Long video-based action segmentation for earthmoving excavators using improved Temporal Convolutional Network models

Xuejian Chen[1], Shuijie Qin[1], Zhongchen Bai[1], Yuanjun Guo[2], Zhile Yang[2], Rui Jiang[2], Chengke Wu[2], Shilong Su[3].

[1] Guizhou University, Guizhou, 550025, China
[2] Shenzhen Institute of Advanced TechnologyChinese Academy of Sciences, Shenzhen, 518055, China
[3] China Construction Science & Technology Group CO., LTD, Shenzhen, 518118, China

Abstract. Earthwork excavator, as an all-terrain and high-efficiency earthwork excavation equipment, has been widely used in earthwork sites. It is very necessary to analyze the work of earthmoving excavator by means of machine vision. In this paper, the action segmentation method based on long video was applied to the analysis and recogniton of the excavator′s action, and compared with other two current best action segmentation models using the real construction site video. Firstly, the sequence features of the excavator′s work video obtained at the cnstruction site was extracted through 3D convolution method, and then two differernt networks with the extracted sequence features were trained and tested. The experimental results showed that the average frame accuracy of MS-TCN model and ASRF model in excavator action segmentation were 82.6490% and 86.1042% respectively. However, for the recognition task under different working environment, the performance of the two models is quite different. The experimental results manifest that the motion segmentation model based on long video reached good results in excavator motion recognition in earthmoving operation. And it is helpful to analyze the long video working behavior sequence of excavator. This research contributes to the identification of critical elements that explains serial action and to the development of a new application scenario for vision-based behavior segmentation network. Additionally, the results of this study were helpful to automatically analyze the working efficiency and monitor the productivity of earthmoving excavators. Using this kind of data-driven decision can improve the work efficiency of earthmoving excavator and promote the project progress.

Keywords
Action segmentation, Earthmoving excavator, Machine vision, Productivity monitoring, Temporal convolutional network.

1 Introduction

Earthmoving operations are necessary parts, such as dam construction, infrastructure work, road construction, and airport construction. In these operations, earth materials, such as soil, are dug from the ground and transferred from one location to another location. Most of the construction activities in these operations are driven by heavy equipment (Mahmood, B *et al.*2022). As a kind of all-terrain

construction machinery, earthmoving excavators can be adapted to all kinds of poor working environment, and now has become an indispensable engineering equipment in many construction sites. Furthermore, an excavation construction site is a complex system with a high dynamics and instantaneity. It is a unique unstructured environment, just like any other construction sites (Naghshbandi, S. N *et al.* 2021). The operation of earthmoving equipment in complex and dynamic excavation sites will cause many safety risks and may result in accidental damages, such as worker casualties, equipment damage or property loss (Francis, R and Bekera, B 2014).To make full use of excavators' capabilities and improve the safety of earthmoving operations, it is essential to recognize their operation types continuously such as 'Bucket Digging', 'Dumping Bucket', 'Turning', 'Moving', and 'Stopping'; through continuous monitoring of excavator operations, field managers can obtain detailed information about their operational efficiency and measure operational indicators such as cycle time , dele time and direct rate of operation (Golparvar-Fard, M *et al.* 2013). Based on these indicators, site managers not only can monitor whether the excavator operates normally, but also can estimate the time and cost required to complete repeated earthmoving operations to improve the efficiency of earthmoving operations (Kim, J *et al.* 2018); for instance, the time required for a given project can be computed based on 1) the average cycle time for excavating and transporting a certain amount of soil and 2) the total amount of soil to be excavated and transported. Then, site managers can make important decisions related to the project such as resource allocation, work planning and scheduling, and operator training based on the quantitative results and the associated analysis (Kim, J *et al.* 2019). This data-driven decision making can provide opportunities to improve operational efficiency and to complete engineering projects.

2 Literature Review

Video has long been used as an effective and inexpensive technology for productivity analysis on construction sites. However, as schedules of modern construction projects become more and more compressed, the fast and efficient characteristics of video analysis are in sharp contrast to the heavy work of intensive manual review process, and there is a greater demand for various production activities of video analysis on construction sites (Gong, J and Caldas, C. H 2010). Therefore, construction sites urgently need an efficient automated supervision method to improve site construction safety and improve work efficiency (Chi, S *et al.* 2009). One of the most widely used technologies today is an Internet-of-Things-based (IoT-based) system. The IoT-based approaches involve attaching electronic sensors to target objects and their components, collecting continuous point locations of the sensors, and analysing the physical motions of the equipment, acceleration, velocity, and orientation (Kim, H *et al.* 2018; Yang, K *et al.* 2017; Ahn, C. R *et al.* 2015; Akhavian, R and Behzadan, A. H 2016; Vahdatikhaki, F *et al.* 2015). However, during excavation, the working device probably collide with soil or rock, which may damage the sensor on excavators.

In recent years, convolutional neural network has made great progress in machine vision and has become a research hotpot in the field of artificial intelligence (Shi, Y *et al.* 2021). In 2018, a model was proposed, which can automatically identify distributed feature representations of data and the problem of manually designing features in conventional machine learning had been solved (Huang, L. W *et al.* 2018). Since the working state of earthmoving excavators is mostly a series of sequential actions, some researchers use action recognition based on pictures or video clips to identify the actions of excavators respectively, and then combine multiple actions according to the sequence logic of the actions of excavators to obtain the combined actions of a long-time sequence (Kim, J *et al.* 2019; Gong, J *et al.* 2011; Chi, S and Caldas, C. H 2011; Chi, S and Caldas, C. H 2012; Kim, H *et al.* 2018). However, the application of long video-based action segmentation technology for excavators' action segmentation can better analyse its temporal action, eliminating the process of multi-action combination according to

the timing logic of action. The traditional method of long video action segmentation mainly divided into two steps, by getting framewise probabilities and then putting them to high-level temporal models, recent approaches use temporal convolutions to directly classify the video frames (Farha, Y. A and Gall, J. 2019). In this research, Because of the excellent performance of these two networks in human motion segmentation and their lightweight models, we applied Multi-Stage Temporal Convolutional Network (MS-TCN++) and Action Segment Refinement Framework (ASRF) to the excavator action segmentation respectively, and compared the performance of the two methods (Li, S. J *et al*. 2020; Ishikawa, Y *et al.* 2021).

3 Research Methodology

In this study, we first adopted the 3D convolution method to preprocess the site video data and extracts its temporal features. Then we input the extracted temporal features into Multi-Stage Temporal Convolutional Network (MS-TCN++) and Action Segment Refinement Framework (ASRF) for training respectively, and compare the performance of the two networks.
To improve the data quality, first we clip the collected video, discarding the missing fragments of the target and too much interference. Then clip the video into segments of one minute to two minutes long, and all video segments contain multiple tag actions. After that, we input the clipped videos into the C3D network to extract its temporal features.

3.1 Features Extraction

In this section, features extraction is needed in order to realized long video process. To extract the

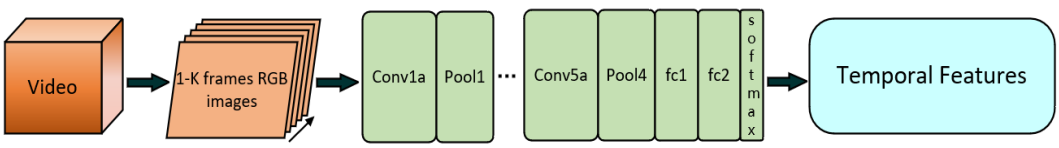

Figure 1. Convolutional 3D(C3D) Network Structure

temporal characteristics of video data, this paper adapted one of the most popular and outstanding object detection models, Convolutional 3D(C3D) (Tran, D *et al.* 2015). This model was proposed by Tran, D and has been used extensively in action recognition. The network structure used for feature extraction of excavators is shown in Fig.1. Compared to 2D ConvNet, 3D ConvNet is implemented through 3D filter and 3D pooling layer, so it can better model temporal information. In 3D ConvNet, convolution is performed simultaneously in space and temporal, while 2D ConvNet only performs convolution in space. The 2D convolution output of one image is an image, and the 2D convolution output of multiple images is an image too. Therefore, 2D ConvNet loses the time information of the input signal, while 3D ConvNet convolutes in temporal while convoluting in space. It retains the temporal information of the input signal, resulting in the input volume.
To extract C3D features, the original long video is divided in to 16-frame long clips with 8 frames overlapping between two consecutive clips. These clips are transferred to C3D network to extract 3D features. These 3D features are averaged to form a 2048-dim video descriptor which is then followed by an L2-normalization. Then we combine the 3D features of these clips according to the timing to obtain the temporal features of the whole video. The shape of the temporal feature of each video is 2048*k (k is the number of frames of the video).

3.2 Excavator Action Segmentation via Temporal Convolutional Network

In this part, we use MS-TCN++ and ASRF to recognize and segment actions of excavator working videos, which are obtained from different perspectives and different working environments. In the work of excavator, our classification of excavators' operation types is mainly divided into five kinds, namely 'digging', 'dumping', 'swinging', 'moving', and 'stopping'. In the actual working process, the 'digging' and 'dumping' are generally connected with the 'swinging'. In order to better segment these two kinds of jointed action, we used the same discriminant method to segment the jointed out of these two kinds of actions in the calibration data set. Our method was that when we segmented and recognized these two kinds of actions, we mainly focused on whether there was contact between the action interaction between the bucket and the excavated material. When there is no contact between the bucket and the excavated material, we define it as the end of the 'digging' and the beginning of the 'swinging'. Similarly, when the bucket releases the excavated material, we defined the moment when the bucket becomes empty as the end of the 'dumping' and the beginning of the next action.

3.2.1 Multi-Stage Temporal Convolutional Network for Excavator's Action Segmentation

First, we trained the MS-TCN++ model with a portion of the temporal features extracted from the video. The MS-TCN++ model consists of single stage TCN(SS-TCN) and multi-stage TCN(MS-TCN). The single stage consists of only temporal convolution layers, and pooling layers are not used in order not to reduce the temporal resolution. Also, to avoid limiting the size of input, the single stage dose not use the fully connected layers. The specific operations are shown in figure 2

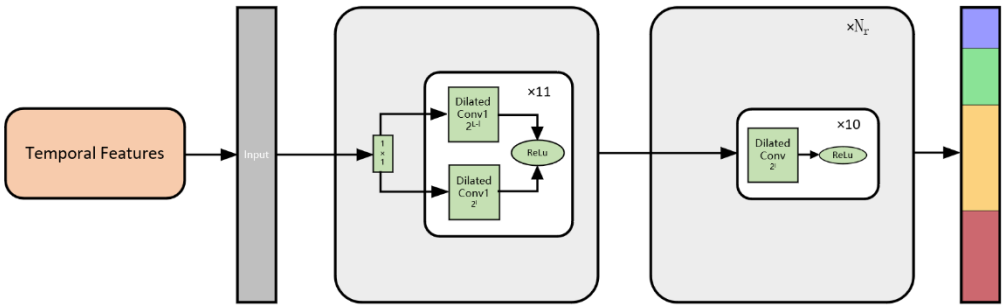

Figure 2. Overview of MS-TCN++

The first layer of SS-TCN is a 1*1 convolution layer, which aims to adjust the dimension of input temporal features to match the number of network feature maps. Then, this layer is followed by several dilated 1D convolution layers whose dilation factor is doubled at each layer. And a ReLU activation layer is added in front of each dilation convolution layer. The specific process can be described as follows:

$$\widehat{H}_l = ReLU(W_d * H_{l-1} + b_d) \quad (1)$$

$$H_l = H_{l-1} + W * \widehat{H}_l + b \quad (2)$$

where H_l is the output of layer l, $*$ denotes the convolution operator, $W_d \in R^{3\times D\times D}$ are the weights of the dilated convolution filters with kernel size 3 and D is the number of convolutional filters, $W \in R^{1\times D\times D}$ are the weights of a 1×1 convolution, and b_d, $b \in R^D$ are bias vectors. The dual dilated convolution in SS-TCN can not only obtain a large perception field with fewer layers but also give lower layers a larger receptive field, which can reduce the complexity of the model and prevent over-fitting in the

training process. The output of SS-TCN is used as the input of MS-TCN, whose effect is an incremental refinement of the predictions from the previous stages.

3.2.2 Action Segment Refinement Framework for Excavator's Action Segmentation

In this section, we introduce the approach for action segmentation named Action Segment Refinement Framework (ASRF) which proposed by Yuchi Ishikawa. The model structure is shown in figure 3. The Action Segment Refinement Framework (ASRF) consists of a long-term feature extractor, an Action Segmentation Branch (ASB) and a Boundary Regression Branch (BRB). The specific process is that the video temporal features are first input into a long-term feature extractor, which provides shared features for the following Action Segment Branch and Boundary Regression Branch. Then the two branches output the frame-based action classification results and boundary probability respectively according to the shared features. First, Long-Term feature extractor is composed of temporal convolutional network with dilation convolution. Such network structure enables it to have full temporal resolution and obtain a larger visual field of perception. Its function is to obtain long-term dependence between action segments and extract rich features of videos for the following branches. Long-term features derived from Long-Term feature extractor are fed into Action Segment Branch to predict frame-based action classes. In order to better capture time dependence, identify action fragments, and prevent over-segmentation errors, simple Multi-Stage TCN is mainly used in Action segment Branch to carry out preliminary refinement of prediction results.

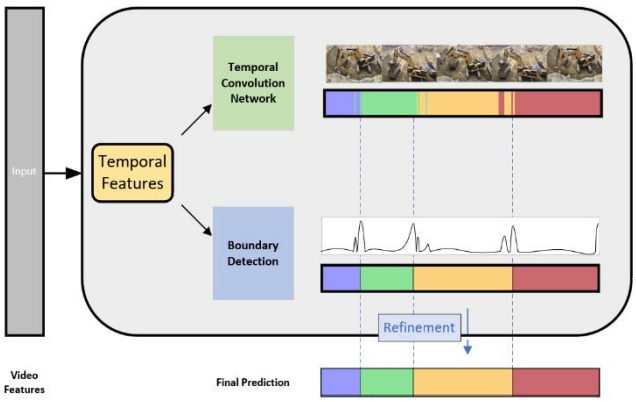

Figure 3. Overview of Action Segment Refinement Framework

Although the superposition of TCNs improves the performance of motion segmentation, over-segmentation errors still exist in the prediction. To solve this problem, we input the long-term features into the Boundary Regression Branch (BRB). The main work of BRB is to return the motion boundary probability in the video, and then refine the motion segmentation result in Action Segmentation Branch (ASB) by using the motion boundary probability.

4 Findings and Discussion

In order to compare the abilities of the two models, we identified three different excavator working scenes with the trained network, namely, the video shooting is stable and the key part of the excavator is not covered, the video shooting is stable but the key part of the excavator is covered, and the video shooting is unstable. From this we can determine which model is more practical. For the authenticity of the experiment, we collected the video of excavator work from the actual earthmoving operation site. All videos are captured using the same smartphone. Since the excavator in the video is carrying out earthwork, the operation type pf each frame of the excavator in the video can be marked. In order to

consider various influences in the actual situation, the videos we collected were divided into the three categories mentioned above. We collected a total of about 60 minutes of video data, and the video ratio of three different excavator working scenes was about 2:1:1. The video resolution is 480×640 pixels and the frame rate is 30 fps for a total of 6480,000 frames of video data. We took 7/10 of the video data of various types as the training set, and the rest as the test set to test the training results of the model.
The specific results and analysis are as follows. Table 1 shows the experimental results of MS-TCN2 and ASRF for excavator action segmentation respectively. In terms of the accuracy of action segmentation, the average recognition accuracy of MS-TCN2 is 82.6490%, edit score is 86.9949, F1 score is {90, 90, 70}. The frame accuracy of ASRF is slightly higher than that of MS-TCN, reaching 86.1042%. The edit score and F1 score of ASRF are slightly lower than those of MS-TCN2, which are 83.0808 and {88.5, 82.6, 75.4} respectively.

Table 1. Comparison of Excavator's Behavior Segmentation Performance between MS-TCN2 and ASRF

Methods	**Evaluations**		
	Acc	**Edit**	**F1@ {10,25,50}**
MS-TCN2	82.6490	86.9949	90, 90, 70.0
ASRF	86.1042	83.0808	88.5, 82.6, 75.4

By comparing the performance of the two models in three different scenes, when the shooting is stable and the key parts of the excavator are not blocked, the performance of the two models in excavator action segmentation is relatively good and the frame accuracy reach over 90%, as shown in Figure 4. When the shooting is stable, but the key part of the excavator is blocked. Compared with ASRF, the performance of MS-TCN is very poor, with more action recognition errors and over-segmentation, as shown in Figure5. In addition, we found that in unstable shooting scenes, both models had action identification errors and over-segmentation, and the accuracy of excavator action segmentation was very low, as shown in Figure 6.

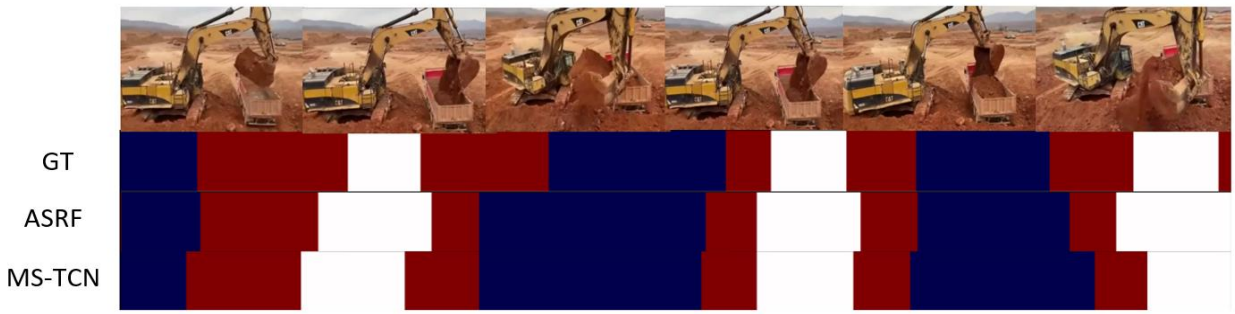

Figure 4. Long video segmentation result when video is stable and excavators are not blocked.

Figure 5. Long video segmentation result when video is stable and excavators are blocked

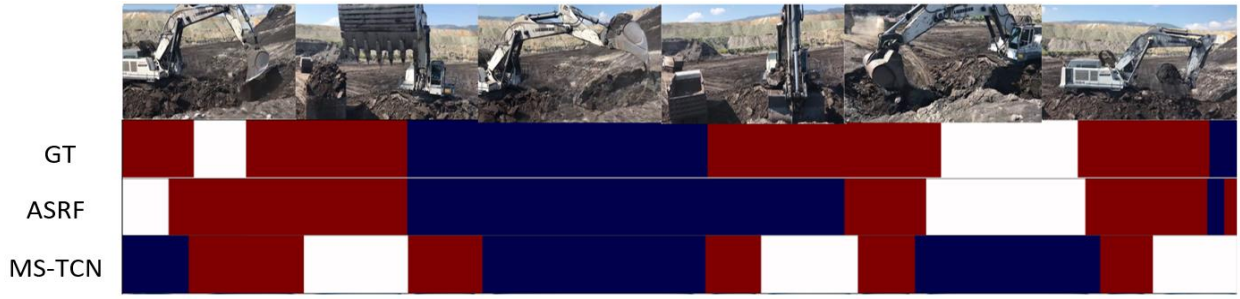

Figure 6. Long video segmentation result when video is unstable

5 Conclusions and Further Research

This paper applies long video action segmentation method to excavator action segmentation and recognition. The performance of two long video action segmentation methods in earthwork excavation is compared. Both methods performed well in the stable shooting with no shelter of the excavator. In addition, the study found that when the video shooting was stable but the excavator was slightly shielded, the ASRF model could still identify the excavator action type well with its advantages of the boundary regression branch. When the excavator is slightly blocked, the efficiency of the MS-TCN model in accurately identifying the action of the earthmoving excavator is reduced. And the two models have higher requirements on the stability of video shooting.

In the actual construction site application, based on the long video to identify the working state of excavator, can automatically calculate the working time of earthmoving excavator of various types of work, analysis of the root cause of productivity deviation. For example, the longer duration of “swing” action may indicate that the excavator rotates at a large angle, which is one of the most critical factors to increase the mining operation cycle and reduce the mining efficiency. In addition, a long “digging” action of excavation may indicate the operating environment, such as high soil strength and great difficulty in excavation. And if it is found that “stopping” action frequently occurs in the work of earthmoving excavator, it may indicate that the arrangement of muck transport vehicle needs to be adjusted. These identification results can help the construction managers to train operators, arrange work and adjust the whole project strategy in time to improve the efficiency of earthmoving excavators.

Acknowledgement

This paper is financially supported by National Science Foundation of China under grants 52077213, 62003332, Project of Outstanding Young Scientific and Technological Talents of Guizhou Province QKEPTRC[2019]5650. In addition, the authors would like to thank the handling editor and the reviewers for their constructive comments.

References

Mahmood, B., Han, S., & Seo, J. (2022). Implementation experiments on convolutional neural network training using synthetic images for 3D pose estimation of an excavator on real images. *Automation in Construction*, 133, 103996.

Naghshbandi, S. N., Varga, L., & Hu, Y. (2021). Technologies for safe and resilient earthmoving operations: A systematic literature review. *Automation in Construction*, 125, 103632.

Francis, R., & Bekera, B. (2014). A metric and frameworks for resilience analysis of engineered and infrastructure systems. *Reliability Engineering & System Safety*, 121, 90-103.

Golparvar-Fard, M., Heydarian, A., & Niebles, J. C. (2013). Vision-based action recognition of earthmoving equipment using spatio-temporal features and support vector machine classifiers. *Advanced Engineering Informatics*, 27(4), 652-663.

Kim, J., Chi, S., & Seo, J. (2018). Interaction analysis for vision-based activity identification of earthmoving excavators and dump trucks. *Automation in Construction*, 87, 297-308.

Kim, J., & Chi, S. (2019). Action recognition of earthmoving excavators based on sequential pattern analysis of visual features and operation cycles. *Automation in Construction*, 104, 255-264.

Gong, J., & Caldas, C. H. (2010). Computer vision-based video interpretation model for automated productivity analysis of construction operations. *Journal of Computing in Civil Engineering*, 24(3), 252-263.

Chi, S., Caldas, C. H., & Kim, D. Y. (2009). A methodology for object identification and tracking in construction based on spatial modeling and image matching techniques. *Computer‐Aided Civil and Infrastructure Engineering*, 24(3), 199-211.

Kim, H., Ahn, C. R., Engelhaupt, D., & Lee, S. (2018). Application of dynamic time warping to the recognition of mixed equipment activities in cycle time measurement. *Automation in Construction*, 87, 225-234.

Yang, K., Ahn, C. R., Vuran, M. C., & Kim, H. (2017). Collective sensing of workers' gait patterns to identify fall hazards in construction. *Automation in construction*, 82, 166-178.

Ahn, C. R., Lee, S., & Peña-Mora, F. (2015). Application of low-cost accelerometers for measuring the operational efficiency of a construction equipment fleet. *Journal of Computing in Civil Engineering*, 29(2), 04014042.

Akhavian, R., & Behzadan, A. H. (2016). Smartphone-based construction workers' activity recognition and classification. *Automation in Construction*, 71, 198-209.

Vahdatikhaki, F., Hammad, A., & Siddiqui, H. (2015). Optimization-based excavator pose estimation using real-time location systems. *Automation in Construction*, 56, 76-92.

Shi, Y., Zhu, Y. Y., Fang, J., & Li, Z. S. (2021). Pose Measurement of Excavator Based on Convolutional Neural Network. *Journal of Network Intelligence*, 6(2), 392-400.

Huang, L. W., Jiang, B. T., Lv, S. Y., Liu, Y. B., & Li, D. Y. (2018). Survey on deep learning based recommender systems. *Chinese Journal of Computers*, 41(7), 1619-1647.

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*, 25(4), 771-782.

Chi, S., & Caldas, C. H. (2011). Automated object identification using optical video cameras on construction sites. *Computer‐Aided Civil and Infrastructure Engineering*, 26(5), 368-380.

Chi, S., & Caldas, C. H. (2012). Image-based safety assessment: automated spatial safety risk identification of earthmoving and surface mining activities. *Journal of Construction Engineering and Management*, 138(3), 341-351.

Kim, H., Bang, S., Jeong, H., Ham, Y., & Kim, H. (2018). Analyzing context and productivity of tunnel earthmoving processes using imaging and simulation. *Automation in Construction*, 92, 188-198.

Farha, Y. A., & Gall, J. (2019). Ms-tcn: Multi-stage temporal convolutional network for action segmentation. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3575-3584)*.

Li, S. J., AbuFarha, Y., Liu, Y., Cheng, M. M., & Gall, J. (2020). Ms-tcn++: Multi-stage temporal convolutional network for action segmentation. *IEEE transactions on pattern analysis and machine intelligence*.

Ishikawa, Y., Kasai, S., Aoki, Y., & Kataoka, H. (2021). Alleviating over-segmentation errors by detecting action boundaries. *In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 2322-2331)*.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. *In Proceedings of the IEEE international conference on computer vision (pp. 4489-4497)*.