2023 European Conference on Computing in Construction
40th International CIB W78 Conference
Heraklion, Crete, Greece
July 10-12, 2023

ENHANCING SINGLE-STAGE EXCAVATOR ACTIVITY RECOGNITION VIA KNOWLEDGE DISTILLATION OF TEMPORAL GRADIENT DATA

Ali Ghelmani, Amin Hammad
Concordia Institute for Information Systems Engineering, Concordia University, Montreal, Canada
ali.ghelmanirashidabad@concordia.ca
amin.hammad@concordia.ca

Abstract

Vision-based single-stage construction entity activity recognition methods that simultaneously analyze spatiotemporal information have been gaining popularity within the construction domain. However, a main disadvantage of these methods is their relatively low per-frame performance. Thus, necessitating additional post-processing to link the per-frame detection results and construct the corresponding action tubes. To address this problem, this study proposes DIGER, which stands for knowledge DIstillation of temporal Gradient data for Excavator activity Recognition. DIGER is built upon the You Only Watch Once activity recognition method and improves its performance by designing an auxiliary backbone to exploit the complementary information present in the temporal gradient data using knowledge distillation. The proposed method achieved an activity recognition accuracy of 93.6%, compared to the YOWO performance of 87.9% (5.7% improvement), on a large custom dataset of 1,060 videos.

Introduction

Conventionally, monitoring the activities of excavators and other earthmoving equipment on construction sites is done manually by on-site superintendents. However, manual monitoring can be very taxing, and prone to inaccuracies particularly on large construction sites (Chen et al., 2020; Roberts and Golparvar-Fard, 2019). Consequently, over the years many automated monitoring routines have been developed to provide project managers with crucial information on productivity and safety (Chen et al., 2020).

Automated activity recognition frameworks proposed in the construction domain can be broadly divided into non-vision-based and vision-based methods (Chen et al., 2020; Jung et al., 2022). Non-vision-based methods use various sensors, such as Global Positioning System (GPS) (Pradhananga and Teizer, 2013) or Ultra-Wideband (UWB) (Zhang et al., 2012) to determine activities based on equipment type, location, and movement information. However, due to the inherent limitations of location data, these methods may not accurately detect a wide range of activities. Furthermore, sensor installation can be expensive and time-consuming.

Traditional Computer Vision (CV)-based automated monitoring methods typically relied on manually designed features to extract useful information for activity recognition from images and videos (Gong et al., 2011; Soltani et al., 2017; Zhu et al., 2017). However, advancements in deep learning techniques have shown their superiority over traditional hand-crafted methods in various applications such as object detection (Redmon et al., 2016) and activity recognition (Donahue et al., 2015). Thus, leading to a corresponding shift in the use of deep learning-based techniques in the construction domain.

Convolutional Neural Networks (CNNs) are the fundamental component in most CV-based deep learning methods. In recent years, numerous 3D CNN-based methods for recognizing construction equipment activities have been proposed. 3D CNN-based methods incorporate the spatiotemporal data extraction into a single architecture, leading to more efficient and effective information extraction. Lou et al. (2020) proposed a multi-stage framework in which workers were first detected using the You Only Look Once (YOLOv3) network. The detected workers were then tracked in consecutive frames and their activities were classified using a 3D CNN architecture. Wang et al. (2021) also proposed a multi-stage framework using object detection and multiple-object tracking for progress monitoring of precast wall installation. In this method, the detected and tracked walls were considered installed if their displacement was less than a certain threshold after a given time interval. Although these frameworks can potentially extract more informative spatiotemporal features using 3D CNN architectures, their multi-stage approach still limits their accuracy. The main limitations of multi-stage methods include not being fully optimized, and the propagation of errors from earlier stages to the later ones, which results in the degradation of the

performance of the entire framework (Jung et al., 2022; Torabi et al., 2022).

In the CV domain, many single-stage activity recognition methods have been proposed, which alleviate the above-mentioned limitations. For instance, Tran et al. (2015) proposed one of the first end-to-end activity recognition methods using a 3D CNN-based architecture named C3D for simultaneous extraction of spatiotemporal data. Diba et al. (2017) proposed the Temporal 3D CNN (T3D) method employing 3D convolutional kernels with variable temporal length in their design to recognize short, mid, and long-term activities. To further improve performance, some works have proposed utilizing multiple modalities for activity recognition. For example, Simonyan and Zisserman (2014) proposed a two-stream method comprising of spatial and temporal networks to process RGB frames and optical flow data to extract appearance and motion features, respectively. Wang et al. (2016) proposed the Temporal Segment Network (TSN) by combining a spatial CNN for processing of the RGB data and a temporal CNN for processing the temporal gradient (TG) data, which is the difference between consecutive RGB frames.

Although incorporating TG or optical flow data modalities may enhance activity recognition, their use requires more computation to extract and process the additional modalities. To this end, some studies have leveraged knowledge distillation to improve model performance (Stroud et al., 2020; Xiao et al., 2022). Knowledge distillation refers to the transfer of information from a typically larger and more complex model to a smaller model to improve its performance while retaining the computational efficiency and ease of deployment (Gou et al., 2021). In the context of activity recognition, the knowledge transfer can also come from other sources of information such as TG and optical flow modalities. For instance, Stroud et al. (2020) proposed the Distillation 3D Network (D3D) consisting of two separate CNNs for processing RGB and optical flow data. To enhance the performance of the RGB CNN, the authors employed knowledge distillation to transfer knowledge from the optical flow network to its RGB counterpart.

Recently, inspired by the advances in the CV domain, some single-stage activity recognition methods have also been proposed in the construction domain. For instance, Jung et al. (2022) proposed a 3D CNN-based single-stage method for simultaneous detection of multiple construction equipment and recognizing their activities by using a 3D attention module and feature pyramid network in a single-stream architecture. Torabi et al. (2022) also proposed a single-stage method based on the You Only Watch Once (YOWO) method called YOWO53 for joint detection and classification of construction workers' activities by improving the 2D backbone of the YOWO method. Despite the advantages, the main limitation of these methods is their relatively low per-frame activity recognition performance. Therefore, requiring additional post-processing to link the per-frame detection results and construct the corresponding action tubes. Thereby, placing a major computational bottleneck on the real-time applicability of these methods.

To overcome the abovementioned limitation, the objective of this work is to improve the per-frame performance of the YOWO activity recognition method, hence eliminating the need for the post-processing linking stage. To this end, this work proposes DIGER, which stands for knowledge DIstillation of temporal Gradient data for Excavator activity Recognition. To improve activity recognition, an auxiliary backbone is designed to incorporate the complementary information present in the TG data using knowledge distillation. It should be noted that the TG and knowledge distillation are employed only during training, with the TG backbone discarded during inference. As a result, no extra computation or delay is required during inference.

Proposed Framework

The overall framework of the proposed method is shown in Figure 1. DIGER is comprised of two main components: (I) the original YOWO architecture including the 2D CNN and 3D CNN RGB backbones and the Channel Fusion and Attention Module (CFAM), and (II) the 3D CNN TG backbone and the modules added to perform knowledge distillation, such as Multi-Layer Projection (MLP) and the knowledge distillation loss function. As a result, the architecture of the proposed method consists of three branches. Two 3D CNN branches, which are used for processing of the RGB and TG data, and one 2D CNN branch which is used to process the last frame of the input clip to improve the localization accuracy. During training, knowledge distillation is used to transfer the information learned by the TG network to its RGB counterpart, thus improving its performance. A detailed description of each of these components and the training procedure is presented in the following sections.

YOWO

YOWO (Köpüklü et al., 2021) is a spatiotemporal activity recognition and localization method, which uses two branches in its architecture (Figure 1(I)). The 3D CNN branch extracts the spatiotemporal information from the input clips, while the 2D CNN branch is used to extract more accurate spatial features from the last frame of the input clip. YOWO uses the Darknet19 network in the 2D CNN branch, which is the backbone of the YOLOv2 (Redmon and Farhadi, 2017) object detection method. Since the Darknet19 network takes images as input, the shape of the input is of the form $[C \times H \times W]$, where C is equal to 3 RGB channels and H and W are the height and width of the input frame, respectively. The shape of the output feature map is of the form $[\mathrm{C}' \times \mathrm{H}' \times \mathrm{W}']$, where

C' is the number of output channels, $H' = H/32$, and $W' = W/32$.

The ShuffleNetV2_2.0 (Köpüklü et al., 2019) network is used as the backbone in the 3D CNN branch. The input to this Branch is a clip of the form $[C \times D \times H \times W]$, where D is the number of frames in the input clip, and C, H, and W are the frame dimensions similar to the input to the 2D CNN branch. Furthermore, the design of the ShuffleNetV2_2.0 backbone is modified in YOWO (Köpüklü et al., 2021) to result in an output of the form $[C^{"} \times D' \times H' \times W']$, where $C^{"}$ is the number of output channels, $D' = 1$, $H' = H/32$, and $W' = W/32$. By default, the output of the 3D CNN branch is 4-dimensional, while the output of the 2D CNN branch is 3-dimensional. Considering that the outputs of these two branches are combined before being input into the CFAM module, the size of their corresponding outputs should be compatible. As a result, the 3D CNN branch is designed to have $H' = H/32$, $W' = W/32$, and a reduced depth component ($D' = 1$), which can be dropped and hence become three-dimensional in effect.

The main component providing the performance boost for the YOWO model is the CFAM module, which operates on the output of the 2D CNN and 3D CNN branches. To this end, the outputs of these two branches are concatenated along the channel dimension before being input into the CFAM module to include both the spatiotemporal and the refined spatial information. The CFAM module uses attention mechanism to capture the inter-channel dependencies. Finally, YOWO uses the focal loss (Lin et al., 2017) for activity classification and the smooth L_1 loss (Girshick, 2015) for bounding box regression.

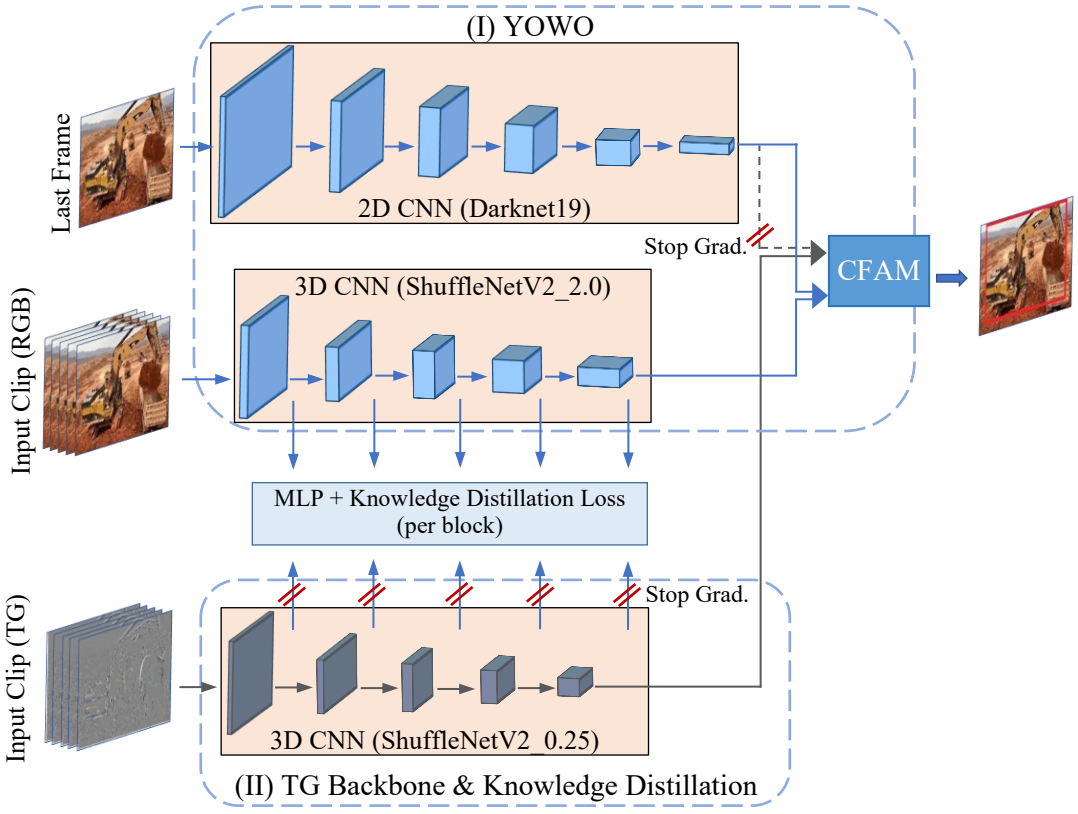

Figure 1. Overall framework of the proposed DIGER method

Temporal Gradient

In this work, a separate 3D CNN branch is added to the original YOWO model to process and extract features from the TG data and consequently improve the activity recognition performance of the model. TG is obtained by calculating the difference between two RGB frames in a video and represents the dynamic changes in the temporal information. When selecting a backbone for processing the TG modality, two factors should be considered. Firstly, the TG modality primarily captures dynamic changes rather than spatial information, so the backbone should be designed to prioritize the extraction of temporal information. Secondly, integrating a smaller auxiliary backbone for extracting temporal information can improve efficiency while allowing the larger backbone to focus on the extraction of spatial information (Feichtenhofer et al., 2019; Xiong et al., 2021). In this work, the ShuffleNetV2_0.25 network was chosen as the TG backbone due to its relatively small size, with about 0.64×10^6 parameters compared to the 5.56×10^6 parameters of the ShuffleNetV2_2.0 network used as the 3D CNN RGB backbone. Table 1 presents the details of the architectures of the two networks along with the

output sizes of different blocks for an input clip of size 3×16×224×224 ($[C \times D \times H \times W]$).

Table 1. Architecture of the 3D RGB and TG backbones

Layer block	Output size	
	ShuffleNetV2_2.0	ShuffleNetV2_0.25
Stem_block	24×8×56×56	24×8×56×56
block_1	224×4×28×28	32×4×28×28
block_2	488×2×14×14	64×2×14×14
block_3	976×1×7×7	128×1×7×7
Final_block	2048×1×7×7	256×1×7×7

As stated earlier, the 3D CNN RGB and TG backbones extract complementary information from the different modalities of the same input clip. In order to prevent any potential divergence in their respective training procedure from affecting the semantics of the extracted features, an approach similar to YOWO was adopted in this work. This involved training the TG backbone in conjunction with the 2D CNN backbone and the CFAM module. However, the TG backbone is mainly concerned with extracting temporal information, while the 2D CNN backbone is designed to extract spatial information. As a result, to allow the effective and efficient extraction of the desired information by both backbones, a stop gradient operation is applied on the 2D CNN backbone to prevent the TG data from affecting it. The stop gradient operation is a mechanism used to prevent the backpropagation of gradients through a given branch and consequently, prevent the updating of the affected weights. In this case, a stop gradient operation was applied to the 2D CNN backbone to prevent the gradients generated during the training of the TG backbone from affecting its weights. Thus, ensuring that the 2D CNN backbone is only trained in conjunction with the 3D CNN RGB backbone.

Knowledge Distillation

The utilization of an additional backbone and the increased computational demands of the TG data may result in slower inference performance. To overcome this challenge, knowledge distillation is used in this work to transfer the knowledge learned by the TG backbone to the corresponding 3D CNN RGB backbone during training, while the TG backbone is discarded during inference. There are numerous different approaches to perform knowledge distillation. However, assigning a loss value to measure the dissimilarity between the output of the corresponding blocks of the two networks is the most common. In this work, the cross-entropy loss is used to measure the dissimilarity.

It should be noted that, performing knowledge distillation in the high dimensional space of the output of different blocks would require a large amount of data for training of the model to converge. To address this problem, Multi-Layer Projection (MLP) is used to map the outputs of the different blocks to a lower dimensional projection space for calculating the loss function. In this regard, the dimension of the projection space plays a vital role in the effectiveness of the knowledge transfer approach, and consequently, the convergence of the knowledge distillation loss. It should be noted that separate MLP layers are utilized for each block involved in knowledge distillation. Figure 1 (II) gives a comprehensive illustration of the different modules used in knowledge distillation for each block of the two networks.

Finally, one crucial aspect of the knowledge distillation approach is to prevent a degenerate loop in which the 3D CNN RGB backbone learns from and then later teaches the TG backbone. To this end, a stop gradient operation is used to prevent the TG backbone from receiving any gradient (as explained above) from the knowledge distillation loss. This ensures the gradients would only flow in the direction of the 3D CNN RGB backbone. Furthermore, given that the 3D CNN RGB backbone has more spatial information than the TG backbone, the use of the stop gradient would enable the TG backbone to only focus on the extraction of fine-grained motion features and not to be disturbed by the RGB model.

Experiments

Dataset Description

The video clips used in creating the custom dataset used in this work were manually collected from various sources including videos posted on websites such as YouTube, and also videos used in similar research works, which were made publicly available (Roberts and Golparvar-Fard, 2019). Each video clip contains one or more excavators performing three types of activities: digging, swinging, and loading the trucks. To add to the diversity of the collected dataset, the videos are collected from 25 different construction sites, incorporating various site conditions, such as different camera angles, illuminations, occlusions, weather conditions, and video resolutions. Table 2 provides the statistics of the collected dataset. Some video clips contain more than one excavator with each excavator involved in a different activity. As a result, there is discrepancy between the sum of the number of frames and clips reported for each individual activity, and the total reported values in Table 2.

Table 2. Statistics of the collected dataset

Activity type	Number of video clips	Number of frames	Average clip length (sec)
Digging	295	64,436	7.28
Swinging	476	51,441	3.60
Loading	321	51,632	5.36
Total	1,060	163,295	5.13

Implementation Details

In this work, Stochastic Gradient Descent (SGD) algorithm with the momentum value of 0.9 is used as optimizer during training. The learning rate is linearly warmed-up in the first five epochs followed by a half-period cosine annealing learning rate scheduling strategy without restarts (Loshchilov and Hutter, 2016). All models are trained with a batch size of 128 on three RTX A6000 GPUs in Ubuntu 20.04 and Python 3.8 environment and PyTorch 1.12. 80% of the videos in the dataset were randomly selected for training, 10% were selected for validation, and the remaining 10% were used for testing. The two ShuffleNetV2_x 3D CNN networks were pre-trained on the large-scale Kinetics-600 dataset (Carreira et al., 2018). All layers of these networks are fine-tuned on the excavator dataset using their corresponding RGB and TG data modalities. The 2D CNN network Darknet19 is pre-trained on the COCO dataset (Lin et al., 2015). All layers of the 2D CNN network are also fine-tuned on the excavator dataset.

Experimental Results

Table 3 presents the results of the proposed DIGER method for activity recognition on the test dataset, along with the results of the original YOWO method for comparison. It should be noted that these results present the per-frame performance of both methods without the post-processing linking stage. The reported classification accuracy indicates the activity recognition performance of the model. To further investigate the effectiveness of the proposed DIGER method, Table 3 also presents the effect of various design choices in performing knowledge distillation on the final model performance. It can be seen that in all cases, adding the TG backbone and utilizing knowledge distillation improves the activity recognition performance of the model compared to the original YOWO model.

Table 3 also shows the impact of different number of blocks used for knowledge distillation on the performance of the model for different sizes of the projection space dimension. It can be seen that changing the dimension of the projection space in which knowledge distillation loss is calculated has a significant impact on the final performance of the model for all combination of distillation blocks (Table 3). In particular, a projection space size of 64 appears to be too small to distinguish between the mappings of the outputs of different blocks of the two networks for different input data, leading to ineffective knowledge transfer. Conversely, a projection space size of 512 results in increased number of parameters and a larger projection space, making it challenging for the model to transfer knowledge effectively given the amount of available training data. The optimal results are obtained with projection spaces of size 128 or 256, depending on the number of blocks used for knowledge distillation. It can be seen in Table 3 that the best performance is achieved by using three blocks for knowledge distillation in a projection space of size 128, resulting in a 93.6% activity recognition accuracy, which corresponds to a 5.7% improvement over the original YOWO.

Table 3. Comparison of the per-frame performance for YOWO and DIGER

Model	Distill blocks	Projection dimension	Classification accuracy (%)
YOWO	---	---	87.9
DIGER	Final_block	64	88.6
		128	91.1
		256	91.8
		512	89.6
	Final_block, block_3	64	87.8
		128	88.4
		256	91.5
		512	88.0
	Final_block, block_3, block_2	64	90.2
		128	**93.6**
		256	91.4
		512	88.6

Interestingly, the results in Table 3 demonstrate that while the model performance slightly declines when including two blocks for knowledge distillation, including three blocks leads to the best activity recognition performance. It should be noted that obtaining improved performance when including more blocks in knowledge distillation is consistent with the results reported in previous studies (Xiao et al., 2022). However, the reason for the inconsistency in the improvement when including two blocks can be attributed to the structure of the ShuffleNetV2_x backbone used in this work. In particular, the inclusion of earlier blocks is intended to enforce the similarity between the two networks at different semantic stages of feature extraction. However, there is only one 3D CNN layer between block_3 and the Final_block (Table 1). As a result, there is not much difference between the semantics of the extracted features of these two blocks in the ShuffleNetV2_x backbone. Thus, the additional number of parameters of the MLP module for needed the second knowledge distillation block outweighs the minor benefits of its inclusion.

Conclusions and Future Work

This study proposes DIGER, a novel method for automated excavator activity recognition on construction sites. DIGER is built upon the YOWO activity recognition method and improves its performance by employing the TG data modality and knowledge distillation. The proposed method improves the activity recognition accuracy by designing an auxiliary backbone to process the complementary information present in the TG data modality and transfering its knowledge using knowledge distillation. DIGER achieved excellent performance on a large custom dataset of 1060 videos, with an activity recognition accuracy of 93.6% compared to the YOWO performance of 87.9% (5.7%

improvement). It should be noted that TG data and knowledge distillation are only used during training. As a result, the proposed method can be deployed in real-time applications without any extra computation or delay during inference. Furthermore, it should be noted that while in this study only excavator activities are considered, similar improvements can be expected using the proposed approach to recognize the activities of other types of construction equipment as well.

Considering the current challenges in the development of a general construction activity recognition method, a possible direction for further research can focus on improving the localization accuracy of the YOWO model, in addition to the activity recognition accuracy improved in this work. Improving the localization will enable the development of activity recognition methods for simultaneous detection of the activities of multiple construction entities on a per-frame basis. Such a system is essential for facilitating automated interaction between different construction entities. For instance, the interaction between workers and construction equipment, which requires accurate localization to ensure workers' safety when working in close proximity of heavy equipment.

References

Carreira, J., Noland, E., Banki-Horvath, A., Hillier, C., Zisserman, A., 2018. A Short Note about Kinetics-600. arXiv:1808.01340 [cs].

Chen, C., Zhu, Z., Hammad, A., 2022. Critical Review and Road Map of Automated Methods for Earthmoving Equipment Productivity Monitoring. Journal of Computing in Civil Engineering 36, 03122001. https://doi.org/10.1061/(ASCE)CP.1943-5487.0001017

Chen, C., Zhu, Z., Hammad, A., 2020. Automated excavators activity recognition and productivity analysis from construction site surveillance videos. Automation in Construction 110, 103045. https://doi.org/10.1016/j.autcon.2019.103045

Diba, A., Fayyaz, M., Sharma, V., Karami, A.H., Arzani, M.M., Yousefzadeh, R., Van Gool, L., 2017. Temporal 3D ConvNets: New Architecture and Transfer Learning for Video Classification. https://doi.org/10.48550/arXiv.1711.08200

Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T., 2015. Long-term recurrent convolutional networks for visual recognition and description, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2625–2634.

Feichtenhofer, C., Fan, H., Malik, J., He, K., 2019. SlowFast Networks for Video Recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision. Presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6202–6211. https://doi.org/10.1109/ICCV.2019.00630

Girshick, R., 2015. Fast R-CNN. Presented at the Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448.

Gong, J., Caldas, C.H., Gordon, C., 2011. Learning and classifying actions of construction workers and equipment using Bag-of-Video-Feature-Words and Bayesian network models. Advanced Engineering Informatics, Special Section: Advances and Challenges in Computing in Civil and Building Engineering 25, 771–782. https://doi.org/10.1016/j.aei.2011.06.002

Gou, J., Yu, B., Maybank, S.J., Tao, D., 2021. Knowledge Distillation: A Survey. International Journal of Computer Vision 129, 1789–1819. https://doi.org/10.1007/s11263-021-01453-z

Jung, S., Jeoung, J., Kang, H., Hong, T., 2022. 3D convolutional neural network-based one-stage model for real-time action detection in video of construction equipment. Computer-Aided Civil and Infrastructure Engineering 37, 126–142. https://doi.org/10.1111/mice.12695

Köpüklü, O., Kose, N., Gunduz, A., Rigoll, G., 2019. Resource efficient 3d convolutional neural networks, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops.

Köpüklü, O., Wei, X., Rigoll, G., 2021. You Only Watch Once: A Unified CNN Architecture for Real-Time Spatiotemporal Action Localization. arXiv:1911.06644 [cs].

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollar, P., 2017. Focal Loss for Dense Object Detection. Presented at the Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollár, P., 2015. Microsoft COCO: Common Objects in Context. arXiv:1405.0312 [cs].

Loshchilov, I., Hutter, F., 2016. SGDR: Stochastic Gradient Descent with Restarts, in: International Conference on Learning Representations. pp. 1–16. https://doi.org/10.48550/arXiv.1608.03983

Luo, X., Li, H., Yu, Y., Zhou, C., Cao, D., 2020. Combining deep features and activity context to improve recognition of activities of workers in groups. Computer-Aided Civil and Infrastructure Engineering 35, 965–978. https://doi.org/10.1111/mice.12538

Pradhananga, N., Teizer, J., 2013. Automatic spatio-temporal analysis of construction site equipment operations using GPS data. Automation in Construction 29, 107–122. https://doi.org/10.1016/j.autcon.2012.09.004

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 779–788.

Redmon, J., Farhadi, A., 2017. YOLO9000: Better, Faster, Stronger. Presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7263–7271.

Roberts, D., Golparvar-Fard, M., 2019. End-to-end vision-based detection, tracking and activity analysis of earthmoving equipment filmed at ground level. Automation in Construction 105, 102811. https://doi.org/10.1016/j.autcon.2019.04.006

Simonyan, K., Zisserman, A., 2014. Two-Stream Convolutional Networks for Action Recognition in Videos, in: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems. Curran Associates, Inc.

Soltani, M.M., Zhu, Z., Hammad, A., 2017. Skeleton estimation of excavator by detecting its parts. Automation in Construction 82, 1–15. https://doi.org/10.1016/j.autcon.2017.06.023

Stroud, J., Ross, D., Sun, C., Deng, J., Sukthankar, R., 2020. D3D: Distilled 3D Networks for Video Action Recognition. Presented at the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 625–634.

Torabi, G., Hammad, A., Bouguila, N., 2022. Two-Dimensional and Three-Dimensional CNN-Based Simultaneous Detection and Activity Classification of Construction Workers. Journal of Computing in Civil Engineering 36, 04022009. https://doi.org/10.1061/(ASCE)CP.1943-5487.0001024

Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M., 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. Presented at the 2015 IEEE International Conference on Computer Vision (ICCV), IEEE Computer Society, pp. 4489–4497. https://doi.org/10.1109/ICCV.2015.510

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L., 2016. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition, in: Leibe, B., Matas, J., Sebe, N., Welling, M. (Eds.), Computer Vision – ECCV 2016. Springer International Publishing, Cham, pp. 20–36.

Wang, Z., Zhang, Q., Yang, B., Wu, T., Lei, K., Zhang, B., Fang, T., 2021. Vision-Based Framework for Automatic Progress Monitoring of Precast Walls by Using Surveillance Videos during the Construction Phase. Journal of Computing in Civil Engineering 35, 04020056. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000933

Xiao, J., Jing, L., Zhang, L., He, J., She, Q., Zhou, Z., Yuille, A., Li, Y., 2022. Learning from Temporal Gradient for Semi-Supervised Action Recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3252–3262. https://doi.org/10.1109/CVPR52688.2022.00325

Xiong, B., Fan, H., Grauman, K., Feichtenhofer, C., 2021. Multiview Pseudo-Labeling for Semi-Supervised Learning from Video, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7209–7219. https://doi.org/10.1109/ICCV48922.2021.00712

Zhang, C., Hammad, A., Rodriguez, S., 2012. Crane Pose Estimation Using UWB Real-Time Location System. Journal of Computing in Civil Engineering 26, 625–637. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000172

Zhu, Z., Ren, X., Chen, Z., 2017. Integrated detection and tracking of workforce and equipment from construction jobsite videos. Automation in Construction 81, 161–171.