Hindawi
Advances in Civil Engineering
Volume 2020, Article ID 8812928, 14 pages
https://doi.org/10.1155/2020/8812928

Research Article

A Deep Learning-Based Approach to Enable Action Recognition for Construction Equipment

Jinyue Zhang,[1] Lijun Zi,[2] Yuexian Hou,[3] Mingen Wang,[3] Wenting Jiang,[2] and Da Deng[3]

[1]*College of Management and Economics, Tianjin University, Tianjin, China*
[2]*Guangzhou Metro Design and Research Institute Co., Ltd., Guangzhou, China*
[3]*College of Intelligence and Computing, Tianjin University, Tianjin, China*

Correspondence should be addressed to Yuexian Hou; yxhou@tju.edu.cn

Received 10 June 2020; Revised 21 October 2020; Accepted 22 October 2020; Published 5 November 2020

Academic Editor: Jia-Rui Lin

In order to support smart construction, digital twin has been a well-recognized concept for virtually representing the physical facility. It is equally important to recognize human actions and the movement of construction equipment in virtual construction scenes. Compared to the extensive research on human action recognition (HAR) that can be applied to identify construction workers, research in the field of construction equipment action recognition (CEAR) is very limited, mainly due to the lack of available datasets with videos showing the actions of construction equipment. The contributions of this research are as follows: (1) the development of a comprehensive video dataset of 2,064 clips with five action types for excavators and dump trucks; (2) a new deep learning-based CEAR approach (known as a simplified temporal convolutional network or STCN) that combines a convolutional neural network (CNN) with long short-term memory (LSTM, an artificial recurrent neural network), where CNN is used to extract image features and LSTM is used to extract temporal features from video frame sequences; and (3) the comparison between this proposed new approach and a similar CEAR method and two of the best-performing HAR approaches, namely, three-dimensional (3D) convolutional networks (ConvNets) and two-stream ConvNets, to evaluate the performance of STCN and investigate the possibility of directly transferring HAR approaches to the field of CEAR.

1. Introduction

Human action recognition (HAR) is one of the most popular research areas in computer vision (CV), and the outcomes have been extensively applied in video surveillance and human-machine interaction for a variety of application scenarios such as safety monitoring and control [1]. A basic methodology for recognizing human actions is to detect a human in a video/photograph, segmenting and mapping the body attributes followed by the result [2]. A number of methods for achieving HAR have been developed in the past few decades. Poppe [3] and Purohit and Chauhan [4] reviewed and classified different recognition algorithms for human actions. The deep learning (DL) method has attracted more attention in the area of CV following the success of AlexNet [5] in using convolutional neural networks (CNN) for image classification, gradually replacing the traditional HAR approaches based on the spatial and temporal structure of body movements. With advances in action recognition and the popularization of high-end hardware such high-resolution surveillance cameras, HAR is increasingly applied at work (e.g., in video surveillance at airports for security purposes) and at home (e.g., in health monitoring for elderly people).

In the construction industry, digital twin has been a well-recognized concept for smart construction. Beside the analysis and optimization supported by a digital representation of the facility itself, there is also a need to enhance the performance of workers and equipment, which is facilitated by automated perception and analysis of on-site activities. This requires the computer to not only identify workers and equipment but also recognize their locations and actions [6]. Many potential benefits could be realized if HAR could be applied for the monitoring and control of construction

equipment. With a detailed analysis of equipment actions and a continuous optimization of equipment operation, a construction project can be better managed: for example, equipment productivity would be improved by reducing the idle time, the environment would be protected by lowering carbon emissions, and accidents would be avoided by coordinating the movement of workers and equipment [7]. Currently, the analysis of construction machinery activity is mainly performed by human workers, a process that can be time-consuming, costly, and prone to errors. A low-cost, reliable construction equipment action recognition (CEAR) method can enable automated analysis of many scenarios at construction sites and can support smart construction applications.

Researchers have long studied the recognition of construction equipment actions and shown promising results. The CEAR methods can be roughly divided into sensor-based methods and visual-based methods. In the works that use sensors for recognizing the activity of construction equipment, of note are real-time location systems (RTLS) [8, 9], audio signals [10, 11], inertial measurement units (IMU) [12, 13], and so forth. In the works using visual-based methodology, many studies have adopted image processing-based approaches. Zou and Kim [14] used image color space (hue, saturation, and value) as the basis for image segmentation and tracing algorithms to identify the changing centroid coordinates of an excavator in successive images taken at fixed time intervals to achieve recognition of equipment movement. Gong et al. [15] proposed a visual learning approach to classify actions of construction workers and equipment using the Harris 3D interest point detector as the feature detector, local histograms as the feature representation, a bag-of-words model as the feature model, and Bayesian network models as the learning mechanism. Using a similar method, Golparvar-Fard et al. [7] studied single actions of construction equipment used for earthmoving. In this method, a video is initially represented as a collection of spatiotemporal visual features by extracting space-time interest points and describing each feature with a histogram of oriented gradients (HOG). The algorithm automatically learns the distributions of the spatiotemporal features and action categories using a multiclass support vector machine (SVM) classifier.

Many existing CEAR approaches rely on hand-crafted features, which is necessary to manually segment the time-series data to extract statistical features. This feature extraction process will be limited by human knowledge and can only extract shallow features specified by humans. Thus, they will only work well for simple actions that can be easily distinguished. The results will no longer be reliable once the viewpoint has changed, more background noise is present, and the views become blocked [3], which will be inevitable due to the dynamic nature of a construction site. Furthermore, these traditional methods require manual intervention at multiple points to adjust parameters, which limits the application of these methods in cases where end-to-end output is desired. Moreover, nearly all current methods for action recognition are data driven; that is, they require a large amount of data in order to train the recognition model to achieve a better recognition rate. As such, a large-scale dataset is of considerable significance. Golparvar-Fard et al. [7] developed the first comprehensive video dataset for action recognition of excavators and dump trucks. However, it is rare to find similar datasets for other types of construction equipment.

As a part of this research, a video dataset for excavators and dump trucks is first established to expand the size of the existing datasets on construction equipment, along with a corresponding optical flow dataset that can be used for related algorithm studies. Second, a deep learning-based CEAR model was proposed by taking advantage of deep learning, which has good generalization performance, is capable of end-to-end training, and requires no feature engineering [16]. Lastly, the deep learning-based method is compared with existing similar methods and some of the best HAR methods to determine whether the proposed method is advanced and whether advanced HAR algorithms could be directly transferred to CEAR. Related research works in HAR and CEAR are discussed in the next section, along with their limitations. Following that, the dataset development is illustrated in terms of camera arrangement, data acquisition method, and data processing method, and a detailed presentation of proposed deep learning-based CEAR model is provided. The experimental results and a comparison of the results of the proposed model to those of two HAR algorithms are discussed. The final section presents the conclusions that can be made based on the results of this research study.

2. Materials and Methods

2.1. Related Works. In recent years, numerous research efforts have been made to apply action recognition technologies in the construction industry. Some of these studies used construction workers as their research subjects and focused on worker health and safety, while other studies used construction equipment as the research subject and aimed to improve productivity and reduce costs for the construction project. According to different data sources, the approaches used in these studies can be divided into methods based on visual data and methods based on sensors data. In the current construction industry, cameras are generally installed for surveillance and other purposes. This makes the acquisition of visual data more convenient and cost-effective. Moreover, visual data can usually provide richer information. As such, this research focuses on determining the activities of construction equipment from visual data. Previous works that use vision-based methods, known as computer vision (CV), are discussed in this section. Seo et al. [17] reviewed previous attempts to apply CV for construction safety and health monitoring from both a technical perspective and a practical perspective. They categorized previous studies into three groups—object detection, object tracking, and action recognition—based on the type of information required to evaluate unsafe conditions and actions. However, in the current status of CV application in construction sites, even the most advanced theories are faced with challenges such

as a lack of high-quality datasets and the slow development of algorithms. At the same time, the development of deep learning technology, the increased image processing power provided by a graphics processing unit (GPU), and the decrease in price of specialized cameras bring about new opportunities for adopting CV-based applications in the construction industry.

2.1.1. Human Action Recognition. Many research studies have investigated CV-based human action recognition (HAR) systems, including both traditional hand-crafted and learning-based action representation approaches. The difference between these two approaches lies mainly in the method used to extract features from images. A traditional hand-crafted representation-based approach relies on the expert-designed feature detectors and descriptors such as Hessian3D, scale-invariant feature transform (SIFT), HOG, enhanced speeded-up robust features (ESURF), and local binary pattern (LBP). On the contrary, a learning-based representation approach uses a trainable feature extractor that automatically learns features from the raw data, eliminating the need for manual assignment and enabling end-to-end learning.

The traditional hand-crafted action representation approach has been popular in the HAR community and has achieved remarkable results when using various well-known public datasets [18]. This approach includes four techniques: the space/time-based method [19], the appearance-based method [20, 21], the LBP-based method [22], and the fuzzy logic-based method [23]. Using these methods, the important features from a sequence of image frames are extracted to build the feature vector prior to classification by a trained classifier. For example, dense trajectory (DT) uses trajectories to capture the local motion information of the video, and a dense representation guarantees good coverage for capturing foreground motions as well as the surrounding context [24]. Wang and Shmid [25] proposed an improved DT (iDT) approach to improve the performance of video representation by making corrections that take camera motion into account. Currently, researchers are aiming to increase the quantity and quality of the dataset for human action recognition. However, most successful hand-crafted representation methods are based on local densely sampled descriptors, which will result in a higher computational cost.

The appropriate and efficient representation of data is the key to HAR. Unlike the above-mentioned approaches, where an action is represented by hand-crafted feature detectors and descriptors, learning-based representation approaches have the ability to learn a feature automatically from the raw data, thus introducing the concept of end-to-end learning, which refers to transformation from the pixel level to an action classification and is not limited by human knowledge [18]. Some learning-based approaches are based on genetic programming [26] and dictionary learning [27], while others employ deep learning-based models for action representation.

Deep learning is an important area of machine learning which aims to achieve learning at multiple levels of representation and abstraction in order to make sense of data such as speech, images, and text. The research of Karpathy et al. [28] showed the potential of CNN for large-scale video classification tasks. Simonyan and Zisserman [29] proposed a two-stream convolutional networks (two-stream ConvNets) architecture that incorporates spatial and temporal networks and demonstrated that a ConvNet trained on multiframe dense optical flow is able to achieve very good performance despite the limited amount of available training data. Tran et al. [30] argued that deep three-dimensional (3D) ConvNets trained on a large-scale supervised video dataset are effective for spatiotemporal feature learning. The 3D ConvNets build on two-dimensional (2D) ConvNets but include a time dimension; thus, they solve the issue of the inability of CNNs to extract temporal features. Carreira and Zisserman [31] introduced a new two-stream inflated 3D (I3D) ConvNet, where filters and pooling kernels of very deep image classification ConvNets are expanded into 3D, making it possible to learn seamless spatiotemporal feature extractors from video while leveraging successful ImageNet architecture designs and even their parameters. Varol et al. [32] made identifications from video representations using neural networks with long-term temporal convolutions (LTC) and demonstrated that LTC-CNN models with increased temporal extents improve the accuracy of action recognition. Ng et al. [33] employed a recurrent neural network that uses long short-term memory (LSTM) cells that are connected to the output of the underlying CNN. This LSTM-CNN approach exhibits significant performance improvement over previously published results on the Sports 1 million dataset and the UCF101 dataset [34]. Donahue et al. [35] developed a novel recurrent convolutional architecture suitable for large-scale visual learning, which is end-to-end trainable, and they demonstrated the value of these models for benchmark video recognition tasks. Sevilla-Lara et al. [36] investigated the impact of different flow algorithms and input transformations to better understand how these would affect a state-of-the-art action recognition method, and they recommended a better way of using optical flow in the future.

Recently, the research community has paid a great deal of attention to deep learning-based approaches, mainly due to their excellent performance as compared to hand-crafted action representation approaches. However, some of the best learning-based methods still rely on hand-crafted features. The main reason is the lack of huge datasets for action recognition which are required to train the feature extractors. As no huge dataset such as ImageNet in the field of object recognition exists, the HAR community is working on the development of useful datasets. HMDB [37] is an action video database with 51 action categories, which in total contain around 7,000 manually annotated clips. UCF101 [34] consists of 101 action classes in a total of 13,320 clips of video data. More recently, the development of a large-scale dataset called ActivityNet [38] provides samples from 203 activity categories with an average of 137 untrimmed videos per class and 1.41 activity instances per video, for a total of 849 hours of video. YouTube-8M is the largest multilabel video classification dataset [39] and it includes about 8

million videos (approximately 500,000 hours of video), annotated with a vocabulary of 4,800 visual entities.

2.1.2. Construction Equipment Action Recognition. Although many research studies have investigated the use of HAR in the construction industry for health/safety monitoring and for the control of construction workers, the study of construction equipment action recognition (CEAR) is still premature. The first work that can be found in CEAR was conducted by Gong and Caldas [40], who developed an intelligent video computing method to interpret videos of cyclic construction operations and translate the images automatically into productivity information through the recognition of actions of a concrete bucket in the concrete pour process. Akhavian and Behzadan [41] used built-in smartphone sensors as ubiquitous multimodal data collection and transmission nodes in order to detect detailed construction equipment activities, which can ultimately contribute to the process of simulation input modeling. In a case study of front-end loader activity recognition, certain key features are extracted and are used to train supervised machine learning classifiers. Cao et al. [42] proposed a classification algorithm based on acoustics processing for four types of excavation equipment. They developed new acoustic statistical features (short frame energy ratio, concentration of spectrum amplitude ratio, truncated energy range, and pulse interval) to characterize acoustic signals; then, based on the probability density distributions of these acoustic features, a novel classifier was proposed. This approach has a great potential to be generalized. Han and Golparvar-Fard [43] investigated current strategies for leveraging emerging big visual data in construction performance monitoring from the standpoints of reliability, relevance, and speed, and they structured a road map for research in visual sensing and analytics for construction. Roberts and Golparvar-Fard [44] presented a new benchmark dataset consisting of ten videos that can be used to detect, track, and analyze construction work activities of excavators and dump trucks. They also gave an action recognition framework composed of detection module, tracking module, and recognition module. This method can automatically identify excavators and dump trucks from per-frame of the video sequence and track their activities and finally identify their construction actions. Rashid and Louis [45] proposed a data-augmentation framework for generating synthetic time-series training data for RNN-based deep learning networks. Their research results show that deep learning framework outperformed the shallow network regarding model accuracy and generalization, and the data-augmentation methodology has the ability to correctly simulate real-world dataset. Kim and Chi [46] proposed a vision-based action recognition framework that considers the sequential working patterns of earthmoving excavators. The framework includes three main processes: excavator detection, excavator tracking, and excavator action recognition. Among them, the action recognition process used CNN-DLSTM model. The framework demonstrates good generalization performance and proves the important positive impact of sequential pattern modeling on recognition performance. Rashid and Louis [13] researched the use of activity-specific equipment motions instead of vibration for action recognition. The study showed that using inertial measurement unit (IMU) data of different articulated elements can significantly improve the activity recognition results.

Previous CV-based research in the construction industry mainly focused on either identifying/tracking workers and equipment or recognizing workers' actions. Although deep learning-based action recognition algorithms have dominated the field of HAR, they are not widely investigated for CEAR. The challenges in doing so are threefold.

The most significant challenge is the lack of comprehensive datasets for CEAR. Deep learning-based approaches largely rely on large numbers of high-quality raw video clips to train the feature extractor for a better performance in the later classification task. Although the number of available datasets for HAR is increasing, as mentioned earlier, this is not the case in the field of CEAR. A comprehensive dataset for CEAR needs to include video clips from a variety of working environments, from different viewing angles, and with various amounts of background clutter.

Currently, there is a shortage of algorithms that can be applied at real construction sites. Most existing researches of CEAR in the construction industry use pattern recognition methods and those algorithms rely on hand-crafted features and will be limited by human knowledge. When they are applied to actual cases, their recognition performance will be greatly affected by human factors.

At the present time, there is no clear benchmark for CEAR. The number of studies in CEAR is fewer than that for HAR and, due to the shortage of standard video datasets for construction equipment, it is hard to compare different approaches that are based on the same dataset.

Focusing on these three aspects, in this study, a new dataset for excavators and dump trucks was initially developed. This dataset was used to expand the existing dataset for the same types of equipment [7] and to train/verify the feature extractor using the deep learning-based method proposed in this research. The action recognition method is based on deep learning theory, which can automatically extract high-level features from raw data without feature engineering. The possible problems of manual operation features can be avoided. Comparison with existing similar CEAR method [46] proves that it has comparable performance. By comparing the results after applying two advanced deep learning-based HAR approaches to the same dataset, this research study investigates the possibility of transferring some of the best HAR algorithms to CEAR and broadens the path of researching CEAR methods.

2.2. Dataset Development

2.2.1. Dataset Collection. The construction equipment considered in the new dataset includes an excavator and a dump truck. For these two pieces of equipment, there are five activities in all, as listed in Table 1 and shown in Figure 1.

Table 1: Construction equipment activities.

Equipment	Activities
Excavator	Digging, swinging, dumping
Dump truck	Moving forward, moving backward

To take advantage of the recent improvements to cameras in smartphones, this research used smartphones with a 12-megapixel rear camera to collect raw data. Video footage was collected at a resolution of 720 progressive scan (720p) and at 25 frames per second (fps). The selection of camera placement needs to be taken into account to prevent possible obstructions and to ensure a good proportion of construction equipment in the picture. Based on the site survey, it was found that the front views and the side views of construction equipment can provide more effective information about their actions than other views. The front view refers to the projection view from the direction that the driver faces when driving the device normally moving forward. Therefore, in order to capture the various views of the construction equipment and their actions from different perspectives, four cameras were used to collect videos within a 180-degree range around the construction equipment, as shown in Figure 2. Cameras 1, 2, and 3 were used to capture the front view, the side view, and the rear view, and Camera 4 was placed at a position in between Camera 1 (front view) and Camera 2 (side view), as can be seen in Figure 2, in order to provide supplemental information. This camera configuration guarantees a sufficient number of action views while using a minimum amount of video capturing resources. Finally, an annotation document was created based on the five action types for construction equipment. Table 2 shows the annotations for the action types, the total number of video clips of each type, and the numbers of video clips for various subsets of data.

2.2.2. Data Processing. DivX, a video codec developed by DivX, LLC, was used to transcode the original videos to MPEG-4 format with a resolution of 480 pixels by 360 pixels. The transcoded videos were classified according to the four recording angels as shown in Figure 2; then videos were cut into shorter clips to ensure that each video clip, which had a duration ranging from 3 to 20 seconds, would contain only one complete action of a single piece of equipment. The format for the file-naming convention for the shorter video clips is "v_equipment_action_ID#" (e.g., v_evacuator_swing_001.MPEG). Finally, in the category of each recording angle, video clips were classified according to the five action types, and stratified sampling was performed according to a ratio of 6 : 2:2 to form a training set, a test set, and a verification set.

For the processed data, corresponding optical flow dataset was developed by the method of Lucas-Kanade algorithm [46]. This dataset will be used to verify the performance of the HAR algorithm used for CEAR and can be used as a complement to the development of CEAR datasets. The examples of the optical flow dataset are shown in Figure 3. The video dataset developed in this research was of open source. Everybody can obtain the dataset from https://github.com/hnpyn/CEAR_dataset. The authors will update and maintain this project regularly.

2.3. Development of a Simplified Temporal Convolutional Network (STCN). Ng et al. [33] proposed a method for modeling video frames into ordered frame sequences by using a recurrent neural network (RNN), which connects the LSTM units to the output of the CNN. The CNN structure is based on GoogLeNet [47], while the RNN adopts a deep LSTM structure [48] with five LSTM layers. This model performed very well in action recognition as compared to the best approaches available at the time. Similarly, Donahue et al. [35] also employed a method of directly connecting the RNN to the CNN structure and found that when nonlinearity is incorporated into the network status updates, learning of long-term dependencies is possible. As such, temporal dynamics and convolutional perceptual representations can be learned by jointly training the RNN and the CNN. It has been proved that the joint architecture of the RNN and the CNN is effective and feasible. Inspired by these studies, the authors have proposed a CEAR process (described in Figure 4) that can automatically extract video data features and perform end-to-end training for action recognition. The core of this process is the neural network shown in the dashed frame, which combines a CNN for extracting features from video clips and a LSTM for extracting the temporal dynamics. A fully connected layer is employed to connect the CNN and the LSTM, and a softmax layer is used to determine the classification of the equipment action.

2.3.1. Deep Learning. As an emerging research direction in the field of machine learning [49], deep learning was first proposed by Hinton et al. [50]. The merit of deep learning lies in the additional levels of nonlinear operation that it encompasses [51], and deep learning is able to form more abstract high-level representations or features by combining low-level features, thereby displaying the hierarchical feature representation of the data. Therefore, deep learning is able to automatically learn to obtain the hierarchical feature representations [16] that are more conducive to classification tasks. Traditional machine learning and pattern recognition methods require the manual extraction of features. The model itself only classifies or predicts according to the features, and the hand-crafted features, to a large extent, will determine the quality of the method. It requires both professional knowledge and enough time to allow for manual extraction of features. Thus, it will be limited by human knowledge, and most of the models can only obtain shallow features. This is the intrinsic driving force of this research to develop action recognition algorithm based on deep learning.

2.3.2. Convolutional Neural Network. A convolutional neural network (CNN) refers to a class of feedforward neural networks (FNNs) having a convolutional structure, and this

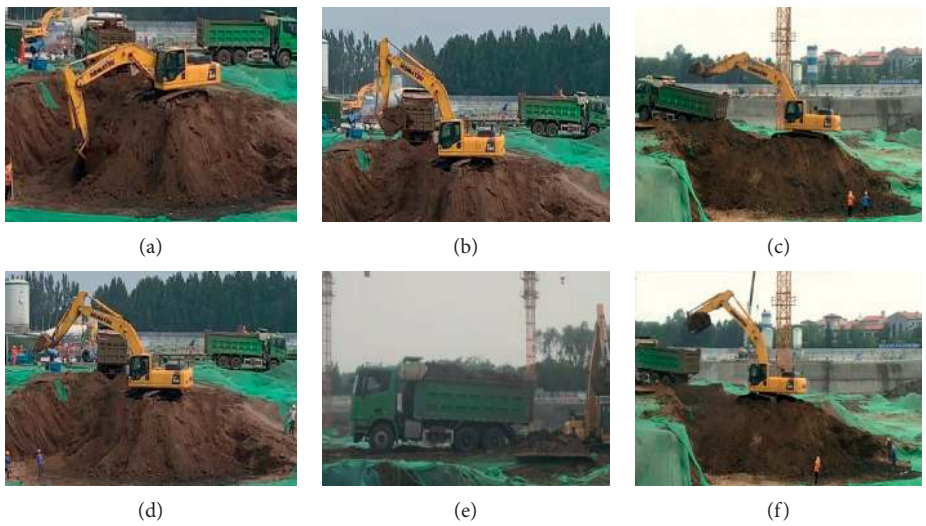

Figure 1: Examples of the activities of an excavator and a dump truck: (a) excavator digging, (b) excavator swinging, (c) excavator dumping, (d) excavator swinging, (e) dump truck moving forward, and (f) dump truck moving backward.

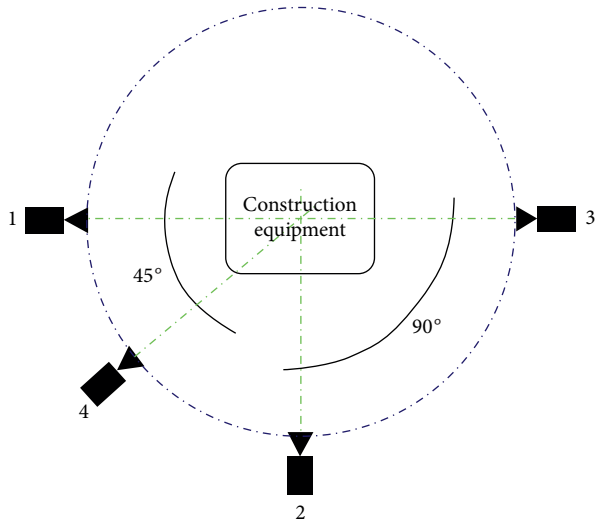

Figure 2: Camera setup used for recording the movement of construction equipment.

Table 2: Dataset composition and action annotations.

Equipment	Action type	Type annotation	Videos of this type	Videos in training set	Videos in test set	Videos in validation set
Excavator	Digging	0	498	298	100	100
	Dumping	1	513	307	103	103
	Swinging	2	940	565	187	188
Dump truck	Moving backward	3	54	32	11	11
	Moving forward	4	59	34	13	12

type of network is one of the representative algorithms in deep learning. The CNN uses the idea of sparse connection and weight sharing to solve the parameter explosion in ordinary FNNs, and it adopts convolution and pooling operations to obtain local features and reduce the dimension of feature space. The CNN generally completes the final classification through softmax by connecting to a fully connected layer.

CNN can use the original data as the input characteristics, which avoids the complex feature extraction process in a traditional machine learning algorithm, and it reduces the number of weights in the weight-sharing structure, thus

(a) (b) (c)

Figure 3: Examples of the optical flow dataset.

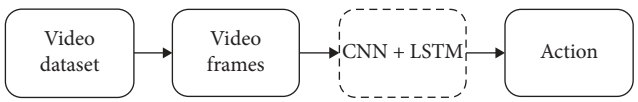

Figure 4: Flow chart for the process of construction equipment action recognition.

decreasing the complexity of the model. At the same time, the feature map is subsampled by using the principle of local correlation in images at the subsampling stage, which effectively reduces the amount of data processing required while retaining useful structural information [49]; because of this, CNNs have been widely used in CV-related tasks in recent years.

LeNet-5 [52] was a milestone in early CNN development, and it has been used to determine the basic structure of a CNN—which contains convolutional layers, pooling layers, and fully connected layers. Later CNNs have basically followed this same structure, with less or more optimization and improvement. AlexNet [5] adopted a structure consisting of five convolutional layers and three fully connected layers, which helped it to succeed in the ImageNet competition. The success of AlexNet indicates that deep learning is a reliable method, and it lays the foundation for using deep learning in image classification and object recognition tasks.

2.3.3. Long Short-Term Memory. Long short-term memory (LSTM) [53] is a variant of RNN, which is a feedback neural network that not only inherits most features of the RNN model but also solves the issue of vanishing gradients in regular RNNs [54]. As a nonlinear model, it can be used to build larger and more complex deep neural networks. A common LSTM architecture is composed of a cell (the memory part of the LSTM unit) and three "gates" of the flow of information inside the LSTM unit: an input gate, an output gate, and a forget gate, as shown in Figure 5. These gates are used to either remove or add information to the cell. Cells are circularly connected to each other, replacing the hidden unit in the regular recurrent network. The state unit has a linear self-loop structure whose weight is controlled by the forget gate.

LSTM has two states, the cell state and the hidden state. The cell state changes slowly with time, while the hidden state can vary widely at different times. The gate mechanism

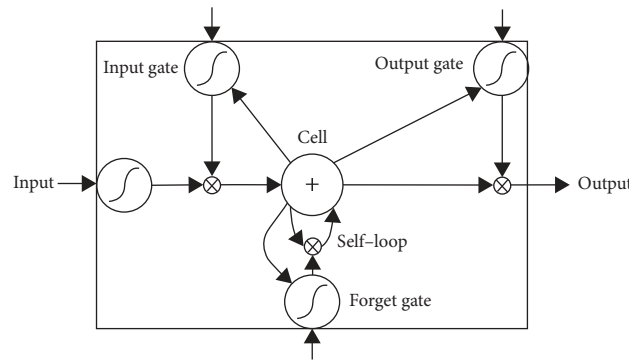

Figure 5: Cell structure in long short-term memory (LSTM).

can adjust the focus of memory according to the training target and then recode to control the trade-off between the input at one time and the input at a subsequent time. Therefore, LSTM can remember the information that needs to be remembered for a long time, while forgetting any relatively unimportant information. With LSTM, it is easier to learn long-term dependency than in the simple RNN architecture, and it is useful for dealing with problems that are highly related to time series, such as the video sequences used for CEAR.

2.3.4. Simplified Temporal Convolutional Network for CEAR. Considering that video sequences contain dynamic images that include information on both space and time, that it is difficult to extract temporal features using a simple CNN, and that a regular RNN is not able to extract image features well, a combination of CNN and LSTM is proposed in this research for CEAR [54], where the CNN is used to extract the image features of the video frame sequences and LSTM is used to extract the temporal features. In this process, the probabilities for all frames generated by the softmax layer are averaged, and the label with the highest probability is selected as the final classification result. This proposed method is called a simplified temporal convolutional network (STCN), and the structure of the STCN is shown in Figure 6.

Theoretically, a model with more parameters will have a higher complexity, and when the amount of training data is insufficient, a complex model is easy to be subjected to overfitting [55]. At present, considering the small amount of data in the field of CEAR, a less complex CNN model

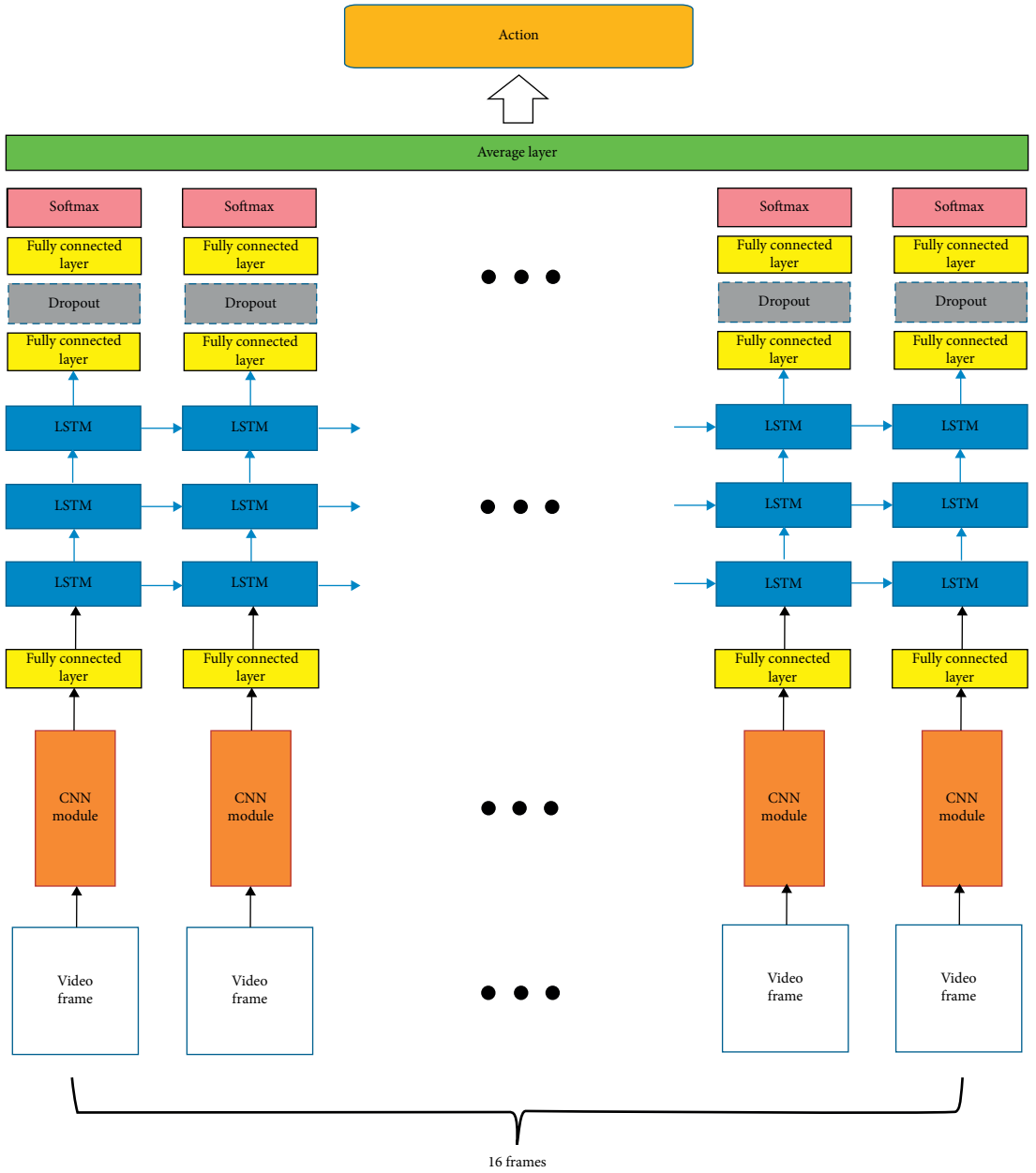

Figure 6: Structure of the simplified temporal convolutional network (STCN).

consisting of five convolutional layers is employed in this research, as shown in Figure 7. In order to obtain a sufficient number of effective receptive fields [56], the CNN model uses 7×7 and 5×5 convolution kernels in the Conv1 and Conv2 layers, respectively, and 3×3 convolution kernels for the other convolutional layers. In order to ensure the stability of data distribution in each layer to improve the training efficiency, batch normalization is used in all convolutional layers [57]. Because the maximum pooling operation can reduce the estimated mean shift caused by parameter errors in the convolutional layers while maintaining translation invariance, thereby minimizing the number of parameters and reducing model complexity [58], a max pooling layer is added at the end of each convolutional layer, and 2×2 convolution kernels are used in all pooling layers.

There is no pretraining in STCN because, based on the research result of Glorot et al. [59], the performance of a rectified linear unit (ReLu) network is far better than those of other activation function networks even without pretraining. Sparsity can be introduced to the ReLu activation function to allow each neuron to fully play its screening effect—those values matching the median value of a certain feature will be amplified, while the outlying values will be abandoned. Since the ReLu activation function only needs to perform the calculation for the maximum value, it is also superior in terms of calculation speed. As a result, ReLu is chosen as the activation function in the proposed model.

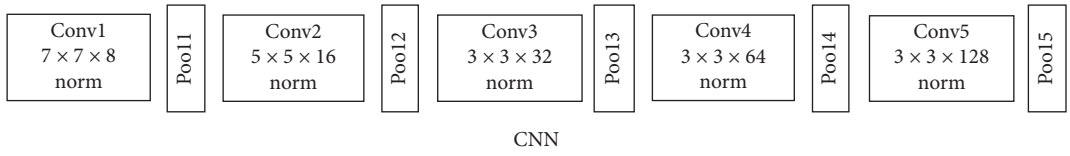

Figure 7: Structure of the convolutional neural network (CNN) used in this study.

A fully connected layer is used to transmit the output of the CNN model to the three subsequent layers of the LSTM network. LSTM units for all continuous image subsequences were connected to each other. The probability of each action category is then generated using the fully connected layers and the softmax function. In order to generate the action prediction label of a given video clip, an average layer is employed to take the average probability of all image frames in the video clip, and the category label with the maximum probability is used as the result for action recognition.

3. Results and Discussion

3.1. Results. Using the dataset introduced in Section 2.2, the authors extracted 16 image frames from each video clip to train the algorithm discussed in Section 2.3. For video clips that are less than 4 seconds, between 5 and 10 seconds, and longer than 10 seconds, image frames were extracted at an interval of 5 frames, 10 frames, and 15 frames. If there were less than 16 frames that can be extracted based on this rule for a very short video clip, the last frame was repeated to complement 16 frames. After that, all images were adjusted to a resolution of 320 pixels by 240 pixels, and a random cropping with a size of 224 pixels by 224 pixels was used as data augmentation. Finally, all these frames were sent into the STCN model in a chronological order.

All the work was carried out by PyTorch on a workstation with a 6-core 3.8 GHz Intel processor, 16 GB memory, a GTX1060 graphics card with 6 GB memory, and Windows 10 operating system. In order to quantify the performance of the action recognition algorithm, three common performance metrics are employed, that is, precision, recall, and F-1 score. Calculating the precision and recall rate of the model can assess the costs associated with misclassification. The F-1 score is the harmonic mean of accuracy and recall, which takes into account both precision and recall. Their mathematical equations are shown as follows:

$$\begin{aligned}\text{Precision} &= \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \times 100\%,\\ \text{Recall} &= \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \times 100\%,\\ F1\ \text{score} &= 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.\end{aligned} \quad (1)$$

As mentioned above, STCN model is mainly composed of two parts, that is, CNN module and LSTM module. The inputs to the model are the video clip frames. Experiments were conducted on the parameter configuration of the STCN model. In the experiment for weight initialization, this study performed Xavier initialization [60] on the CNN module. The results showed that the model with weights initialization has a better F-1 score (2.31% higher) than a scenario where no parameter initialization was performed. Then there was an experiment on the parameter settings of the CNN module. This experiment mainly studies the influence of the size of the convolution kernels on the recognition performance. The experimental results are shown in Table 3. Four sets of convolution kernels configurations were used for control experiments. The results indicated that this model has the best performance, when Conv1 and Conv2 used 7×7 and 5×5 convolution kernels and the other convolution layers used 3×3 convolution kernels.

Next, the experimental research on LSTM parameter settings showed that the number of LSTM layers has a greater impact on recognition performance of the dataset developed in this research. In this study, the LSTM model architecture used is shown in Figure 5. Experiment shows that the three-layer LSTM model has the highest F-1 score, which is at least 2% higher than the F-1 score of other LSTM layer settings. The result is shown in Figure 8. In contrast, it was found that the number of hidden units in the LSTM had little effect on the F-1 score. The authors tested the STCN model with 64, 128, and 256 hidden units in the LSTM module, and the difference in F-1 score was within 1%, as shown in Table 4. Therefore, in order to reduce the complexity of the model to save computational cost and reduce overfitting, this model employs 64 hidden units.

In the experiment of this research, all models use Adam optimization algorithm [61] as the optimizer, the learning rate is set at 0.001, and the other parameter settings are $\beta 1 = 0.9$, $\beta 2 = 0.999$, and $\varepsilon = 10e - 8$. Under the above settings, the STCN model achieves a precision of 77.55%, a recall rate of 75.00%, and an F-1 score of 76.25%, as shown in Table 5.

In addition, this research also studied the recognition effect of the STCN model for each action category. The experimental results are shown in Table 6. In this experiment, accuracy was used as the main evaluation metric, that is, the ratio of the number of examples that are correctly predicted to the number of all examples that are predicted to be that category for each action category. The results indicate that in general this model can recognize the action of excavators better than dump truck.

3.2. Discussion. In order to compare the performance of the STCN method and investigate the possibility of transferring some of the best HAR methods for use in CEAR, the dataset developed in this research was used to examine the

Table 3: *F*-1 scores of STCN models under different size convolution kernels configurations.

Conv1	Conv2	Conv3	Conv4	Conv5	*F*-1 score (%)
3 × 3	3 × 3	3 × 3	3 × 3	3 × 3	75.26
5 × 5	5 × 5	3 × 3	3 × 3	3 × 3	74.88
7 × 7	5 × 5	3 × 3	3 × 3	3 × 3	**76.25**
7 × 7	7 × 7	3 × 3	3 × 3	3 × 3	74.33

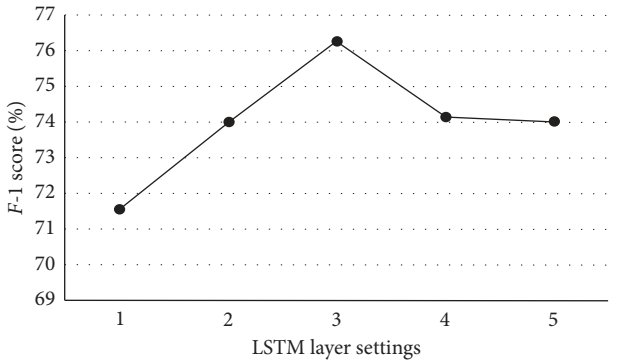

Figure 8: The *F*-1 score of different LSTM layer settings in STCN.

Table 4: *F*-1 scores of STCN models with different LSTM hidden units.

The number of hidden units	*F*-1 score (%)
64	**76.25**
126	76.09
256	76.28

Table 5: Performance measures of different action recognition methods.

Method	Precision (%)	Recall (%)	*F*-1 score (%)
CNN	71.26	68.01	69.60
LSTM	44.48	44.43	44.46
CNN-DLSTM	76.40	74.14	75.25
STCN	77.55	75.00	76.25
3D ConvNets	75.45	71.74	73.55
Two-stream ConvNets	77.80	74.79	76.26
I3D	78.13	75.02	76.54

performance of the CNN-DLSTM model used by Kim and Chi [46] for CEAR, 3D ConvNets proposed by Tran et al. [30], two-stream ConvNets proposed by Simonyan and Zisserman [29], and I3D ConvNets proposed by Carreira and Zisserman [31]. The results, which are summarized in Table 5, indicate that the STCN method has a performance comparable to those of the method proposed by Kim and Chi, the 3D ConvNets method, and the two-stream ConvNets in precision, recall, and *F*-1 score.

The results of using CNN module and LSTM module for CEAR, respectively, are also presented in Table 3. Their performance when working alone is not very satisfactory, and it is also verified that extracting key visual features and extracting context features have a significant positive impact on CEAR performance. The CNN-DLSTM model refers to the method used by Kim and Chi [46] in the action recognition process. Its CNN consists of 10 convolutional layers and 5 pooling layers, and the sequential patterns learning module consists of two LSTM models.

The 3D ConvNets model was reproduced as shown in Figure 9. The convolutional and pooling operations of the 3D ConvNets model were performed in the temporal dimension. All 3D convolution filters are $3 \times 3 \times 3$ in size with a stride of $1 \times 1 \times 1$. The first 3D pooling layer is $1 \times 2 \times 2$ with a stride of $1 \times 2 \times 2$, and the remaining 3D pooling layers are $2 \times 2 \times 2$ with a stride of $2 \times 2 \times 2$. This design preserves the temporal information in the early stages. The model can simulate both appearance information and action information simultaneously, and it produces excellent results in HAR tasks. Using the same dataset, the 3D ConvNet model achieved *F*-1 score of 73.55%.

Two-stream ConvNets were also examined using the same dataset. As shown in Figure 10, the two-stream ConvNet consists of two convolutional network structures: a spatial stream and a temporal stream. The spatial stream convolution network is essentially an image classification network that acquires static appearance features using the input of a single frame image. The temporal stream network uses multiframe optical flow images as input, as shown in Figure 11. In order to employ the two-stream ConvNets, a corresponding optical flow dataset was created to extract temporal features. An action classification result was obtained by taking the average of the classification scores. The two-stream ConvNets model is in a leading position in HAR tasks in terms of its performance, and it achieved an *F*-1 score of 76.26% in CEAR in this research.

By comparing the results for the STCN method, the CNN-DLSTM method, the 3D ConvNets method, the two-stream ConvNets method, and the I3D ConvNets method, it is found that the STCN method proposed in this study exhibits a performance in CEAR tasks which is comparable to those of other deep learning-based methods that have proven to have good results. Figure 12 shows the time consumption of training these methods. It can be seen that, in the case of close performance, STCN requires the shortest training time, so it has a certain speed advantage. As what has been mentioned before that the two-stream ConvNets model achieved a slightly better *F*-1 score than STCN (76.26% versus 76.25%) but the two-stream ConvNets model and I3D ConvNets model consumed much more computing time (9 h 28 m and 17 h 22 m versus 7 h 43 m), the STCN model is still better in general. In addition, the study demonstrates the feasibility of directly transferring some HAR methods to the CEAR field, as the 3D ConvNets model and the two-stream ConvNets model both achieved an acceptable rate of accuracy when using the same dataset as input.

Table 6: Construction equipment action recognition results per activity.

Equipment	Activity	Precision (%)	Recall (%)	F-1 score (%)
Excavator	Digging	76.38	74.06	75.20
	Dumping	83.78	80.57	82.14
	Swinging	78.58	76.10	77.32
Dump truck	Moving backward	56.27	55.13	55.69
	Moving forward	54.48	53.72	54.10

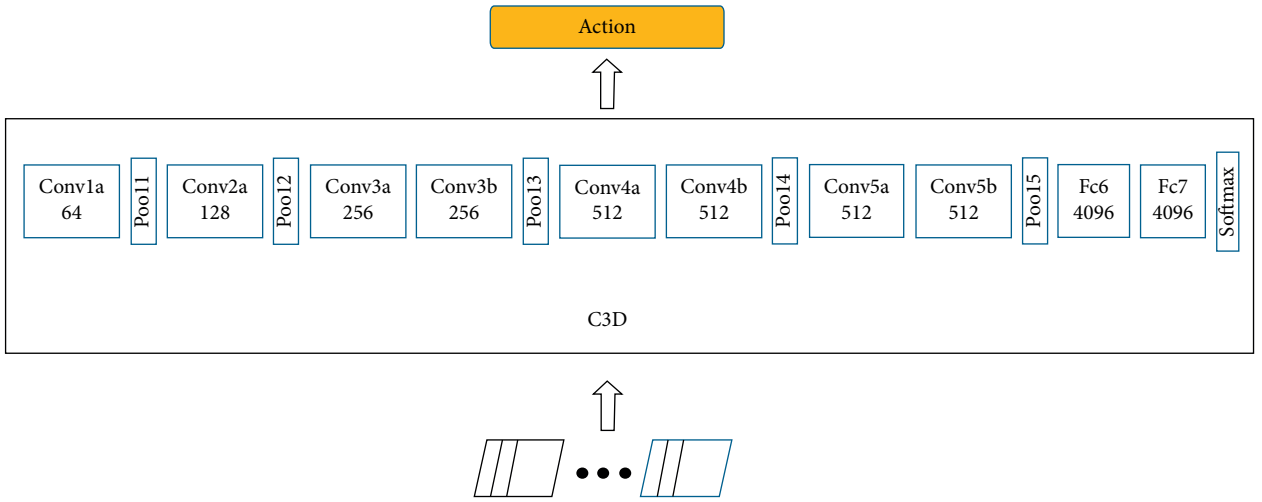

Figure 9: Structure of the 3D ConvNets model.

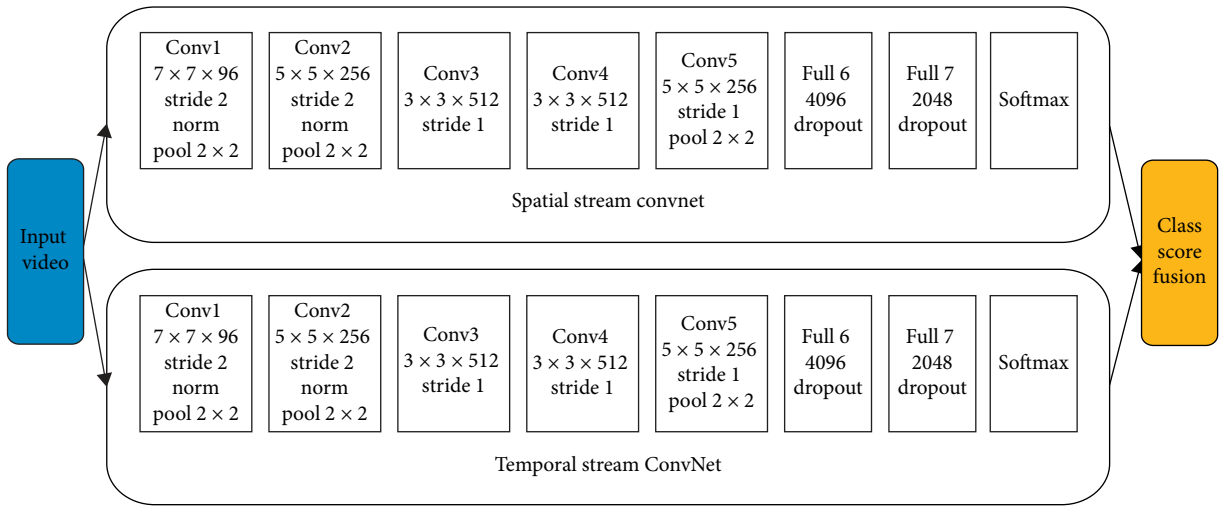

Figure 10: Two-stream architecture for action recognition.

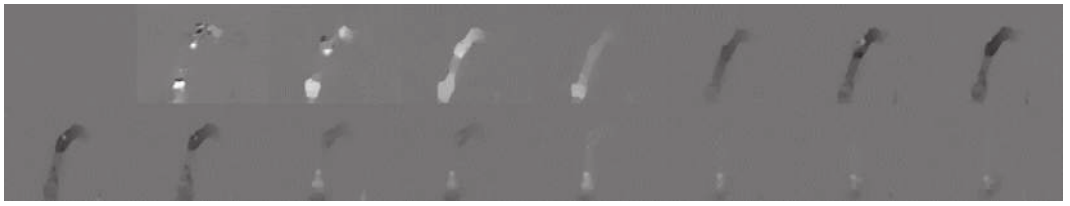

Figure 11: An example of optical flow images for the input of the two-stream ConvNets.

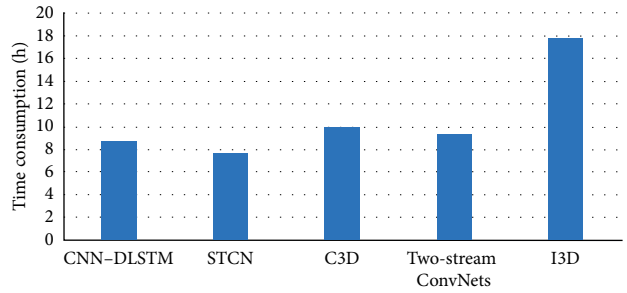

Figure 12: Time consumption of different action recognition methods.

4. Conclusions

Digital twin is not only for the facility to be built but also for the workers and construction equipment on a construction site. Research in CEAR is no less significant than that of HAR in terms of smart construction, because the perception of a construction activity needs to understand the what (object identification), where (object tracking), and how (action recognition) for both construction workers and equipment. Much of the research in the field of HAR can be applied in the construction industry. However, the research in the field of CEAR is very limited to date, due to a lack of high-quality datasets, the complexity of the application environment at construction sites, and the difficulty to benchmark the performance in CEAR.

In this research, the authors developed an open-source video dataset of 2,064 video clips with five action types for excavators and dump trucks, including a corresponding optical flow dataset. This dataset supplements the existing datasets for CEAR research and could potentially be used by other researchers in later studies. One major reason that hinders the research in the area of CEAR is the lack of high-quality datasets, and the authors encourage other researchers to share their datasets. Another contribution of this research in the field of CEAR is a new deep learning-based approach, STCN, which combines CNN and LSTM—where CNN is used to extract the image features from the video frame sequences and LSTM is used to extract the temporal features. The STCN proposed in this research achieved an F-1 score of 76.25% for the dataset developed earlier. In order to evaluate the performance of the STCN, a similar CEAR method and three of the best-performing deep learning-based approaches in the field of HAR, namely, the 3D ConvNets method, the two-stream ConvNets method, and I3D ConvNets method, were examined using the same dataset as input, and they achieved the F-1 scores of 75.25%, 73.55%, 76.26%, and 76.54, respectively. Those four methods either underperformed in comparison to STCN method or had a similar performance but needed significantly higher computing time. The time consumption of training the STCN is the shortest. This comparison indicates not only the advantage of the STCN but also the possibility of directly transferring some HAR methods to the field of CEAR.

There are some limitations in this research. First, the dataset developed in this study is still relatively insufficient compared to datasets available in other application areas using deep learning-based solutions. It is expected that a better accuracy rate would be achieved once a much larger dataset is developed. With a limited dataset, one possible solution is to pretrain the recognition algorithm using datasets for HAR. Second, this research only studied the recognition of actions for two types of construction equipment (excavators and dump trucks) and did not investigate whether different types of equipment (e.g., equipment with apparent joints such as excavators and equipment with no joints such as dump trucks) require different recognition algorithms in order to achieve better action recognition.

There is much important work to be done in the future. There exist hundreds of types of construction equipment in the construction industry, and each type of equipment has multiple actions. It is important to develop more comprehensive video datasets for various types of equipment under different conditions, for example, camera motion or disruptive weather (e.g., heavy winds or rain). In addition, it is important to study the perception of activities where multiple types of construction equipment are moving at the same time. In many cases, a surveillance camera will capture a scene of a construction site containing several pieces of equipment; this becomes a challenge when different types of equipment in the same frame require different action recognition algorithms to achieve better recognition performance.

Data Availability

The dataset used to support the findings of this study may be released upon application to Tianjin University, which can be contacted at jinyuezhang@tju.edu.cn.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Authors' Contributions

Jinyue Zhang, Lijun Zi, and Yuexian Hou contributed equally to this paper.

Acknowledgments

This work was supported in part by the National Key R&D Program of China (2018YFC0406900 and 2017YFE0111900), the National Natural Science Foundation of China (61876129), and the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie (721321).

References

[1] R. Zhang, Q. S. Li, and J. Chu, "Human action recognition algorithm based on 3D convolution neural network," *Computer Engineering*, vol. 45, no. 1, pp. 259–263, 2019.

[2] Y. Zhu, J. K. Zhao, N. Yi, and B. B. Zheng, "A review of human action recognition based on deep learning," *Acta Automatica Sinica*, vol. 42, no. 6, pp. 848–857, 2016.

[3] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. 976–990, 2010.

[4] A. Purohit and S. S. Chauhan, "A survey on human action recognition," *IOSR Journal of Computer Engineering*, vol. 19, pp. 43–50, 2017.

[5] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proceedings of the 26th Neural Information Processing Systems*, Lake Tahoe, CA, USA, December 2012.

[6] C. T. Haas, P. M. Goodrum, and C. C. Caldas, "Leveraging technology to improve construction productivity," *Technology Field Trials*, CII, vol. 3, p. 106, Austin, TX, USA, 2010.

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

[8] L. Song and N. N. Eldin, "Adaptive real-time tracking and simulation of heavy construction operations for look-ahead scheduling," *Automation in Construction*, vol. 27, pp. 32–39, 2012.

[9] F. Vahdatikhaki, A. Hammad, and H. Siddiqui, "Optimization-based excavator pose estimation using real-time location systems," *Automation in Construction*, vol. 56, pp. 76–92, 2015.

[10] C.-F. Cheng, A. Rashidi, M. A. Davenport, and D. V. Anderson, "Activity analysis of construction equipment using audio signals and support vector machines," *Automation in Construction*, vol. 81, pp. 240–253, 2017.

[11] C. A. Sabillon, A. Rashidi, B. Samanta, C.-F. Cheng, M. A. Davenport, and D. V. Anderson, "A productivity forecasting system for construction cyclic operations using audio signals and a Bayesian approach," in *Proceedings of the Construction Research Congress*, pp. 295–304, New Orleans, LA, USA, April 2018.

[12] N. Mathur, S. S. Aria, T. Adams, C. R. Ahn, and S. Lee, "Automated cycle time measurement and analysis of excavator's loading operation using smart phone-embedded IMU sensors," in *Proceedings of the International Workshop on Computing in Civil Engineering*, pp. 215–222, Austin, TX, USA, June 2015.

[13] K. M. Rashid and J. Louis, "Automated activity identification for construction equipment using motion data from articulated members," *Frontiers in Built Environment*, vol. 5, p. 144, 2020.

[14] J. Zou and H. Kim, "Using hue, saturation, and value color space for hydraulic excavator idle time analysis," *Journal of Computing in Civil Engineering*, vol. 21, no. 4, pp. 238–246, 2007.

[15] J. Gong, C. H. Caldas, and C. Gordon, "Learning and classifying actions of construction workers and equipment using bag-of-video-feature-words and bayesian network models," *Advanced Engineering Informatics*, vol. 25, no. 4, pp. 771–782, 2011.

[16] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[17] J. Seo, S. Han, S. Lee, and H. Kim, "Computer vision techniques for construction safety and health monitoring," *Advanced Engineering Informatics*, vol. 29, no. 2, pp. 239–251, 2015.

[18] A. Sargano, P. Angelov, and Z. Habib, "A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition," *Applied Sciences*, vol. 7, no. 1, 2017.

[19] D. Dawn and S. H. Shaikh, "A comprehensive survey of human action recognition with spatio-temporal interest point (stip) detector," *The Visual Computer*, vol. 32, no. 3, pp. 289–306, 2015.

[20] H. Li and M. Greenspan, "Multi-scale gesture recognition from time-varying contours," in *Proceedings of the Tenth IEEE International Conference on Computer Vision*, IEEE, Beijing, China, October 2005.

[21] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer Vision and Image Understanding*, vol. 104, no. 2-3, pp. 249–257, 2006.

[22] T. Ojala, M. Pietikainen, and D. Harwood, "Performance evaluation of texture measures with classification based on Kullback discrimination of distributions," in *Proceedings of the 12th IAPR International Conference on Pattern Recognition*, Jerusalem, Israel, October 1994.

[23] S. Sadek, A. Al-Hamadi, B. Michaelis, and U. Sayed, "An action recognition scheme using fuzzy log-polar histogram and temporal self-similarity," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, no. 1, 2011.

[24] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision*, vol. 103, no. 1, pp. 60–79, 2013.

[25] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proceedings of the Third IEEE International Conference on Computer Vision*, pp. 3551–3558, Sydney, Australia, June 2013.

[26] L. Liu, L. Shao, X. Li, and K. Lu, "Learning spatio-temporal representations for action recognition: a genetic programming approach," *IEEE Transactions on Cybernetics*, vol. 1, no. 46, pp. 158–170, 2016.

[27] F. Zhu, L. Shao, J. Xie, and Y. Fang, "From handcrafted to learned representations for human action recognition: a survey," *Image and Vision Computing*, vol. 55, pp. 42–52, 2016.

[28] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and F. F. Li, "Large-scale video classification with convolutional neural networks," in *Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1725–1732, Columbus, OH, USA, June 2014.

[29] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proceedings of the 28th Neural Information Processing Systems*, pp. 568–576, Montreal, Canada, December 2014.

[30] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proceedings of the 15th IEEE International Conference on Computer Vision*, pp. 4489–4497, Santiago, Chile, December 2015.

[31] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, July 2017.

[32] G. Varol, I. Laptev, and C. Schmid, "Long-term temporal convolutions for action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1510–1517, 2017.

[33] Y. H. Ng, M. Hausknecht, S. Vijayanarasimhan et al., "Beyond short snippets: deep networks for video classification," in *Proceedings of the 28th IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4694–4702, IEEE, Boston, MA, USA, June 2015.

[34] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: a dataset of 101 human actions classes from videos in the wild," 2012, http://www.crcv.ucf.edu/data/UCF101.php.

[35] J. Donahue, L. A. Hendricks, M. Rohrbach et al., "Long-term recurrent convolutional networks for visual recognition and description," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 677–691, 2016.

[36] L. Sevilla-Lara, Y. Y. Liao, F. Güney et al., "On the integration of optical flow and action recognition," in *Proceedings of the 5th Global Conference on Psychology Researches*, pp. 281–297, Istanbul, Turkey, February 2018.

[37] H. Kuehne, H. Jhuang, E. Garrote, T. A. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2556–2563, IEEE, Barcelona, Spain, November 2011.

[38] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "ActivityNet: a large-scale video benchmark for human activity understanding," in *Proceedings of the 28th IEEE Conference on Computer Vision and Pattern Recognition*, pp. 961–970, IEEE, Boston, MA, USA, June 2015.

[39] S. Abu-El-Haija, N. Kothari, J. Lee et al., "YouTube-8M: a large-scale video classification benchmark," 2016, http://arxiv.org/abs/1609.08675v1.

[40] J. Gong and C. H. Caldas, "An intelligent video computing method for automated productivity analysis of cyclic construction operations," in *Proceedings of the 2009 ASCE International Workshop on Computing in Civil Engineering*, pp. 64–73, Austin, TX, MSA, June 2009.

[41] R. Akhavian and A. H. Behzadan, "Construction equipment activity recognition for simulation input modeling using mobile sensors and machine learning classifiers," *Advanced Engineering Informatics*, vol. 29, no. 4, pp. 867–877, 2015.

[42] J. Cao, W. Wang, J. Wang, and R. Wang, "Excavation equipment recognition based on novel acoustic statistical features," *IEEE Transactions on Cybernetics*, vol. 42, no. 99, pp. 1–13, 2016.

[43] K. K. Han and M. Golparvar-Fard, "Potential of big visual data and building information modeling for construction performance analytics: an exploratory study," *Automation in Construction*, vol. 73, pp. 184–198, 2016.

[44] D. Roberts and M. Golparvar-Fard, "End-to-end vision-based detection, tracking and activity analysis of earthmoving equipment filmed at ground level," *Automation in Construction*, vol. 105, 2019.

[45] K. M. Rashid and J. Louis, "Times-series data augmentation and deep learning for construction equipment activity recognition," *Advanced Engineering Informatics*, vol. 42, Article ID 100944, 2019.

[46] J. Kim and S. Chi, "Action recognition of earthmoving excavators based on sequential pattern analysis of visual features and operation cycles," *Automation in Construction*, vol. 104, pp. 255–264, 2019.

[47] C. Szegedy, W. Liu, Y. Jia et al., "Going deeper with convolutions," in *Proceedings of the 28th IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, Boston, MA, USA, June 2015.

[48] A. Graves, A. R. Mohamed, and G. E. Hinton, "Speech recognition with deep recurrent neural networks," in *Proceedings of the 38th International Conference on Acoustics, Speech, and Signal Processing*, pp. 6645–6649, Vancouver, Canada, May 2013.

[49] B. C. Yin and L. C. Wang, "Review of deep learning," *Journal of Beijing University of Technology*, vol. 41, no. 1, pp. 48–59, 2015.

[50] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[51] Y. Bengio, "Learning deep architectures for AI," *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[52] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[53] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[54] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[55] Z. H. Zhou, *Machine Learning*, Tsinghua University Press, Beijing, China, 2016.

[56] W. J. Luo, Y. J. Li, R. Urtasun, and R. S. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Proceedings of the 30th Neural Information Processing Systems*, pp. 4898–4906, ACM, Barcelona, Spain, December 2016.

[57] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning*, pp. 448–456, Lille, France, July 2015.

[58] Y. L. Boureau, F. Bach, Y. Lecun, and J. Ponce, "Learning mid-level features for recognition," in *Proceedings of the 23rd IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2559–2566, IEEE, San Francisco, CA, USA, June 2010.

[59] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, pp. 315–323, Lauderdale, FL, USA, April 2011.

[60] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," in *Proceedings of the 2nd International Conference on Learning Representations*, San Diego, CA, USA, May 2015.

[61] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," *Journal of Machine Learning Research*, vol. 9, pp. 249–256, 2010.