the authors intend to classify the value-added activities of the construction equipment in the future.

There are some limitations in this research: (1) The activity recognition performance of the model is affected by the detection results in the presence of two or more excavators. (2) The light conditions of the construction site video during the operation impact on the activity recognition results. When the light was too bright or low during the construction operation, it is difficult to recognize the moving features in video frames. (3) The diversity of data sets used in this study is limited, which may have affected the activity classification performance. Considering the limitations and applications of the research, the future goals of this study are to (1) improve the robustness of the activity recognition under varying light conditions by using some filters and implementing multiple cameras for better visual results in moving frames. (2) Various data should be collected from construction sites under different light conditions with various activities of excavators being performed. Furthermore, the authors intend to classify the value-added activities of the construction equipment in the future.

Author Contributions: The authors have contributed as follows. Conceptualization, J.K., J.S. and D.-E.L.; methodology, I.-S.K. and K.L.; software, K.L. and A.S.; validation, I.-S.K. and A.S.; formal analysis, I.-S.K., K.L. and A.S.; investigation, A.S. and J.S.; resources, J.K. and J.S.; data curation, I.-S.K. and K.L.; writing—original draft preparation, I.-S.K., K.L. and A.S.; writing—review and editing, J.S. and J.K.; visualization, I.-S.K., K.L. and A.S.; supervision, J.K. and J.S.; project administration, J.S.; funding acquisition, J.S. and D.-E.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the <National Research Foundation of Korea (NRF)> grant funded by the Korean government (MSIT) [No. NRF-2018R1A5A1025137 and No. NRF-2019R1A2C2006577]. Kamran Latif is extremely thankful to the <Higher Education Commission of Pakistan> for a Human Resource Development Initiative Universities of Engineering Science and Technology scholarship.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The author declares no conflict of interest.

Abbreviations

SVM	Support Vector Machine
SURF	Speeded-up Robust Features
HOG	Histogram of Oriented Gradients
ILSVRC	ImageNet Large Scale Visual Recognition Competition
RNN	Recurrent Neural Network
CNN	Convolutional Neural Network
RPN	Region Proposal Network
ROC	Receiver Operating Characteristic (ROC)
R-FCN	Region-based Fully Convolutional Network
IFaster R-CNN	Region based Convolutional Neural Networks
LRCN	Long-term Recurrent Convolutional Network
LSTM	Long Short-Term Memory (LSTM)
CNN-DLSTM	CNN based Double-layer Long Short-Term Memory
CNN-DBLSTM	CNN based Deep Bidirectional Long Short-Term Memory
CNN-BiLSTM	CNN based Bidirectional Long Short-Term Memory

References

1. Navon, R.; Shpatnitsky, Y. Field Experiments in Automated Monitoring of Road Construction. *J. Constr. Eng. Manag.* **2005**, *131*, 487–493. [CrossRef]
2. Altuntas, S.; Dereli, T.; Kemal Yılmaz, M. Evaluation of excavator technologies: Application of data fusion based multimoora methods. *J. Civ. Eng. Manag.* **2015**, *21*, 977–997. [CrossRef]
3. Apanavičienė, R.; Juodis, A. Construction projects management effectiveness modelling with neural networks. *J. Civ. Eng. Manag.* **2003**, *9*, 59–67. [CrossRef]
4. Yousefi, V.; Haji Yakhchali, S.; Khanzadi, M.; Mehrabanfar, E.; Šaparauskas, J. Proposing a neural network model to predict time and cost claims in construction projects. *J. Civ. Eng. Manag.* **2016**, *22*, 967–978. [CrossRef]
5. Kim, J.; Chi, S.; Seo, J. Interaction analysis for vision-based activity identification of earthmoving excavators and dump trucks. *Autom. Constr.* **2018**, *87*, 297–308. [CrossRef]
6. Kim, J.; Chi, S. Action recognition of earthmoving excavators based on sequential pattern analysis of visual features and operation cycles. *Autom. Constr.* **2019**, *104*, 255–264. [CrossRef]
7. Zou, J.; Kim, H. Using hue, saturation, and value color space for hydraulic excavator idle time analysis. *J. Comput. Civ. Eng.* **2007**, *21*, 238–246. [CrossRef]
8. Sonmez, R.; Sözgen, B. A support vector machine method for bid/no bid decision making. *J. Civ. Eng. Manag.* **2017**, *23*, 641–649. [CrossRef]
9. Rezazadeh Azar, E.; McCabe, B. Part based model and spatial-temporal reasoning to recognize hydraulic excavators in construction images and videos. *Autom. Constr.* **2012**, *24*, 194–202. [CrossRef]
10. Golparvar-Fard, M.; Heydarian, A.; Niebles, J.C. Vision-based action recognition of earthmoving equipment using spatio-temporal features and support vector machine classifiers. *Adv. Eng. Inform.* **2013**, *27*, 652–663. [CrossRef]

11. Chen, C.; Zhu, Z.; Hammad, A. Automated excavators activity recognition and productivity analysis from construction site surveillance videos. *Autom. Constr.* **2020**, *110*, 103045. [CrossRef]
12. Fang, Q.; Li, H.; Luo, X.; Ding, L.; Luo, H.; Rose, T.M.; An, W. Detecting non-hardhat-use by a deep learning method from far-field surveillance videos. *Autom. Constr.* **2018**, *85*, 1–9. [CrossRef]
13. Ding, P.; Zhang, Y.; Deng, W.J.; Jia, P.; Kuijper, A. A light and faster regional convolutional neural network for object detection in optical remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2018**, *141*, 208–218. [CrossRef]
14. Ding, L.; Fang, W.; Luo, H.; Love, P.E.D.; Zhong, B.; Ouyang, X. A deep hybrid learning model to detect unsafe behavior: Integrating convolution neural networks and long short-term memory. *Autom. Constr.* **2018**, *86*, 118–124. [CrossRef]
15. Chen, Y.; Wang, Z.; Peng, Y.; Zhang, Z.; Yu, G.; Sun, J. Cascaded Pyramid Network for Multi-person Pose Estimation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7103–7112. [CrossRef]
16. Luo, H.; Wang, M.; Wong, P.K.Y.; Cheng, J.C.P. Full body pose estimation of construction equipment using computer vision and deep learning techniques. *Autom. Constr.* **2020**, *110*, 103016. [CrossRef]
17. Gong, J.; Caldas, C.H.; Gordon, C. Learning and classifying actions of construction workers and equipment using Bag-of-Video-Feature-Words and Bayesian network models. *Adv. Eng. Inform.* **2011**, *25*, 771–782. [CrossRef]
18. Luo, H.; Xiong, C.; Fang, W.; Love, P.E.D.; Zhang, B.; Ouyang, X. Convolutional neural networks: Computer vision-based workforce activity assessment in construction. *Autom. Constr.* **2018**, *94*, 282–289. [CrossRef]
19. Donahue, J.; Hendricks, L.A.; Rohrbach, M.; Venugopalan, S.; Guadarrama, S.; Saenko, K.; Darrell, T. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 677–691. [CrossRef]
20. Ullah, A.; Ahmad, J.; Muhammad, K.; Sajjad, M.; Baik, S.W. Action Recognition in Video Sequences using Deep Bi-Directional LSTM With CNN Features. *IEEE Access* **2018**, *6*, 1155–1166. [CrossRef]
21. Zhu, Z.; Ren, X.; Chen, Z. Integrated detection and tracking of workforce and equipment from construction jobsite videos. *Autom. Constr.* **2017**, *81*, 161–171. [CrossRef]
22. Azar, E.R.; Kamat, V.R. Earthmoving equipment automation: A review of technical advances and future outlook. *J. Inf. Technol. Constr.* **2017**, *22*, 247–265.
23. Anand, B. Face recognition using SURF features. *Int. J. Electron. Eng. Res.* **2009**, *8*, 749628.
24. Fang, W.; Ding, L.; Zhong, B.; Love, P.E.D.; Luo, H. Automated detection of workers and heavy equipment on construction sites: A convolutional neural network approach. *Adv. Eng. Inform.* **2018**, *37*, 139–149. [CrossRef]
25. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
26. Kibria, S.B.; Hasan, M.S. An analysis of Feature extraction and Classification Algorithms for Dangerous Object Detection. In Proceedings of the 2017 2nd International Conference on Electrical & Electronic Engineering (ICEEE), Rajshahi, Bangladesh, 27–29 December 2017; pp. 1–4. [CrossRef]
27. Siami-Namini, S.; Tavakoli, N.; Namin, A.S. The Performance of LSTM and BiLSTM in Forecasting Time Series. In Proceedings of the 2019 IEEE International Conference on Big Data, Big Data 2019, Los Angeles, CA, USA, 9–12 December 2019; pp. 3285–3292.
28. Liu, Z.; Zhang, D.; Luo, G.; Lian, M.; Liu, B. A new method of emotional analysis based on CNN–BiLSTM hybrid neural network. *Cluster Comput.* **2020**, *23*, 2901–2913. [CrossRef]
29. Schoene, A.M.; Turner, A.P.; Dethlefs, N. Bidirectional dilated LSTM with attention for fine-grained emotion classification in tweets. *CEUR Workshop Proc.* **2020**, *2614*, 100–117.
30. Huang, Y.; Jiang, Y.; Hasan, T.; Jiang, Q.; Li, C. A topic BiLSTM model for sentiment classification. In Proceedings of the 2nd International Conference on Innovation in Artificial Intelligence, ICIAI '18, Shanghai, China, 9–12 March 2018; ACM Press: New York, NY, USA; pp. 143–147.
31. Yang, J.; Shi, Z.; Wu, Z. Vision-Based Action Recognition of Construction Workers Using Dense Trajectories. *Adv. Eng. Informatics* **2016**, *30*, 327–336. [CrossRef]
32. Luo, X.; Li, H.; Cao, D.; Dai, F.; Seo, J.; Lee, S. Recognizing Diverse Construction Activities in Site Images via Relevance Networks of Construction-Related Objects Detected by Convolutional Neural Networks. *J. Comput. Civ. Eng.* **2018**, *32*, 04018012. [CrossRef]
33. Zhou, Q.; Chen, G.; Jiang, W.; Li, K.; Li, K. Automatically Detecting Excavator Anomalies Based on Machine Learning. *Symmetry* **2019**, *11*, 957. [CrossRef]
34. Liu, Q.; Feng, C.; Song, Z.; Louis, J.; Zhou, J. Deep Learning Model Comparison for Vision-Based Classification of Full/Empty-Load Trucks in Earthmoving Operations. *Appl. Sci.* **2019**, *9*, 4871. [CrossRef]
35. Bhokare, S.; Goyal, L.; Ren, R.; Zhang, J. Smart Construction Scheduling Monitoring Using YOLOv3-Based Activity Detection and Classification. *J. Inf. Technol. Constr.* **2022**, *27*, 240–252. [CrossRef]
36. Cheng, M.-Y.; Cao, M.-T.; Nuralim, C.K. Computer Vision-Based Deep Learning for Supervising Excavator Operations and Measuring Real-Time Earthwork Productivity. *J. Supercomput.* **2022**. [CrossRef]
37. Chen, C.; Xiao, B.; Zhang, Y.; Zhu, Z. Automatic Vision-Based Calculation of Excavator Earthmoving Productivity Using Zero-Shot Learning Activity Recognition. *Autom. Constr.* **2023**, *146*, 104702. [CrossRef]

38. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. *arXiv* **2014**, arXiv:1409.4842.
39. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.